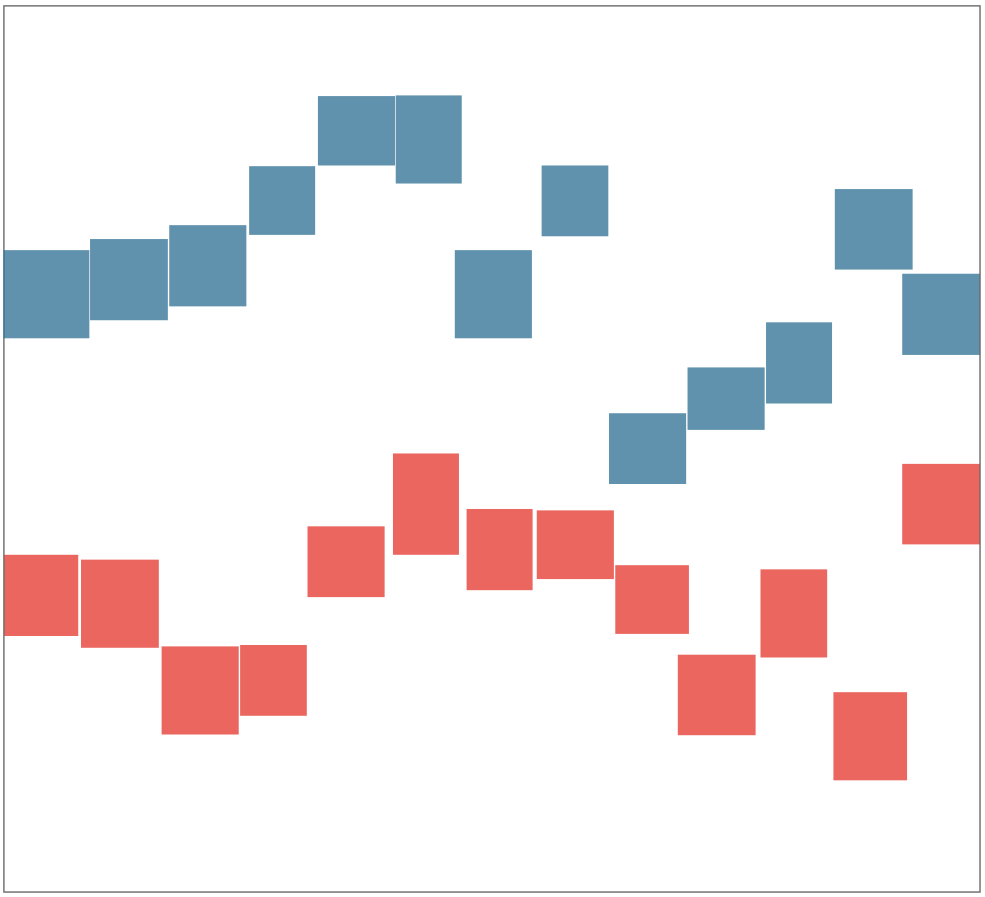
# Data mining on the B cell receptor and antibody repertoire

Masterproef ingediend tot het bekomen van de graad van Master in de Biologie, afstudeerrichting cel- en systeembiologie

Mattijs Beirinckx



Promotor: Dr. Pieter Meysman

Coach: Sofie Gielis

Faculteit Wetenschappen

Departement Biologie

Academiejaar 2018-2019

University of Antwerp

**ABSTRACT**

The adaptive immune system aims to recognize foreign antigens through specialized receptors on the surfaces of B and T cells. As a result of genetic recombination and diversification mechanisms, these immune receptor repertoires show incredible diversity and specificity. The rapid developments in high-throughput sequencing (HTS) of these repertoires are revolutionizing our understanding of lymphocyte biology and have wide implications for the development of preventive and therapeutic measures against infectious and autoimmune diseases as well as cancer. In one part of this study we apply machine learning algorithms on features derived from the amino acid sequence of the antibody heavy chain complementary determining region 3 (HCDR3) to predict antibody-epitope interactions for chemically distinct antigens. We also use an unsupervised clustering technique on a dataset of HCDR3 sequences manually collected from the literature to gain insights into how similar HCDR3s are within and between antigen specific repertoires as defined by a Levenshtein distance metric (also known as the edit distance). Our classifier initially showed poor predictive quality to distinguish antibodies specific to epitopes from different types of antigen. Predictions improved however with the addition of a larger distinct set of antigen specific HCDR3s underlining the importance of high quality data. Unsupervised clustering of antigen specific HCDR3s showed that B cell responses to one antigen show substantial diversity and that different antigen specific repertoires can share similar CDR3 sequences. This clustering approach can be used to investigate and extract relevant information from different immunosequencing datasets and the results are an initiative for further research.

**SAMENVATTING**

Ons adaptieve immuunsysteem probeert vreemde antigenen te herkennen met behulp van gespecialiseerde receptoren op B- en T-cellen. Als een gevolg van genetische recombinatie en diversificatie mechanismen vertonen deze immuunreceptoren een ongelofelijke diversiteit en specificiteit. De snelle ontwikkelingen in *high-throughput sequencing (HTS)* van deze repertoires leveren een enorme bijdrage aan de kennis van het immuunsysteem en hebben implicaties voor de ontwikkeling van preventieve en therapeutische middelen in het bestrijden van infectie- en auto-immuunziekten alsook kanker. In het eerste deel van deze studie gebruiken we *machine learning* algoritmen om via eigenschappen van aminozuren in de *complementary determining region 3* (HCDR3) van de zware keten in antilichamen te voorspellen of antilichamen zullen binden met epitopen van chemisch niet verwante antigenen. We gebruiken ook een clusteringtechniek op een dataset van HCDR3s die we verzamelden uit de literatuur om inzage te krijgen in hoeverre HCDR3s op elkaar gelijken zowel binnen eenzelfde als tussen verschillende antigen specifieke repertoires gedefinieerd door de Levenshtein afstand. Onze modellen slaagden er eerst niet in om met voldoende nauwkeurigheid antilichamen te onderscheiden die specifiek zijn voor een bepaald type antigen. Desalniettemin nam de nauwkeurigheid toe na het toevoegen van een grotere meer afgetekende set HCDR3s wat het belang voor data van een hoge kwaliteit onderstreept. Clustering van antigen specifieke HCDR3s toont dat een B-cel respons tegen een antigen een grote diversiteit vertoont en dat verschillende antigen specifieke repertoires zeer gelijkende CDR3 sequentie kunne delen. Deze clusteringstechniek kan gebruikt worden om *immunosequencing* datasets te bestuderen en er relevante informatie uit te onttrekken en geeft een aanzet voor verder onderzoek.

**ABSTRACT in Layman's terms**

Our immune system depends on B and T cells to recognize parts of pathogens like viruses, bacteria and parasites in order to fight them and eradicate them from our bodies. The parts are recognized by specialized receptors on the B and T cells in a similar way as a key fits a lock. In this analogy the key is the pathogen and the lock is the receptor on the B or T cell. Our body is able to produce enormous amounts of different locks that each require their own key to 'open' i.e. to activate a response and get rid of the key's source (the pathogen). By using the newest technologies, it is possible to look inside the mechanism of the locks, in a similar way a locksmith would do, and make predictions of what key fits that lock. In this study we focussed on B cells and used a computer to make difficult calculations that allowed us to predict if a certain key fits a certain lock. Unfortunately, the calculations of our computer did not succeed in making accurate predictions for key and lock pairings. We also collected detailed information on a series of known key-lock pairings to investigate if similar locks can be opened by similar keys or different keys. It turns out that similar locks can be opened by keys that resemble each other to a certain degree. Out of the analogy this means that certain pathogens, like a family of closely related Viruses are detected by the same and similar locks. Even though not all our results were positive, our findings may be useful in the further understanding of the key-lock mechanism in human immunity and its applications in human biology and medicine.

# Table of contents

## List of abbreviations

| | |
|---|---|
| Ab | Antibody |
| AID | Activation-Induced cytidine Deaminase |
| APC | Antigen Presenting Cells |
| AUC | Area Under the Curve |
| AUROC | Area Under the Receiver Operating Characteristic |
| BCR | B cell receptor |
| BER | Base Excision Repair |
| CD4 | Cluster of Differentiation 4 |
| CD8 | Cluster of Differentiation 8 |
| cDNA | complementary Deoxyribonucleic Acid |
| CDR | Complementary Determining region |
| CH | Constant heavy chain region |
| CL | Constant light chain region |
| CSR | Class Switch Recombination |
| CSV | Comma Seperated Values |
| dbGAP | database of Genotypes and Phenotypes |
| DBSCAN | Density-based spatial clustering of applications with noise |
| DNA | Deoxyribonucleic acid |
| ELISA | Enzyme-Linked Immunosorbent Assay |
| EMBL-EBI | European Molecular Biology Laboratory - European Bioinformatics Institute |
| Fab | Fragment antigen-binding region |
| Fc | Fragment crystallizable region |
| FN | False Negatives |
| FP | False Positives |
| FPR | False Positive Rate |
| FR | Framework Region |
| Fv | variable Fragment |
| H Chain | Heavy Chain |
| HCDR3 | Heavy chain CDR3 |

| | |
|---|---|
| HIV | Human Immunodeficiency Virus |
| HTS | High-throughput sequencing |
| IEDB | Immune Epitope Database |
| Ig | Immunoglobulin |
| IGH | Immunoglobulin Heavy Chain |
| IGL | Immunoglobulin Light Chain |
| IMGT | ImMunoGeneTics (information system) |
| kNN | k Nearest Neighbours |
| L Chain | Light Chain |
| LN | Lymph Nodes. |
| LPS | Lipopolysaccharide |
| MHC | Major Histocompatibility Complex |
| MMR | Mismatch Repair Mechanisms |
| mRNA | messenger Ribonucleic Acid |
| MSTA | Multiple Structure Alignment |
| NCBI | National Center for Biotechnology Information |
| PAM | Point Accepted Mutation |
| PBMC | Peripheral Blood Mononuclear Cell |
| PCR: | Polymerase Chain Reaction |
| PDB | Protein Data Bank |
| PIRD | Pan immune repertoire database |
| PR | Precision-recall curve |
| PRAUC | Precision Recall Area Under the Curve |
| RABA | Radioactive antigen-binding assay (RABA |
| RAG | Recombination-Activating-Genes |
| REP-seq | Repertoire sequencing |
| RF | Random Forest |
| ROC | Receiver Operating Characteristic |
| RSV | Respiratory Syncytial Virus |
| RT | Reverse Transcriptase |
| SHM | Somatic Hypermutation |
| SLN | Secondary Lymphoid Nodes |

| | |
|---|---|
| SLO | Secondary Lymphoid Organs |
| SN | Sensitivity |
| TBAdb | T cell, B cell and Antibody database |
| TCR | T Cell Receptor |
| TN | True Negatives |
| TP | True Positives |
| TPR | True Positive Rate |
| TT | Tetanus Toxoid |
| V(D)J | V D J gene recombination |
| VH | Variable Heavy Chain |
| VL | Variable Light Chain |

# 1. INTRODUCTION

## 1.1 The immune system in short

Vertebrates are constantly exposed to and threatened by potentially harmful microorganisms and have evolved systems of immune defence to eliminate infective pathogens in the body. The mammalian immune system is comprised of two branches: innate and adaptive immunity. The innate immune system is considered the first line of defence and is mediated by phagocytes including macrophages and dendritic cells. Adaptive immunity is involved in elimination of pathogens in the later phases of infection and generation of immunological memory.

The adaptive immune response involves two types of lymphocytes, T cells and B cells, which carry receptors on their surfaces that are able to recognize structural sequence motifs called epitopes within molecules usually related to pathogens called antigens. B cell can also express their receptor in a secreted form called antibody or immunoglobulin. Initially, specificity for epitopes develops by clonal selection from a vast repertoire of lymphocytes bearing antigen receptors which are generated via germline gene rearrangements and additional random somatic mutations in the case of B cells. In theory, this receptor diversity is able to recognize virtually any antigen. During the clonal selection process lymphocytes normally survive only when their receptor is able to distinguish self from non-self structures. However, as a result of a defective elimination or control mechanisms, self-reactive lymphocyte receptors can persist and trigger immune responses directed against the individual in what is referred to as autoimmunity.
Lymphocytes are activated when they are presented with an epitope for which they show high affinity. Consequently, they proliferate to create a clonal population sharing the same receptor making detection of the epitope more likely and increasing the chance of eradication of the pathogen.

The adaptive immune response has a humoral and cell-mediated component which allows detection of circulating as well as cell invading intracellular pathogens. CD8+ T cells recognize antigen that has been processed intracellularly and their fragments

presented by major histocompatibility complex I (MHC) on the surface of cells. CD4+ T cells on the other hand recognize antigen originating from the extracellular environment presented on MHC II molecules on the surfaces of specialized antigen presenting cells (APC). B cells themselves are also APCs. Recognition of epitopes by T cells triggers a cascade of downstream immunological events in order to eradicate the pathogen from which the epitope was derived. This includes activation of B cells. Upon activation, B cells can differentiate into plasma cells which can secrete large amounts of antibody. These antibodies circulate the blood and lymphatic system where they mark pathogens for ingestion by phagocytes or activate effector cells that will ultimately result in destruction of the invading microbe. Activation of B and T cells leads to the formation of immunological memory where receptor populations that were able to successfully eradicate a pathogen become long lived and are able to initiate a similar response once the same pathogen is re-encountered. In what follows we dive into greater details of B cell immunity, the subject of this thesis.

### 1.2 BCR and antibody structure

The BCR is composed of a membrane immunoglobulin (Ig) domain and a short intracellular domain. The immunoglobulin domain is linked with a Igα/Igβ heterodimer which is responsible for signalization. Antibodies are the soluble form of the BCR and require binding to antibody receptors in order to mediate their function. Antibodies have two distinct functions: one is to bind specifically to their target antigens, the other is to elicit an immune response against the bound antigen by recruiting other cells and molecules. The antibody-antigen association involves various non-covalent interactions between the epitope (binding site on the antigen) and the paratope (binding site on the antibody).

Antibodies consist of four polypeptide chains (figure 1): two identical heavy (H) chains (μ, α, γ, δ or ε) and two identical light (L) chains (κ or γ). The light and heavy chains are linked by disulphide bonds to form the arms of a Y-shaped structure, each arm is known as a fragment antigen-binding region (Fab) (Edelman and Benacerraf 1962). Each Fab is composed of two variable domains (VH and VL respectively) and two constant domains (CH1 and CL).

In the pairing of light and heavy chains, the two variable domains dimerize to form the variable fragment (Fv) which contains the antigen binding site. Within each of the variable fragments lie six hypervariable loops; three in the light chain (L1, L2 and L3) and three in the heavy chain (H1, H2 and H3), which are embedded in a conserved framework region (FR). The six hypervariable loops within the variable domains are commonly termed complementarity determining regions (CDRs) which are thought of to be the most important in defining antibody-antigen specificity and affinity (figure 3, Kuroda et al. 2008, Barrios et al. 2004, Sela-culang et al. 2013). The light and heavy variable domains are folded in such a way that brings the hypervariable domains together in near proximity to create the antigen binding site. Finally, the fragment crystallizable region (Fc) consisting of two domains on the heavy chain (CH2 and CH3) is responsible for mediating the biological activity of the antibody molecule. Connection of the CH2 domain in the Fc region with the CH1 domain of the Fab region is facilitated by the hinge region which characteristics define the degree of flexibility between the two Fab regions and which differs between isotypes (Janda et al. 2016).

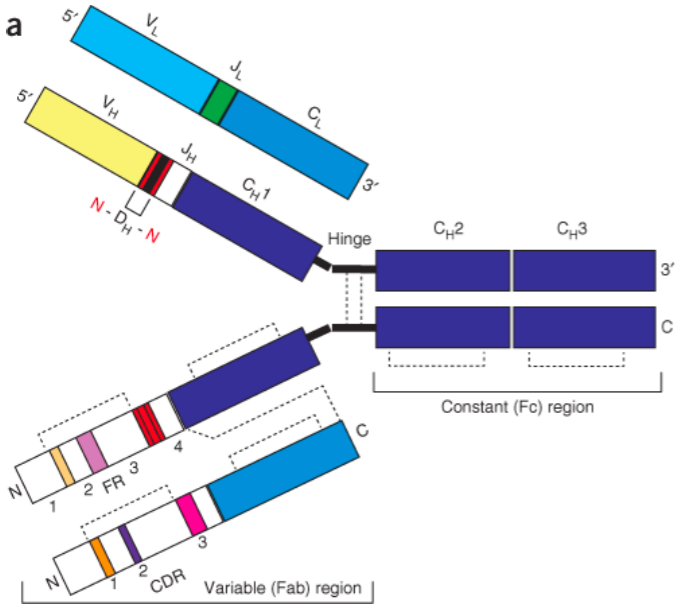



**Figure 1: Antibody structure**: Schematic of the predominant antibody IgG. In the top chains the domains encoded from V, D, J and C gene segments are indicated. In the bottom chains framework (FR) and complementary determining regions (CDRs) are indicated. The dashed lines represent disulfide bonds (figure from Georgiou et al. 2014)

## 1.3 B cells and diversity of the BCR and antibody repertoire

B cells or B lymphocytes are white blood cells which function in the humoral component of the adaptive immune system. In mammals, B cells are generated in the bone marrow and migrate to the spleen where they differentiate into naïve B cells. Subsequently they re-enter circulation and move on to the secondary lymphoid organs (SLO).

### *1.3.1 Gene rearrangements in the BCR locus*

In the bone marrow and before antigen encounter, BCRs are produced by a developmentally ordered series of gene rearrangements (figure 2). These events are mediated through recombination-activating-genes (RAG) and are the basis for an enormous antibody diversity, theoretically at least $10^{16}$ possible variants (Briney et al. 2019). Before B cells encounter antigen (naïve B cells) the total BCR diversity is the result of allelic diversity in BCR gene segments, combinatorial diversity introduced during somatic recombination, junctional diversity caused by imprecisions of the recombination process, pairing of IgH and IgL polypeptide chains and receptor editing.

B cells can recognize a wide variety of antigen ranging from proteins, lipids, carbohydrates to drugs and antibiotics. For a humoral immune response to initiate, rare antigen-reactive naïve B cells have to come in to contact with their cognate antigen. These encounters predominantly occur in lymph nodes (LNs). If no antigen is encountered, B cells eventually exit lymphoid tissues and travel to other lymphoids to continue their surveillance program (Cyster et al. 2012).

Once antigen is encountered, signalling via the B cell receptor (BCR) initiates B cell activation. Depending on what antigen is encountered, B cells may or may not need help of T cells to get activated. Antigens that do require T help are referred to as T-dependent, whereas antigens that don't are T-independent. In the case of T-dependent antigens, BCR triggering is followed by internalization of the BCR and its bound antigen. B cells can process and present the antigen in association with MHC

class II molecules, thereby recruiting specific CD4+ T-cell help which will determine their next differentiation state.

For the B cell response to progress, proliferation has to occur necessary for the generation of a clonal population of the daughter cell. During this proliferation the second diversification mechanism comes in to play (figure 2). These include somatic hypermutations of the variable regions and class switch recombination. Both are mediated by activation-induced cytidine deaminase (AID).

### *1.3.2 Somatic hypermutation*

During somatic hypermutation, AID deaminates cytidine residues in transcribed V(D)J and VJ regions. Base excision repair (BER) and mismatch repair (MMR) mechanisms then convert cytidine lesions to point mutations and a small frequency of insertions and deletions (Hwang et al. 2015). Higher activity of AID in complementary determining regions (CDRs) is partially explained by the enrichment of a consensus AID motif in these regions (Yeap et al. 2015). B cells carrying beneficial mutations are then selected at the expense of their neighbours for their continued participation in the response as a result of having an increased capacity to capture antigens (clonal selection). This selection leads to B cells producing antibodies with increased affinity for an antigen during the course of the immune response (affinity maturation). As a result of SHM, responses to a specific antigen produce a diverse population of antigen-specific B cells and are not limited to the cells producing the antibodies with the highest affinity (Finney et al. 2018). This is crucial in achieving protection from rapidly evolving pathogens that might quickly escape recognition (Baumgarth et al. 2013).

### *1.3.3 Class switch recombination*

After activation, B cells also may undergo class switch recombination (CSR) where the constant regions of a BCR are changed. This switches the isotype from IgM and IgD on the naïve B cell to IgG, IgA, or IgE on fully differentiated cells. Antibody isotypes differ in several respects (Janda et al. 2016). These include the number and location of interchain disulphide bonds, number of attached oligosaccharides, number of C domains and the length of the hinge region. Different functional roles are ascribed to

antibodies with different isotypes, many of them originating from variations in binding to Fc receptors, but they also have been shown to influence antigen specificity and affinity (Dodev et al. 2015).

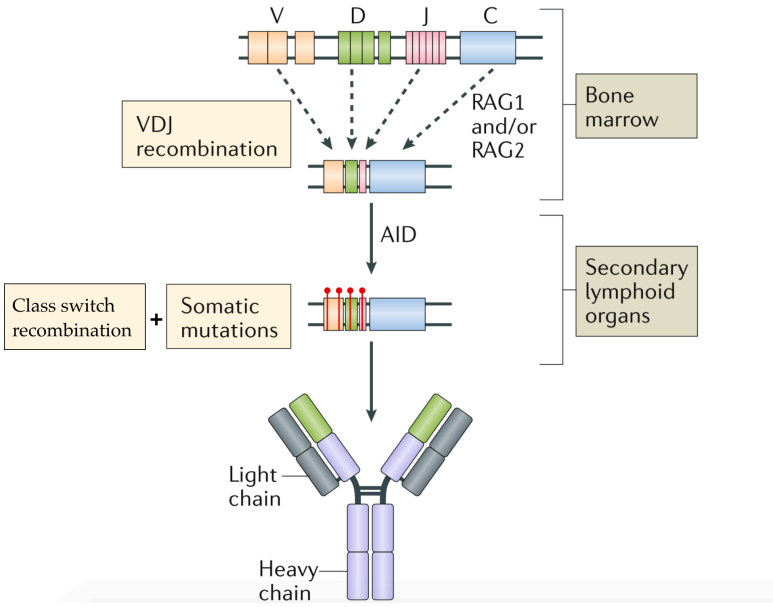



**Figure 2: Diversification mechanisms of the BCR and antibody repertoire**
The diversity of antibody repertoires is primarily generated by two genetic mechanisms: Before antigen encounter (in the bone marrow) the process of V (variable), D (diversity) and J (joining) (VDJ) gene recombination through recombination-activating-genes (RAG) is at the root of the variability of the immune repertoire by introducing combinatorial and junctional diversity in the variable regions of the antibody. After contact with cognate antigen, which usually occurs in secondary lymph organs (SLO) and in the presence of helper T cells, somatic point mutations are introduced throughout the heavy variable and light variable gene regions through activation induced cytidine deaminase (AID). At the same time BCR may undergo class switch recombination (CSR) and change their isotype. (Figure adapted from Kanyavuz et al. 2019)

### 1.4 CDR3

The CDR3 region is been recognized as the most variable region within an antibody. It is situated at and includes the junction of the V, D and J genes. Furthermore, because of the potential diversity of BCRs it is highly improbable that BCRs converge on the same CDR3 sequence, making each CDR3 sequence unique for a B cell clone. Together with CDR1 and CDR2 it is the region in closest proximity to the antigen and thus defines antibody-antigen specificity to a great extent (figure 3). Interestingly, it turned out that residues that are directly involved in the interaction with the antigen are in general the most variable ones (Padlan et al. 1995). A major focus in analysing the structural basis for antigen recognition has been in identifying the exact boundaries of the CDRs in a given antibody. Several numbering schemes have been published (but also withdrawn) reflecting the interest in this subject in antigen recognition (Al-Lazikani et al. 1997, Honegger and Pluckthun. 2001). Nonetheless, different CDR identification methods often identify radically different sequence stretches which can make it hard to compare findings between studies.

While CDRs are used to determine paratopes, not all residues within the CDR contact the antigen (figure 3). At certain positions within the CDR3 none or only a small percentage of the antibodies contact the antigen. Residues within the CDRs could also be important for maintaining structural conformations of the CDRs and not necessarily for recognition of the antigen (MacCallum et al. 1996).

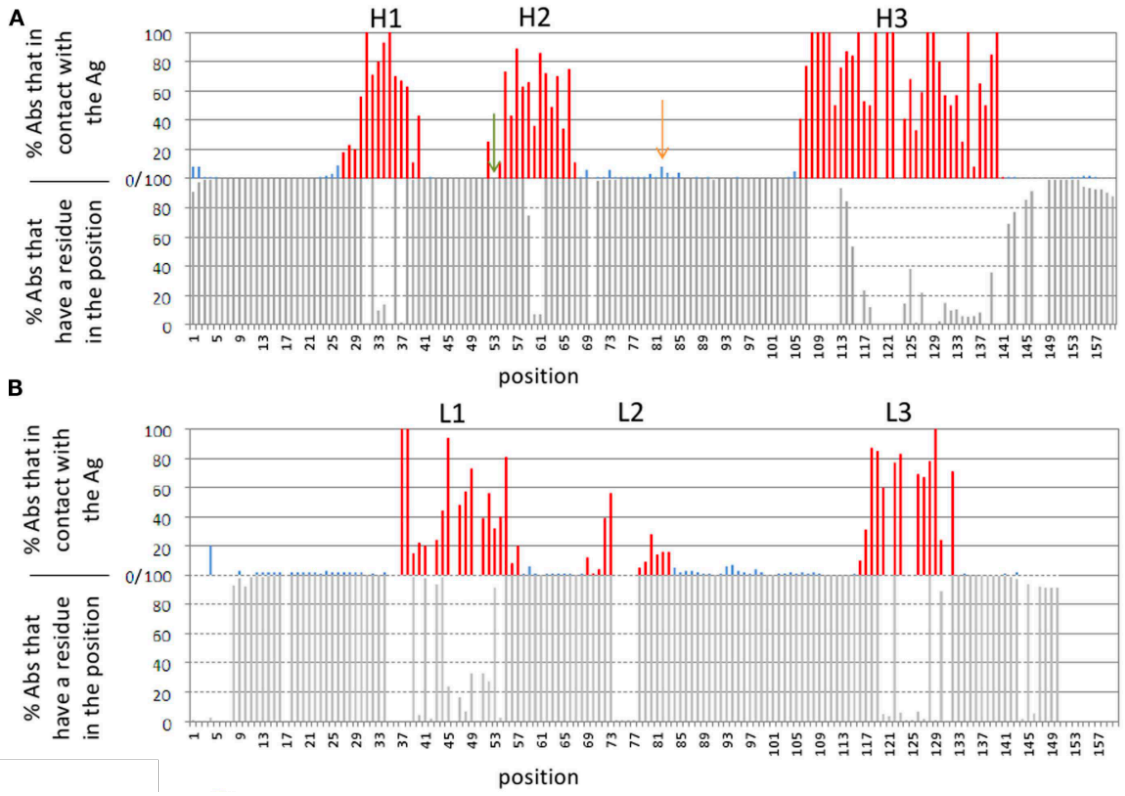



**Figure 3: Antibody residues that contact the antigen**. The lower graphs show the percentage of antibodies with known 3D structure that have a residue in a given position (i.e., in other Abs there is a gap in the MSTA in that position). The upper graphs show the percentage of Abs that contact the Ag out of those Abs that have a residue in that position. (A) Depicts the heavy chain and (B) depicts the light chain. (Adapted from Sela-Culang 2013)

### 1.5 High throughput sequencing of immune repertoires

Despite the huge theoretic diversity of the antibody repertoire, only a subset of this diversity is expected to be physiologically present in the repertoire of an individual. Furthermore, the actual repertoires differ from the theoretical figures due to various limitations and biases (Briney et al. 2019, Khass et al. 2016 ). High-throughput DNA sequencing technologies, either using germline DNA or cDNA , to determine the BCR and antibody repertoire encoded by B cells (Ig-seq, BCR-seq, Rep-seq) have been advancing at a rapid pace and are providing valuable information in our understanding of the humoral immune response (Friedensohn et al. 2017, Wardemann and Busse 2017).

### *1.5.1 Immunosequencing strategies*

Immunosequencing is a multiplex PCR-based method that allows bulk lymphocyte sequencing by amplifying the rearranged BCR sequences for a BCR locus. Initially, due to limitations in read lengths, sequencing was focused on the analysis of the V(D)J joint in heavy chain CDR3s which harbour the highest degree of diversity, but current methods provide full length Ig gene information. The multiplex PCR mixture usually contains forward primers in each V segment and reverse primers in each J or C segment and as such allows for amplification by reverse transcription PCR (RT) and sequencing of the recombined Ig genes (figure 4)

This classical setup however suffers from multiple problems which limits the use and the interpretation of the data obtained and methods should be carefully considered depending on the study goal (for a review Wardemann and Busse 2017). Alongside the advancing technologies, development of informatics analysis solutions is helping to extract the most pertinent biological information out of these experiments (Bolotin et al. 2015, Marcou et al. 2018, Greiff et al. 2015).

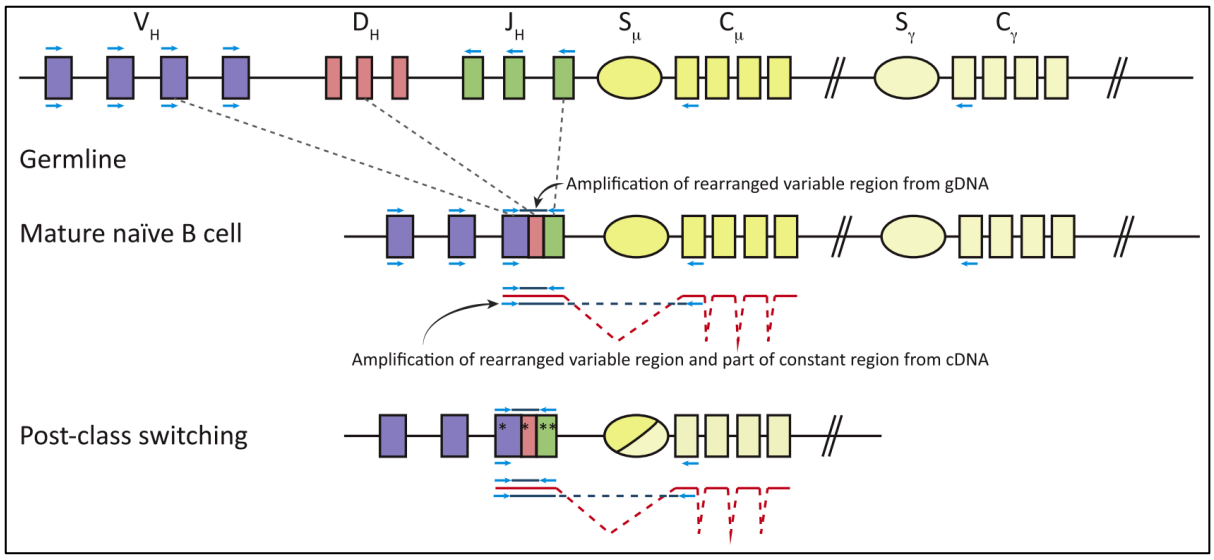



**Figure 4: Classical multiplex PCR repertoire sequencing approach**.
Schematic representation of an IGH locus. The top row shows the locus in its germline configuration. The primers indicated as the blue arrows will not produce amplicons because the distance between binding sites is too large. The middle line represents the locus in a mature naïve B cell, i.e. a B cell where V(D)J rearrangements have occurred. The red line indicates the mature mRNA produced from this locus after excision of introns. The bottom line represents the locus after class switch recombination and somatic hypermutations. V/C primer pairs here are informative regarding the isotype but will include the switched constant region intron. Rearrangements of the light chain loci occur accordingly except for the fact these loci lack D segments and do not undergo CSR (Figure from Wardemann and Busse 2017).

One important consideration to make is the source of the B cells being studied. Most sequencing studies start off with collecting peripheral blood mononuclear cell (PBMC) from blood samples. However, it is estimated that only 2% of all B cells in the human body are present in peripheral blood, compared to 28% in lymph nodes, 23% in the spleen and on mucosal surfaces and 17% in the red bone marrow (Glanville et a. 2009). The antibody repertoire in peripheral B cells therefore could only provide a narrow view of the humoral response to antigen challenge. Also, one study reported adenoid-derived antibodies to show higher binding affinities and neutralization potencies compared to antibodies isolated from peripheral blood (Shehata et al. 2019). These observations should be taken into consideration when drawing conclusions in repertoire studies.

### *1.5.2 Immunosequencing applications*

Through sequencing repertoires of groups that are differently immune challenged, one can gain insights into how much these states are reflected in the immune repertoire and differ between groups. Furthermore, individuals might share B cell clones with the same sequences specific for a given pathogen. Such shared clones are referred to as public clones whereas private clones are those that remain unique to the individual (Greiff et al. 2015). Adaptive immune receptor sequences that are present in an affected group but not in a healthy control can even be used as signatures for the specific condition (De Neuter et al. 2018, Skowera et al. 2015). Furthermore, individuals each carry different alleles and some Ig gene variants have been associated with increased susceptibilities to various conditions (Ray and Hachochen et al. 2015). Immunosequencing will without a doubt provide more knowledge to this field. Sequencing also allows detection of cancerous B cells. Because it is usually a single rogue cell that spawns the entire cancer, the receptor sequences on these cancerous cells can be detected and used as a molecular tag for the cancer.

Next to identification of condition related sequences, repertoires sequencing can be used the follow up on the dynamics of adaptive immune responses. Clonal expansions and contractions and central memory formation can be tracked in vivo over time in response to different immune triggers. This aspect is especially valuable in the design, evaluation and reengineering of vaccination strategies where acquired immunity through immune memory is the desired outcome.

### *1.5.3 Immunosequencing: What does a given sequence do?*

The reason why sequences are related to specific conditions often remain elusive. Evidently, one great interest is identifying the epitope targets of these signature receptor sequences. This is a tedious process that involves several repeated screening and binding experiments in the case where the suspected interacted epitope is actually known which is not necessarily the case. In the latter case for example, one has to scan epitopes using large mutagenesis libraries.

### *1.5.4 Machine Learning to the rescue?*

Alternatively, outside of the wet lab, machine learning methods are being used as a means to translate sequence information into predictions of antigen receptor function. Recently, de Neuter et al. 2018 were able to predict TCR-epitope interactions for linear peptidic epitopes using a random forest machine learning model based on features derived from the CDR3 region in the beta chain of the TCR. Adding to this Meysman et al. 2018 investigated how dissimilar these TCR sequences can be before they no longer bind the same epitope using the unsupervised clustering technique DBSCAN. Based on the findings from the beforementioned studies, Gielis et al. 2018 developed a promising webtool to analyse TCR sequences and predict the likelihood that they target specific epitopes.

## 1.6 What is machine learning?

Machine learning is a collection of data-analytical techniques wherein computers are programmed to learn patterns from data often aiming to develop a predictive model based on statistical associations among features from a given dataset. It provides higher level analysis in complex datasets generating new perspectives and hypothesis. There are two categories of machine learning methods. Unsupervised approaches are used when the labels on the input data are unknown whereas supervised methods are applied when labels are available for the input data.

The input to a supervised machine-learning algorithm typically consists of data, features and targets. Features are the measurements or variables across all samples. They can be used either in their raw form or transformed. Targets are what the machine-learning model aims to predict and can be of various nature including binary responses, categorical labels or continuous values. When the target of the model is a label, category or class, a model can also be referred to as a classifier.

*1.6.1 The basic machine learning workflow*

In its simplest form, the machine learning workflow is to process the input data, extract features, fit and train the underlying model and evaluate the model with new unseen data (figure 5). The dataset on which a machine learning method will be applied holds experimental data with variables corresponding to a certain label or target. First relevant features are included or derived from the variables in the dataset. Next the dataset is split into training and testing subsets (the testing subset recommended to be around 20-30% of the original size of the dataset). The dataset to test the model should match the general properties of the data used to train the model and should be processed using the same pipeline. Then a machine learning method is used to create a model based on the ground truths of the training data (features and corresponding labels). Performance of the model is evaluated using the left out testing subset where the model tries to predict the correct label for the features of these testing instances. The output of the model is then compared to the actual labels of the testing subset.

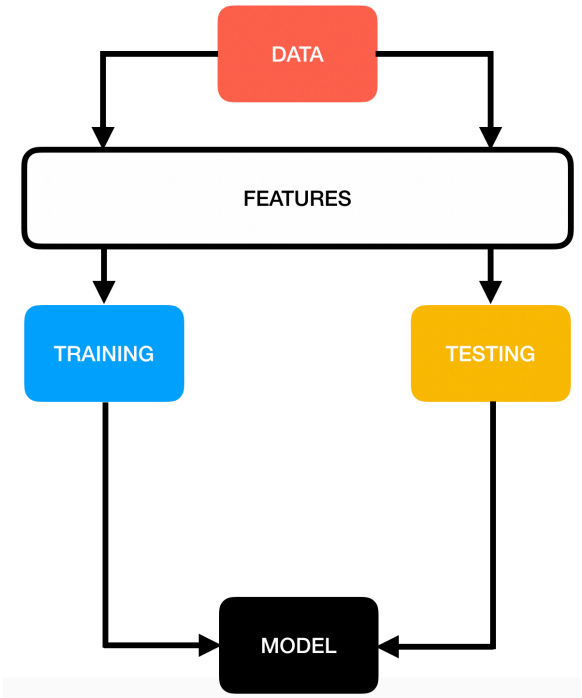



**Figure 5: Schematic representing the basic machine learning workflow.**

### *1.6.2 Evaluation of a machine learning model*

#### *1.6.2.1 Cross validation*

Machine learning models need to be trained, tested on independent data sets to avoid overfitting and assure that the model will generalize unseen data. For smaller dataset a common strategy is to use cross-validation (figure 6). Evaluation with *k* fold cross-validation first randomly splits the dataset in *k* equally sized, non-overlapping subsets. Training and test subsets are created by using a single subset as test set while training the classifier on the remaining subsets. For each fold the model is trained and tested with different subsets. Overall performance for the model is then the average performance on the fold's test subsets over all folds.

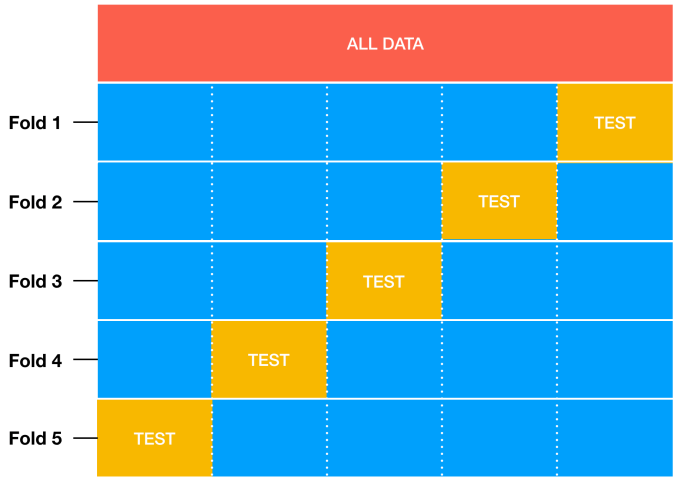



**Figure 6: Splitting of the data in train and test sets during cross validation**. In cross validation the dataset is randomly split in *k* equally sized, non-overlapping subsets. Training and test subsets are created by using a single subset as test set while training the classifier on the remaining subsets. For each fold the model is trained and tested with different subsets. Overall performance for the model is then the average performance on the fold's test subsets over all folds.

*1.6.2.2 Confusion matrix*

In machine learning performance of a model is typically evaluated using the confusion matrix (figure 7). This matrix layout compares the models predicted labels of the test set versus their actual labels. Several metrics can be derived from the matrix and are used to describe the performance of the model. The true positive rate (TPR) or recall is the proportion of actual positives that are correctly classified as positives. Similarly, the False positive rate (FPR) is the proportion of negatives instances that are wrongly classified as positives. Accuracy is the ratio of correctly classified instance on the total number of instances. Precision is defined as the fraction of true positives on all positive predictions, i.e. how many of the predicted positives are true positives.

| | **Actual positive** | **Actual negative** |
|---|---|---|
| **Predicted positive** | True positive | False positive |
| **Predicted negative** | False negative | True negative |

Accuracy = (TP + TN) / (TP + TN + FN + FP)

Precision = TP / (TP + FP)

Recall, TPR, SN = TP / (TP + FN)

FPR = FP / (TN + FP

**Figure 7: The confusion matrix.** The matrix layout places the predicted labels against their actual labels. Several metrics can be derived from the matrix; TP: True positives, TN: True negatives, FP: False positives, FN: False negatives, FPR: False positive rate, TPR: True positive rate, SN: Sensitivity.

*1.6.2.3 ROC curve*

By plotting the true positive rate versus the false positive rate for different cutoff values you become the receiver operating characteristic curve (ROC). This curve displays how well a model can classify binary outcomes. To see how the ROC curve can be interpreted we take a look at an example:

Most models allow output of the probability (also 'scores') of an instance in the training data to belong to a class label (see explanation models). For an example with good separability, plotting the density of probabilities for each class label would look something like the plots in left column of figure 8. The red curve represents probabilities for the positive class, the blue curve those for the negative class. On the top row, when using a high positive cutoff value, the classifier correctly predicts some positive instances, but some are considered as false negatives. On the other hand, no positive predictions are made for negative instances (no false positives). The red dot on the ROC curve on the right represents the TP and FP ratio for the threshold on the density plot on the left. In the middle graphs, for the classical threshold of 0.5, the classifier correctly predicts most of the positive instances, but still some of the positive instances will turn out to be false negatives. For this threshold, some negative instances will be wrongly assigned the positive label making them false positives, but most of them will be correctly assigned as true negatives. For the small threshold all positive instances will be correctly predicted as positives (consequently there will be no false negatives), but also the majority of negatives will be wrongly predicted as positives and only a smaller fraction of negatives is correctly predicted.

AUC

The area under the ROC curve (AUC) represents the degree of separability between the two classes. The higher the AUC the better a model will distinguish two classes. Lower AUCs represent cases where the probability density plots for the two classes have significantly more overlap, i.e. the model can't clearly separate both labels.

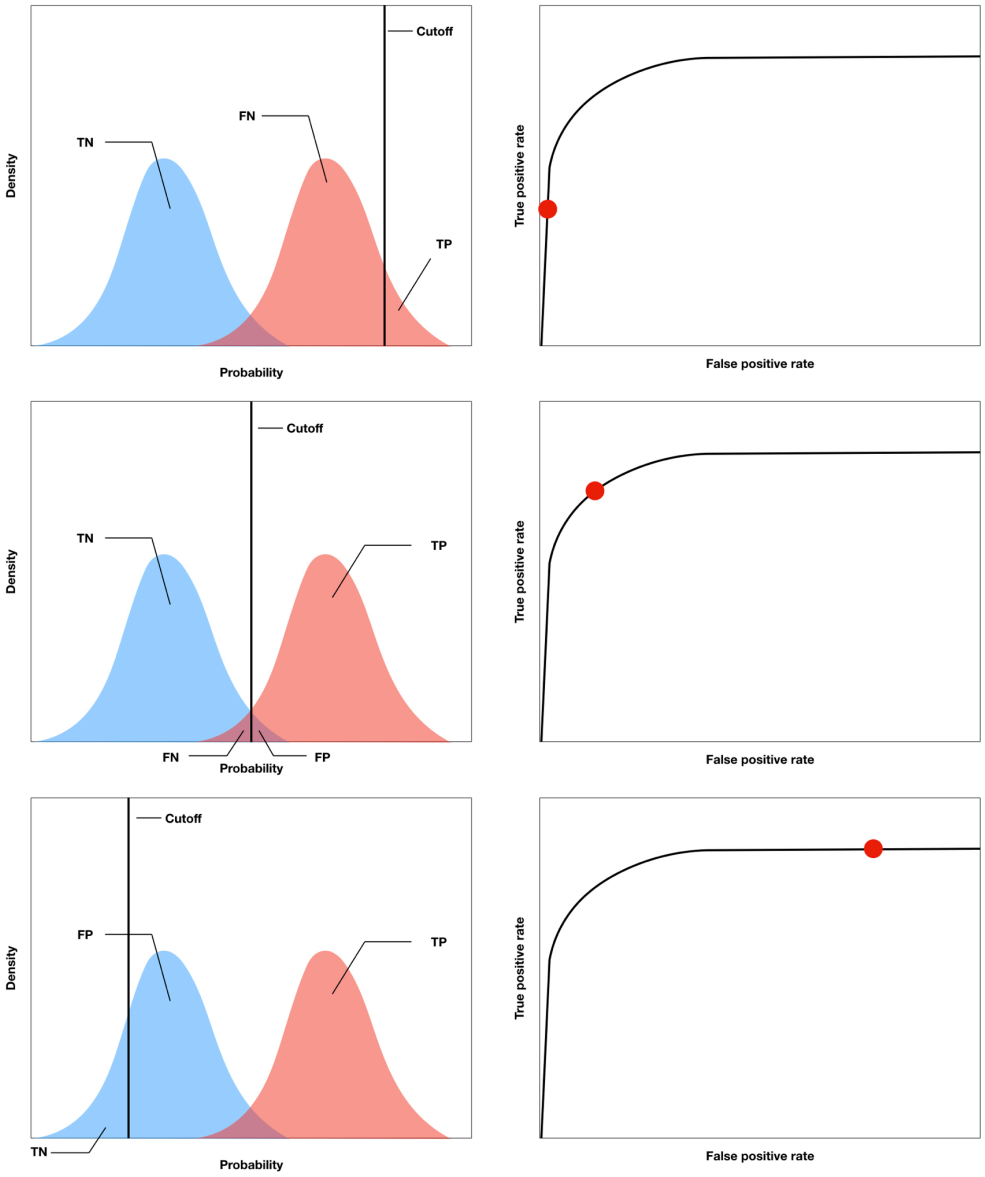



**Figure 8: The receiver operating characteristic curve**: Plots in the left column represent the probability density plots for instances of each class. Red is the positive class, blue is the negative class. The corresponding ROC curves are in the right column where the red point on the curve represents the TPR and FPR for the cut off used in the left column. TP: True positives, TN: True negatives, FP: False positives, FN: False negatives

*1.6.2.4 PR curve*

The precision-recall (PR) curve shows the trade-off between precision and recall for different thresholds. It is generated in the same way as the ROC curve except now precision is plotted versus recall. As is the case for ROC curves, models that are able to clearly separate both classes have a higher PR AUC than models that can't. Important to note is that PR curves are sensitive to class imbalances in the dataset.

*1.6.3 The random forest*

De neuter et al. 2018 used a Random forest (RF) to predict the epitope target of a given TCR CDR3 region. Random forest classifiers are a supervised machine learning method used to train a model to recognise patterns that are predictive for data labels. Random forest are referred to as ensemble learners as they combine multiple independent machine-learning models (decision trees) into a single predictive model (the forest) to obtain better predictive performance.

The RF algorithm for a binary classification problem can be represented as 3 main steps (figure 9). In the first step, random subsets of features and samples are drawn from the training data (bagging). Next, for each subset drawn, a decision tree is fitted which tries to optimally split both labels. Predictions are made by running each instance from the testing data through the collection of decision trees where the predicted label output by the RF model is the majority vote of all individual decision trees. The model is then evaluated by placing the predicted labels versus their actual labels and performance of the model is described as mentioned above.

Probabilities for instances belonging to a class or not are based on the number of votes received for the instance to belong to the class. For example, an instance for which all decision trees have the same positive class vote, has the highest probability to belong to that positive class according to the model. Logically the model will predict this instance as a true positive. Where the voting is more ambiguous, probabilities are smaller, and the model has less of an evidence basis to assign the correct label to that instance.

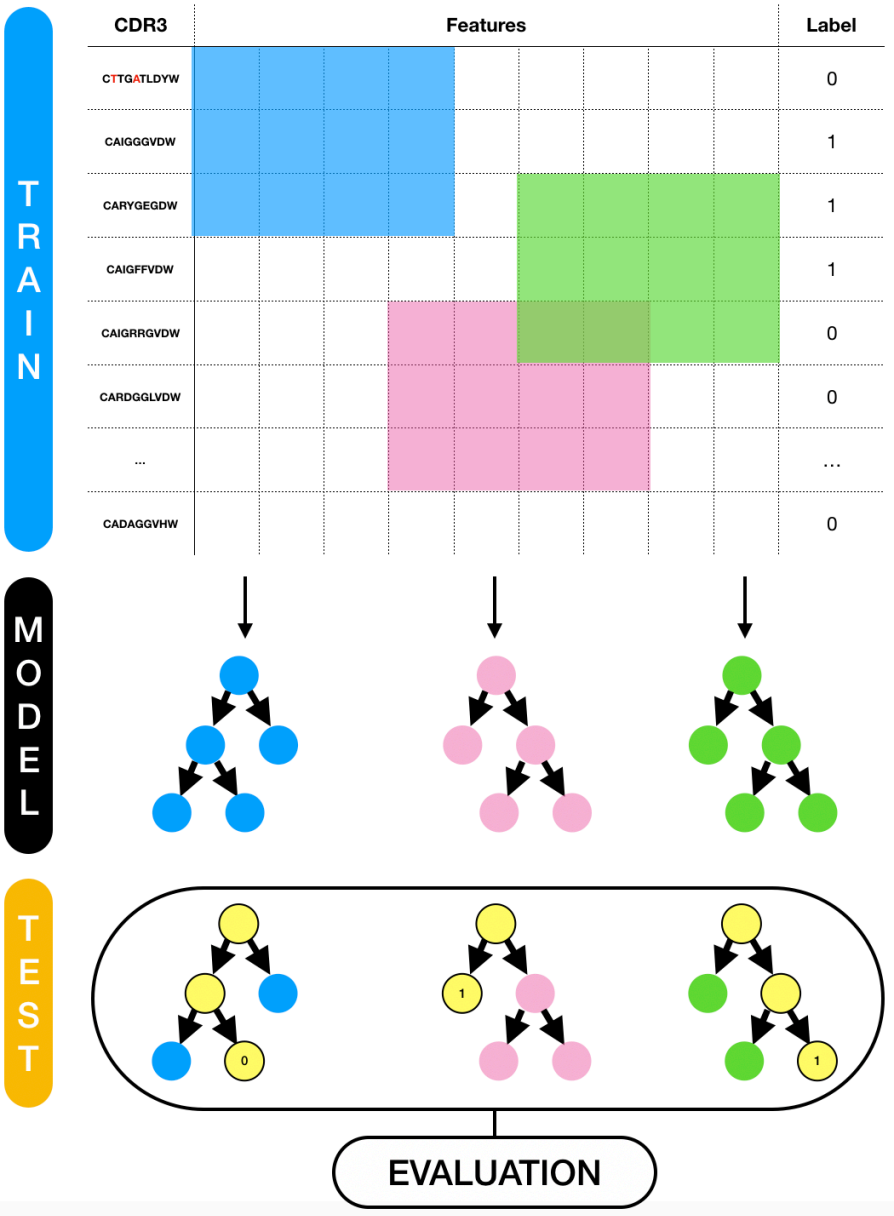



**Figure 9: Schematic representation of the random forest algorithm.** The colours on the left side of the figure refer to the colours used in figure 5 where the basic machine learning workflow was explained. First random subsets of samples and features are drawn from the training data. For each subset a decision tree is fitted which tries to optimally split both labels (0 and 1). Predictions are then made by running each instance from the testing set through the collection of decision trees (the model) where the predicted label is decided by the majority vote of all the individual decision tree outputs. To evaluate the model the predicted labels are placed against their ground truths and model performance is described using metrics derived from the confusion matrix.

*1.6.4 DBSCAN: Density-based spatial clustering of applications with noise*

Meysman et al. 2018 investigated how dissimilar TCR sequences can be before they no longer bind the same epitope. They used the unsupervised technique DBSCAN to cluster TCR sequence based on different similarity measures and checked if the formed clusters targeted the same epitopes. They showed that clustering based on a Levenshtein distance matrix already was capable of clustering TCRs targeting the same epitope. The advantage of the DBSCAN algorithm is that it does not require one to specify the number of clusters in advance. Thus, there is no need to introduce any prior knowledge on the number of clusters that may be presented in the dataset.

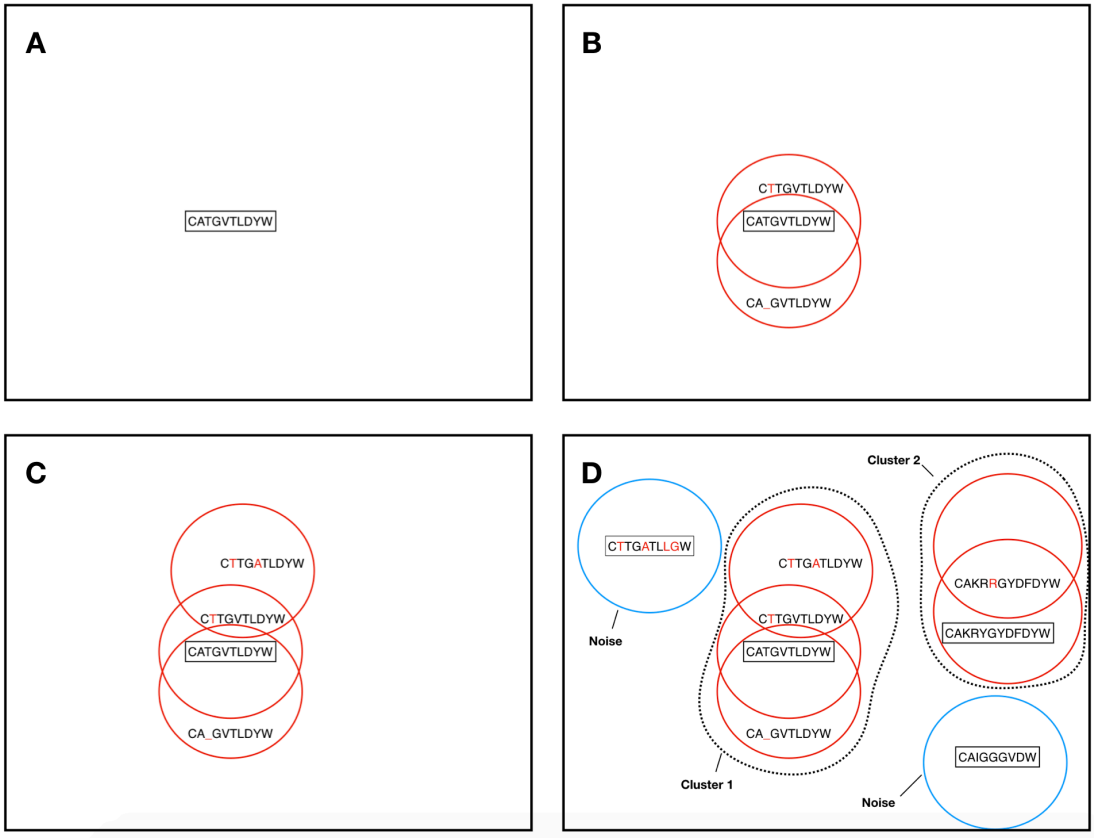



**Figure 10: Representation of the DBSCAN algorithm used to cluster similar sequences based on Levenshtein distance:** see text below for details.

The DBSCAN algorithm used on a Levenshtein distance matrix works as follow: First a random sequence is drawn (figure 10A). Using a threshold of 1 Levenshtein distance, the algorithm looks to find all sequences within 1 Levenshtein distance away from the first drawn sequence. Next, the same is done for the sequences that fulfilled this criterion, i.e. sequences within one Levenshtein distance away from these previously added sequences are also added to the cluster (figure 10B). This propagation of the cluster carries on until no more sequences can be added (figure 10C). Clusters are only assigned when their size exceeds the user defined minimal cluster size which in this case was 2 as this many similar CDR3s represent the minimal amount of information of interest. When cluster formation is ended, a new random unseen point is being drawn and the process repeats itself (figure 10D). When no similar sequence within the threshold distance can be found for a new point, the point is assigned the noise tag and a new point is selected once again. Important to note is that sequences from one end of a cluster are not necessarily within 1 Levenshtein distance away from sequences at the other end of the cluster. It is maybe better to say that two sequences at the extremities of a cluster are 'connected' through steps of 1 Levenshtein distance.

### 1.7 Research goal

The aim of this thesis was to investigate the possibility of using a similar machine learning approach to identify BCR sequences that target a specific epitope. We searched the literature for relevant data and trained a model to identify BCR sequences targeting chemically distinct antigens. We also manually collected a database of antigen specific HCDR3 sequences from the literature. On this dataset we used the DBSCAN algorithm to gain insights into how similar HCDR3s are within and between antigen specific repertoires as defined by a Levenshtein distance metric.

# 2. METHODS

## 2.1 Searching for relevant data

Several data repositories exist for sharing immune receptor repertoire sequencing data (table 1). These can be used to query specific sequences or subsequences against the whole database of sequences, return sequences fulfilling specific criteria or compare and combine multiple studies on the same or a different topic against each other. These repositories provide valuable information for researchers and enhances the value of adaptive immune receptor-seq data for improving biomedical research and patient care. Many of the repositories store data that are minimally processed where sequences associated with infected patients are not always true signature sequences specific to the infection compared to a healthy cohort. Often additional processing is required to isolate sequences that are differential between multiple donor groups. Unfortunately, also many studies deposit raw sequencing data in repositories like the European Nucleotide Archive (EMBL-EBI) or the dbGAP and Genbank (NCBI), which requires processing and analysis to be redone, or do not put in the effort of making their data public (unless upon reasonable request). For antibody annotated data, several databases exist that are primarily based on the antibodies available in the Protein Data Bank (PDB). These databases automatically extract the PDB files for newly added antibodies and store them following their own protocol. Because they all rely on the same PDB-derived antibodies, they more or less share the same information.

**Table 1: Listing of the most important repositories and databases for Ig-seq and structural annotated antibody data.**

| **Repositories and Databases** | | | |
|---|---|---|---|
| Name | Description | Reference | URL |
| iReceptor | Sequence search, metadata queries, combines different repositories into one. | Corrie et al. 2018 | ireceptor.irmacs.sfu.ca |
| immuneACCESS | Analysis of repertoires and raw sequencing data | DeWitt et al. 2016 | clients.adaptivebiotech.com/immuneacess |
| ImmPort | Data archiving, dissemination, analyses, and reuse. | Bhattacharya et al. 2018 | www.immport.org |
| JingleBells | A repository of standardized immune-related single cell rna-seq datasets for analysis and visualization at the single cell level. | Ner-Gaon et al. 2017 | jinglebells.bgu.ac.il |
| VDJServer | Integrated in iReceptor (see above) | Christley et al. 2018 | vdjserver.org |
| **Antibody databases** | | | |
| Name | Description | Reference | URL |
| SAbDab | Consistent annotation of antibody structures, unbound and bound Abs, sequence search, CDR search, therapeutic Abs | Dunbar et al. 2014 | opig.stats.ox.ac.uk/webapps/sabdab-sabpred/ |
| DIGIT | Annotated Ig variable sequences, includes unbound Abs structures | Chailyan et al. 2012 | http://circe.med.uniroma1.it/digit |
| AbDb | Non redundant Abs associated with their cognate antigen, different numbering schemes | Ferdous and Martin 2018 | www.bioinf.org.uk/abs/abdb/ |
| IMGT-3D | 3D structures of B cell epitopes with Ig annotation | Ehrenmann et al. 2009 | www.imgt.org/ |
| IEDB | Manually curated epitope-antibody annotated database | Vita et al. 2018 | www.iedb.org |
| PIRD (TBAdb) | Collects and stores annotated TCR and BCR sequencing data with online analysis and visualisation, sequences are specific to antigen or diseases. | Zhang et al. 2018 | db.cngb.org/pird |

To train a similar classifier for the prediction of antibody-epitope interactions one needs a big enough collection of non-identical HCDR3s that target the same epitope. Therefore, the contents of the immune epitope database (IEDB, Vita et al. 2018), which stores annotated antibody and BCR structures along with their respective epitopes, was investigated to see if enough data was available

### 2.2 Manually collected database for annotated CDRs

As the number of BCR in the IEDB database was limited, we searched the literature in the hope to find more experimentally validated data on BCR-epitope interactions to train our classifier with. Initially we looked for other databases (table 1) but as most of these are also dependent on antibody structures already contained in the PDB these were shown to provide redundant information: these databases contain references to the original PDB ID, which we could check also being included in the IEDB. The PIRD TBAdb however contains CDRs that are specific for a given antigen or are specific to a disease or condition.

We searched the literature for the following terms: “Ig/immunoglobulin/BCR-seq”, “antibody/BCR sequencing”, “B cell response/repertoire” (all terms separated by “/” were searched for). Table 2 summarizes the studies from which we could collect annotated CDR data to create our own database. In these studies, mostly CDR3 sequences are provided, underlining their importance to the field. We also collected other CDR sequences in our database when these were provided.

The CDRs from these studies can be specific to a given antigen studied in the context of a disease, infection or pathogen and their affinity towards the antigen is often validated by ELISA. Other CDRs are found in patients suffering from a disease and which are not found in healthy controls. Differences between pre and post-vaccination repertoires are also included where the CDRs are presumed to be specific for the pathogen towards vaccination was aimed.

Many of the CDRs had to be manually copied from tables in the articles themselves or from supplementary tables and spreadsheets. As some studies provided more information, for example isotypes, these were also collected. For the scope of this thesis we will only focus on the HCDR3. Collectively our database is comprised of

3225 HCDR3s of which 570 originate from the PIRD database and 2655 are manually curated from 20 different studies.

**Table 2: Overview of the studies collected in our curated database.** RABA: Radio Antigen Binding Assay), ELISA: enzyme-linked immunosorbent assay, TT: Tetanus toxoid, Ab: antibody, NA: Not Applicable

| **Authors** | **B cell /Abs description** | **Donors** | **Affinity assay** | **Size** |
|---|---|---|---|---|
| Croote et al. 2018 | Single cell sequencing BCR of IgE B cells in peripheral blood | 6 Individuals with food allergies | NA | 973 |
| Chen et al. 2017 | Anti-desmoglein Ab from human serum | 6 patients with pemphigus | ELISA | 526 |
| Ellebrecht et al. 2018 | Anti-desmoglein Abs in peripheral blood using phage display | 4 Pemphigus patients | ELISA | 149 |
| Parameswaran et al. 2013 | Acute Dengue infection vs non Dengue | 60 Dengue patients | NA | 130 |
| Myhrinder et al. 2013 | Chronic lymphocytic leukemia cell lines vs normal B cells | 17 CLL patients | NA | 107 |
| Scheid et al. 2009 | Abs against the HIV gp140 protein in human serum | 6 HIV infected patients | ELISA | 68 |
| Stettler et al. 2016 | 119 Abs against Zika Virus NS1 and E proteins and Zika neutralizing Abs | 4 Zika Virus infected patients | ELISA | 67 |
| Reason et al. 2009 | Protective antigen-specific human monoclonal Ab | 7 human anthrax vaccinated donors | ELISA | 64 |
| Smith et al. 2014 | Monoclonal Abs against *S.pneumoniae* polysaccharides | 4 human Pneumovax23 vaccinated donors | ELISA | 63 |
| Shehata et al. 2019 | RSV F-specific Abs from paired adenoid and peripheral blood samples | 4 young children | Cytometry sorting | 60 |
| Meffre et al. 2016 | HIV envelope gp140 CD4–binding site–sorted B cells from serum | 3 individuals with chronic HIV viremia | ELISA | 58 |
| Zhou et al. 2004 | Human Ab repertoire specific for the capsular polysaccharide (PS) of *S.pneumoniae* 6B | 6 donors vaccinated with PS vaccine | RABA | 55 |
| De Kruif et al. 2009 | Ab repertoire against Tetanus Toxoid | 2 humans with TT vaccine | ELISA | 53 |
| McCarthy et al. 2018 | Influenza hemagglutinin directed | 3 influenza vaccinated individuals | Bead assay | 51 |

| | | | | |
|---|---|---|---|---|
| Tsioris et al 2015 | West Nile Virus E protein-specific memory B cells and Abs | 11 Human WNV infection phenotypes | Immunoblot | 44 |
| Zhou et al. 2002 | Abs specific for the capsular polysaccharide of *S.pneumoniae* 23F | 7 vaccinated donors | RABA | 27 |
| Chi et al. 2015 | Abs Targeting Anthrax Protective Antigen following Vaccination | 1 individual | ELISA | 23 |
| Lee et al. 2016 | Ab repertoire before and after vaccination with trivalent seasonal influenza vaccine | 4 young adults | / | 17 |
| Lavinder et al. 2014 | BCR repertoires following tetanus toxoid (TT) booster vaccination | 2 individuals | Anti-TT ELISA | 15 |
| Williams et al. 2018 | Anti-MPER of HIV-1 gp41 Abs | HIV chronically infected individual | MPER probe cell sorting | 13 |

### 2.3 Predicting BCR epitope interactions

Data

Data for the classifiers was obtained from the IEDB database which provides annotated BCR sequences along with their respective targeted epitopes. The csv file can be found under 'CSV Metric Reports' under 'Database Export'. The file used in this manuscript was downloaded on 14/04/2019. The first classifier was trained on non-redundant HCDR3s with all species included. The second classifier, with epitopes derived from LPS and polysaccharides as the positive class, includes only mouse and human HCDR3s. The latter dataset was enriched with polysaccharide specific HCDR3s from 3 different studies on *Pneumoccocus* in humans (table 2).

Labels

To predict epitope binding to a BCR the classifier was tasked with correctly assigning whether a CDR3 binds to either a peptidic, the negative class, or non peptidic epitope, the positive class. In the follow up classifier the classifier was tasked with correctly assigning whether a CDR3 binds to either an epitope in LPS and polysaccharide antigens, the positive class, or an epitope not located on these kinds of antigens, the negative class. For each classifier duplicate CDR-label pairings were filtered out.

Features

The following properties of the BCR heavy chain CDR3 were encoded as numeric features: sequence length; the absolute count of each individual amino acid in the sequence; the total mass of the amino acids in the sequence according to molecular weights and the average CDR3 basicity, hydrophobicity, helicity, isoelectric point, and mutation rate. The average CDR3 mutation rate was calculated by taking the average of the mutation rate for each amino acid in the CDR3 sequence, where the mutation rate of an amino acid is obtained from the diagonal of a PAM250 substitution matrix. Physicochemical amino acid property values were used as described in MS2PIP Degroeve et al. 2013 and Elias et al. 2004. Positional features were added for each CDR3 residue position. Due to the variable length of the sequences present in the data, amino acid positions were translated into numerical positions by assigning each position an index value relative to the center of the sequence. For example, a sequence of length 3 would be encoded by the positions –1, 0 and 1 whereas a sequence of length 4 would be encoded as –2, –1, 1 and 2. For each position encoding generated in such a way, a binary feature (0 or 1) representing the presence or absence of an amino acid at that position was encoded. In addition, numerical features encoding individual amino acid basicity, hydrophobicity, helicity, isoelectric point, and mutation stability were also created for each position. Features were always created prior to model training and evaluation.

Models

Peptide binding was predicted using a random forest classifier (Breiman 2001) consisting of 300 trees and a *k* nearest neighbours classifier with *k* set equal to 3. Both models were used as implemented in sci-kit learn (Pedregosa et al. 2011).

The kNN algorithm is here used as a supervised clustering technique. The algorithm starts off with storing the feature space of the training data (figure 11). To reduce dimensionality, the multidimensional feature space can be transformed to the two first principle components as defined by principle component analysis (PCA). To make predictions for the instances of the test data, the *k* nearest neighbours in the stored feature space determine the label assigned to those test instances (as indicated by

the dotted lines in figure 11.). For $k$ = 3, if the 3 nearest neighbours to the instance from the testing set are labelled 1, the predicted label by the kNN model will be 1 for that instance. Where the 3 nearest neighbours do not share the same label, the label assigned by the model is that of the majority vote, i.e if 2 labels are negative and one is positive, the predicted label is negative. The predicted labels of the test set are then again compared to their actual labels using the confusion matrix.

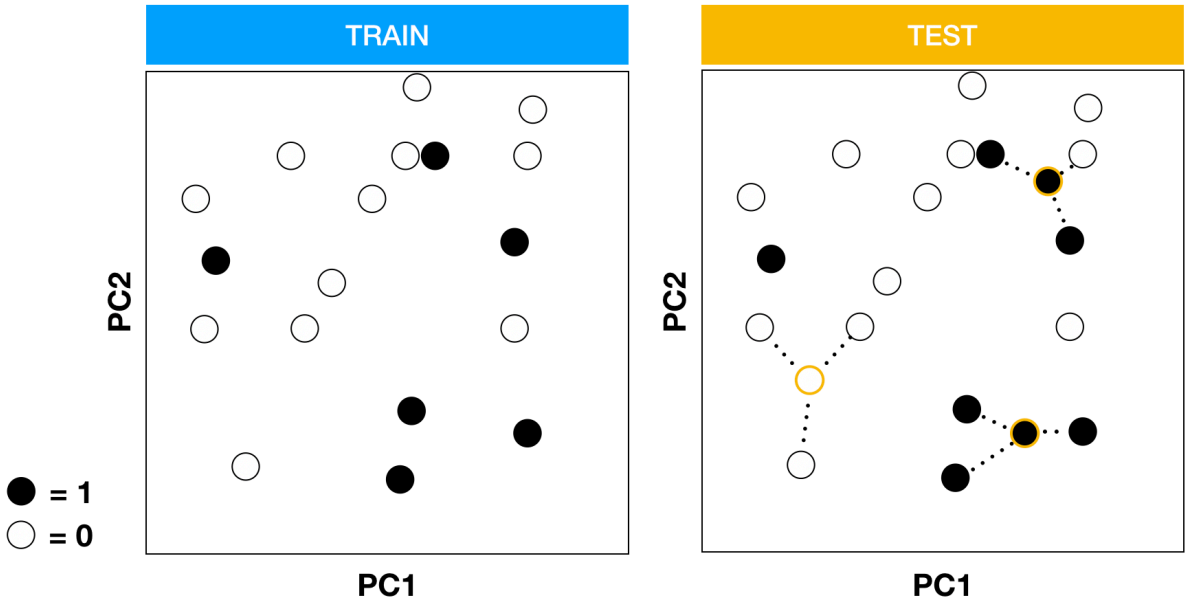



**Figure 11: Representation of the k Nearest Neighbours algorithm:** See text for details.

Model evaluation

We used Repeated Stratified K fold cross validation with 5 folds and 4 repeats. This translates to splitting the data in 5 equally sized, non-overlapping subsets that hold the same positive-negative label ratio as found in the full dataset. Training and test subsets were then created by using a single subset as test set while training the classifier on the remaining subsets. This splitting, training and testing of the data was repeated 4 times in which the subsets between repeats were different each repeat.

The following validation measures were calculated on the held-out test data: prediction accuracy, AUC and PR. Overall classifier performance was evaluated in terms of prediction accuracy, AUC values and mean PR values averaged over 20 folds (5 folds times 4 repeats). The receiver operating characteristic curve and precision-recall curve were drawn up for each model as well. For our first classifier, precision is interpreted as how many of the HCDR3s predicted to bind non-peptidic epitopes actually bind a non-peptidic epitope and recall as how many of the HCDR3s that bind non-peptidic epitopes are predicted to bind a non-peptidic epitope. The evaluation measures were reported as their mean over the number of subsampled executions (cross-validation) ± their standard deviation.

### 2.4 Unsupervised clustering of repertoire specific CDR3 sequences

Data

Data used to perform the repertoire clustering was collected from the existing literature (table 2). Duplicate CDR3 were filtered out as these would've ended up in the same clusters because the Levenshtein distance between them is zero. Many studies dealt with different antigens for the same infectious agent or microorganism, therefore, for a more general overview, we housed different antigens for the same organism under the same repertoire. For example, Stettler et al. 2016 studied antibodies against NS1 and E proteins of Zika virus but both antigen specific antibodies were put under the Zika virus repertoire. When the numbering scheme used by the studies to determine the CDR3 was uncertain or they used an alternative numbering like the Kabat scheme these CDR3s were left out of the analysis. The correct numbering scheme is critical in identifying similar CDR3s using the Levenshtein distance as incorrect alignments will increase the distance measured.

DBSCAN parameters

The DBSCAN clustering algorithm was applied on a distance matrix in which for each CDR3 sequence the Levenshtein distance (also edit distance) was calculated against all other CDR3s in the dataset. The Levenshtein distance is defined as the amount of substitutions, duplications or deletions that are required to correctly align one sequence versus the other. For identical sequences Levenshtein distance is evidently

zero. When the difference between two sequences is one amino acid substitution the Levenshtein distance is 1. When the difference is a substitution and a deletion the distance is 2 etc. (Figure 12)

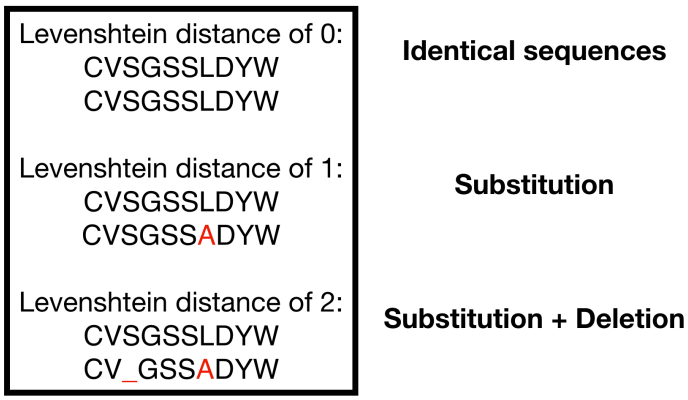



**Figure 12: Examples of Levenshtein distances for different sequence alignments.**

The distance matrix was used to calculate mean distances and mean lengths within and across all repertoires. The DBSCAN algorithm as implemented in *sklearns* (Pedregosa et al. 2011) requires two arguments; one being the threshold, the other the minimum samples required to form a cluster. For both thresholds used in this paper, namely 1 and 2, the minimum cluster size was kept at 2.

All code and analysis were done in Python and code is available on github: https://github.com/BeirinckxM/Masterthesis

# 3. Results

## 3.1 Database inspection and parsing

The full immune receptor database was downloaded from IEDB as a large csv file and inspected. The file contains 31429 T cell and 3379 B cell entries. The epitope sequence is provided together with the name of the organism and antigen it is derived from, but not all the time. BCR entries are annotated with VDJ genes for both heavy

and light chains and the organism from which the BCR originates from. Organism names however are numerically notated according to NCBI Taxonomy and require conversion by downloading another file from the IEDB website which holds the database's internal taxonomy. Most B cell entries contain information on the three complementing determining regions (CDRs) according to the IMGT numbering scheme. Also, the full sequence for the Fab domain is provided. As the focus of this thesis is on B cell epitopes and its interaction with the BCR the file was first filtered for B cell entries.

The antibodies originate from a broad range of animals with the top 3 consisting of Mouse (*Mus musculus*), Human (*Homo sapiens*) and Llama (*Lama glama*) BCRs. Epitopes in the receptor database are not characterized in a way that allows clear identification of the type of epitope. However, when linking the IDs of entries in the receptor database with the same IDs in the reference database (also available for download from the IEDB website) it is possible to retrieve more information. Given the fact that one ID in one database matches with multiple entries in the other, this becomes a bit of a difficult exercise and it's easy to lose track. Anyhow, the epitopes would only be labelled with one of the following; discontinuous peptide (on multichain or not), linear peptide or non-peptidic.

Based on the notation of the epitopes however there was a possibility to make a narrower labelling. Searching for the presence or absence of specific string characters or notations lead to the labelling of the epitopes in several more distinct groups (figure 13 and table 3). The non-peptidic label is assigned to structures that are non-peptidic but in its own is very diverse as it contains structure like lipids, carbohydrates, hormones, antibiotics etc. In total, the IEDB database contains 2031 unique epitopes with at least one annotated B cell CDR associated (usually more than one either heavy or light chain) (figure 13).

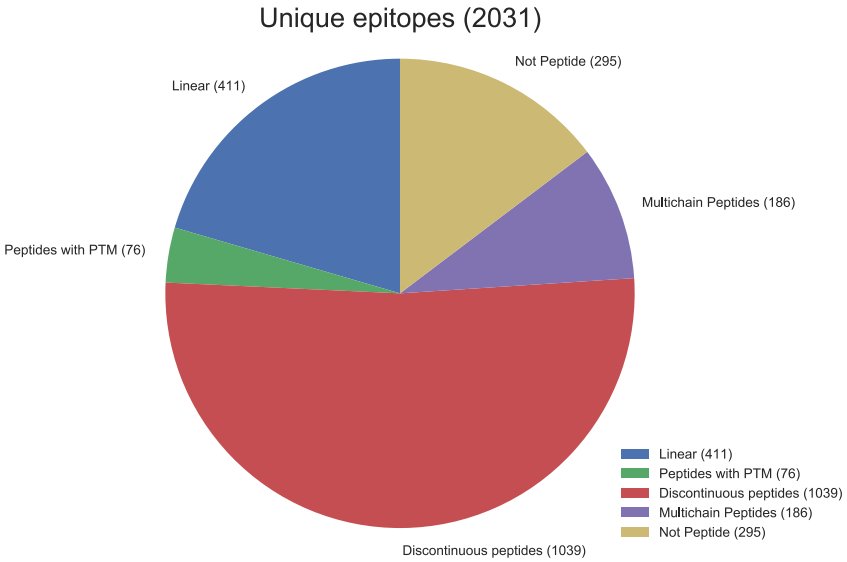



**Figure 13: Pie chart showing the sizes of the different unique epitopes for all species contained in the B cell dataset** (PTM: Post Translational Modification).

**Table 3: Representative examples for different labels assigned to epitopes in the dataset** (PTM: Post Translational Modification).

| Label | Representative example |
|---|---|
| Linear peptide | NWFDITNWLWYIK |
| Linear peptide with PTM | ARTKQTARKSTG + METH(K4) |
| Discontinuous peptide | K42, F43, M44, D45, Y47, Q48, R49, Y51, K127, C130 |
| Discontinuous peptide with PTM | H129, N166, D173, K174, K175, K177 + GLYC(N162, N166) |
| Multichain peptide | A: S63, S64, D65, Y66, R69; B: L97, Y108, H109, M110, N111, P149, I150 |
| Multichain peptide with PTM | HA1: E341; HA2: E360, G361, I363, D364, R370, E375, T377, G378, Q379, A380, A381, L383, N491, E495 + GLYC(2:N499) |
| Alternative peptide notation | Ac-Gln-D-Phe-His-D-Pro-NH2 |
| Non-peptidic | 1,2-dihexanoyl-sn-glycero-3-phospate |

## 3.2 Predicting BCR interactions with chemically distinct epitopes

### *3.2.1 BCR interactions with peptidic and non peptidic epitopes*

Not many unique BCRs targeted the same epitope and it became clear that not enough data was available to train a similar classifier as Gielis and collegues (Gielis et al. 2018, De Neuter et al. 2018). Additionally, because unlike the strictly linear peptides presented by MHC I to T cells, the BCR recognizes peptidic epitopes in their native form that can be linear, discontinuous and even range over multiple peptide chains (Kringelum et al. 2013). Our labelling efforts however uncovered several distinct groups of epitopes. Because non peptidic epitopes often represent functional groups or chemical characteristics we are not likely to find in peptidic epitopes (unless through post translation modifications) we asked the question if we could predict accurate BCR epitope interactions to either peptidic or non-peptidic epitopes based on amino acid properties of the CDRs. This question arises from the assumption that some of the characteristics for non-peptidic epitopes, like negatively charged phosphatic groups, longer polysaccharides, fatty acids with long aliphatic chains and larger organic structures like steroids and quinons, require different structural configurations and characteristics of amino acids in the HCDR3 compared to the BCRs specific for peptidic epitopes. From here on out we refer to the heavy chain CDR3 as CDR3 unless otherwise mentioned

To train our classifier, first we assigned labels to the CDR3s; either the CDR3 targets a peptidic epitope and is given the negative label ‘0’, or the BCR targets a non-peptidic epitope and is given the positive label ‘1’. We included BCRs from all species. We first tested a one vs one random forest and K neighbour classifier where the classifier attempts to predict which type of epitope is most likely to be bound by a given CDR3 sequence. We included the kNN as we expected both labelled groups, especially the positive labelled group (non peptidic epitopes), to be quiet heterogenous. The reasoning is that within the heterogenous groups, nearest neighbours to a given CDR3 sequence are distinct for a specific subgroup within the positive label. For example, fatty acids and polysaccharides are both non-peptidic but are not exactly alike and CDR3s targeting these epitopes could reflect this dissimilarity.

Because peptidic epitopes in our dataset outnumber the non peptidic epitopes our dataset is considered imbalanced. Input features for the classifiers were derived from CDR sequences and include positional and physicochemical properties (See methods). We used a repeated cross validation approach to validate our classifier in which part of the data is left out as an independent test set for the trained classifier.

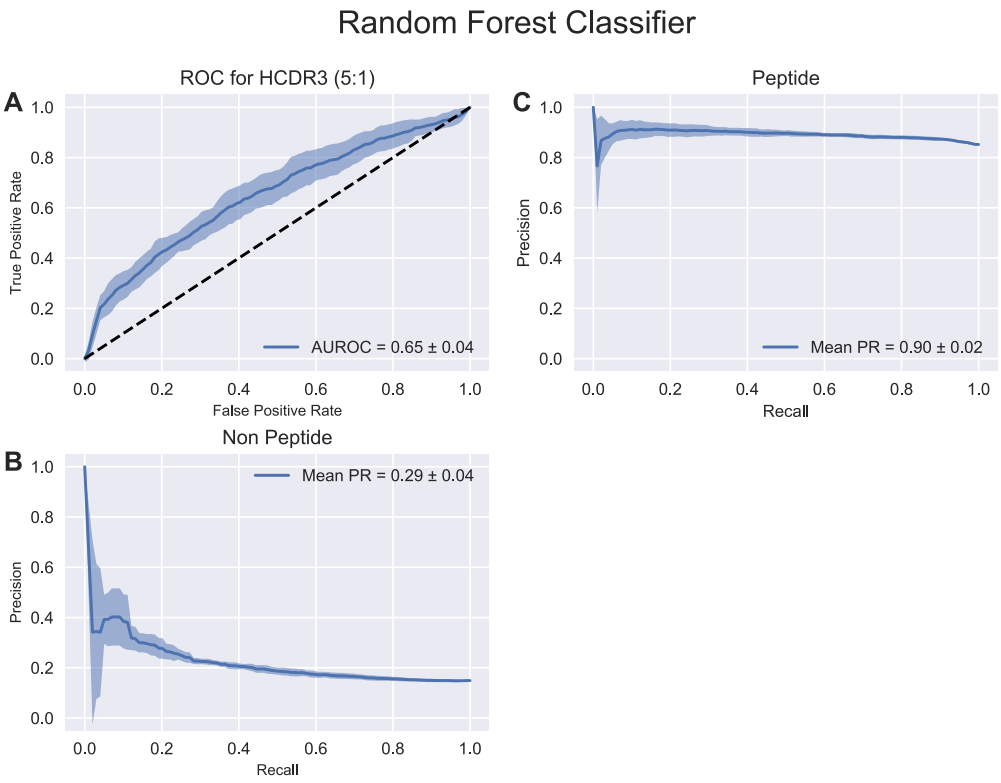



**Figure 14: Predictions for peptide specific CDR3s versus non peptide specific CDR3s**: ROC and PR curves for the Random Forest Classifier using features derived from the heavy chain CDR3 region sequence. ROC curve: (true positive rate versus false positive rate). Averaged values were plotted as a single line while the surrounding area indicates the standard deviation as observed during cross-validation. The (0,0) to (1,1) striped diagonal is illustrative for a random classifier assigning random labels to BCRs. The more north-west from the diagonal a ROC is situated, the higher the performance of the classifier (higher true positive rate for a lower false positive rate). Precision recall curves: Averaged values were plotted as a single line while the surrounding area indicates the standard deviation as observed during cross-validation. In the perfect case, the PR curve is a horizontal line with precision always equal to 1, representing the lack of false-positive predictions over the entire recall range.

Mean accuracy for the average random forest classifier was 84 ± 1%, with a mean area-under-the-receiver-operating-characteristic-curve (AUROC) of 0.65 ± 0.04 and a mean precision over a recall range of 0 to 1 (PR) of 0.29 ± 0.04 (positive) and 0.90 ± 0.02 (negative) on the independent test data (figure 14). The nearest neighbour classifier had respectively a mean accuracy of 81 ± 2 %, a mean AUC of 0.61 ± 0.03 and PRs of 0.22 ± 0.03 (positive) and 0.88 ± 0.01 (negative). AUROC values range from 1 (perfect predictions) to 0 (no correct predictions) with a value of 0.5 representing a model that randomly predicts interactions. Even though AUROC values are significantly better ($p<0.001$) from the AUROC of 0.5 of a random classifier, much of this is due to the more correct false negative predictions in the imbalanced dataset: a smaller number of false positive predictions divided by the total negatives in the majority class returns a small 'rate' whereas a same number of true positive prediction divided by the total positives in the minority class return a larger 'rate'.

The high accuracies for both models make believe this is an accurate classifier with high rates of true positives and true negatives. However, our dataset is imbalanced (5:1 ratio negatives/positives) and accuracy for imbalanced datasets is not a good performance measure (Saito and Rehmsmeier 2015, He 2009). For an imbalanced dataset with a N/P ratio of 19/1, correctly predicting the label for the most abundant class will result in a high overall accuracy of 95%, but tells us nothing about the quality of predictions for the minority class.

Overall both classifiers will only reach a high TPR at the cost of a high FPR indicating that they can't easily distinguish CDR3s specific for non-peptidic epitopes from the CDR3s that target peptidic epitopes. PR for the positive class crashes rapidly likely due to the fact that a small fraction of false positives already severely impacts precision for the positive class.

### *3.2.2 BCR interactions with epitopes in LPS*

Further inspection of the non-peptide epitope group taught us that the biggest fraction here were epitopes studied as part of lipopolysaccharides (LPS). This is not surprising because of their vital role in the structural integrity of bacterial cell walls. The non peptidic epitopes were already expected to be very noisy with several groups chemically different than peptidic epitopes, but also very different to themselves like lipids and carbohydrates. Initially, we used BCRs originating from all species in the database. Different species however express different BCRs (Muyldermans and Smider 2016). Especially BCRs from Artiodactyls like bovines, camels and llamas can differ dramatically. For example; in camelids heavy chain antibodies bind antigen with only a single heavy chain variable region in the absence of light chains and in bovines, ultralong HCDR3 regions form an independently folding minidomain which can protrude from the surface of the antibody. In order to decrease noise in our dataset, we filtered out the BCRs originating from humans and mice that, even though differences have been reported between the two, recognize antigen using the relatively flat binding surface formed by CDRs on the variable regions of the heavy chain/light chain heterodimer.

To test our hypothesis on a cleaner dataset we tried to see if our classifier could distinguish BCRs specific for LPS and derived polysaccharide structures from the rest of the database. Some but not all LPS in the database however have specified LPS as their antigenic structure. Looking at the notation of these LPS labelled epitopes, it was obvious that not all LPS or LPS-derivatives were labelled as such in the database. Substrings that are contained in the epitope labelled as LPS were found in other structures as well and further investigation of these substrings taught us they were LPS specific or at least represented polysaccharide or glycan structures (table 4). However again on the basis of notation and the presence of certain substrings related to LPS and polysaccharides we were able to extract most of these structures (most of them contain more than one substring). The resulting liposaccharide group was inspected, and mislabelled epitopes were not labelled as LPS if not sure. A total of 96 unique epitopes were eventually identified as being epitopes as part of LPS and polysaccharide antigens.

**Table 4: A selection of substrings and their relevance to LPS and derived polysaccharide structures.**

| **Examples of LPS and derived polysacharide notations** | |
|---|---|
| alpha-D-Kdo-(2->8)-[alpha-D-Kdo-(2->4)]-alpha-D-Kdo-(2->4)-alpha-D-Kdo',<br>alpha-D-GalpNAc-(1->3)-[alpha-L-Fucp-(1->2)]-beta-D-Galp,<br>3-deoxy-alpha-D-manno-oct-2-ulopyranosonic acid | |
| **Substrings** | |
| Kdo: 3-Deoxy-D-manno-oct-2-ulosonic acid | Used by bacteria in LPS synthesis |
| Rhamno, Rha: rhamnose | Cell wall components in some bacteria |
| Mannan | Linear polymer of mannose |
| GalpNAc: N acetyl-D-Galactopyranosyl | Part of peptidoglycans in the bacterial cell walls |
| GlcNAc: N-acetyl-D-glucosamine | |
| Araf: alpha-D-arabinofuranosyl | Unit of arabinan polymers in LPS |
| GalNAcAN:2-acetamido-2-deoxy-D-galacturonamide | Repeating unit in LPS derived polysaccharides |

As we want to train a model that can predict interaction with LPS based solely on the sequence of a CDR3, we added CDR3s that were known to interact with LPS from the database we manually collected from the literature (see table 2). These HCDR3 sequences were numbered according to Kabat numbering where the HCDR3 region starts at position 3 in the IMGT numbering scheme. In other words, the Kabat numbering misses 2 amino acids at the start of the CDR3 sequence compared to its IMGT numbering (Dondelinger at al. 2018). We could not add these to the dataset as features for the Kabat numbered HCDR3 will shift two places to the left neither could we add mean feature values for the two missing amino acids as these would most likely be falsely picked out as specificity determining features by the classifier. Both would have resulted in a CDR3 which would not reflect the true features that constitute LPS specificity. However, one solution remained where we changed the IMGT numbered CDR3s from the IEDB database to the Kabat numbering. This required us to leave out the first two amino acids of the IMGT numbered CDR3s. This way all CDR3s hold ‘natural’ features that could reflect their true binding specificity in our

classifier. To train the classifier we assigned new labels to the CDR3s; Either the CDR3 targets an epitope as part of LPS and is given the positive label '1', or the BCR doesn't target an epitope as part of LPS and is given the negative label '0'. The latter labelled group is from now on referred to as the 'rest' group. We filtered the dataset for unique CDR label pairings and this time included only human and mouse BCRs.

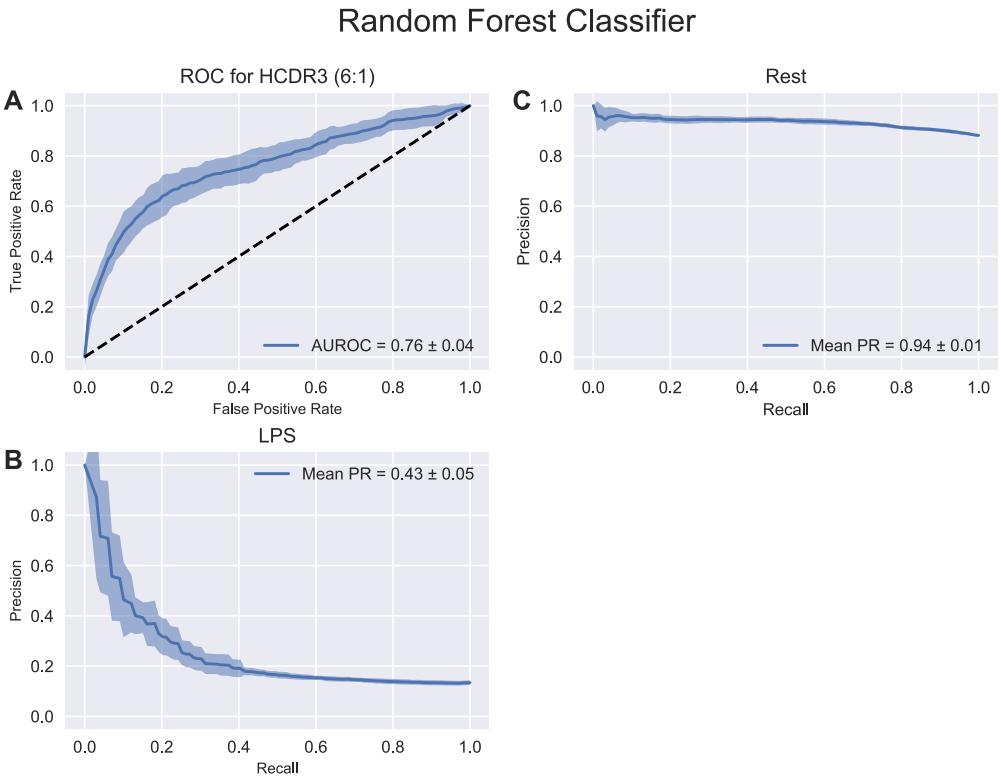



**Figure 15: Predictions for LPS epitopes versus rest**: ROC and PR curves for the Random Forest Classifier using features derived from the heavy chain CDR3 region sequence. ROC curve: (true positive rate versus false positive rate). Averaged values were plotted as a single line while the surrounding area indicates the standard deviation as observed during cross- validation. The (0,0) to (1,1) striped diagonal is illustrative for a random classifier assigning random labels to BCRs. The more north-west from the diagonal a ROC is situated, the higher the performance of the classifier (higher true positive rate for a lower false positive rate). Precision recall curves: Averaged values were plotted as a single line while the surrounding area indicates the standard deviation as observed during cross-validation. In the perfect case, the PR curve is a horizontal line with precision always equal to 1, representing the lack of false-positive predictions over the entire recall range. A dropping PR is indicative of a loss of precision to reach a higher recall.

With a mean AUROC of 0.76 ± 0.04 our classifier seems to benefit from the decrease in noise as a result of including only human and mouse BCRs and a narrower labelled epitope field (Figure 15). Predictive quality however is still not strong (PR = 0.43 ± 0.05). The classifier is again only able to retain high precision for the positive class for a small recall. The ROC curve is slightly steeper at the start which suggest some higher ranked instances can be retrieved with high precision. However, the ROC quickly levels off. PR immediately drops but more gradually compared to the PR curve of the first classifier again suggesting some instances can be predicted with decent precision. To reach higher recalls, precision is sacrificed again likely due to class imbalance. The negative class, which includes all epitopes except the ones as part of LPS is almost always correctly classified. This is a bit counterintuitive as this class is supposed to be noisier than earlier when it contained only peptidic epitopes.

### 3.3 Unsupervised clustering of repertoire specific CDR3 sequences

The CDR3 sequences in the collected database are not suited for a classifier predicting antibody-epitope interactions as the epitope they bind cannot be pinpointed within their respective antigens. At least these CDR3s are specific for the antigen and repertoire they bind and belong to. This database gave us an opportunity to test this hypothesis i.e are CDR3 sequences unique to a given antigen and repertoire or can a CDR3 sequence be shared between repertoires and in response to different antigens? In a way we are looking for CDR3s that can possibly cross-react which is the phenomenon whereby antibody shows specificity for an epitope that is different from the immunogen (the epitope that activated the B cell that produced the antibody).

#### *3.3.1 Analysis of the distance matrix*

A distance matrix was generated in which for each CDR3 sequence the Levenshtein distance was calculated against all other CDR3s in the dataset. Figure 16 shows a heatmap of the entire clustering matrix. The (0,0) to (1,1) diagonal represents the sequences compared to themselves (with a distance of zero). The lack of clear visible patterns is indicative of high sequence dissimilarity within aswell as between repertoires. One exception is the Dengue virus repertoire in which sequences appear to resemble each other to a higher degree.

Indeed, the mean distance in the Dengue virus repertoire was the smallest around 6 whereas the mean distance in other repertoires was 11 and higher. The study from which thes CDR3s were collected was a study comparing CDR3s from Dengue infected versus non Dengue infected patients. Importantly, this study had a sample size of 60 greatly exceeding that of the other studies in the database (table 2). Additionally this study has excluded BCR sequences lacking VD and DJ junctional bases out of their analysis to cancel out indivdual variability and focus only on “in frame” convergent BCR signatures. Therefore it’s hard to conclude that the higher similarity within these sequences is either the result of an immune response to a limited set of targeted epitopes or either the consequence of the larger sample size and filtering steps used in the study methods. Nevertheless, the larger sample size should have influenced the diversity as BCRs towards the same epitope are rarely the same across individuals and more individuals sequenced will have resulted in a greater diversity for the same epitopes.

For most repertoires a sequence on average needs a total of 11 deletions, mutations or substitutions to align with any other sequence in that repertoire. Given a mean length for all CDR3 sequences of 17 ± 3 amino acids, this reflects a high dissimilarity and because sequences belonging to the same clonal family have identical CDR3 length (Greiff et al. 2017), this indicates multiclonal responses to antigens in the repertoires. Sequences appear to show varying resemblance to other sequences as visualised by alternating blue and green lines throughout the matrix but some sequences appear to be more unique in respect to all other CDR3s as can be concluded by the presence of continuous green and yellow lines across the matrix. This is the case for the CDR3s in the HIV repertoire, a virus exhibiting extreme variability and high evolution rates and notorious for evading the immune system itself (Santoro et al. 2013). On closer inspection however, it turns out that the green lines represent sequences with longer CDR3s lengths and occur in repertoires where the mean distance is greater compared to the majority of the repertoires. Aligning longer sequences with shorter sequences increases the edit distance measured between them simply by taking account of the deletions of the shorter versus the longer sequence.

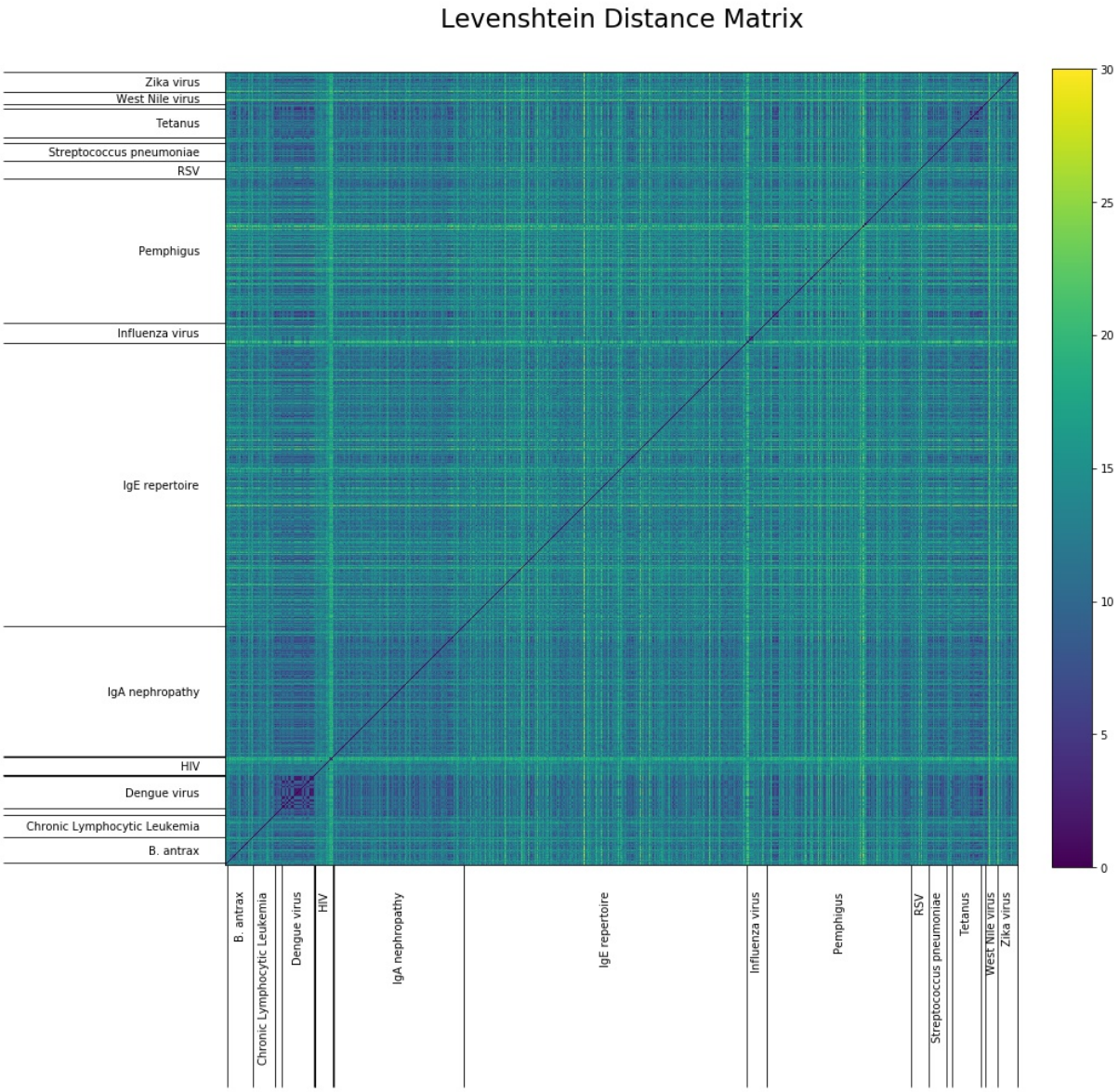



**Figure 16. Heatmap of the entire clustering matrix**. Darker blue colours represent lower Levenshtein distances between sequences, green to yellow represent higher distances. Repertoires are indicated on both axes but don't include all repertoire names due to lack of space. For one sequence in a repertoire, distances to all other sequences were calculated. Along the (0,0) to (1,1) diagonal, sequences are compared to themselves and Levenshtein distance is zero, hence its blue colour. The matrix is symmetrical with respect to this diagonal.

*3.3.2 Clustering similar CDR3 sequences*

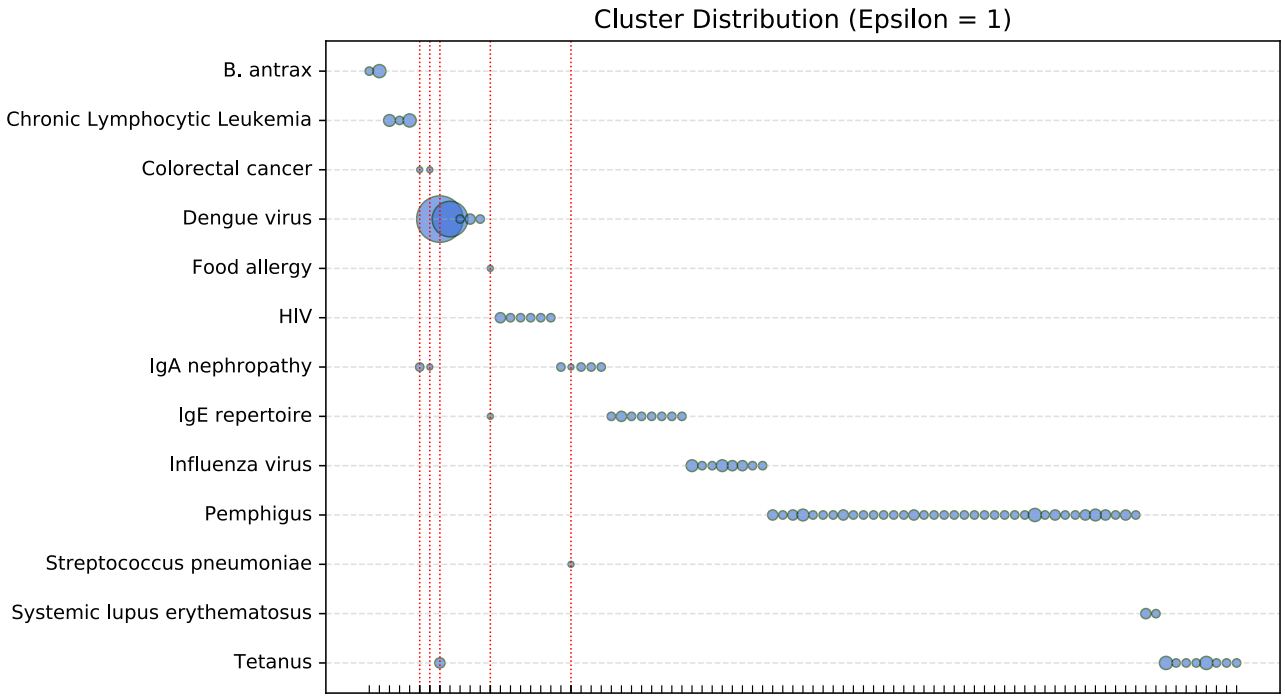
To see if HCDR3 sequences are specific for a given repertoire, we used the DBSCAN clustering algorithm with a threshold of one to see if very similar CDR3s cluster together and if sequences can be shared between repertoires. With a minimal cluster size of 2, only 11% of all sequences is distributed over a total of 87 different clusters. The other 89% were considered noise points and thus have a Levenshtein distance to the sequences in clusters that exceeds 1. Infact, RSV, West Nile and Zika Virus and hepatitis C virus repertoires are left out of clusters completely indicating that these repertoires don't contain very similar CDR3 sequences.

Figure 17 shows the cluster distribution within each repertoire for a threshold of 1. Overall cluster sizes are small holding only 2 or 3 sequences. The clusters in the Dengue virus repertoire are larger as could've been expected from it being the repertoire with the lowest mean distance. Out of the total of 87 clusters, 5 of them are heterogenous and contain sequences belonging to different repertoires. Increasing the clustering threshold to 2 clustered 16% of all sequences and was associated with an increased number of clusters (121), and a larger set of heterogenous clusters (15) (figure 18). Also repertoires that fell out of clustering with a threshold of 1 were now included. Most of the heterogenous clusters consist of sequences from two different repertoires and remain small for most cases. Some clusters show considerably larger sizes. For some of these, this is due to the high degree of similarity in the Dengue virus repertoire, but cluster 2 reached a size of 29 as a result of clustering sequences from six different repertoires.

To investigate the significance of the heterogenous clusters, we searched the literature for potential correlations and cross-reactivity between the members of those clusters. For colorectal cancer, the IgE and IgA nephopathy repertoire significance of CDR3s contained herein is difficult as these repertoires are more systemic rather than antigen specific. However, in IgA nephropathy (a chronic kidney disease) a dysregulated mucosal immune system with defective immune tolerance to commensal or commonly encountered pathogens may be a factor triggering this disease (Penfold et al. 2018). Given the fact that the human colon is a place where commensals are

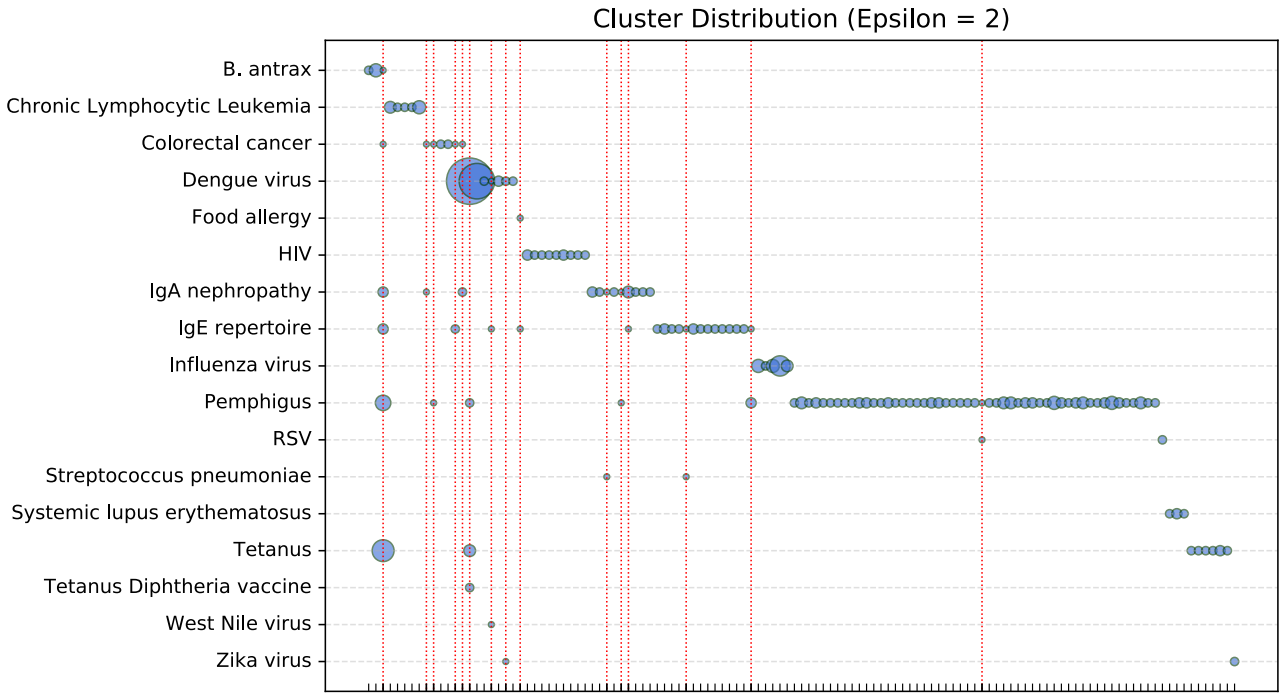
present in abundance it is maybe not surprising that the two repertoires end up in the same cluster multiple times. Interestingly there is a relation between IgA and Streptococcal infection of the upper respiratory tract which explains CDR3s from these repertoires clustering together (Meng et al. 2014, Schmitt et al. 2010). Clusters containing CDR3s from the IgE repertoire could be indicative for IgE antibodies binding to the antigen from the repertoire that is part of the same cluster. A clear example is the pairing of Food allergy with IgE given the importance of IgE antibody in allergic disease (Galli et al. 2012).

Dengue and Zika virus pairings can be explained because of the fact that a study specifically looking for cross-reactivity between these related species is included in the database (Stettler et al. 2016). West Nile virus, like Dengue and Zika also a member of the Flavivirus family is also reported to cross react with Dengue (Papa et al. 2011). Heterogenous clusters containing CDR3s from Pemphigus repertoires, which holds autoantibodies against desmoglein are harder to interpret and so are clusters containing tetanus toxoid reactive CDR3s. For neither of the two correlations or cross reactivities seem to have been reported.




| **Clusters sharing CDR3s from different repertoires (5)** | | |
|---|---|---|
| Cluster number | Repertoires | Size |
| Cluster 5 | Colorectal cancer, IgA nephropathy | 3 |
| Cluster 6 | Colorectal cancer, IgA nephropathy | 2 |
| Cluster 7 | Dengue virus, Tetanus | 66 |
| Cluster 12 | Food allergy, IgE repertoire | 2 |
| Cluster 20 | IgA nephropathy, Streptococcus pneumoniae | 2 |

**Figure 17: DBSCAN Cluster distributions using a threshold of 1 for every repertoire and description of the heterogenous clusters**. For every repertoire, the size of the blue bubbles on the horizontal dotted lines is representative for the number of sequences included in a cluster from that repertoire. Indicators on the x axis represent clusters. The red vertical dotted lines indicate heterogenous clusters and bubbles with no vertical lines passing through them are homogenous clusters. The table sums up the heterogenous clusters with the repertoires included and their respective size.




| **DBSCAN Clusters sharing CDR3s from different repertoires (15)** | | |
|---|---|---|
| Cluster number | Repertoires | Size |
| Cluster 2 | *B. antracis*, Colorectal cancer, IgA nephropathy, IgE repertoire, Pemphigus, Tetanus | 29 |
| Cluster 8 | Colorectal cancer, IgA nephropathy | 2 |
| Cluster 9 | Colorectal cancer, Pempigus | 2 |
| Cluster 13 | Colorectal cancer, IgE repertoire | 3 |
| Cluster 14 | Dengue virus, Tetanus and Diphtheria vaccine, Pemphigus, Tetanus | 71 |
| Cluster 17 | Dengue virus, IgE repertoire, West Nile virus | 3 |
| Cluster 19 | Dengue virus, Zika virus | 3 |
| Cluster 21 | Food allergy, IgE repertoire | 2 |
| Cluster 33 | IgA nephropathy, Streptococcus pneumoniae | 2 |
| Cluster 35 | IgA nephropathy, Pemphigus | 2 |
| Cluster 36 | IgA nephropathy, IgE repertoire | 5 |
| Cluster 44 | IgE repertoire, Streptococcus pneumoniae | 2 |
| Cluster 53 | IgE repertoire, Pemphigus | 4 |
| Cluster 85 | Pemphigus, RSV | 2 |

**Figure 18: DBSCAN Cluster distributions using a threshold of 2 for every repertoire and description of the heterogenous clusters**. For every repertoire, the size of the blue bubbles on the horizontal dotted lines is representative for the number of sequences included in a cluster. Indicators on the x axis represent clusters. The red vertical dotted lines indicate heterogenous clusters and bubbles with no vertical lines passing through them are homogenous clusters. The table sums up the heterogenous clusters with the repertoires included and their respective size.

## 4. Discussion

The recent successful developments in predicting TCR-epitope interactions instigated us to investigate if a similar approach can predict interactions between the BCR and its antigens (De Neuter et al. 2018, Gielis et al. 2018). We started off with a search for annotated antibody and BCR data that bind a certain epitope. Several databases exist (table 1), but all more or less depend on structures deposited in the Protein Data Bank (PDB). The immune epitope database (IEDB) provided this information in the most comprehensive way and makes it publicly available for download from their website. The limited amount of data required us to adjust our initial research question. B cells recognize not only protein antigens but also lipid, carbohydrate and basically any other structure or combination for which a BCR happens to be specific (Dowds et al. 2014). Because of the distinct chemical and structural differences between these antigens (Cobb and Kasper 2008, Taylor et al. 1986), the question was raised if also differences exist in BCRs sequences targeting epitopes on these different antigens.

First, we built a classifier to predict BCR interactions to peptidic and non-peptidic epitopes based on features derived from the heavy chain CDR3 region. Next, we predicted interactions with a more distinct group of epitopes that are part of lipopolysaccharide and polysaccharides structures and also added some stringency by including only mouse and human antibodies. The predictive quality of the random forest and nearest neighbour classifiers remained poor as true positive rates could only be achieved with high false positive rates. Nonetheless, features derived from heavy chain CDR3s and light chain CDR3s in general showed better results than features derived from CDR2 and CDR1 on the respective chains confirming the importance of the CDR3 region in the subject of antigen recognition (data not shown). Several remarks can be made to explain the poor predictive quality of the classifiers. First, both the non-peptidic and peptidic epitope groups contained a collection of epitopes that in general might share chemical properties but could still be very diverse in terms of epitopes targeted. We expected the kNN classifier to deal with this kind of noise better, but performance didn't improve maybe also due to the high dimensionality of the feature matrix (Raeisi Shahraki et al. 2017). Additionally, within the non peptidic group some structures might contain smaller peptide fragments as

part of glycoproteins. This diversity makes it harder for the classifiers to look for commonalities in the CDRs, if any, that are specific for epitopes in the two groups. Second, most non-peptidic antigens are T cell independent antigens which are generally bound by antibody with modest affinity (partly due to low degree of SHM (Kreutzmann et al. 2003)) and often require antibody cross-linking and additional signals from Toll like receptors (TLRs) to activate B cells (Sun et al. 2016, Bello-gil et al. 2019, Dowds et al. 2014). Random forest predictions for interactions with LPS and polysaccharides vs the 'rest' group showed better results but still struggled to reach higher recalls with decent precision. The presence of some higher ranked instances in this case might be attributed to the addition of *Pneumococcus* polysaccharide specific human CDR3s to this dataset making them overrepresented in the positive class. This could have led to the classifier specifically trained toward recognizing these CDR3s. Third, even though we used only mouse and human BCR data to train our classifier for LPS and polysaccharide epitopes, also mouse and human immune systems have significant differences (Mestas et al. 2004). They don't all affect BCR-epitope recognition, but some discrepancies exist that could have influenced our results. For example, the human HCDR3 region is generally longer and more diverse than its murine counterpart as a result of a higher number of amino acid changes due to somatic hypermutations and more abundant N-region addition (Zemlin et al. 2003, Shi et al. 2014). Lastly, recognition determinants in only one CDR alone may not be sufficient to predict accurate epitope binding specificities for antibodies using our approach (Adams et al. 2016).

We also collected HCDR3 sequences from different repertoires and clustered them using a generated Levenshtein distance matrix. It is important to note that the repertoires can represent a limited view of a B cell response to a particular antigen. The reason is that many researchers focus on and publish only the most neutralizing antibodies or antibodies with the highest affinities. Moreover, antibodies with lower affinities sometimes go unnoticed (Bobrovnik et al. 2010, Havenar-Daughton 2018) and repertoires also depend on what subset of B cells is studied, the origin of the sample and the individual studied (Shehata et al. 2019, Mroczek et al. 2014, Imkeller and Wardemann 2018, Allman et al. 2005). Because methods and protocols

varied between studies we collected data from, the basis of our findings is restricted to only the sequences themselves and the Levenshtein distance between them.
Within most repertoires a broad range of dissimilarity existed between CDR3s as was evident from the high mean distances in the repertoires. This is in line with previous findings that observed less than 50% sequence similarity in antigen specific CDR3s from different individuals (Wu et al. 2011, Prabakaran et al. 2012). Because Ig molecules can recognize nearly any antigenic structure and epitopes therein, either soluble or in membrane-bound form, through interactions with conformational or linear epitopes, this is not surprising. Additionally, the probability of finding identical antibody sequences in different individuals is reported to be extremely low even in monozygotic twins (Glanville et al., 2011). The thresholds used in the CDR3 clustering resulted in only a smaller fraction of sequences clustered, but most clusters are homogenous with respect to the repertoire and antigen they belong to. Also, the heterogenous clusters using both thresholds seem to cluster relevant sequences according to correlations and cross-reactivities published in the literature (see results).

Even though we couldn't sufficiently predict BCR-epitope interactions, the improved results for the classification of polysaccharide specific CDR3s is an indication that this machine learning approach using BCR sequence information is possible and will most certainly improve over the next couple of years as more data becomes available. The same goes for the clustering approach as indicated by the Dengue repertoire which showed clear separation in the distance matrix likely as a result of the large sample size of this study. Chances are that more insights into the way specific epitopes are recognized at the BCR-epitope interface could introduce new relevant features to be included in the models and alter the way the models are implemented.
Some of the heterogeneous repertoire clusters are difficult to interpret without any experimental evidence. There are also reasons to be to be careful with drawing conclusions based on CDR3 sequences alone. The CDR3 heavy chain region is only one determining region of an entire BCR but not the only one. Furthermore, identical CDR3 sequences can originate from rearrangements of different V and D genes and two antibodies with identical CDR3 regions can have different specificities (D'angelo et al. 2018). Still, most of the repertoires included were assessed using ELISA

methods with the antigen of interest. The findings here could be a starting point to test affinities of a single repertoire on different antigens according to the contents of the heterogeneous clusters reported here.

## 5. Conclusion

Our classifier initially showed poor predictive quality to distinguish CDR3s specific to epitopes from different types of antigen. However, quality improved in the follow up classifier with the addition of a larger distinct set of antigen specific CDR3s suggesting that performance for this kind of classifier could improve when more relevant data comes available. Unsupervised clustering of antigen specific HCDR3s showed that B cell responses to one antigen show substantial diversity and that different antigen specific repertoires can share similar CDR3 sequences. This clustering approach can be used to investigate and extract relevant information from different immunosequencing datasets and the results are initiatives for further research.

# 6. References


**Adams** RM, Mora T, Walczak AM, Kinney JB. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. Elife. 2016;5:e23156. Published 2016 Dec 30. doi:10.7554eLife.23156

**Al-Lazikani** B, Lesk AM, Chothia C. Standard con- formations for the canonical structures of Allman D, Miller JP. B cell development and receptor diversity during aging. Curr Opin Immunol. 2005;17:463-467.

**Barrios** Y, Jirholt P, Ohlin M. Length of the antibody heavy chain complementarity determining region 3 as a specificity- determining factor. JMol Recognit (2004) 17:332–8. doi:10.1002/jmr. 679106.

**Baumgarth**, N. (2013). How specific is too specific? B-cell responses to viral infections reveal the importance of breadth over depth. Immunol. Rev. 255, 82–94.

**Bello-gil,** D., Audebert, C., Olivera-ardid, S. & Pérez-cruz, M. The Formation of Glycan-Specific Natural Antibodies Repertoire in GalT-KO Mice Is Determined by Gut Microbiota. 10, 1–13 (2019).

**Bhattacharya**, S. et al. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. Sci. Data 5, 1–9 (2018).

**Bobrovnik**, S. A., Demchenko, M., Komisarenko, S. & Stevens, F. Traditional ELISA methods for antibody affinity determination fail to reveal the presence of low affinity antibodies in antisera: an alternative approach. J. Mol. Recognit. 23, 448–456 (2010).

**Bolotin**, D. A. et al. MiXCR: Software for comprehensive adaptive immunity profiling. Nat. Methods 12, 380–381 (2015).

**Breiman** L. Random forests. Mach Learn. 2001;45(1):5–32.

**Briney**, B., Inderbitzin, A., Joyce, C., and Burton, D.R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature 566, 393–397.

**Chailyan**,A., Tramontano,A. and Marcatili,P. (2012) A database of immunoglobulins with integrated tools: DIGIT. Nucleic Acids Res., 40, D1230–D1234.

**Chen**, J. et al. Proteomic Analysis of Pemphigus Autoantibodies Indicates a Larger, More Diverse, and More Dynamic Repertoire than Determined by B Cell Genetics. Cell Rep. 18, 237–247 (2017).

**Chi**, X. et al. Generation and Characterization of Human Monoclonal Antibodies Targeting Anthrax Protective Antigen following Vaccination with a Recombinant Protective Antigen Vaccine. Clin. Vaccine Immunol. 22, 553–560 (2015).

**Christley**, S. et al. VDJServer: A Cloud-Based Analysis Portal and Data Commons for Immune Repertoire Sequences and Rearrangements . Frontiers in Immunology 9, 976 (2018).

**Cobb** BA, Kasper DL. Characteristics of carbohydrate antigen binding to the presentation protein HLA-DR. Glycobiology. 2008;18(9):707–718. doi:10.1093/glycob/cwn050

**Corrie**, B. D. et al. iReceptor : A platform for querying and analyzing antibody / cell and T- - cell receptor repertoire data across federated repositories. 24–41 (2018). doi:10.1111/imr.12666

**Croote**, D., Darmanis, S., Nadeau, K.C., and Quake, S.R. (2018). High-affinity allergen-specific human antibodies cloned from single IgE B cell transcrip- tomes. Science 362, 1306–1309

**Cyster**, J.G., and Schwab, S.R. (2012). Sphingosine-1-phosphate and lympho- cyte egress from lymphoid organs. Annu. Rev. Immunol. 30, 69–94.

**D'Angelo** S, Ferrara F, Naranjo L, Erasmus MF, Hraber P, Bradbury ARM. Many Routes to an Antibody Heavy-Chain CDR3: Necessary, Yet Insufficient, for Specific Binding. Front Immunol. 2018;9:395. Published 2018 Mar 8. doi:10.3389/fimmu.2018.00395

**de Kruif**, J. et al. Human Immunoglobulin Repertoires against Tetanus Toxoid Contain a Large and Diverse Fraction of High-Affinity Promiscuous V H Genes. J. Mol. Biol. 387, 548–558 (2009)

**De Neuter**, N. et al. Memory CD4+ T cell receptor repertoire data mining as a tool for identifying cytomegalovirus serostatus. Genes Immun. 1 (2018). doi:10.1038/s41435-018-0035-y

**De Neuter**, N. et al. On the feasibility of mining CD8+ T cell receptor patterns underlying immunogenic peptide recognition. Immunogenetics 1–10 (2017). doi:10.1007/s00251-017-1023-5

**Degroeve**, S., Martens, L. & Jurisica, I. MS2PIP: A tool for MS/MS peak intensity prediction. Bioinformatics 29, 3199–3203 (2013).

**DeWitt**, W. S. et al. A public database of memory and naive B-cell receptor sequences. PLoS One 11, 1–18 (2016).

**Dodev**, T. S., Bowen, H., Shamji, M. H., Bax, H. J., Beavil, A. J., McDonnell, J. M., et al. (2015). Inhibition of allergen-dependent IgE activity by antibodies of the same specificity but different class. Allergy 70, 720–724. doi: 10.1111/all.12607

**Dondelinger**, M. et al. Understanding the Significance and Implications of Antibody Numbering and Antigen-Binding Surface/Residue Definition. Front. Immunol. 9, 1–15 (2018).

**Dowds** CM, Kornell SC, Blumberg RS, Zeissig S. Lipid antigens in immunity. Biol Chem. 2014;395(1):61–81. doi:10.1515/hsz-2013-0220

**Dunbar**, J. et al. SAbDab: The structural antibody database. Nucleic Acids Res. 42, 1140–1146 (2014).

**Edelman** GM, Benacerraf B. On structural and functional rela- tions between antibodies and proteins of the gamma-system. Proc Natl Acad Sci U S A (1962) 48:1035–42. doi:10.1073/pnas.48. 6.1035

**Ehrenmann**, F., Kaas, Q. & Lefranc, M. P. IMGT/3dstructure-DB and IMGT/domaingapalign: A database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MHcSF. Nucleic Acids Res. 38, 301–307 (2009).

**Elias**, J. E., Gibbons, F. D., King, O. D., Roth, F. P. & Gygi, S. P. Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat. Biotechnol. 22, 214–219 (2004).

**Ellebrecht**, C. T. et al. Autoreactive IgG and IgA B Cells Evolve through Distinct Subclass Switch Pathways in the Autoimmune Disease Pemphigus Vulgaris Article Autoreactive IgG and IgA. CellReports 24, 2370–2380 (2018).

**Ferdous, S. & Martin**, A. C. R. Original article AbDb : antibody structure database — a database of PDB-derived antibody structures. 1–9 (2018). doi:10.1093/database/bay040

**Finney**, J., Yeh, C.H., Kelsoe, G., and Kuraoka, M. (2018). Germinal center responses to complex antigens. Immunol. Rev. 284, 42–50.

**Friedensohn**, S., Khan, T. A. & Reddy, S. T. Advanced Methodologies in High-Throughput Sequencing of Immune Repertoires. Trends Biotechnol. 35, 203–214 (2017).

**Galli** SJ, Tsai M. IgE and mast cells in allergic disease. Nat Med. 2012;18(5):693–704. Published 2012 May 4. doi:10.1038/nm.2755

**Georgiou**, G. et. al. The promise and challenge of high-throughput sequencing of the antibody repertoire. 32, 158–168 (2014).

**Gielis**, S., Moris, P., Neuter, N. De, Bittremieux, W. & Ogunjimi, B. TCRex : a webtool for the prediction of T-cell receptor sequence epitope specificity. 6–9 (2018). doi:10.1101/373472

**Glanville** J, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci USA. 2009; 106:20216– 20221.

**Glanville**, J. et al. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proc. Natl. Acad. Sci. 108, 20066–20071 (2011).

**Greiff**, V. et al. Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires. J. Immunol. ji1700594 (2017). doi:10.4049/jimmunol.1700594

**Greiff**, V., Miho, E., Menzel, U. & Reddy, S. T. Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. Trends Immunol. 36, 738–749 (2015).

**Havenar-Daughton**, C., Abbott, R. K., Schief, W. R. & Crotty, S. When designing vaccines, consider the starting material: the human B cell repertoire. Curr. Opin. Immunol. 53, 209–216 (2018).

**He**, H. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21, 1263–1284 (2009).

**Honegger** A, Pluckthun A. Yet another numbering scheme for immunoglobulin variable domains: an automatic mod- eling and analysis tool. J Mol Biol (2001) 309:657–70. doi:10.1006/jmbi.2001.4662

**Imkeller, K. & Wardemann**, H. Assessing human B cell repertoire diversity and convergence. Immunol. Rev. 284, 51–66 (2018).

**Janda**, A., Bowen, A., Greenspan, N. S. & Casadevall, A. Ig Constant Region Effects on Variable Region Structure and Function. Front. Microbiol. 7, (2016).

**Kanyavuz**, A., Marey-Jarossay, A., Lacroix-Desmazes, S. & Dimitrov, J. D. Breaking the law: unconventional strategies for antibody diversification. Nat. Rev. Immunol. (2019). doi:10.1038/s41577-019-0126-7

**Khass** M, Blackburn T, Burrows PD, Walter MR, Capriotti E, Schroeder HW Jr. VpreB serves as an invariant surrogate antigen for selecting immunoglobulin antigen-binding sites. Sci Immunol. (2016) 1:aaf6628. doi: 10.1126/sciimmunol.aaf6628

**Kringelum** JV, Nielsen M, Padkjær SB, Lund O. Structural analysis of B-cell epitopes in antibody:protein complexes. Mol Immunol (2013) 53:24–34. doi:10.1016/j.molimm. 2012.06.001

**Kruetzmann** S, Rosado MM, Weber H, et al. Human immunoglob- ulin M memory B cells controlling Streptococcus pneumoniae infec- tions are generated in the spleen. J Exp Med. 2003;197:939-945

**Kuroda** D, Shirai H, Kobori M, Nakamura H. Structural classifica- tion of CDR-H3 revisited: a les- son in antibody modeling. Pro- teins (2008) 73:608–20. doi:10. 1002/prot.22087

**Lavinder**, J. J. et al. Identification and characterization of the constituent human serum antibodies elicited by vaccination. Proc. Natl. Acad. Sci. 111, 2259–2264 (2014).

**Lee**, J. et al. Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. Nat. Med. 22, 1456–1464 (2016).

**MacCallum** RM, Martin AC, Thornton JM. Antibody-antigen interactions: contact analysis and binding site topography. JMol Biol (1996) 262:732–45. doi:10.1006/ jmbi.1996.0548

**Marcou**, Q., Mora, T. & Walczak, A. M. High-throughput immune repertoire analysis with IGoR. Nat. Commun. 9, (2018).

**McCarthy**, K. R. et al. Memory B Cells that Cross-React with Group 1 and Group 2 Influenza A Viruses Are Abundant in Adult Human Repertoires. Immunity 48, 174-184.e9 (2018).

**Meffre**, E. et al. Maturational characteristics of HIV-specific antibodies in viremic individuals. JCI Insight 1, 1–17 (2016).

**Meng** T, Li X, Ao X, et al. Hemolytic Streptococcus may exacerbate kidney damage in IgA nephropathy through CCL20 response to the effect of Th17 cells. PLoS One. 2014;9(9):e108723. Published 2014 Sep 29. doi:10.1371/journal.pone.0108723

**Mestas**, J. & Hughes, C. C. W. Of mice and not men: differences between mouse and human immunology. J. Immunol. 172, 2731–8 (2004).

**Meysman**, P. et al. Sequence analysis On the viability of unsupervised T-cell receptor sequence clustering for epitope preference. 1–8 (2018). doi:10.1093/bioinformatics/bty821

**Mroczek** ES, Ippolito GC, Rogosch T, et al. Differences in the composition of the human antibody repertoire by B cell subsets in the blood. Front Immunol. 2014;5:96

**Muyldermans, S. & Smider,** V. V. Distinct antibody species: structural differences creating therapeutic opportunities. Curr. Opin. Immunol. 40, 7–13 (2016).

**Myhrinder**, A. et al. Molecular characterization of neoplastic and normal “sister” lymphoblastoid B-cell lines from chronic lymphocytic leukemia. Leuk. Lymphoma 54, 1769–1779 (2013).

**Ner-Gaon**, H., Melchior, A., Golan, N., Ben-Haim, Y. & Shay, T. JingleBells: A Repository of Immune-Related Single-Cell RNA–Sequencing Datasets. J. Immunol. 198, 3375 LP – 3379 (2017).

**Padlan** EA, Abergel C, Tipper JP. Identification of specificity- determining residues in antibodies. FASEB J (1995) 9:133–9

**Papa**, A., Karabaxoglou, D. & Kansouzidou, A. Acute West Nile virus neuroinvasive infections: cross-reactivity with dengue virus and tick-borne encephalitis virus. J. Med. Virol. 83, 1861–1865 (2011).

**Parameswaran**, P. et al. Convergent antibody signatures in human dengue. 13, 691–700 (2014).

**Pedregosa** et al. Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011

**Penfold** RS, Prendecki M, McAdoo S, Tam FW. Primary IgA nephropathy: current challenges and future prospects. Int J Nephrol Renovasc Dis. 2018;11:137–148. Published 2018 Apr 12. doi:10.2147/IJNRD.S129227

**Prabakaran** P, Zhu Z, Chen W, Gong R, Feng Y, Streaker E, Dimitrov D. Origin, diversity, and maturation of human antiviral antibodies analyzed by high-throughput sequencing. Front Microbiol. 2012; 3:277

**Reason**, D. C. et al. Domain specificity of the human antibody response to Bacillus anthracis protective antigen. Vaccine 26, 4041–4047 (2008).

**Saito, T. & Rehmsmeier**, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10, 1–21 (2015).

**Santoro, M. M. & Perno**, C. F. HIV-1 Genetic Variability and Clinical Implications. ISRN Microbiol. 2013, 1–20 (2013).

**Scheid**, J. F. et al. Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. Nature 458, 636–640 (2009)

**Schmitt** R, Carlsson F, Mörgelin M, Tati R, Lindahl G, Karpman D. Tissue deposits of IgA-binding streptococcal M proteins in IgA nephropathy and Henoch-Schonlein purpura. Am J Pathol. 2010;176(2):608–618. doi:10.2353/ajpath.2010.090428

**Sela-culang**, I., Kunik, V. & Ofran, Y. The structural basis of antibody-antigen recognition. Front. Immunol. 4, 1–13 (2013).

**Shahraki**, H., Pourahmad, S. & Zare, N. K Important Neighbors: A Novel Approach to Binary Classification in High Dimensional Data. Biomed Res. Int. 2017, 1–9 (2017).

**Shehata**, L. et al. Systematic comparison of respiratory syncytial virus-induced memory B cell responses in two anatomical compartments. Nat. Commun. doi:10.1038/s41467-019-09085-1

**Shi** B, Ma L, He X, et al. Comparative analysis of human and mouse immunoglobulin variable heavy regions from IMGT/LIGM-DB with IMGT/HighV-QUEST. Theor Biol Med Model. 2014;11:30. 2014.

**Smith**, K. et al. Fully human monoclonal antibodies from antibody secreting cells after vaccination with Pneumovax®23 are serotype specific and facilitate opsonophagocytosis. Immunobiology 218, 745–754 (2013).

**Stettler**, K. et al. Specificity, cross-reactivity, and function of antibodies elicited by Zika virus infection. Science (80-. ). 353, 823–826 (2016).

**Sun**, L., Middleton, D. R., Wantuch, P. L., Ozdilek, A. & Avci, F. Y. Carbohydrates as T-cell antigens with implications in health and disease. Glycobiology 26, 1029–1040 (2016).

**Taylor**, W. R. The classification of amino acid conservation. J. Theor. Biol. 119, 205–218 (1986).

**Tsioris**, K. et al. Neutralizing antibodies against West Nile virus identified directly from human B cells by single-cell analysis and next generation sequencing. Integr. Biol. 7, 1587–1597 (2015).

**Vita**, R. et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. 47, D339–D343 (2019).

**Wardemann, H. & Busse**, C. E. Novel Approaches to Analyze Immunoglobulin Repertoires. Trends Immunol. 38, 471–482 (2017).

**Williams**, L. D. et al. Potent and broad HIV-neutralizing antibodies in memory B cells and plasma. 2, (2018).

**Wu** X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, Chen X, Longo N, Louder M, McKee K, et al. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. Science. 2011; 333:1593–1602

**Yeap**, L.S., Hwang, J.K., Du, Z., Meyers, R.M., Meng, F.L., Jakubauskait _e, A., Liu, M., Mani, V., Neuberg, D., Kepler, T.B., et al. (2015). Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. Cell 163, 1124–1137.

**Zemlin**, M., M. Klinger, J. Link, C. Zemlin, K. Bauer, J. A. Engler, H.W. Schroeder, Jr., and P. M. Kirkham. 2003. Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. J. Mol. Biol. 334:733

**Zhang** W, Wang L, Liu K, Wei X, Yang K, Du W, Wang S, Guo N, Ma C, Luo L, et al. PIRD: Pan immune repertoire database. bioRxiv (2018) doi:10.1101/399493

**Zhou**, J., Lottenbach, K. R., Barenkamp, S. J. & Reason, D. C. Somatic hypermutation and diverse immunoglobulin gene usage in the human antibody response to the capsular polysaccharide of Streptococcus pneumoniae type 6B. Infect. Immun. 72, 3505–3514 (2004).

**Zhou**, J., Lottenbach, K. R., Barenkamp, S. J., Lucas, A. H. & Reason, D. C. Recurrent variable region gene usage and somatic mutation in the human antibody response to the capsular polysaccharide of Streptococcus pneumoniae type 23F. Infect. Immun. 70, 4083–4091 (2002).