Rationally designed mimotope library for profiling of the human IgM repertoire

Anastas Pashov[1, *], Velizar Shivarov[2], Maya Hadzhieva[1], Victor Kostov[1,3], Dilyan Ferdinandov[3], Karen-Marie Heinz[5], Shina Pashova[1,4], Milena Todorova[1], Tchavdar Vassilev[1], Thomas Kieber-Emmons[6], Leonardo A. Meza-Zepeda[5], Eivind Hovig[5]

[1] Department of Immunology, Stephan Angeloff Institute of Microbiology, BAS, Sofia, Bulgaria
[2] Laboratory of Clinical Immunology and Department of Clinical Hematology, Sofiamed University Hospital, Sofia, Bulgaria; ORCID: 0000-0001-5362-7999
[3] Neurosurgery Clinic, St. Ivan Rilsky Hospital, MU, Sofia, Bulgaria
[4] Department of Molecular Immunology, Institute of Biology and Immunology of Reproduction, BAS, Sofia, Bulgaria
[5] Department of Tumor Immunology, Oslo University Hospital, Oslo, Norway
[6] Winthrop P. Rockefeller Cancer Research Center, UAMS, Little Rock, AR, USA

[*] Corresponding author.

E-mail: a_pashov@microbio.bas.bg (AP)
Phone: +359 897 944628

Abstract

A large part of the IgM repertoire is characterized by few somatic mutations, physiological autoreactivity with housekeeping function and polyspecificity. It can be a versatile source of biomarkers for immune monitoring consisting of multidimensional antibody profiles. Probing this repertoire with a set of immunodominant self-proteins provides only coarse information on protein reactivity patterns. In contrast, here we describe a deep panning approach for the rational selection of a set of peptide mimotopes that represents optimally the diversity of human IgM reactivities using a relatively small library of probes. A 7-mer random peptide phage display library was panned on pooled human IgM. A non-exhaustive set of 224087 mimotopes was defined using next generation sequencing of the selected phage. A Kullback–Leibler divergence based GibbsCluster algorithm grouped them in 790 sequence clusters. A set of 594 mimotopes representative of the most significant clusters was used to test the hypothesis that this approach samples uniformly the IgM repertoire. When probed with sera from patients with or without different brain tumors in an oriented peptide array, this set produced a higher dynamic signal as compared to random peptides, to random peptides purged of mimotope sequence profiles and to mimotopes from only 10 out of 790 clusters. In this respect, the representative library is considered an optimal probe of the human IgM diversity. Proof of principle predictors based on the optimized library demonstrated its capacity to produce well separable data for the randomly selected diagnoses. Thus, an optimized library of IgM mimotopes is found to address very efficiently the dynamic diversity of the human IgM repertoire, providing informationally dense, structurally interpretable, profiles of the repertoire's reactivities.

Introduction

The repertoire of human IgM contains a considerable proportion of moderately autoreactive antibodies characterized by low intrinsic affinity/ low specificity, functioning as a first line of defense [1], as scavengers of senescent cells and debris [2-6], and even in tumor surveillance [7]. It is becoming increasingly clear that the human antibody repertoire has an organization similar to that of its murine counterpart [8-12]. About one fourth of the murine splenic B lymphocytes that respond to lipopolysaccharide have B cell receptors which are moderately autoreactive. Affected very little by somatic mutations and follicular evolution, the physiological self-reactivities largely overlap with germline-encoded polyspecific antibodies [13-15]. Eighty percent of murine serum IgM falls in this category and is referred to as natural antibodies (nAbs) [16, 17]. Apart from the polyspecific splenic B cells, the source of nAbs in mice seems to be mostly a population of B1-related IgM^+ plasma cells residing in a unique IL-5 dependent bone marrow niche [18]. By interacting with structures of self, this part of the antibody repertoire can reflect changes in the internal environment and, thus, it can be a source of biomarkers if probed as a natural biosensor.

Our working hypothesis is that an essential part of the human IgM repertoire involved in homeostasis can be probed by a set of mimotopes, the size of which can be tailored to the diagnostic goals by optimization. The existing approaches for immunosignature [19, 20] or immunomic [21] analysis of the immunoglobulin repertoires focus mostly on IgG and have used arrays of either 10^2 proteins or 10^4-10^5 random peptides. The IgM repertoire has been previously probed by protein arrays [22] containing a biologically determined representative set of autoantigens which is a structurally coarse approach. We set out to explore the feasibility of a method that, similar to the self-protein “homunculus” arrays, targets a small set of rationally selected probes, but also preserves the structural interpretability of peptides in a format applicable for routine diagnostics.

Materials and methods

Deep panning

Human IgM was isolated from a sample of IgM enriched IVIg - IgM-Konzentrat (Biotest AG, Dreieich, Germany, generously provided by Prof. Srini Kaveri), while human monoclonal IgM paraprotein was isolated from an IgM myeloma patient's serum selected from the biobank at the Center of Excellence for Translational Research in Hematology at the National Hematology Hospital, Sofia (with the kind cooperation of Dr. Lidiya Gurcheva). In both cases, IgM was purified using affinity chromatography with polyclonal anti-μ antibody coupled to agarose (A9935, SIGMA-ALDRICH, USA). A 7-mer random peptide library (E8100S, Ph.D. -7, New England Biolabs, USA) was panned overnight at 4°C on pooled human IgM adsorbed on polystyrene plates at a concentration of 0.1 mg/ml, washed, eluted with glycine buffer at pH 2.7 and immediately brought to pH7. The eluate was transferred to a plate coated with monoclonal IgM and incubated according to the same protocol, but this time the phage solution was collected after adsorption and amplified once, according to Matochko et al. [23]. Briefly, the phage DNA was extracted and the peptide-coding fragment amplified by PCR. The amplicons were subjected to deep sequencing using the Next Seq platform (Illumina, USA), performed at the Sequencing Core Facility of Oslo University Hospital.

Mimotope selection

The reads from the deep sequencing experiment were processed using the script provided by Matochko et al. [23]. Reads found in 3-10 copies per million (CPM) were selected and subjected to clustering using the GibbsCluster-2.0 method [24]. The number of clusters was optimized using the Kulback-Leibler distance (KLD) from the background model of random sequences [24]. Position weighted matrices (PWM) were defined for each cluster using pseudo counts as follows:

$$PWM_{k,j} = log_2\left(\frac{\sum I\left(X_{i,j} = k\right) + b_k\sqrt{N}}{(N + \sqrt{N})b_k}\right)$$

where i=1..N are the rows of the alignment, j=1:7 are the columns of the alignment, I(aa=k) is the indicator function used to count the occurrences of amino acid k in column j and b_k is the background frequency of amino acid k in the phage display library [25].

Using the PWMs, the median of the log odds (LO) scores of the peptides in each cluster was calculated. Next, the probability of the occurrence of peptides with LO greater than the respective median in a set of random peptides 10-fold larger than the analyzed library was determined empirically. Using this estimate, the probability of the chance occurrence of as many peptides with scores higher than the median score as observed in each cluster was calculated using the binomial distribution. The probabilities, thus found for each cluster, were used as a criterion for their relevance to the mimotope library being generated.

The clustering of the mimotopes was visualized using the t-sne (t-Distributed Stochastic Neighbor Embedding) algorithm [26] in its implementation using the faster Barnes-Hut algorithm (Rtsne package) with theta parameter default value of 0.5 and a maximum of 1500 iterations [27]. Lower theta values would yield more reproducible maps, but they slow down (may be up to 10 fold) the algorithm which is very hard to parallelize. The amino acid residues were described using the 5-dimensional scale of amino acid residue properties published by Hellberg et al. [28]. The amino acid residue properties quantification used was the same as for the pepStat binding normalization [29]. The 5 scales (z1-z5) were extracted as the latent variables describing the major factors underlying the variability of amino acids in the space of 26 physicochemical parameters. Thus, each peptide was represented by a vector of 35 scores corresponding to its seven positions. For a comparison, the same number of random 7-mer peptides generated using the amino acid background frequencies of Ph.D.-7 were clustered using the same parameters. The clusters in the t-sne plot were labeled using k-mean clustering.

Patients' sera

Sera from randomly selected patients with glioblastoma multiforme (GBM), low grade glioma (G), brain metastases of breast (MB) or lung (ML) cancers, as well as non-tumor bearing patients (C) (herniated disc surgery, trauma, etc.) of the Neurosurgery Clinic of St. Ivan Rilski University Hospital, Sofia acquired according to the rules of the ethics committee of the Medical University in Sofia were analyzed on the sets of peptides defined in microarray format. The sera were aliquoted and stored at -20°C. Before staining the sera were thawed, incubated for 30 min at 37°C for dissolution of IgM complexes, diluted 1:100 with PBS, pH 7.4, 0.05% Tween 20 with 0.1% BSA, further incubated for 30 min at 37°C and filtered through 0.22µm filters before application on the chips.

Peptide microarray

The customized microarray chips were produced by PEPperPRINT^{TM} (Heidelberg, Germany) by synthesis in situ as 7-mer peptides attached to the surface through their C-terminus and a common spacer GGGS. The layout was in a format of a single field of up to 5500 or five fields of up to 600 peptides in randomly positioned duplicates. The chips were blocked for 60 minutes using PBS, pH 7.4, 0.05% Tween 20 with 1% BSA on a rocker, washed 3x1 min with PBS, pH 7.4, 0.05% Tween 20 and incubated with sera in dilutions equivalent to 0.01 mg/ml IgM (approx. 1:100 serum dilution) on a rocker overnight at 4°C. After 3x1 minute washing the chips were incubated with secondary antibodies at RT, washed, rinsed with distilled water and dried by spinning in a vertical position in empty 50 ml test tubes at 100 x g for 2 minutes.

Microarray data treatment

The microarray images were acquired using a GenePix 4000 Microarray Scanner (Molecular Devices, USA). The densitometry was done using the GenePix® Pro v6.0 software. All further analysis was performed using publicly available packages of the R statistical environment for Windows (v3.4.1) (Bioconductor –

Biostrings, limma, pepStat, sva, e1071, Rtsne, clvalid, entropy, RankProd, multcomp) as well as in house developed R scripts (https://github.com/ansts/IgMimoPap1 and https://github.com/ansts/IgMimoPap2).

Local normalization

The spot intensities of a non-treated chip were subtracted from the stained chip spot intensities. Treatment with secondary antibody only did not result in any binding. Next, the local background had to be inferred due to the dense spot layout. To that end an approximated background was smoothed by support vector regression and the result passed to the backgroundCorrect [30] function for local normalization using the normexp (mle) method. The initial background approximation is described in the supplement file 1.

Global normalization

The peptides with missing values (flagged "bad") in some of the patients were removed. Log transformed locally normalized microarray data were next normalized for amino acid composition dependent binding using the ZpepQuad method of pepStat package [31]. This step is considered indispensable because of the strong effect of amino acid composition on binding due mostly to electrostatic interactions. The amino acid residue properties were quantified using the 5 dimensional descriptor z1-z5 of Sandberg et al. (1998) [29]. This was followed by global normalization using normalizeCyclicLoess (method affy) from the package limma [30] and subjected to batch effect compensation using the ComBat [32] function from package sva whenever necessary (for different chip batches and for different channels). For batch effect compensation the groups were balanced by selecting a subset of the patients with relatively even representation of each tested diagnosis. Finally, each peptide binding intensity was represented by the mean of its duplicates.

Because of a lack of a suitable negative control, the baseline binding was determined from the data. The clean data used for the testing of the diagnostic potential was subjected to dimensionality reduction of

the peptide reactivities using t-sne [26], which clearly outlined a group of peptides with uniformly low reactivities, (Supplement fig.1) that was considered background binding. The mean of the background binding intensity was subtracted from the data before testing for significant reactivities.

A general linear model, followed by Tukey contrast, was used to compare the expression of reactivities in the different tested libraries. The total correlation between the different peptide reactivity profiles of each library with 10 patients' serum IgM was used as a measure of the redundancy of the library. The profiles were viewed as points in a 10-mer space with each dimension corresponding to a patient. Total correlation was calculated as KLD from the theoretical maximal entropy joint distribution of the profile data, binned appropriately. To keep the number of bins comparable to the number of peptides, the intensity values for each patient were discretized just to "high" and "low" relative to the median for the studied library, effectively centering each library's data and, thus, removing the effect of the median intensity of each library. To further limit the number of bins, randomly selected bins (the same for all libraries) were aggregated in groups of 8, yielding 128 bins for the calculation of the baseline and the actual joint distributions. To equalize the size of the libraries, random samples of 380 peptides were used from each library, except for pepneglo which was excluded from this analysis due to its small size. The mean KLD values were determined in a bootstrap procedure with the described sampling repeated with replacement 100 times for the bin aggregation and 30 times for the library resampling producing 3000 samples. KLD for each library was thus calculated using the entropy::KL function.

The same approach could not be used for the transposed matrix, due to the disproportionately high number of bins necessary to describe the distribution in high/low values of 500-1000 different peptides. Instead, the mean correlation coefficient between the patient profiles across the peptides was used. To compare the mean correlation values between libraries by GLM, the correlation values were converted to z-scores.

Another test of the optimal sampling of the mimotope space was the mean nearest neighbor distance (NND) per library which was normalized (nNND) relative to the theoretical mean distance between the points in the 10-dimensional data cloud for the different libraries:

$$nNND_L = \frac{NND_L}{\left(\frac{V}{N_L}\right)^{\frac{1}{k}}},$$

where L=1..8, N_L is the number of peptides in each library, k is the dimensionality (10 since the peptides are compared on the basis to their reactivity to 10 patients' sera) and V is the volume of the data cloud approximated as a k-dimensional ellipsoid [33]:

$$V = \frac{2\pi^{\frac{k}{2}}}{k\Gamma\left(\frac{k}{2}\right)} \prod_{i=1}^{k} (2 * \sigma_i),$$

where σ_i is the standard deviation of the data along the i_{th} dimension (the values of the i_{th} serum) and Γ is the gamma function. The logarithms of nNND were compared by general linear models (GLM).

Together with median intensity per library as a measure of the intensity of the signal, total correlation, mean correlation between patients' sera reactivities and mean nearest neighbor distance were combined to measure the optimization of the libraries. To make all four criteria positively correlated with the desired qualities of the libraries, the sign of the correlation measures was changed. The rank product method (p value for consistent "high expression" across the four tests) was used to test the optimization of the libraries.

Feature selection algorithm

For the design of a practical algorithm for extracting information about a particular diagnosis, a feature selection approach was used based on recursive feature elimination using the quality of clustering of the predetermine diagnostic groups as a criterion. The cluster separation of the cases of interest was measured using a combined criterion:

$$Crit = \frac{100 * Dunn(d,c) * BH\gamma(d,c)}{Conn(d,c)}$$

Where Dunn(d,c) is the Dunn's clustering criterion [34, 35] which is based on single inter-cluster and intra-cluster distances (extreme case), BHγ(d,c) is the Baker-Hubert Gamma index [36] which gives the overall agreement of distances and cluster assignment based on all the data, Conn is the connectivity validation measure for a given clustering partitioning based on 10 nearest neighbors [34], which emphasizes the agreement of distances and partitions in the vicinity of each case; *d* is the matrix of distances between the cases, and *c* is the diagnosis code for the cases. Let F be the set of features used - the union of features of significant expression in any diagnosis (n=380). The recursive elimination features selection from F was performed multiple times in a leave one out setting, both as a direct validation as well as a bootstrap scheme used to improve further the generalization of the proof of principle predictor.

The following algorithm was used for predicting GBM as contrasted to the rest of the cases. Let N be the number of cases. Each of the N bootstrap samples contained the binding data for $N-1$ cases. The recursive elimination feature selection applied to each of those samples ultimately produced an intersecting family of N feature sets. To ensure better generalization, only the sets of features common for at least 2 of those sets were analyzed. The common features were grouped and pooled by the number of sets they were common for (or "commonality").

Results

Selection of 7-mer mimotopes

We chose to pan a commercially available 7-mer random peptide phage display library of diversity 10^9. Thus, the size of the mimotopes would be in the range of the shorter linear B cell epitopes in the IEDB database (http://www.iedb.org/). At the same time, the complete diversity - 20^7=1.28x10^9 different peptides of sequences of that length could be interrogated. The phage eluted from the IgM pool representative of the repertoire of at least 10 000 healthy donors were adsorbed on a monoclonal IgM to

filter out non-specific binding of phage to the constant regions, and thereby focus binding only on the mimotopes (Fig. 1). The peptide inserts were amplified and deep sequenced using the approach described by Matochko et al. (2012) [23]. Two separate experiments starting with 20% of the original phage library were performed (experiments A and B), while in a third one (C), a preamplified 20% sample of the original phage library was used. The yield was 688 860 (experiment A), 518 533 (experiment B) and 131 475 (experiment C) unique reads. The Phred quality score cut off used was 32 with a probability for an erroneous base call of $6x10^{-4}$. The unique reads from experiments A and B were pooled so that only one copy of a sequence existed in the final set (n=1 114 845). The criterion for retaining a sequence was an occurrence in 3-10 CPM, yielding 224 087 unique sequences. Considering that 79% of the single base changes lead to a change in the encoded amino acid [37], the used Phred score led to about $1.1\ 10^4$ sequences with one erroneous amino acid residue if there is only one wrong base call per sequence and fewer, if the very rare multiple errors per sequence are considered. Since the criteria for inclusion is at least 3 CPM, the probability for the error to yield one sequence already present in 2 CPM (so that it is included in the selected set, there are 182 424 sequences at 2 CPM) is 7% and less than 0.3% for 2 and more incorrect sequences (binomial). This error rate could lead also to a loss of about 80 potentially selected sequences when it affects a sequence presented at 3 CPM (n=89 167). So, this Phred score can slightly restrict the selected sequences (false negatives), but produces practically no false positives. The probabilities for a change at the higher threshold (10 CPM) are negligible.

The limit of 10 CPM on the high copy numbers was applied after comparison of the distributions of the number of clones by CPM between the original and the preamplified library (Fig. 2). It showed that 10 CPM was the threshold discriminating the original and preamplified libraries, with diversity in the latter skewed towards highly proliferating clones. This fact was interpreted in view of the observation that the affinity selection seemed to favor low CPM clones. Therefore, the high CPM clones were excluded to avoid

a possible contamination with non-selected clones having an advantage when they are highly proliferating. This restriction led to the exclusion of 9.96% of the reads.

To optimally probe the mimotope structural space, the library was clustered based on the sequences. Thus, the set of 224 087 unique 7-mer mimotope sequences (from experiment A and B) were subjected to clustering using the GibbsCluster-2.0 method [24] originally applied for inferring the specificity of multiple peptide ligands tested on multiple MHC receptors. The algorithm was used in its stand-alone version, which allowed an unlimited number of clusters and the use of large data sets. The number of clusters was optimized using the Kullback-Leibler distance from the background model of random sequences [24] in the range of 100 to 2500 clusters. This criterion indicated optimal clustering in 790 clusters (Fig. 3). PWMs were calculated from each cluster and used further to find the most relevant peptide sequence as a representative of each cluster (Supplement file 2).

Structural properties of the mimotope clusters

The frequencies of the amino acids residues in the mimotopes selected from the phage library showed a skewing in favor of the amino acid residues G,W,A,R,T,H,M,P,Q and against C,F,N,Y,I,L and S (Fig. 4A). The evidence of selection stood out in the overall PWM, showing higher divergence of the frequencies in the N-terminal part (Fig. 4B). L,F and C were not only underrepresented, but they also were particularly reduced in frequency in the first position. The most pronounced deficit was for C being even further selected against with respect to its already reduced representation in the phage display library (accounted for by the background probabilities used to calculate the log odds). Interestingly, proline was avoided at the first position, but was highly selected for in the second positions, and in the C-terminus. G,W,A,T,M and E were also highly represented in the N-terminus.

As expected, the residues found selected for or against in particular N-terminal positions were also involved in shaping the majority of the cluster features. The log odds (LO) forming the position weighted matrices for the different clusters were further combined in a linearized form in an overall matrix of 140

positions/amino acids by 790 clusters. This matrix was used to determine the amino acid residue in each position defining the clusters. Biclustering of the overall LO matrix (Supplement Figure 2) again showed clearly that P, W and H in all positions stand out as the most informative residues that determine to a great extent the formation of most of the 790 prototype mimotope clusters. Cysteine and several avoided residues belong also to this cluster of informative residues due to their extremely low LO values. A small number of mimotope clusters are defined by the presence of C, which is practically lacking from the rest of the clusters.

Generation of different mimotope libraries

The maximal optimized library at this stage would consist of these 790 representative sequences as they evenly sample the mimotope sequence space, as ensured by the GibbsCluster algorithm. This algorithm looks for an association between a peptide and a profile (cluster) versus other profiles. The final clusters vary with respect to the probability of occurrence of groups of that size and PWM in a random collection of peptides. The respective probability (based on binomial distribution) was used to rank the profiles by significance. Only the top 594 clusters (with $p<1e-4$) were retained, and only the mimotope with the top LO score from each cluster was kept as a mimotope prototype for the profile. This library was labeled peppos. Next, each of $2.3x10^6$ random 7-mer peptide sequences was tested against each of the PWM of mimotope clusters defined. For each random peptide, only the score of the top scoring cluster was retained, and the peptides were ranked in the ascending order of these scores. The lowest ranking peptides represented random sequences that were the least related to the selected library and the first 753 were, thus, used as a negative control (library pepnegrnd).

Other libraries were also generated based on: 1) low probability for relevance to the cluster based on KLD scores of the peptides in each cluster (pepneg and pepneglo); 2) 2 groups of 5 highest scoring clusters (pep5 and pepother5); 3) random 7-mer peptides predicted to belong to 5 highest scoring clusters by calculating the respective scores relative to the PWM of the clusters (pep5pred) and 4) random 7-mer

peptides (peprnd) (see Table 1 for description of all libraries). The number of peptides per library was constrained by the size of the chip.

Comparison between libraries

Sera from patients with glioblastoma multiforme (GBM), brain metastases of breast cancers (MB) as well as non-tumor bearing neurosurgery patients (C) (herniated disc surgery, trauma, etc.) were analyzed on the sets of peptides synthesized in an oriented (C-terminus attached) planar microarray format. In the first round of experiments, the 8 different libraries defined were compared based on the IgM reactivity in the sera from 10 patients (Supplement Fig. 3 and 4). The data on the mean serum IgM reactivity of each peptide across the different sera were grouped by library and the libraries were compared for their overall reactivity using GLM (Fig. 5A). The proposed optimized library (peppos) had significantly higher ($p<0.001$) average reactivity than pepneg, peprnd or pepnegrnd. Interestingly, the library theoretically purged of relevant reactivities (pepnegrnd) had indeed the lowest reactivity, significantly lower than both the weakly clustering peptides (pepneg) and the random sequences (peprnd) (Suppl. Table 1).

Next, the capacity of the different libraries to sample optimally the space of mimotopes was tested by comparing the total correlation of the IgM reactivities of the peptides mapped on the 10 different sera (Fig. 5B). The total correlation is a KLD based multidimensional generalization of mutual information. High total correlation would signify redundancy in the library with many peptides sharing similar reactivity profiles. On the contrary, the minimal total correlation would be found in a library where all reactivity profiles are close to orthogonal to each other. The library peppos had the lowest total correlation, while pepnegrnd, and especially pep5pred had the top total correlation, indicating redundancy in the information each mimotope carries about the patients, while the pools of 5 clusters – pep5 and pepother5 – had relatively low correlation. All differences, except between the top two libraries (the library purged of sequences similar to the clusters and the library of predicted sequences similar to 5 highly significant clusters), were significant (Supplement, Table 2).

Another way to test the optimal probing of the mimotope space in the different libraries is to compare the mean nearest neighbor distance (MNND) of the scaled and centered data of mimotope staining intensity mapped to the 10 patients' sera IgM. Theoretically, the projections of the mimotopes of an optimal library in the space of multiple individual sera/repertoires should be widely and evenly spaced. The MNND of peppos ranked second only to pepneglo (Fig. 5D) and was significantly higher than MNND of all the other libraries (Supplement Table 4.).

Alternatively, the correlation of the serum profiles based on the different mimotopes (transposing the matrices of the previous tests) can be viewed as a criterion for the capacity of the libraries to extract as much information as possible from the IgM repertoire. Due to the extreme multidimensionality of this feature matrix and the very small sample size (10 patients), the total correlation approach could not be applied. Instead, the mean correlation between patient pairs was used to compare the libraries after z transformation to allow comparison by linear models (Fig. 5C). Again, when tested against the peppos library, the sera exhibited the lowest correlation, being significantly lower compared to the reactivities to all the other libraries except for pepnegrnd and pepneglo (Supplement. Table 3.)

Finally, all four criteria were summarized using a rank product test, which proved that reactivity with peppos stands out from all the other tested libraries as the optimal among them for probing the IgM repertoire (Table 2).

Visualization of the Mimotope Space

T distributed stochastic neighbor embedding (t-sne) based on the Barnes-Hut algorithm was used to visualize the structure of the mimotope space as represented by the general mimotope library produced by deep panning (n=224 087). The sequences were represented by converting each amino acid residue to a 5-dimensional vector of physical property scores as described in Materials and methods. Thus, each 7-mer peptide is represented by a 35-dimensional vector. The correlation dimension showed that the

intrinsic dimension of the peptide space in this mapping was 11.25 for random peptides and 11.05 for the phage selected mimotope library. Principle component analysis indicated that the first 14 principle components account for approximately 75% of its variance and the remaining components had each a sharply lower contribution than the first 14 (Supplemental Figure 5). The t-sne mapping was done after reducing the 35 dimensions to 14 by PCA (the initial.dims parameter of the rtsne function). Figure 6 shows the maps for the mimotope library and for an equal number of random 7-mer peptides constructed using the residue background frequencies of the phage display library. Mapping also the five most significant clusters defined by GibbsCluster (the pep5 library) on the t-sne map shows that the two clustering approaches yield different results with only limited correlation between them. Nevertheless, the mapping of the optimized library (peppos) showed that even on this map it covers quite uniformly the mimotope space (Figure 7). To compare in more detail the mimotope space to the overall random peptide space, equally sized random samples of 50 000 mimotopes and random peptides together with the pepnegrnd library were merged. After removing the duplicate sequences, the mixture contained 49964 mimotopes, 49950 random peptides and 684 peptides from library pepnegrnd. This set of peptides was plotted again using the same approach and parameter values as for the whole library (Figure 8). While most of the clusters were uniformly represented by the mimotope library, the rest contained fewer mimotopes, although the structural space in them was still sampled uniformly. The random peptide sequences selected for their minimal similarity to the mimotope clusters defined in the previous section (library pepnegrnd) mapped to the areas of the low density of mimotopes. The t-sne representation of the mixture of peptides was further clustered using k-means clustering in 350 clusters as a means of binning neighboring points. The number of random peptides and mimotopes were counted in each cluster (Supplement Figure 6). This plot allowed to identify the parts of the peptide space which were underrepresented by the mimotopes. According to this classification of the peptides, approximately 42 % of the random peptides, 14% of the mimotopes and 85% of the pepnegrnd library were in the

underrepresented areas (Chi square, $p<0.0001$). The underrepresented areas had very similar sequence profiles to the normally represented areas, except for less abundant charged residues (Suppl. Data file "t-sne cluster profiles").

Diagnostic potential of a rationally designed mimotope library

A universal mimotope library sampling optimally the public IgM reactivities would have multiple applications both in the theoretical research of antibody repertoires, as well as in the design of theranostic tools. Having support for the hypothesis that the mimotope library peppos, sampling major structural clusters, is optimal, we next studied its diagnostic potential using a larger set of patient sera (n=34). Obviously, the "proof of principle" set of data used here is by far insufficient for generating a reliable predictor, e.g. - one based on machine learning. The main goal was rather to test its capacity to provide feature subsets suitable for building predictors. The composition of the library underwent a small change - 75 of the lowest scoring sequences from pepnegrnd library were added to the library as a negative control. Because of the size limit of the final library, this addition was done by replacement of 75 of the selected positive peptides. The distribution of patients by diagnosis is shown in Table 3. The data was acquired in 2 different batches using the green (batch G) and the red channel (batch R). Eleven cases were part of an earlier experiment with multiple libraries, which included the peppos peptides and that subset of the data could be used also in this assay (batch P). After cleaning, local and global normalization with batch effect compensation, and excluding 2 breast cancer patients found only in batch "R", as well as 4 GBM patients from the same batch to balance the groups, the reactivity data represented 28 patients' serum IgM binding to 586 peptides. The criterion for inclusion of the 5 GBM patients from batch "R" was the minimal difference of the mean and coefficient of variation from the mean and CV of the whole group of GBM in that batch. Having balanced groups warranted the use of ComBat for batch compensation [38]. The comparison between the staining intensities of the mimotopes (or features) in the patients' diagnostic groups yielded overlapping sets of reactivities significantly expressed in each diagnosis - 290 features for

GBM, 263 for ML, and 204 for C. Overall, 380 features showed significant reactivity in at least one of the diagnostic groups. The "negative" peptides (library pepnegrnd) predominated in the non-significant features. They represented 49/206 non-significant and 24/380 significant reactivities (χ^2, $p<0.0001$). The "negative" peptides are rather a set of sequences with low probability for reactivity determined using a small sample of sera. The finding of individuals with IgM reactive for some of them when testing a larger group is not a surprise. That is why the background reactivity was considered more reliably determined by the data analysis, rather than on the mean level of the pepnegrnd library.

A straightforward projection of the cases on the positive reactivities by multidimensional scaling (MDS) showed no separation (Suppl. Fig.7). The feature space is highly multidimensional. The peptide library is not targeted to any particular pathology, but represents a universal tool for IgM repertoire studies, and reflects the population diversity with respect to different pathological and physiological states. Therefore, a feature selection step is needed for each potential predictor of a given state. A recursive elimination algorithm was applied for that purpose, removing at each step the feature, the removal of which improved maximally the quality of clustering of the predetermined diagnostic groups (see Materials and methods section for details). Supplement Figure 8 shows the typical traces of the recursive elimination algorithm with the optimal set of features corresponding to the maximum of the clustering criterion. Furthermore, this feature selection strategy was included in a leave one out (LOOV) setting, both as a validation test for the feature selection and also as a bootstrap approach that allowed for a study of the variability of the diagnostic feature sets and further improvement of the feature selection.

Using this approach, we tested the capacity of the selected feature sets to separate dichotomously each of the diagnoses from the rest. As expected, LOOV based on a very small training set did not show significant prediction. In an attempt to compensate for the small training set and to increase the generalization of the predictors, next we used the LOOV as a bootstrap scheme. The feature sets derived from each of these bootstrap samples contained recurring features, and the number of sets a within which

a feature was found was considered its “commonality”. The selected features were regrouped, uniting features with commonality greater than a threshold, i.e. – the group labeled 1 included all features, group n included features found in at least n bootstrap sets and group 28 contained the features found in all bootstrap sets. Next, we tested the predictor’s efficiency as a function of the feature set ‘s commonality. Multidimensional scaling was used to extract only the 2 most informative factors from the feature sets - the maximal acceptable dimensionality for the small size of the training set of sera. The predictor was constructed with the resulting two-dimensional data using a radial basis function (Gaussian) kernel-based support vector machine. The performance of the model was measured using the Matthew’s correlation coefficient (MCC). As can be seen in Fig. 9, the most efficient predictors were composed on the basis of feature sets with intermediate commonality. Figure 10 shows the distribution of the features of commonality 50% derived by applying this algorithm to each of the three diagnoses separately.

Among the set of features selected to differentiate the GBM cases, we used further those of 50% commonality (n=55) to predict the diagnoses of the cases of batch “R”. By including cases omitted before the batch compensation, this batch served both as a source of a validation set, as well as a control for the non-confounding effect of the batch compensation. As can be seen from Fig. 11, the feature set of intermediate commonality (50%) predicted correctly the validation cases using an optimal support vector machine model (the color shade of the plot stands for the probabilities of each point to belong to the GBM or non-GBM set). The Supplemental Figure 9 shows a schematic representation of this algorithm.

Discussion

The introduction of high-throughput omics screening methods has expanded the knowledge base and the potential for diagnosis. All these approaches have led to the identification of biomarkers as profiles extracted from a particular dynamic diversity – proteome, genome, glycome, secretome, etc. A less explored source of biomarker profiles is the IgM antibody repertoire. IgM antibodies appear early in the course of an infection. However, they fall relatively fast, even after restimulation, providing a dynamic signal. Constitutive natural IgM levels, on their part, vary as a consequence of their homeostatic function. Consequently, IgM antibodies have gained interest as biomarkers of physiological or pathological processes [39-43]. IgM antibodies are still underused as immunodiagnostics, although their interactions with sets of antigens have been studied in a range of platforms [39, 42-45]. These studies indicate a significant, but untapped, potential in systems level screening of patients' blood serum using arrays of appropriately designed targets for the discovery of (natural) IgM signatures, enabling diagnosis and prognosis.

The use of the antibody repertoire as a source of biomarkers has been defined and approached in multiple ways. First came the technically minimalistic, but conceptually loaded, semiquantitative immunoblotting, developed 20 years ago. This technique served as no less than a paradigm setter for systems immunology [46-51]. The further development produced methods that have been referred to as functional immunomics [21] in terms of protein reactivities, as immunosignaturing [19] in terms of random peptide libraries, or described as a deep panning technique [52] and in terms of IgOme of mimotopes selected from random phage display libraries. Here we describe the design of the first mimotope library for the analysis of the human IgM repertoire of reactivities recurrent in most individuals [12, 53, 54]. **T**he development of a rationally designed IgM mimotope library presented here represents an omics concept-driven study of antibody specificity. The deep panning approach relies on next generation sequencing (NGS), and thus requires balancing between sequence fidelity and diversity. Within a number of reads

limited by the technology (i.e. – on the order of 10^7), high sequencing depths are impossible when the target diversity is itself on the order of 10^5-10^6. Hence our choice of 3 CPM as a lower limit. In addition, highly represented sequences are to be avoided, in order to reduce the effects of the collapse of diversity during phage display amplification. Choosing reads of less than 10 CPM , we obtain a set of 224 087 mimotopes which does not contain any of the parasitic sequences reported by Matochko et al. [55]. The use of monodispersed droplets-compartmentalized phage amplification [56] can overcome the latter problem. But even with diversity affected by discarding sequences of one and two copies per million on the one hand, and overgrowth of phage clones on the other, our strategy still manages to find a general representation of the mimotope space by identifying clusters of mimotopes. This relatively small set of structural classes is hypothesized to be related to the modular organization of the repertoire defined previously [57].

In general structural terms, the reduced sequence entropy of the selected mimotopes, especially of their N-terminus, indicates selection and comes as a proof of the efficiency of the phage selection step. The abundance of W and P is not surprising, since these amino acids are often found in epitopes and carbohydrate mimotopes [58]. The central role of prolines in the natural antibody mimotopes has also been observed previously [59]. Since Tchernychev et al. also used a phage display library, it is possible that the high proline content is related to the structural basis of recognition by antibodies, because prolines are associated with turns and flanking structures. Proline abundance may also be related to a necessity to control the entropic component of the binding, which is important in the case of polyspecific antibodies and peptide ligands.

The mimotope library of diversity 10^5 we derived by deep panning reflects the recurrent (also referred to as public) IgM specificities found in the human population. It can be used as is in large arrays when applicable. Using this library, a generalization of the mimotope space based on Gibbs Clustering was attempted, yielding optimal clustering in 790 sequence clusters. The latter are viewed as broad structural

contexts, each including multiple polyspecificities. This classification was used to produce a smaller and more applicable library for clinical use, of a subset of approx. 600 mimotopes (peppos) by picking representative sequences from the most significant clusters. Thus, this library was designed to optimally represent the mimotopes' main public reactivity patterns. The proposed optimized library could be used as a tool for the study of the IgM public repertoire, as a source of mimotopes for immunotherapeutics design [60-63], but it may serve also as a multipurpose diagnostic tool.

To study the properties of the optimized small library peppos, we compared it to 7 other libraries (Table 1.). Some were derived from the same large mimotope library as peppos, but represented two different subsets of 5 highly significant clusters (pep5 and pepother5), or the peptides with the least certain classification from each cluster (pepneg and pepneglo). Other libraries represented either random peptides with the same amino acid frequencies as the phage display library (peprnd), filtered random peptides scoring low with the PWMs of the 790 clusters (pepnegrnd) or sequences predicted using the PWM of 5 of the defined clusters (pep5pred). When the 8 libraries were compared across patients, the reactivities to the optimized library peppos had the lowest total correlation, while pepnegrnd, and especially the predicted mimotopes pep5pred, had the top total correlation. This measure relates to mutual information and in this case defines the redundancy of the reactivity profiles within each library. When its value is high, many peptides in the library have very similar profiles and basically carry redundant information about the patients, while the goal is to increase the information provided by a library by decreasing this redundancy. Thus, peppos proved the most informative, while the predicted pep5pred had the highest redundancy. The latter library represents similar peptides defined by a rule (PWM) derived on the basis of another library - pep5. Sequence profile captures the central structural theme, but mislead about the existing diversity in the cluster. The other measures used to test the optimality of peppos were two more correlation measures that capture different aspects of this redundancy, as well as the mean intensity to ensure that the library provides a sufficiently strong signal.

As a diagnostic tool, the optimized small library has some key properties that distinguish it from the omic sets. Since it is designed to represent practically ubiquitous public specificities, the diagnostic profiles would be quantitatively, rather than qualitatively, defined by combinations of reactivity to the different mimotopes. Indeed, the sets of features (mimotope reactivities), significantly expressed in the different diagnoses, were overlapping considerably. No single reactivity was correlating strongly with a whole diagnostic group, but subsets of reactivities collectively could separate the diagnoses. Thus, feature selection becomes essential for the design of predictors based on polyspecificities. Using the proposed algorithms, the typical feature set optimal for a dichotomous separation of diagnoses contained between 28 and 111 sequences (median=66). The improvement of generalization by keeping only features recurring in the bootstrap feature selection cycles helped reduce the overfitting of the models, due to the very small number of samples. This approach may be useful also in the case of appropriately sized teaching sets that would include in the order of 10^3 samples. Dependent on their commonality, the common feature sets contained 3-212 features. Those of the optimal 50% minimal commonality contained 55-62. Thus, if the library provides in the order of 500 significant features for most of the diagnoses, the theoretical capacity of this approach is $>10^{70}$ different subsets. These are further increased by at least 6 orders of magnitude when different quantitative profiles are considered if expression is discretized to only low, medium and high, and much more without discretization. Thus, the information provided by a typical experiment with the library is probably enough to describe any physiological or pathological state of clinical relevance provided it is reflected in the IgM repertoire status. Of course, this is just an estimate of the resolution of the method. The number of naturally occurring profiles and their correlation with clinically relevant states will determine the actual capacity. Another obvious pitfall would be the capacity of the predictor to generalize, which would necessitate further development of the feature selection strategies, as well as training with appropriately sized teaching sets.

The novelty of our approach is based on the combination of several previously existing concepts:

First, the antibody repertoire is compartmentalized in various ways (by isotype, B-cell subsets, nature or origin of the antigens, relatedness to immunopathology, etc.). Early studies argued that the physiologically autoreactive natural antibodies comprise such a consistent, organized compartment [48, 51, 64-67]. The consistency of the natural antibody self-reactivity among individuals was considered evidence for the existence of a relatively small set of preferred self-antigens. Such "public reactivities" are most probably related to the germline repertoire of antibodies generated by evolutionarily encoded paratope features and negative/positive selection [22, 68]. This set was targeted using protein microarrays, the utility of which has been previously demonstrated [21, 22, 43, 57]. Recently, the existence of structurally distinct public V-regions has been analyzed using repertoire sequencing [12], noting that they are often found in natural antibodies. If the repertoire should be read as a source of information providing consistent patterns that can be mapped to physiological and pathological states, the public natural autoreactivity seems to be the most suitable compartment. Therefore, the template used for this study was a pooled human IgM preparation, which emphasizes common specificities and dilutes out unique or rare ones. Respectively, IgM seems more suitable than IgG or IgE, but IgA could be also very useful, especially for probing the immune system reflection of the microbiome dynamics.

Second, germline variable regions are optimized for polyspecificity or cross-reactivity with protein and non-protein antigens [14]. This raises the question: Do nominal antigens actually exist for polyspecific germline antibodies, and what is the structure of this convoluted (because it is polyspecific) repertoire of reactivities? It seems that going for epitopes could be a way to approach the said repertoire convolution. Yet, the actual epitopes will be mostly conformational and hard to extract. In similar tasks, mimotopes are often used [69-72], although M.H. Van Regenmortel argues that mimotopes are of little use to structural prediction of the B-cell epitope [71]. Their utility might be rather in the structural study of the repertoire as a whole.

Third, the usage of peptide arrays for the analysis of the antibody repertoire is increasingly popular [52, 73-76]. It involves the use of random peptide arrays for extracting repertoire immunosignatures on the one hand and deep panning of phage display libraries to analyze antibody responses on the other. Since an antibody can often cross-react with a linear epitope that is part of the nominal conformational epitope [71], the 7-residue library offers suitable short mimotopes as compared to typical B-cell epitopes. Furthermore, from the Immunoepitope Database (http://www.iedb.org) collection of linear B cell epitopes, 4821 of 45829 entries are less than 8 residues long.

The study of the IgM repertoire might be expected to give information about interactions that occur mostly in the blood and the tissues with fenestrated vessels since, unlike IgG, IgM cannot easily cross the normal vascular wall. Yet, IgM tissue deposits are a common finding in diverse inflammatory conditions [77-79] and especially in the disorganized vasculature of the tumors, where they are a key element of the innate immune surveillance mechanism [7, 80, 81]. Changes in the IgM repertoire further reflect B cell function affected by antigenic, danger and inflammatory signals, but also by anatomical changes leading to vascular permeability or disruption. Thus, IgM repertoire monitoring has the potential to provide clinically relevant information about most of the pathologies involving inflammation and vascular remodeling, as well as all types of cancer. The library also provides a rich source of mimotopes that can be screened for different theranostic tasks focused on particular targets. On an omics scale, the optimized mimotope library proposed here probes efficiently the relevant repertoire of public IgM reactivities matching its dynamic diversity with potentially over 10^{70} distinct profiles. The major task ahead is designing studies aimed at efficiently extracting specific diagnostic profiles and building appropriate predictors, e.g. – for classifying immunotherapy responders, predicting the risk of malignancy in chronic inflammation, etc.

ACKNOWLEDGEMENTS This work was performed with the support of EEA/Norway Grant BG09/D03-103 and the Bulgarian Fund for Scientific Research Grant D01-11/2016. The authors wish to thank Prof. Radha Nagarajan, Prof. Ivanka Tsakovska and Prof. Soren Hairabedyan for critically reading the manuscript and a number of useful comments.

1 **Tables**

2 **Table 1. Libraries of 7-mer peptides studied.**

3

Library	Description	N
peppos	The sequence with the highest score for the respective position weighted matrix from each significant cluster (significant clusters are those for which the number of sequences with more than median PWM score is greater than the expected number of occurrences of such score in random peptides - p<0.0001 by Binomial test) .	594
pep5	A group of 5 of the 288 clusters with best binomial p<1e- 16 : clusters # 2,6,9,10,11.	600
pepother5	A group of 5 of the 288 clusters with best binomial p<1e- 16 : clusters # 115,61,55,53, 258.	1193
pep5pred	A hundred and fifty random peptides with log odds scores greater than the median score of the respective cluster for each of 5 clusters (# 2,6,9,10,11).	750
pepneg	The lowest scoring sequence (using KLD) from each significant cluster. These peptides are least certain to belong to the clusters.	594
pepneglo	Among the set of the lowest scoring peptides (pepneg) using GibbsCluster's own "Corrected" score - those with score <5 ([24]).	82
pepnegrnd	From a set of $2x10^6$ random 7-mer peptides, get the max score for each peptide when tested against all cluster PWM and take the 753 lowest scoring.	753
peprnd	800 random peptides.	800
	Total	5366

* The random sets are constructed with underlying frequencies in phage display library Ph.D -7 .

4

5

6

Table 2. Rank product test of four criteria for optimal mimotope library:

Library	Rank Products	p Value
pep5	4.864599	0.78057
pep5pred	5.825901	0.924705
pepneg	5.957892	0.937399
pepneglo	2.114743	0.054359
pepnegrnd	6.0548	0.945747
pepother5	3.22371	0.318709
peppos	**1.189207**	**0.001071**
peprnd	4.864599	0.78057

Table 3. Patients tested using the optimized library.

		Batches			
Diagnosis	**Abbr.**	**G**	**P**	**R**	**Total**
Non-tumor bearing (control)	C	1	3	4	8
Glioblastoma Multiforme	GBM	2	4	9(5)*	15 (11)*
Lung Cancer (Brain Metastasis)	ML	2	4	3	9
Breast Cancer (Brain Metastasis)	MB	0	0	2(0)*	2 (0)*
	Total				34

To balance the group sizes between batches, only 5/9 GBM samples from batch “R” were used and the breast cancer cases were omitted before batch compensation using the ComBat function. All cases in batch “R” were used in the validation step.

Table 4. Distribution of peptides among normally represented and underrepresented parts of the 7-mer peptide space.

	Random Peptides	**Mimotopes**	**pepnegrnd**
Normally Represented Region	35 166	43 708	100
Underrepresented Region	14 784	6 256	584

Legends

Figure 1. Schematic representation of the deep panning experiment.

Figure 2. Results from NGS. Plot of reads distribution by CPM (counts per million). Black – sample 1 – selected from original library, Red – sample 2 – selected form a pre-amplified original library. Here the selection caused deviation from the power law (the straight lines) with an emphasis on the lower CPM. The pre-amplification enriches in better fit phages and leads to a considerable shrinkage of the highly diverse compartment of low CPM where most of the targeted diversity lies.

Figure 3. Results from GibbsCluster of the mimotopes. Different predefined number of clusters were screened for the quality of clustering measured by Kullback-Leibler Divergence – KLD. The inset shows amplified scale around the peak KLD values.

Figure 4. Overall distribution of the amino acid residues in the selected 224 087 mimotopes as log odds (LO) relative to the background probabilities. Total abundance (A) and color coded LO by position in the 7-mer (B). There is a clear skewing of the representation of the residues due to the selection by the IgM repertoire which is indicated by the concentration of the non-random distribution in the free N-terminus as compared to the C-terminus which is attached to the PIII protein of the phage.

Figure 5. Statistics testing the libraries' capacity to probe the mimotope space. A) Mean reactivity of each peptide across patients grouped by library. The optimized library peppos has the highest reactivity. For library content see Table 1. B) Total correlation of the peptide profiles grouped by library across 10 patients. The optimized library peppos provides the least redundant information. C) Mean correlation of patient profiles across the peptides in each library compared after z-transformation. The optimized library peppos provides the most diverse characteristics of the patients, which indicates a high potential for discrimination of different states but increases the requirements for the size of the teaching sets to extract models of good generalization. D) Mean nearest neighbor distance of the peptide profiles across 10 patients in each library compared after z-transformation. Again, the optimized library peppos appears to sample the mimotope space evenly. The width of the bars is proportional to the size of the sample.

Figure 6. Visualization of the 7-mer peptide space of the phage display selected mimotopes (A) and equal number random peptides with the same background residue frequencies (B) constructed using t-sne based on the Barnes-Hut algorithm. The peptides in five clusters (included in library pep5) are color coded in (A). The peptide sequences were encoded using a 5-dimensional score reflecting basic biophysical properties of the amino acids. Each sequence is represented by a 35 dimensional vector but the t-sne mapping is performed after initial reduction to 14 dimensions by PCA.There is only moderate correlation between the clustering visualized by t-sne and the GibbsCluster classification. The image is of high resolution and can be zoomed for better detail inspection.

Figure 7. Visualization of the 7-mer peptide mimotopes space with the optimized library peppos marked in red (see fig.8 for details). Although, individual GibbsCluster defined clusters do not coincide with those shown by t-sne, the mapping of the optimized library apparently probes quite uniformly the mimotope space.

Figure 8. Visualization of the 7-mer peptide mimotope space representing a mixture of random sample of 50 000 phage display selected mimotopes (red) and 50 000 random peptides(gray) plus the

pepnegrnd library (blue). A part of the peptide space is represented by mimotopes at a lower density and the peptides unrelated to the defined 790 mimotope clusters map mosty to this area (blue points).

Figure 9. Matthew's correlation coefficient as a measure of the prediction quality for SVM models constructed using GBM predicting feature sets of different minimal commonality. Minimal commonality of n means that the features in the set are found in n or more of the bootstrap sets. The validation set consists of the cases in batch "R" that were omitted from the batch compensated united sets. The model predicts these cases as belonging to the same class as the rest of the respective cases in batch "R". Since the values in batch "R" were not subject to batch compensation the validation also serves as a control against confounding introduced by the ComBat function.

Figure 10. Venn diagram of the overlapping sets of features (peptide reactivities) of 50% commonality calculated separately for each diagnosis. These are the optimal sets for the prediction of each diagnosis based on the proof of principle training set. The figures correspond to the number of features in the corresponding intersection.

Figure 11. Multidimensional scaling plot of cases in batch "R" based on the feature set of minimal commonality of 14 (50%). See figure 9 for details. The encircled points correspond to the validation set.

References

1. Baumgarth N, Herman OC, Jager GC, Brown LE, Herzenberg LA, Chen J. B-1 and B-2 cell-derived immunoglobulin M antibodies are nonredundant components of the protective response to influenza virus infection. J Exp Med. 2000;192(2):271-80.
2. Ochsenbein AF, Fehr T, Lutz C, Suter M, Brombacher F, Hengartner H, et al. Control of early viral and bacterial distribution and disease by natural antibodies. Science. 1999;286(5447):2156-9.
3. Vollmers HP, Brandlein S. Tumors: too sweet to remember? Mol Cancer. 2007;6:78. PubMed PMID: 18053197.
4. Matter MS, Ochsenbein AF. Natural antibodies target virus-antibody complexes to organized lymphoid tissue. Autoimmun Rev. 2008;7(6):480-6. PubMed PMID: 18558366.
5. Avrameas S, Guilbert B, Dighiero G. Natural antibodies against tubulin, actin myoglobin, thyroglobulin, fetuin, albumin and transferrin are present in normal human sera, and monoclonal immunoglobulins from multiple myeloma and Waldenstrom's macroglobulinemia may express similar antibody specificities. Ann Immunol (Paris). 1981;132C(2):231-6.
6. Panda S, Zhang J, Tan NS, Ho B, Ding JL. Natural IgG antibodies provide innate protection against ficolin-opsonized bacteria. EMBO J. 2013;32(22):2905-19. doi: 10.1038/emboj.2013.199. PubMed PMID: 24002211; PubMed Central PMCID: PMCPMC3831310.
7. Vollmers HP, Brandlein S. Natural antibodies and cancer. N Biotechnol. 2009;25(5):294-8. Epub 2009/05/16. doi: S1871-6784(09)00060-0 [pii] 10.1016/j.nbt.2009.03.016. PubMed PMID: 19442595.
8. Prieto JMB, Felippe MJB. Development, phenotype, and function of non-conventional B cells. Comparative immunology, microbiology and infectious diseases. 2017;54:38-44. Epub 2017/09/17. doi: 10.1016/j.cimid.2017.08.002. PubMed PMID: 28916000.
9. Lobo PI. Role of Natural Autoantibodies and Natural IgM Anti-Leucocyte Autoantibodies in Health and Disease. Front Immunol. 2016;7:198. doi: 10.3389/fimmu.2016.00198. PubMed PMID: 27375614; PubMed Central PMCID: PMCPMC4893492.
10. Rothstein TL, Griffin DO, Holodick NE, Quach TD, Kaku H. Human B-1 cells take the stage. Annals of the New York Academy of Sciences. 2013;1285:97-114. doi: 10.1111/nyas.12137. PubMed PMID: PMC4429725.

11. Weller S, Braun MC, Tan BK, Rosenwald A, Cordier C, Conley ME, et al. Human blood IgM "memory" B cells are circulating splenic marginal zone B cells harboring a pre-diversified immunoglobulin repertoire. Blood. 2004. PubMed PMID: 15191950.
12. Greiff V, Weber CR, Palme J, Bodenhofer U, Miho E, Menzel U, et al. Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires. The Journal of Immunology. 2017;199(8):2985-97. doi: 10.4049/jimmunol.1700594.
13. Van Regenmortel MH. Specificity, polyspecificity, and heterospecificity of antibody-antigen recognition. J Mol Recognit. 2014;27(11):627-39. doi: 10.1002/jmr.2394. PubMed PMID: 25277087.
14. Willis JR, Briney BS, DeLuca SL, Crowe JE, Jr., Meiler J. Human Germline Antibody Gene Segments Encode Polyspecific Antibodies. PLoS Comput Biol. 2013;9(4):e1003045. doi: 10.1371/journal.pcbi.1003045.
15. Cohen IR, Young DB. Autoimmunity, microbial immunity and the immunological homunculus. Immunol Today. 1991;12(4):105-10.
16. Avrameas S, Guilbert B, Mahana W, Matsiota P, Ternynck T. Recognition of self and non-self constituents by polyspecific autoreceptors. Int Rev Immunol. 1988;3(1-2):1-15.
17. Avrameas S. Natural autoantibodies: from 'horror autotoxicus' to 'gnothi seauton'. Immunol Today. 1991;12(5):154-9.
18. Reynolds AE, Kuraoka M, Kelsoe G. Natural IgM is produced by CD5- plasma cells that occupy a distinct survival niche in bone marrow. J Immunol. 2015;194(1):231-42. Epub 2014/11/28. doi: 10.4049/jimmunol.1401203. PubMed PMID: 25429072; PubMed Central PMCID: PMCPMC4272881.
19. Hughes AK, Cichacz Z, Scheck A, Coons SW, Johnston SA, Stafford P. Immunosignaturing Can Detect Products from Molecular Markers in Brain Cancer. PLoS ONE. 2012;7(7):e40201. doi: 10.1371/journal.pone.0040201.
20. Stafford P, Halperin R, Legutki JB, Magee DM, Galgiani J, Johnston SA. Physical Characterization of the “Immunosignaturing Effect”. Molecular & Cellular Proteomics. 2012;11(4). doi: 10.1074/mcp.M111.011593.
21. Quintana FJ, Hagedorn PH, Elizur G, Merbl Y, Domany E, Cohen IR. Functional immunomics: microarray analysis of IgG autoantibody repertoires predicts the future response of mice to induced diabetes. Proc Natl Acad Sci U S A. 2004;101 Suppl 2:14615-21. PubMed PMID: 15308778.
22. Merbl Y, Zucker-Toledano M, Quintana FJ, Cohen IR. Newborn humans manifest autoantibodies to defined self molecules detected by antigen microarray informatics. J Clin Invest. 2007;117(3):712-8. PubMed PMID: 17332892.
23. Matochko WL, Chu K, Jin B, Lee SW, Whitesides GM, Derda R. Deep sequencing analysis of phage libraries using Illumina platform. Methods. 2012;58(1):47-55. doi: http://dx.doi.org/10.1016/j.ymeth.2012.07.006.
24. Andreatta M, Lund O, Nielsen M. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. Bioinformatics. 2013;29(1):8-14. doi: 10.1093/bioinformatics/bts621.
25. de Hoon M, Vitkup D. Comparative Systems Biology of the Sporulation Initiation Network in Prokaryotes. In: Eskin E, Ideker T, Raphael B, Workman C, editors. Systems Biology and Regulatory Genomics: Joint Annual RECOMB 2005 Satellite Workshops on Systems Biology and on Regulatory Genomics, San Diego, CA, USA; December 2-4, 2005, Revised Selected Papers. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 62-9.
26. van der Maaten LJP, Hinton G. Visualizing Data using t-SNE. Journ Machine Learning Res. 2008;9:27.
27. van der Maaten L. Accelerating t-SNE using Tree-Based Algorithms. J Machine Learn Res. 2014;15:3221-45.

28. Hellberg S, Sjoestroem M, Skagerberg B, Wold S. Peptide quantitative structure-activity relationships, a multivariate approach. Journal of Medicinal Chemistry. 1987;30(7):1126-35. doi: 10.1021/jm00390a003.
29. Sandberg M, Eriksson L, Jonsson J, Sjostrom M, Wold S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem. 1998;41(14):2481-91. Epub 1998/07/03. doi: 10.1021/jm9700575. PubMed PMID: 9651153.
30. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. doi: 10.1093/nar/gkv007. PubMed PMID: 25605792; PubMed Central PMCID: PMCPMC4402510.
31. Imholte G, Sauteraud R, Gottardo R. Analyzing Peptide Microarray Data with the R pepStat Package. Methods Mol Biol. 2016;1352:127-42. Epub 2015/10/23. doi: 10.1007/978-1-4939-3037-1_10. PubMed PMID: 26490472.
32. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics (Oxford, England). 2007;8(1):118-27. Epub 2006/04/25. doi: 10.1093/biostatistics/kxj037. PubMed PMID: 16632515.
33. Wilson JA. Volume of n-dimensional ellipsoid. Sciencia Acta Xaveriana. 2010;1(1):101-6.
34. Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. Bioinformatics. 2005;21(15):3201-12. Epub 2005/05/26. doi: 10.1093/bioinformatics/bti517. PubMed PMID: 15914541.
35. Dunn JC. Well-Separated Clusters and Optimal Fuzzy Partitions. Journal of Cybernetics. 1974;4(1):95-104. doi: 10.1080/01969727408546059.
36. Baker FB, Hubert LJ. Measuring the Power of Hierarchical Cluster Analysis. Journal of the American Statistical Association. 1975;70(349):31-8. doi: 10.2307/2285371.
37. Abdullah T, Faiza M, Pant P, Rayyan Akhtar M, Pant P. An Analysis of Single Nucleotide Substitution in Genetic Codons - Probabilities and Outcomes. Bioinformation. 2016;12(3):98-104. doi: 10.6026/97320630012098. PubMed PMID: PMC5267951.
38. Nygaard V, Rodland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics (Oxford, England). 2016;17(1):29-39. Epub 2015/08/15. doi: 10.1093/biostatistics/kxv027. PubMed PMID: 26272994; PubMed Central PMCID: PMCPMC4679072.
39. Silverman GJ, Srikrishnan R, Germar K, Goodyear CS, Andrews KA, Ginzler EM, et al. Genetic imprinting of autoantibody repertoires in systemic lupus erythematosus patients. Clin Exp Immunol. 2008;153(1):102-16. Epub 2008/05/31. doi: 10.1111/j.1365-2249.2008.03680.x. PubMed PMID: 18510544; PubMed Central PMCID: PMCPMC2432104.
40. Sharron B-Z, Dror YK, Gittit D, Asaf M, Yifat M, Francisco JQ, et al. Individual and meta-immune networks. Physical Biology. 2013;10(2):025003.
41. Mao J, Ladd J, Gad E, Rastetter L, Johnson MM, Marzbani E, et al. Mining the pre-diagnostic antibody repertoire of TgMMTV-neu mice to identify autoantibodies useful for the early detection of human breast cancer. J Transl Med. 2014;12:121. Epub 2014/06/03. doi: 10.1186/1479-5876-12-121. PubMed PMID: 24886063; PubMed Central PMCID: PMCPMC4022541.
42. Butvilovskaya VI, Popletaeva SB, Chechetkin VR, Zubtsova ZI, Tsybulskaya MV, Samokhina LO, et al. Multiplex determination of serological signatures in the sera of colorectal cancer patients using hydrogel biochips. Cancer Med. 2016. doi: 10.1002/cam4.692. PubMed PMID: 26992329.
43. Merbl Y, Itzchak R, Vider-Shalit T, Louzoun Y, Quintana FJ, Vadai E, et al. A Systems Immunology Approach to the Host-Tumor Interaction: Large-Scale Patterns of Natural Autoantibodies Distinguish Healthy and Tumor-Bearing Mice. PLoS ONE. 2009;4(6):e6053. doi: 10.1371/journal.pone.0006053.
44. Stafford P, Wrapp D, Johnston S. General assessment of humoral activity in healthy humans. Molecular & Cellular Proteomics. 2016. doi: 10.1074/mcp.M115.054601.

45. Campbell CT, Gulley JL, Oyelaran O, Hodge JW, Schlom J, Gildersleeve JC. Serum Antibodies to Blood Group A Predict Survival on PROSTVAC-VF. Clinical Cancer Research. 2013;19(5):1290-9. doi: 10.1158/1078-0432.ccr-12-2478.
46. Haury M, Grandien A, Sundblad A, Coutinho A, Nobrega A. Global analysis of antibody repertoires. 1. An immunoblot method for the quantitative screening of a large number of reactivities. Scand J Immunol. 1994;39(1):79-87. PubMed PMID: 8290896.
47. Stahl D, Yeshurun M, Gorin NC, Sibrowski W, Kaveri SV, Kazatchkine MD. Reconstitution of Self-Reactive Antibody Repertoires of Autologous Plasma IgM in Patients with Non-Hodgkin's Lymphoma Following Myeloablative Therapy. Clinical Immunology. 2001;98(1):31-8. doi: http://dx.doi.org/10.1006/clim.2000.4949.
48. Mouthon L, Haury M, Lacroix-Desmazes S, Barreau C, Coutinho A, Kazatchkine MD. Analysis of the normal human IgG antibody repertoire. Evidence that IgG autoantibodies of healthy adults recognize a limited and conserved set of protein antigens in homologous tissues. J Immunol. 1995;154(11):5769-78. PubMed PMID: 7751627.
49. Lacroix-Desmazes S, Mouthon L, Coutinho A, Kazatchkine MD. Analysis of the natural human IgG antibody repertoire: life-long stability of reactivities towards self antigens contrasts with age-dependent diversification of reactivities against bacterial antigens. Eur J Immunol. 1995;25(9):2598-604. PubMed PMID: 7589132.
50. Nobrega A, Haury M, Grandien A, Malanchere E, Sundblad A, Coutinho A. Global analysis of antibody repertoires. II. Evidence for specificity, self-selection and the immunological "homunculus" of antibodies in normal serum. Eur J Immunol. 1993;23(11):2851-9.
51. Mouthon L, Nobrega A, Nicolas N, Kaveri SV, Barreau C, Coutinho A, et al. Invariance and restriction toward a limited set of self-antigens characterize neonatal IgM antibody repertoires and prevail in autoreactive repertoires of healthy adults. Proc Natl Acad Sci U S A. 1995;92(9):3839-43.
52. Ryvkin A, Ashkenazy H, Smelyanski L, Kaplan G, Penn O, Weiss-Ottolenghi Y, et al. Deep Panning: steps towards probing the IgOme. PLoS One. 2012;7(8):e41469. doi: 10.1371/journal.pone.0041469. PubMed PMID: 22870226; PubMed Central PMCID: PMC3409857.
53. Truck J, Ramasamy MN, Galson JD, Rance R, Parkhill J, Lunter G, et al. Identification of antigen-specific B cell receptor sequences using public repertoire analysis. J Immunol. 2015;194(1):252-61. Epub 2014/11/14. doi: 10.4049/jimmunol.1401405. PubMed PMID: 25392534; PubMed Central PMCID: PMCPMC4272858.
54. Gu H, Tarlinton D, Muller W, Rajewsky K, Forster I. Most peripheral B cells in mice are ligand selected. J Exp Med. 1991;173(6):1357-71.
55. Matochko WL, Cory Li S, Tang SK, Derda R. Prospective identification of parasitic sequences in phage display screens. Nucleic Acids Res. 2014;42(3):1784-98. Epub 2013/11/13. doi: 10.1093/nar/gkt1104. PubMed PMID: 24217917; PubMed Central PMCID: PMCPMC3919620.
56. Derda R, Tang SK, Li SC, Ng S, Matochko W, Jafari MR. Diversity of phage-displayed libraries of peptides during panning and amplification. Molecules. 2011;16(2):1776-803. Epub 2011/02/23. doi: 10.3390/molecules16021776. PubMed PMID: 21339712.
57. Madi A, Hecht I, Bransburg-Zabary S, Merbl Y, Pick A, Zucker-Toledano M, et al. Organization of the autoantibody repertoire in healthy newborns and adults revealed by system level informatics of antigen microarray data. Proceedings of the National Academy of Sciences. 2009;106(34):14484-9. doi: 10.1073/pnas.0901528106.
58. Luo P, Agadjanyan M, Qiu J, Westerink MA, Steplewski Z, Kieber-Emmons T. Antigenic and immunological mimicry of peptide mimotopes of Lewis carbohydrate antigens. Mol Immunol. 1998;35(13):865-79. PubMed PMID: 9839555.
59. Tchernychev B, Cabilly S, Wilchek M. The epitopes for natural polyreactive antibodies are rich in proline. Proc Natl Acad Sci U S A. 1997;94(12):6335-9.

60. Scott JK. Discovering peptide ligands using epitope libraries. [Review]. Trends in Biochemical Sciences. 1992;17(7):241-5.
61. Westerink MA, Giardina PC, Apicella MA, Kieber-Emmons T. Peptide mimicry of the meningococcal group C capsular polysaccharide. Proc Natl Acad Sci U S A. 1995;92(9):4021-5.
62. Kieber-Emmons T. Peptide mimotopes of carbohydrate antigens. Immunol Res. 1998;17(1-2):95-108. PubMed PMID: 9479572.
63. Pashov A, Canziani G, Monzavi-Karbassi B, Kaveri SV, Macleod S, Saha R, et al. Antigenic properties of peptide mimotopes of HIV-1-associated carbohydrate antigens. J Biol Chem. 2005;280(32):28959-65. PubMed PMID: 15955803.
64. Cohen IR. The cognitive paradigm and the immunological homunculus. Immunol Today. 1992;13(12):490-4. doi: 10.1016/0167-5699(92)90024-2. PubMed PMID: 1463581.
65. Cohen IR. Autoantibody repertoires, natural biomarkers, and system controllers. Trends Immunol. 2013;34(12):620-5. doi: 10.1016/j.it.2013.05.003. PubMed PMID: 23768955.
66. Lacroix-Desmazes S, Mouthon L, Pashov A, Barreau C, Kaveri SV, Kazatchkine MD. Analysis of antibody reactivities toward self antigens of IgM of patients with Waldenstrom's macroglobulinemia. Int Immunol. 1997;9(8):1175-83.
67. Mouthon L, Lacroix-Desmazes S, Nobrega A, Barreau C, Coutinho A, Kazatchkine MD. The self-reactive antibody repertoire of normal human serum IgM is acquired in early childhood and remains stable throughout life. Scand J Immunol. 1996;44:243-51.
68. Hardy RR, Hayakawa K. Positive and negative selection of natural autoreactive B cells. Adv Exp Med Biol. 2012;750:227-38. doi: 10.1007/978-1-4614-3461-0_17. PubMed PMID: 22903678.
69. Putterman C, Deocharan B, Diamond B. Molecular analysis of the autoantibody response in peptide-induced autoimmunity. J Immunol. 2000;164(5):2542-9.
70. Pashov A, Monzavi-Karbassi B, Kieber-Emmons T. Immune surveillance and immunotherapy: Lessons from carbohydrate mimotopes. Vaccine. 2009;27(25-26):3405-15. PubMed PMID: 19200843.
71. Van Regenmortel MH. What is a B-cell epitope? Methods Mol Biol. 2009;524:3-20. doi: 10.1007/978-1-59745-450-6_1. PubMed PMID: 19377933.
72. Huang J, He B, Zhou P. Mimotope-based prediction of B-cell epitopes. Methods Mol Biol. 2014;1184:237-43. doi: 10.1007/978-1-4939-1115-8_13. PubMed PMID: 25048128.
73. Weber LK, Palermo A, Kugler J, Armant O, Isse A, Rentschler S, et al. Single amino acid fingerprinting of the human antibody repertoire with high density peptide arrays. J Immunol Methods. 2017;443:45-54. Epub 2017/02/09. doi: 10.1016/j.jim.2017.01.012. PubMed PMID: 28167275.
74. Weiss-Ottolenghi Y, Gershoni JM. Profiling the IgOme: meeting the challenge. FEBS Lett. 2014;588(2):318-25. Epub 2013/11/19. doi: 10.1016/j.febslet.2013.11.005. PubMed PMID: 24239539.
75. Navalkar KA, Johnston SA, Stafford P. Peptide based diagnostics: Are random-sequence peptides more useful than tiling proteome sequences? J Immunol Methods. 2014. Epub 2014/12/17. doi: 10.1016/j.jim.2014.12.002. PubMed PMID: 25497701.
76. Legutki JB, Zhao ZG, Greving M, Woodbury N, Johnston SA, Stafford P. Scalable high-density peptide arrays for comprehensive health monitoring. Nature communications. 2014;5:4785. doi: 10.1038/ncomms5785. PubMed PMID: 25183057.
77. Kulthanan K, Pinkaew S, Suthipinittharm P. Diagnostic value of IgM deposition at the dermo-epidermal junction. Int J Dermatol. 1998;37(3):201-5. Epub 1998/04/29. PubMed PMID: 9556108.
78. Borrelli M, Maglio M, Agnese M, Paparo F, Gentile S, Colicchio B, et al. High density of intraepithelial γδ lymphocytes and deposits of immunoglobulin (Ig)M anti-tissue transglutaminase antibodies in the jejunum of coeliac patients with IgA deficiency. Clinical and experimental immunology. 2010;160(2):199-206. doi: 10.1111/j.1365-2249.2009.04077.x. PubMed PMID: PMC2857942.

79. Chan RK, Ding G, Verna N, Ibrahim S, Oakes S, Austen WG, Jr., et al. IgM binding to injured tissue precedes complement activation during skeletal muscle ischemia-reperfusion. J Surg Res. 2004;122(1):29-35. Epub 2004/11/04. doi: 10.1016/j.jss.2004.07.005. PubMed PMID: 15522311.

80. Hensel F, Hermann R, Schubert C, Abe N, Schmidt K, Franke A, et al. Characterization of glycosylphosphatidylinositol-linked molecule CD55/decay-accelerating factor as the receptor for antibody SC-1-induced apoptosis. Cancer Res. 1999;59(20):5299-306. PubMed PMID: 10537313.

81. Vollmers HP, Brandlein S. The "early birds": natural IgM antibodies and immune surveillance. Histol Histopathol. 2005;20(3):927-37. PubMed PMID: 15944943.

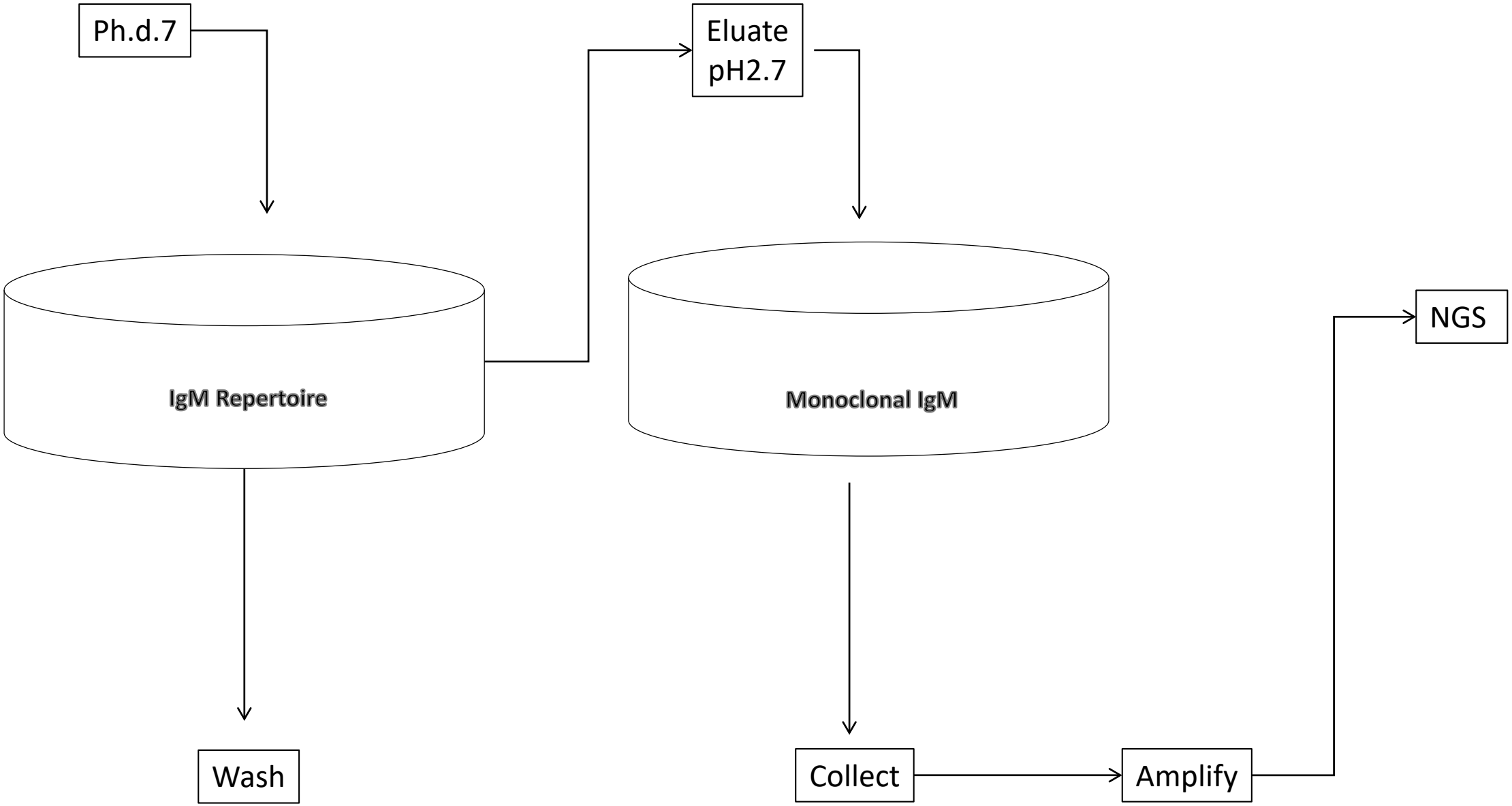

Ph.d.7
Eluate
pH2.7
IgM Repertoire
Monoclonal IgM
NGS
Wash
Collect
Amplify

Fig.1

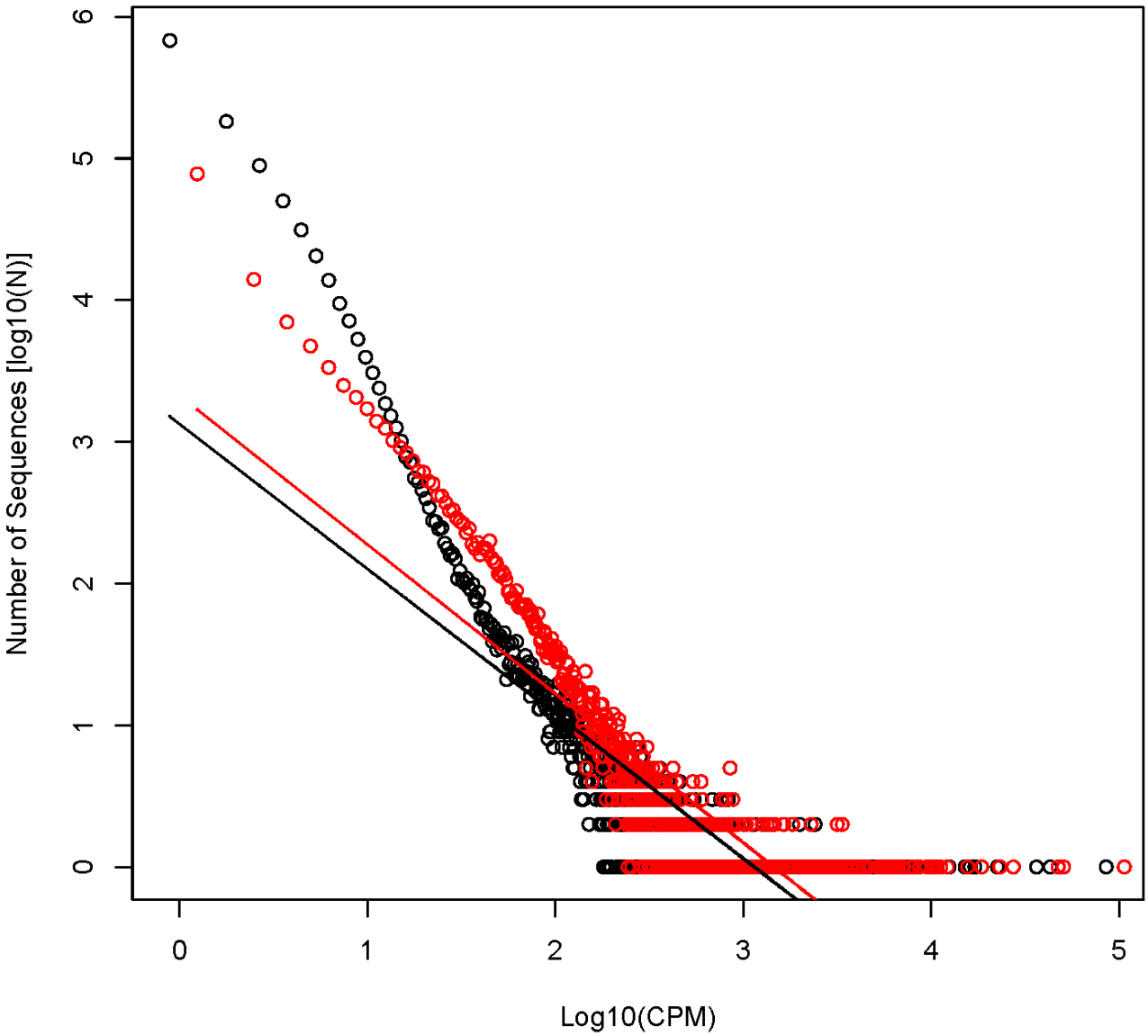


Fig. 2

Fig. 3

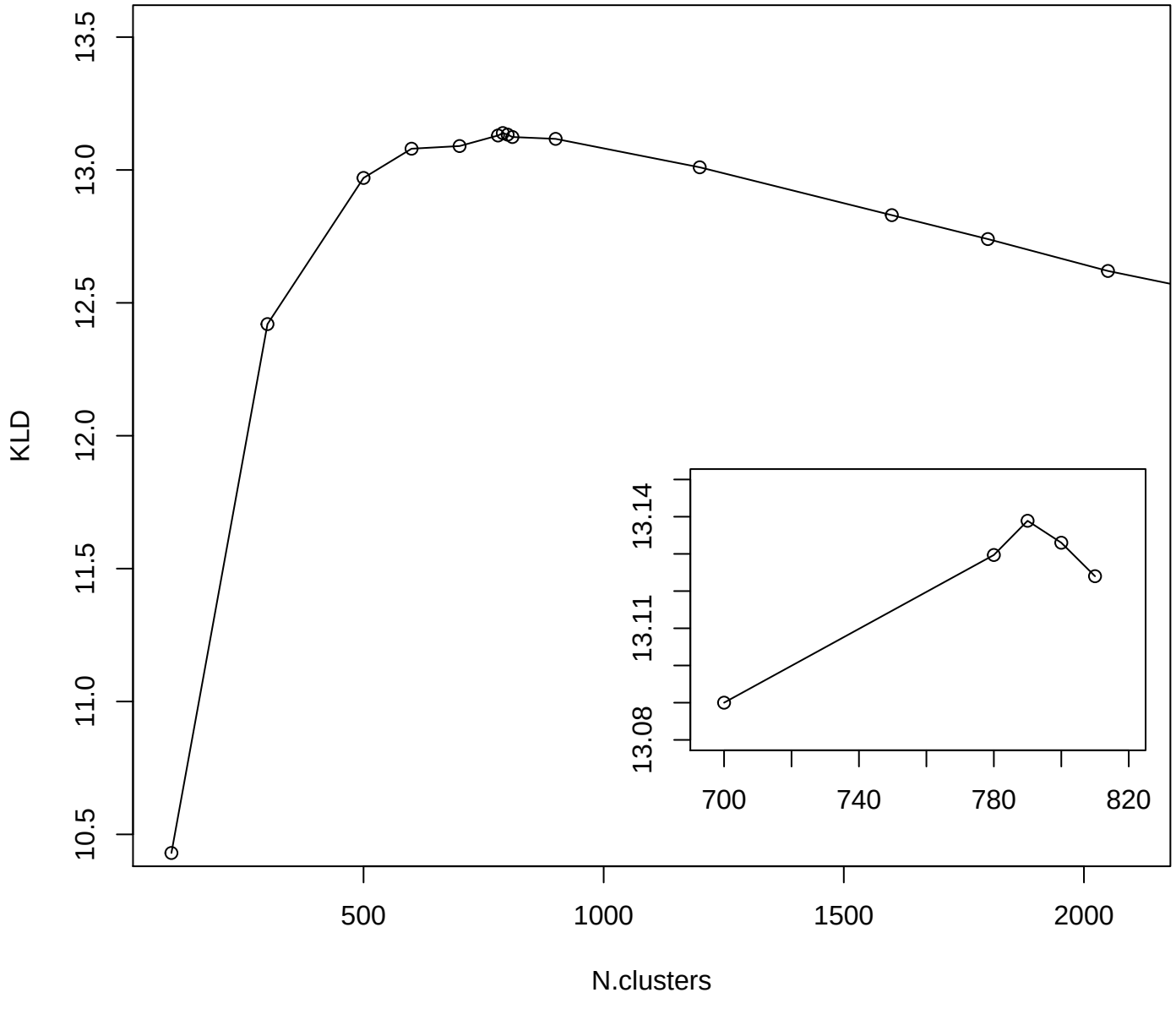

KLD
10.5
11.0
11.5
12.0
12.5
13.0
13.5
500
1000
1500
2000
N.clusters
13.08
13.11
13.14
700
740
780
820

Fig. 4

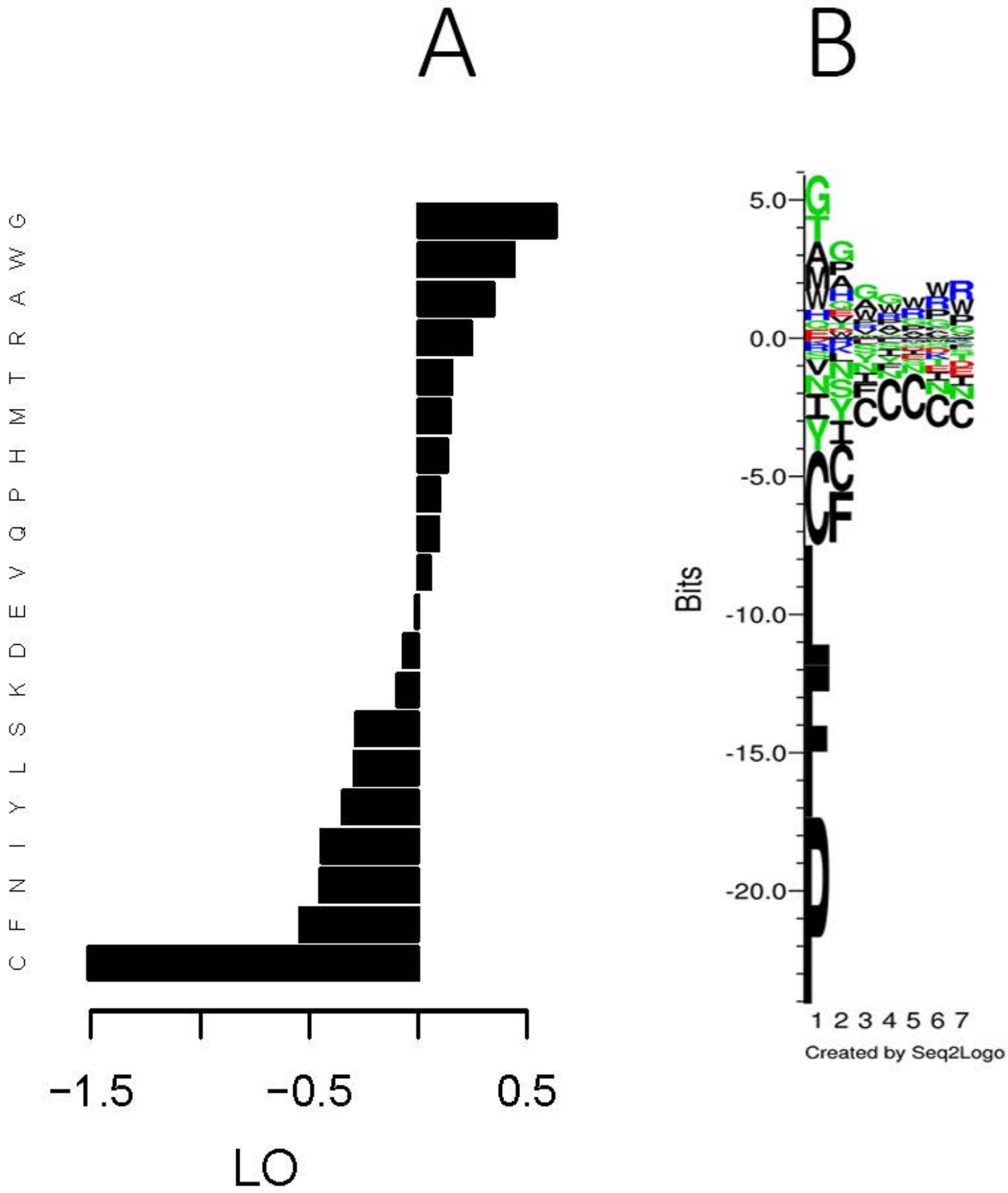

A
B
C F N I Y L S K D E V Q P H M T R A W G
−1.5
−0.5
0.5
LO
5.0
0.0
-5.0
-10.0
-15.0
-20.0
Bits
1 2 3 4 5 6 7
Created by Seq2Logo

Fig. 5

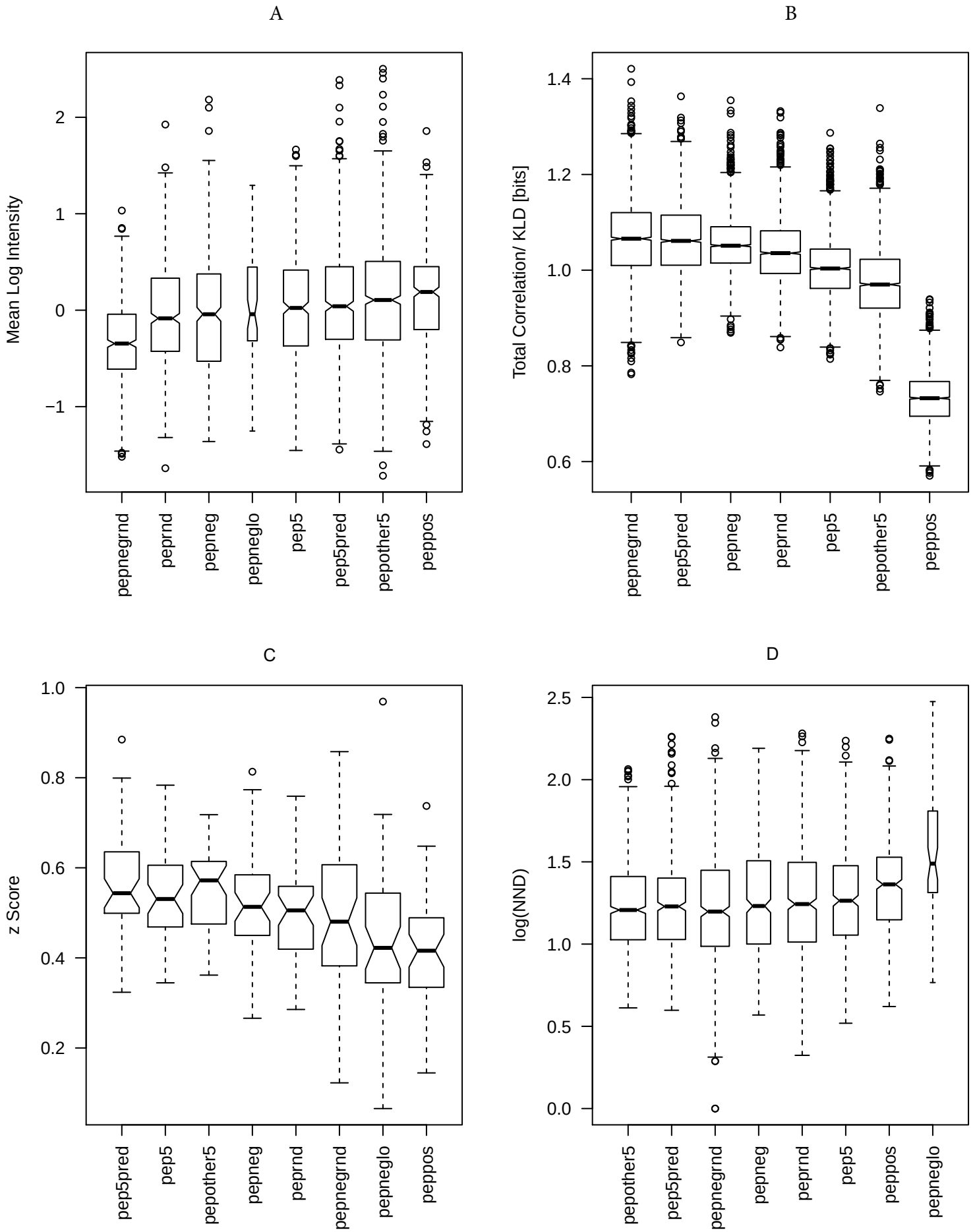
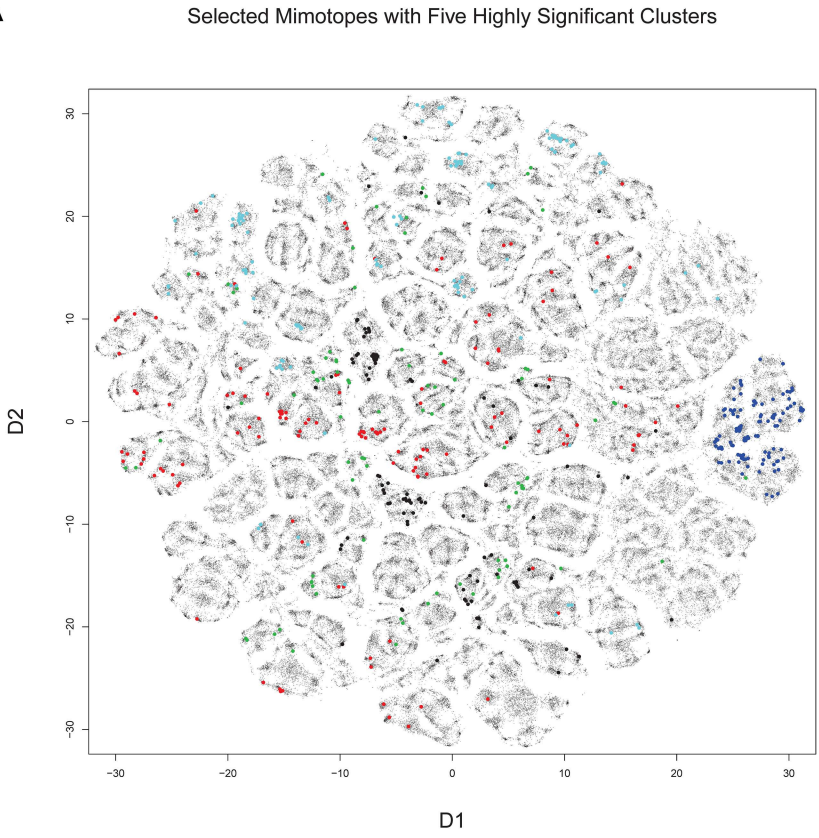
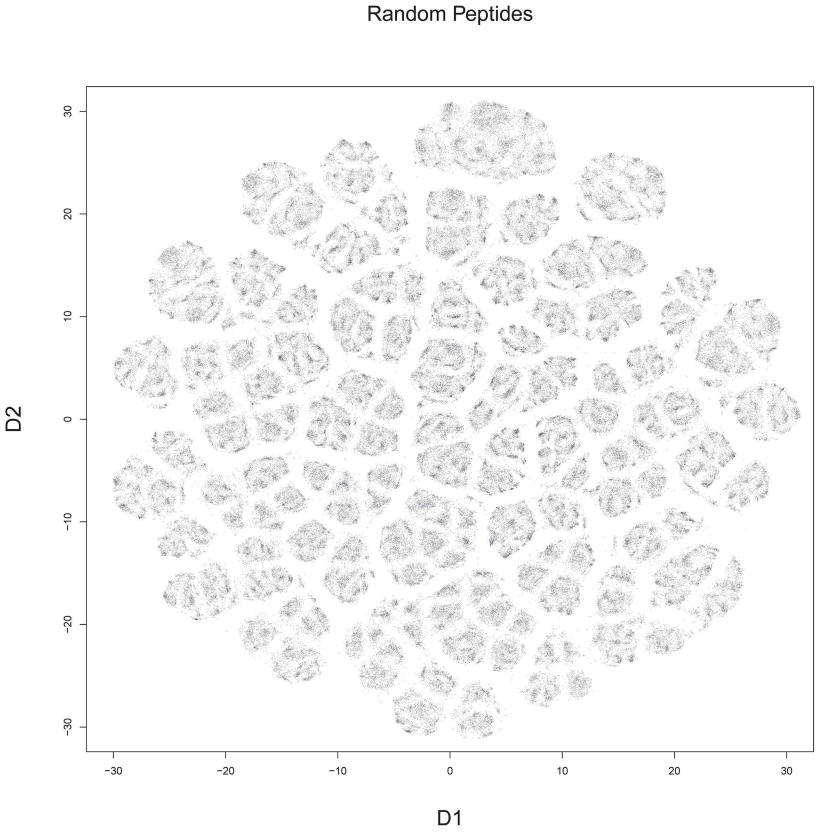

A
Mean Log Intensity
2
1
0
−1
pepnegrnd
peprnd
pepneg
pepneglo
pep5
pep5pred
pepother5
peppos
B
Total Correlation/ KLD [bits]
1.4
1.2
1.0
0.8
0.6
pepnegrnd
pep5pred
pepneg
peprnd
pep5
pepother5
peppos
C
z Score
1.0
0.8
0.6
0.4
0.2
pep5pred
pep5
pepother5
pepneg
peprnd
pepnegrnd
pepneglo
peppos
D
log(NND)
2.5
2.0
1.5
1.0
0.5
0.0
pepother5
pep5pred
pepnegrnd
pepneg
peprnd
pep5
peppos
pepneglo

Fig. 6

A


Selected Mimotopes with Five Highly Significant Clusters
D2
D1
30
20
10
0
−10
−20
−30

B


Random Peptides
D2
D1
30
20
10
0
−10
−20
−30

Selected Mimotopes with the Optimized Library
D2
D1
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

Fig. 7

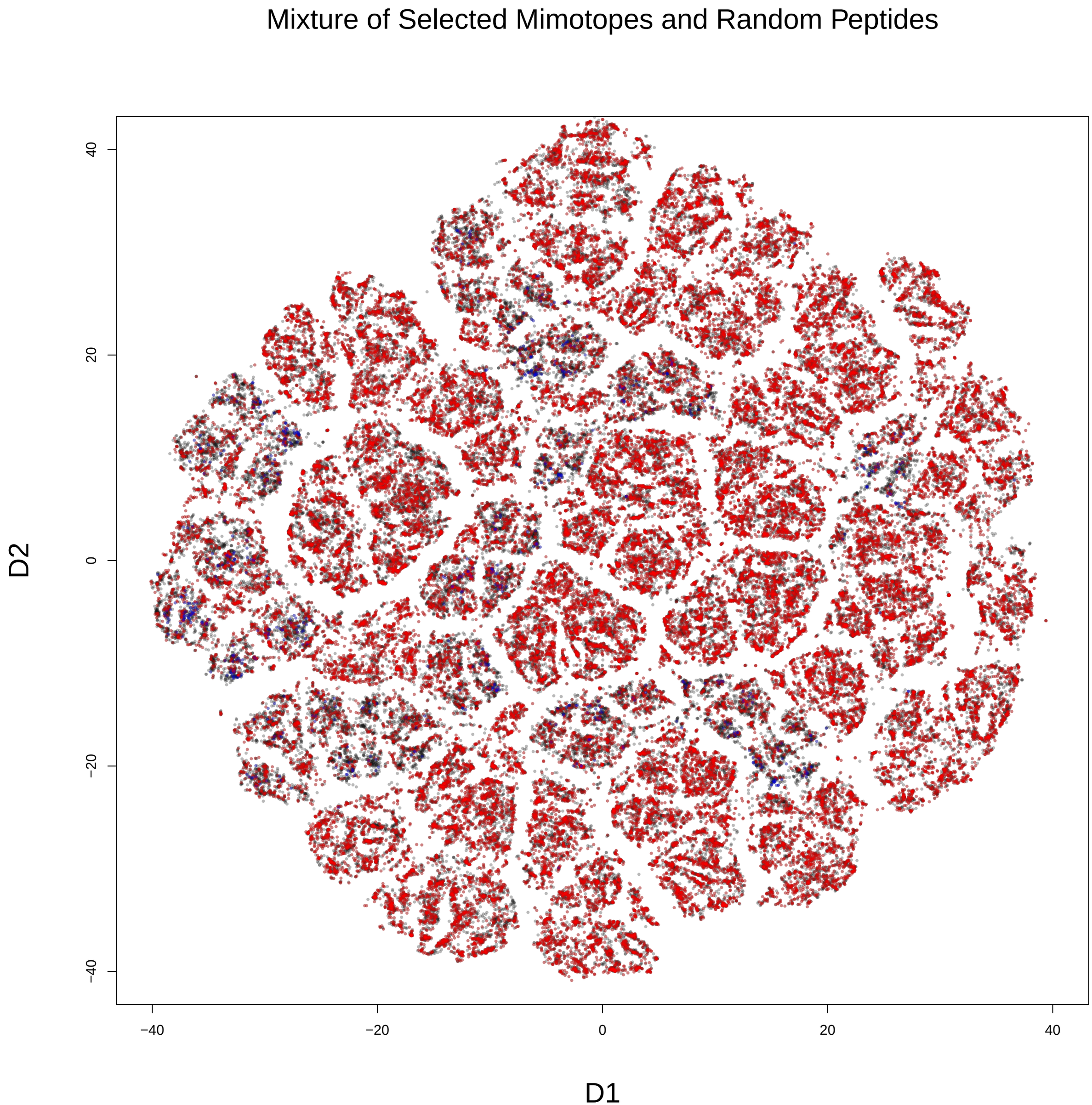
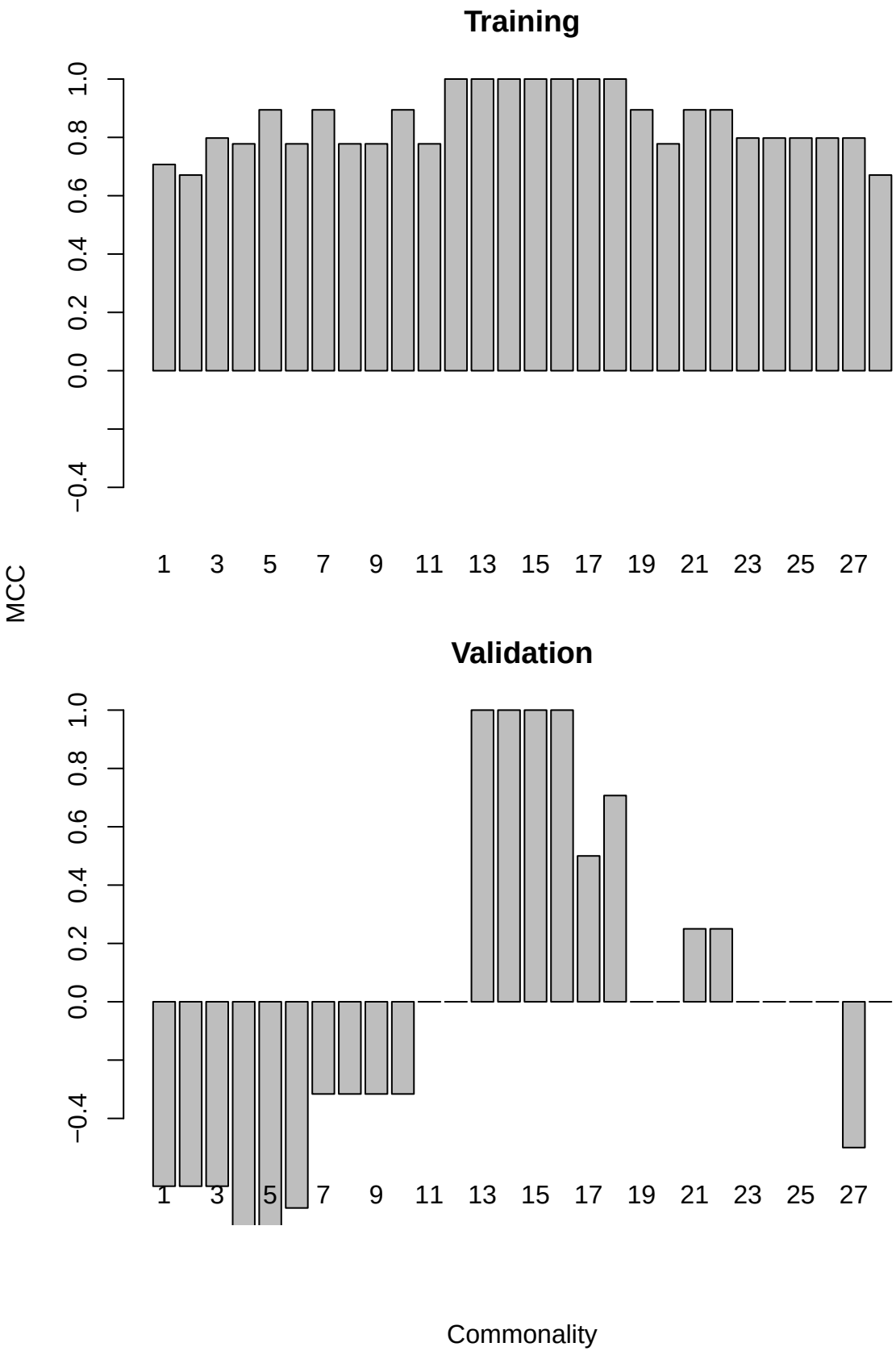

Mixture of Selected Mimotopes and Random Peptides
D2
D1
40
20
0
−20
−40
−40
−20
0
20
40

Fig. 8

Fig. 9


Training
1.0
0.8
0.6
0.4
0.2
0.0
−0.4
1 3 5 7 9 11 13 15 17 19 21 23 25 27
MCC
Validation
1.0
0.8
0.6
0.4
0.2
0.0
−0.4
1 3 5 7 9 11 13 15 17 19 21 23 25 27
Commonality

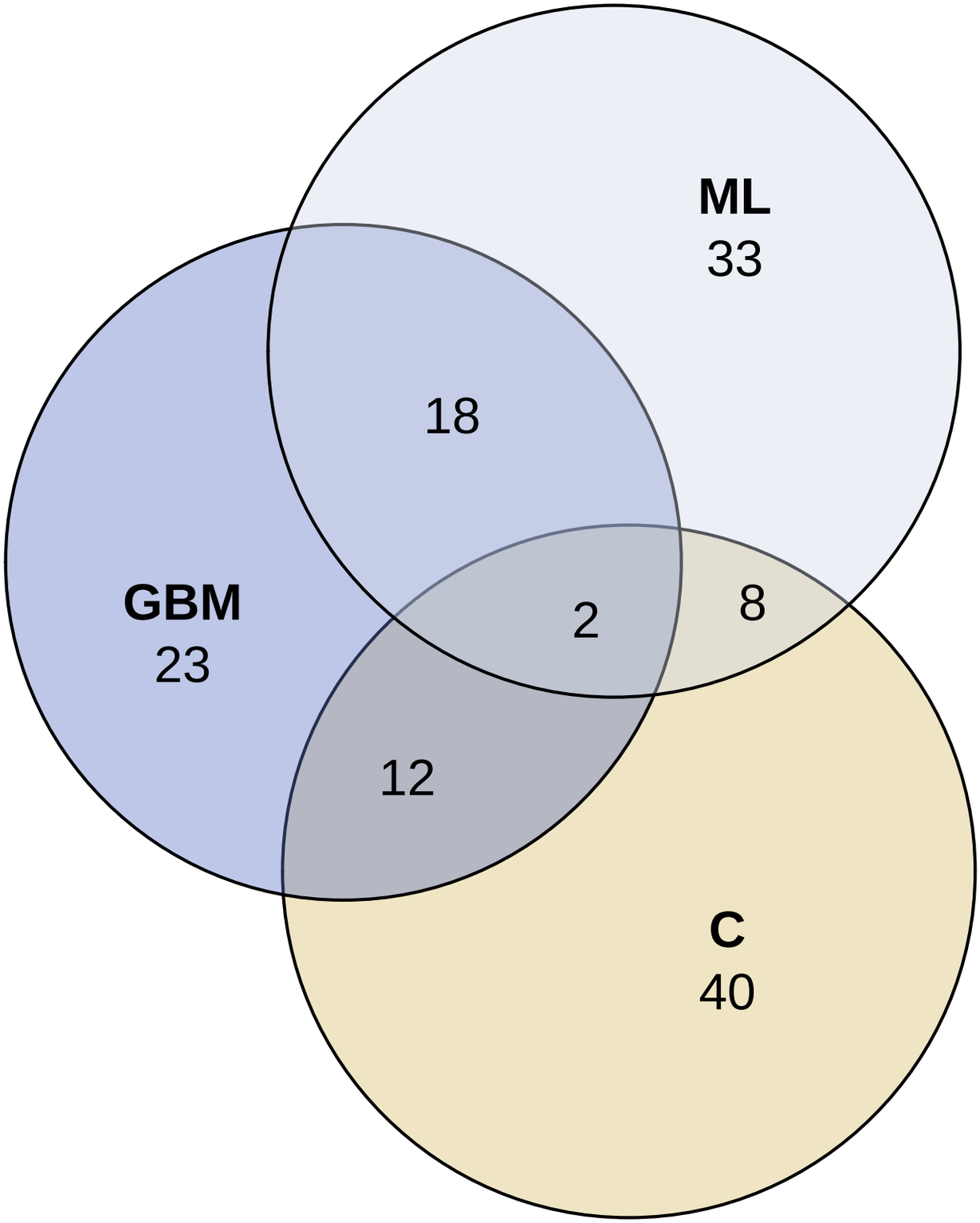


Fig. 10

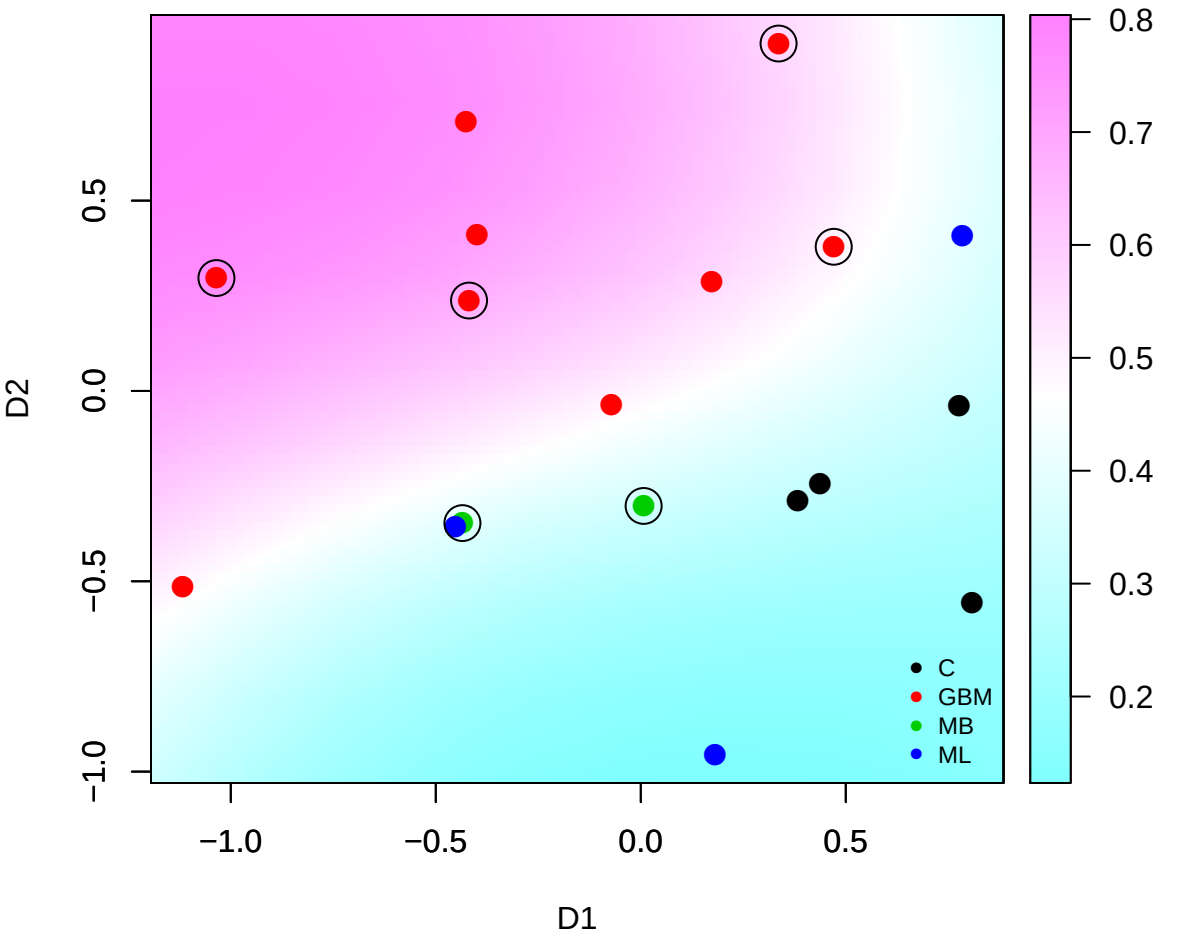



Fig. 11