PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Detection of disease-specific signatures in B cell repertoires of lymphomas using machine learning

Paul Schmidt-Barbo[1,2], **Gabriel Kalweit**[2,3], **Mehdi Naouar**[2,3], **Lisa Paschold**[4], **Edith Willscher**[4], **Christoph Schultheiß**[1], **Bruno Märkl**[5], **Stefan Dirnhofer**[6], **Alexandar Tzankov**[6], **Mascha Binder**[1,2,7☯*], **Maria Kalweit**[2,3☯]

1 Department of Biomedicine, Translational Immuno-Oncology, University Hospital Basel, Basel, Switzerland, **2** Collaborative Research Institute Intelligent Oncology (CRIION), Freiburg, Germany, **3** Neurorobotics Lab, University of Freiburg, Freiburg, Germany, **4** Internal Medicine IV, Oncology/Hematology, Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany, **5** Pathology, University Hospital Augsburg, Augsburg, Germany, **6** Pathology, University Hospital Basel, Basel, Switzerland, **7** Medical Oncology, University Hospital Basel, Basel, Switzerland

☯ These authors contributed equally to this work.
* mascha.binder@unibas.ch

Abstract

The classification of B cell lymphomas—mainly based on light microscopy evaluation by a pathologist—requires many years of training. Since the B cell receptor (BCR) of the lymphoma clonotype and the microenvironmental immune architecture are important features discriminating different lymphoma subsets, we asked whether BCR repertoire next-generation sequencing (NGS) of lymphoma-infiltrated tissues in conjunction with machine learning algorithms could have diagnostic utility in the subclassification of these cancers. We trained a random forest and a linear classifier via logistic regression based on patterns of clonal distribution, VDJ gene usage and physico-chemical properties of the top-n most frequently represented clonotypes in the BCR repertoires of 620 paradigmatic lymphoma samples—nodular lymphocyte predominant B cell lymphoma (NLPBL), diffuse large B cell lymphoma (DLBCL) and chronic lymphocytic leukemia (CLL)—alongside with 291 control samples. With regard to DLBCL and CLL, the models demonstrated optimal performance when utilizing only the most prevalent clonotype for classification, while in NLPBL—that has a dominant background of non-malignant bystander cells—a broader array of clonotypes enhanced model accuracy. Surprisingly, the straightforward logistic regression model performed best in this seemingly complex classification problem, suggesting linear separability in our chosen dimensions. It achieved a weighted F1-score of 0.84 on a test cohort including 125 samples from all three lymphoma entities and 58 samples from healthy individuals. Together, we provide proof-of-concept that at least the 3 studied lymphoma entities can be differentiated from each other using BCR repertoire NGS on lymphoma-infiltrated tissues by a trained machine learning model.

Citation: Schmidt-Barbo P, Kalweit G, Naouar M, Paschold L, Willscher E, Schultheiß C, et al. (2024) Detection of disease-specific signatures in B cell repertoires of lymphomas using machine learning. PLoS Comput Biol 20(7): e1011570. https://doi.org/10.1371/journal.pcbi.1011570

Editor: Stacey D. Finley, University of Southern California, UNITED STATES OF AMERICA

Received: October 5, 2023

Accepted: June 7, 2024

Published: July 2, 2024

Data Availability Statement: The data of our analysis can be accessed at the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB66357 (https://www.ebi.ac.uk/ena/browser/view/PRJEB66357). Corresponding code is available under https://github.com/paulovic96.

Funding: This project was funded by the Mertelsmann Foundation (grant to MB). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Lymphoma, a complex group of malignant blood cancers, poses a significant diagnostic challenge due to its diverse subtypes. Yet, precise classification is crucial for tailored treatment. In our research, we developed a machine learning algorithm and conducted comprehensive validation to discern distinct B cell lymphoma subtypes. We therefore leveraged B cell repertoires of lymphoma-infiltrated tissue, as ascertained through next-generation sequencing. Our data offers three key insights: We detail the creation and training of our machine learning algorithm, explaining how we selected features and designed the model. We demonstrate the algorithm's diagnostic precision using sequencing data from a test-set of patient samples. Moreover, through a deep dive into the most distinguishing aspects of our algorithm, we unveil distinctive disease-related patterns present within the malignant B cell and its surrounding environment. This analysis showed that both the malignant lymphoma cell, but also healthy bystander immune cells contribute to the distinctive architecture that characterizes a specific lymphoma subtype. We hope our work will contribute towards creating tools to diagnose lymphoma more easily and accurately ultimately leading to better outcomes for patients with this type of cancer.

Introduction

B cells are one of the essential pillars of the adaptive immune system that generates highly specific and also long-lasting immunity [1,2]. They originate from hematopoietic precursor cells in the bone marrow and acquire their characterizing feature–the B cell receptor (BCR)–in a multistep recombination and selection process [3,4]. Each BCR consists of a unique configuration of paired immunoglobulin heavy (IGH) and light (IGL) chains that mediate antigen binding specificities and upon engagement trigger, in concert with coactivator molecules, a cascade of signaling events that result in activation and proliferation [5,6]. Activated B cells can then differentiate into plasma cells, which produce and secrete immunoglobulins or antibodies to neutralize cognate antigen, or long-lived memory B cells that are capable to quickly mount high-affinity recall responses [7]. To guarantee an adequate arsenal of binders for the enormous breadth of foreign antigens, the immune system uses the process of immunoglobulin VDJ recombination to generate maximal sequence diversity [3,8]. During VDJ-recombination in developing immature B cells, randomly chosen variable (V), diversity (D) and joining (J) gene segments within the immunoglobulin loci are recombined to chromosomal sequences encoding a functional BCR [8]. This recombination process is facilitated via induced double-strand breaks and DNA repair/ligation mechanisms that may result in additional deletions or insertions that further increase sequence variance of single BCRs [8]. On the repertoire level, most of the immunoglobulin diversity is generated in the complementarity-determining region 3 (CDR3) sequence which spans the joined VDJ regions [9,10]. In addition, BCR diversity is boosted by somatic hypermutation (SHM), an iterative affinity maturation process that is initiated in response to antigen in the germinal centers (GCs) of secondary lymphoid tissues [7,11]. These transient but highly specialized microanatomical structures provide a dynamic environment that enables the proper coordination of repeated SHM and selection cycles to evolve polyreactive low-affinity BCRs into antibodies with maximum epitope selectivity [6,7,11].

Lymphomas represent hematological neoplasms of differentiated lymphocytes which typically originate in lymphatic tissue [12]. Interestingly, 95% of all lymphomas are found in B lineage cells [13]. Lymphomas are classified based on histopathological and clinical features that mostly depend on the putative cell of origin that has undergone malignant transformation [14]. The BCR takes a prominent role as oncogenic driver and mediator of sustained growth in these malignancies [14]. This is not only exemplified by the fact that a specific VDJ-rearrangement characterizes the malignant clonotype in a given patient but also by the acquisition of mutations that mimic chronic active BCR signaling [15]. Moreover, comparative analyses of a large number of prior sequencing studies have revealed prominent repertoire restrictions up to the extent of very similar or even identical CDR3 sequences across different patients with the same disease that are now recognized as subtype-specific sequence features for classification [16–18]. Furthermore, non-malignant bystander cells occupy varying space within the heterogenous tumor microenvironments of different lymphomas [19,20]. Some lymphomas bear very dominant malignant clonotypes, while in other lymphomas the malignant clonotype is less frequently found in a background of non-malignant bystander cells. These bystander immune cells such as T cells and B cells show a diverse range of other T-cell receptor-gene- or VDJ-rearrangements.

It is generally accepted that the correct diagnostic evaluation of lymphomas by the pathologist is complex in light of the numerous WHO-defined–sometimes rare–entities. Since recognition of the patterns of malignant and bystander cells needs a lot of expertise, evaluation by a consultation pathologist represents a standard in most centers. Here, we hypothesized that high-throughput analysis of the B cell architecture of different types of lymphomas may provide an additional basis for diagnosing disease subtypes. In this context, we set out to employ machine learning algorithms, which have demonstrated their capability to enhance the classification of immune states in antigen receptor-repertoire sequencing data, as valuable tools [21–25]. Since BCR next-generation sequencing (NGS) technology relies on unbiased amplification of all VDJ-rearrangements in the tissue of interest, also benign bystander lymphocytes are detected. Here, we present proof-of-concept that a logistic regression model is capable of differentiating between three paradigmatic lymphomas—nodular lymphocyte predominant B cell lymphomas (NLPBL), diffuse large B cell lymphoma (DLBCL), and chronic lymphocytic leukemia (CLL). This data shows the great potential of finding signatures in such large repertoire datasets consisting of the malignant clonotype and benign bystander cells that–clinically–until now remain largely unexploited.

Results

Characteristics of the lymphoma and control cohorts

In our study, we analyzed a total of 620 lymphoma BCR repertoire samples, comprising 90 from NLPBL (formerly nodular lymphocyte predominant Hodgkin lymphoma; NLPBL [26]), 182 from DLBCL, and 348 from CLL cases. We enriched the data with 291 control BCR repertoires derived from the blood of healthy donors (HD). Basic characteristics of the cohort are outlined in Table 1. Detailed information on individual BCR repertoire samples, including corresponding subjects, is provided in S1 Table.

Broad repertoire metrics in the lymphoma cohorts

In a first step, we compared general repertoire metrics between the four different groups. Blood BCR repertoires of HD were the most diverse and less clonal (Fig 1A–1C). Of the lymphoma cases, NLPBL patients showed lowest clonality, followed by patients with DLBCL and CLL (Fig 1A–1C). Of note, some DLBCL and CLL samples contained only the malignant clone and therefore exhibited a clonality of 1. While the term "repertoire" may not be entirely suitable in these cases, for consistency, we maintained this annotation throughout the paper. The percentage of somatically hypermutated clonotypes provides a rough estimate of how

Table 1. Sample characteristics.

Disease entity		
HD	Number of samples Type of samples Median age Sex distribution	291 100% PB 40 y 49% female, 51% male
NLPBL	Number of samples Type of samples Median age Sex distribution	90 2% BM, 85% LN, 1% SPL, 12% TM 37.5 y 30% female, 70% male
DLBCL	Number of samples Type of samples Median age Sex distribution	182 1.6% LV, 94.6% TM, 3.8% LN 71 y 46% female, 54% male
CLL	Number of samples Type of samples Median age Sex distribution	348 100% PB 66 y 30% female, 70% male

- PB = peripheral blood, TM = tumor, LN = lymph node, SPL = spleen, LV = liver, BM = bone marrow

https://doi.org/10.1371/journal.pcbi.1011570.t001

many clonotypes in the repertoire have undergone antigenic selection. The average somatic hypermutation across all BCR repertoires within each cohort was as follows: 19.2% in HD, 18.4% in NLPBL, 44.9% in DLBCL and 22.4% in CLL (Fig 1D). In addition to assessing the overall somatic hypermutation in BCR repertoires, we compared the somatic hypermutation levels of the top 10 clonotypes within each sample (S1 Fig).

Training of machine learning models on BCR repertoires of lymphoma tissue

In our pursuit of accurately classifying lymphoma subtypes based on BCR repertoire characteristics and as an initial benchmarking effort, we developed and trained two machine-learning models. In consideration of interpretability, we opted for two well-established and straightforward models: Logistic regression and random forest. These models were trained using a comprehensive feature set encompassing various aspects of clonotypes, including clonotype fractions, CDR3 sequence lengths, VDJ genes, and Kidera factors [27]. Additionally, we incorporated repertoire metrics such as clonality, Shannon diversity index, richness and the fraction of somatic hypermutation within each repertoire into the feature set. Although some of the metrics were correlated, as shown in the correlation matrix (S2 Fig), we included all relevant metrics to ensure a comprehensive and full understanding of the data, as each metric provides distinct non-overlapping information.

We applied different scenarios in which we challenged the models to discriminate between HD, NLPBL, DLBCL and CLL as well as different combinations thereof. For each scenario we started an individual training phase. For each training phase, the entire dataset comprising the individual classes was partitioned into a training subset (80%) and a test subset (20%). The splitting was performed in a stratified manner, ensuring the proportions of classes were maintained in both subsets, thus guaranteeing an unbiased evaluation of the developed models (Table 2). To account for the imbalance of the classes we used random oversampling, resampling all classes but the majority class.

Within the training phase, we implemented a robust model optimization strategy to avoid model overfitting. This incorporated a stratified k-fold cross-validation approach with k set to 3 folds. The stratification within the cross-validation approach ensured that the ratio of classes

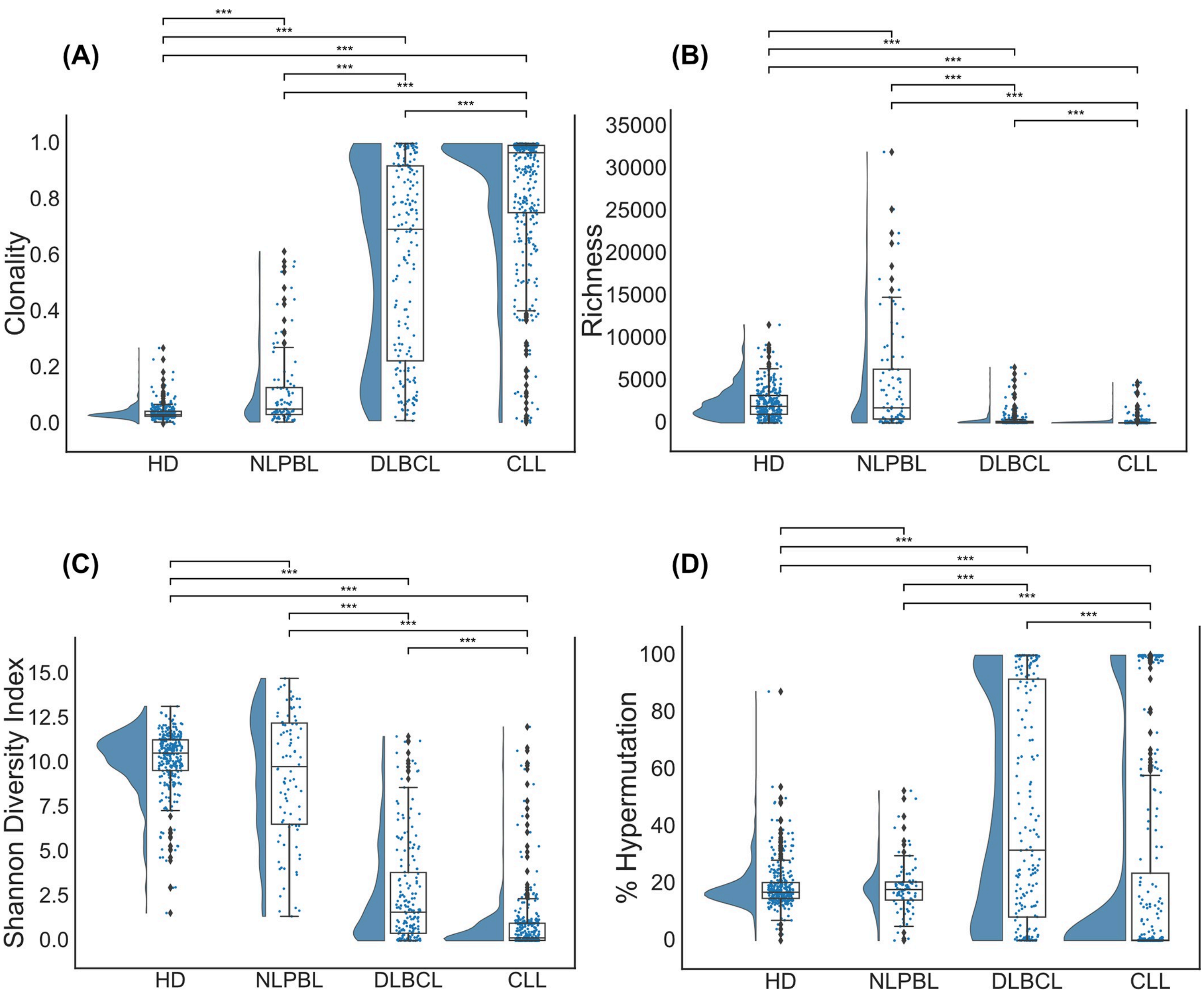

Fig 1. Broad BCR repertoire metrics in lymphoma and control cohorts. In the four panels, the essential repertoire metrics (A) clonality, (B) richness, (C) diversity, and (D) somatic hypermutation rate are shown with corresponding quantiles ($Q_{0.25}$, Median, $Q_{0.75}$). We pairwise performed a two-sided Mann-Whitney-U-Test with $\alpha = 0.05$ (*** $p < 0.001$).

https://doi.org/10.1371/journal.pcbi.1011570.g001

remained consistent across each fold and throughout the entire model training process. A grid search strategy was employed for hyperparameter tuning, traversing a predefined hyperparameter space. A list of resulting hyperparameters and the corresponding search spaces can be found in S2 Table. To assess the impact of bystander cells, we varied the number of top clonotypes considered in the calculation, ranging from 1 to 10, and subsequently extending to 20, 50 or 100 clonotypes.

Table 2. Numbers of BCR repertoires used for training.

	HD	NLPBL	DLBCL	CLL
Training	233	72	145	278
Test	58	18	37	70

https://doi.org/10.1371/journal.pcbi.1011570.t002

Validation of machine learning algorithm

During the training phase, the models were trained on ⅔ and validated on ⅓ of the training data. This form of validation is crucial in affirming the reliability, predictive capacity, and effectiveness of the models in differentiating the subsets based on the featured BCR repertoire [28]. The validation assessment was undertaken through the analysis of key metrics including accuracy, recall, precision and the F1 score. The evaluation of these performance metrics provided a comprehensive understanding of each model's ability to correctly classify the different cases.

Fig 2A illustrates the averaged validation results achieved by the best performing models for different numbers of clonotypes included in the calculation. While both models, the random forest and the logistic regression, appeared to achieve accurate classification of CLL, DLBCL,

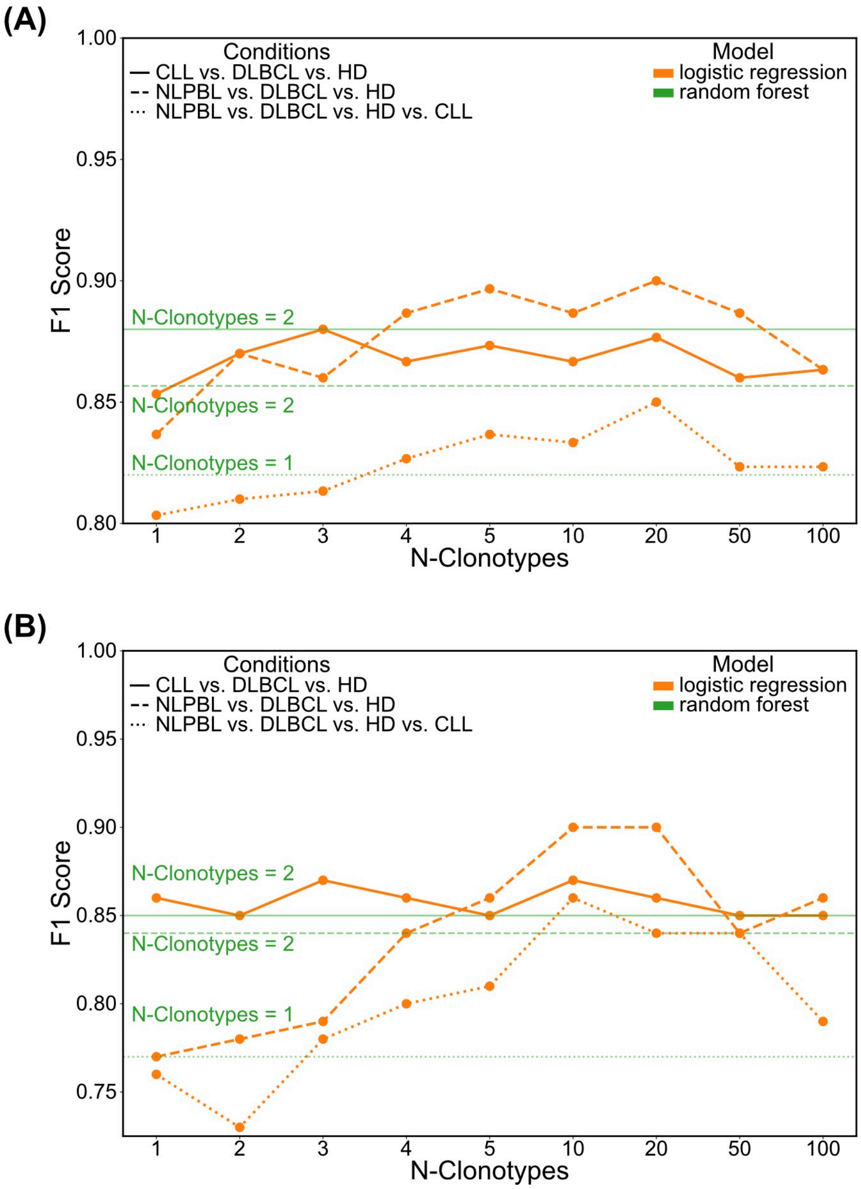

Fig 2. F1 scores of logistic regression models in training and test sets. (A) shows F1 scores averaged over validation folds during training in all three scenarios and (B) those of best validated logistic regression model on the test set in all three scenarios. As a comparison, the performance of the best random forest is displayed in green.

https://doi.org/10.1371/journal.pcbi.1011570.g002

and HD on the validation sets using information from the top two to three clonotypes alone, logistic regression exhibited improved performance with an increasing number of top repertoire clonotypes for all other cases. However, its performance began to decline when utilizing 50 or 100 clonotypes. In all comparisons including NLPBL, logistic regression demonstrated superior performance to random forest when considering information from the top 4–20 clonotypes. The observation that especially discriminating NLPBL benefits from the inclusion of a larger number of top repertoire clonotypes aligns well with the recognized significance of the bystander lymphocyte repertoire in this entity.

Final testing of the machine learning models

Upon completing the training phase, the model that exhibited the highest performance based on validation results was further trained on the entire training dataset and evaluated on a previously unseen test dataset. Table 3 showcases crucial metrics, encompassing accuracy, recall, precision, and the F1 score.

Fig 2B shows the results of the best validated models on the test data not available during the training phase. Similar to the validation results, we saw superior performance of the logistic regression models. While the first scenario of CLL, DLBCL and HD could be accurately classified by a logistic model only using information of the top three clonotypes, in the other scenarios, the logistic regression heavily benefits from the information of the top 4–20 clonotypes. Although the best performing models with respect to the validation results used 20 clonotypes, models using the information of only 10 clonotypes performed equally or even better on the test set.

Data separation for n = 1 to n = 100 clonotypes

To gain a more comprehensive insight into the model's performance, we conducted Principal Component Analysis (PCA) on our feature list while varying the number of top repertoire

Table 3. Validation of best models on the independent test set.

HD vs. DLBCL vs. CLL **Logistic Regression n-clonotypes = 3**	**Precision**	**Recall**	**F1**	**N**
CLL	0.82	0.94	0.88	70
DLBCL	0.85	0.59	0.70	37
HD	0.95	0.97	0.96	58
Accuracy			0.87	
Weighted Avg. F1			0.87	
HD vs. NLPBL vs. DLBCL Logistic Regression n-clonotypes = 20	Precision	Recall	F1	N
NLPBL	0.80	0.67	0.73	18
DLBCL	0.94	0.86	0.90	37
HD	0.91	1.00	0.95	58
Accuracy			0.90	
Weighted Avg. F1			0.90	
HD vs. NLPBL vs. DLBCL vs. CLL Logistic Regression n-clonotypes = 20	Precision	Recall	F1	N
NLPBL	0.80	0.67	0.73	18
DLBCL	0.84	0.57	0.68	37
HD	0.88	1.00	0.94	58
CLL	0.84	0.93	0.88	35
Accuracy			0.85	
Weighted Avg. F1			0.84	

https://doi.org/10.1371/journal.pcbi.1011570.t003 clonotypes included in the analysis. Given the clinical relevance of distinguishing between multiple lymphoma subtypes, our primary focus was on the scenario that encompassed all four groups. When visualizing the data in the first two dimensions, we observed overlapping clusters of lymphoma subtypes across different numbers of top repertoire clonotypes (Fig 3A–3D). Notably, NLPBL BCR repertoires samples exhibited significant overlap with those from HD, whereas DLBCL repertoire samples appeared to overlap more with those from CLL. Examining the various embeddings, it became evident that the problem was not perfectly linearly separable along the dimensions of the first and second principal components. However, we did observe the formation of distinct clusters within these two dimensions. Corroborating our performance findings, the most significant separations were achieved when considering the top 4–20 repertoire clonotypes.

Exploration of the factors contributing most to correct lymphoma classification

Next, we wished to explore, which of the features contributed most to correct lymphoma classification. We show the 20 predictors with greatest coefficient magnitude averaged over all classes in the best performing model for the comparison of NLPBL, DLBCL with CLL and HD (Fig 4A). This overview helps to understand the overall importance and strength of each predictor variable for the model across all classes. We further analyze the contribution of each predictor between pairs of cohorts (Fig 4B).

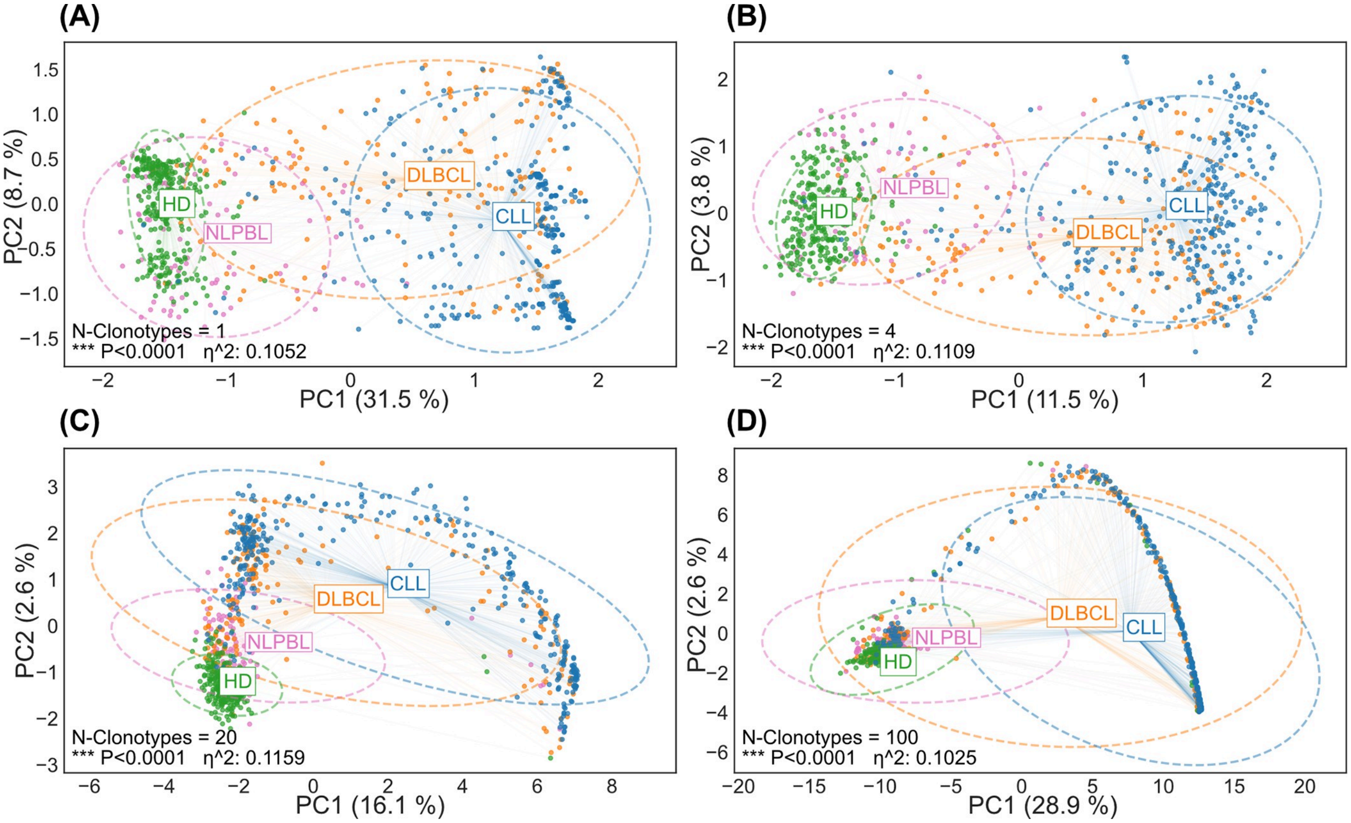

Fig 3. Data separation using n = 1 to n = 100 top repertoire clonotypes. Principal Component Analysis (PCA) was performed on the feature list while varying the number of top repertoire clonotypes included in the analysis. (A) n = 1 clonotype, (B) n = 4 clonotypes, (C) n = 20 clonotypes, (D) n = 100 clonotypes. We compared sample means using a multivariate analysis of variance (MANOVA).

https://doi.org/10.1371/journal.pcbi.1011570.g003

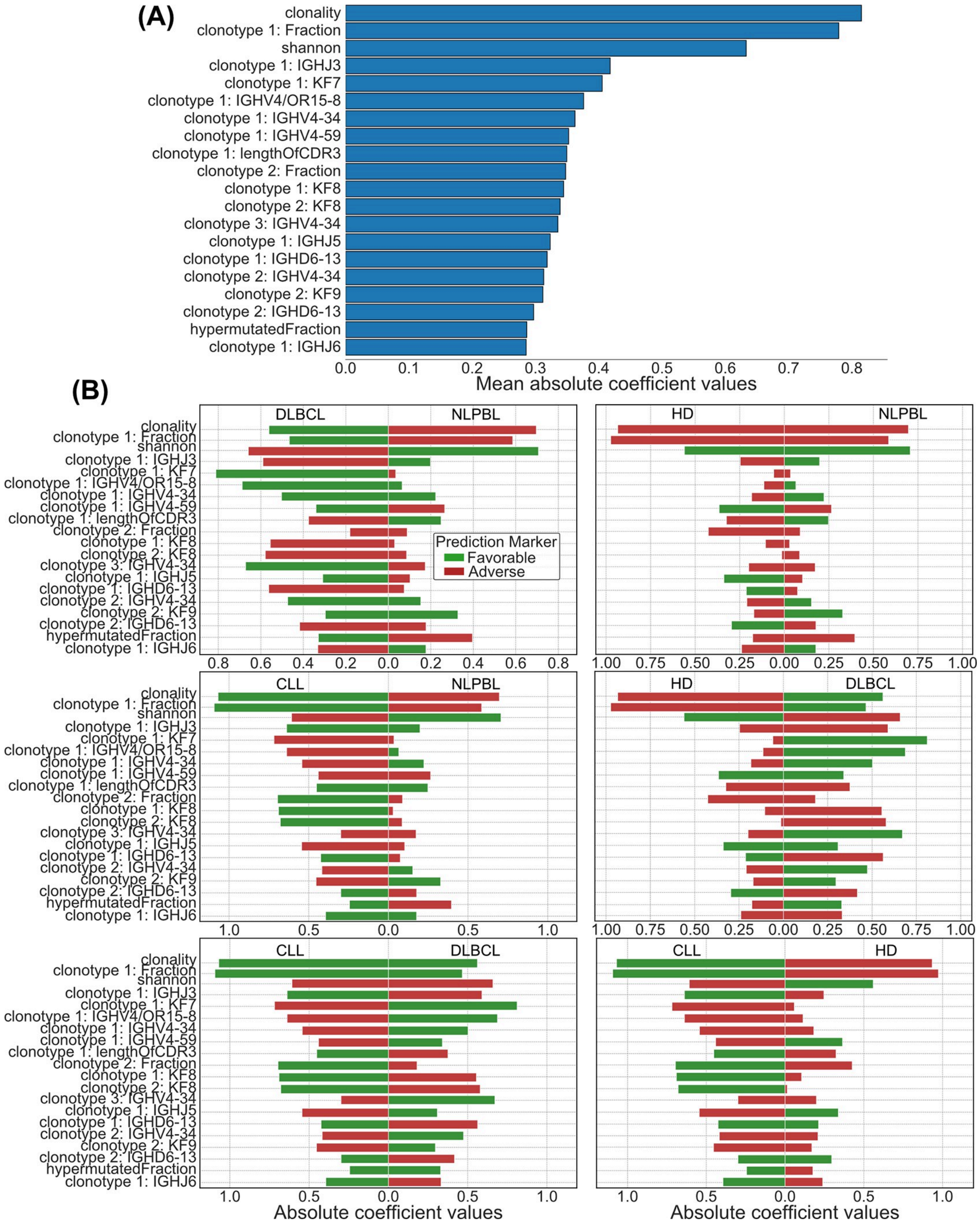

Fig 4. 20 lymphoma subset predictors with greatest coefficient magnitude. (A) Predictors were averaged over all classes in the best performing model for discrimination of HD vs. NLPBL vs. DLBCL vs. CLL. (B) Contribution of each predictor to the discrimination between pairs of cohorts.

https://doi.org/10.1371/journal.pcbi.1011570.g004

The three most important predictors of the logistic regression were the frequency of the most abundant clonotype in the repertoire and crude repertoire metrics (clonality, Shannon diversity). While the fraction of somatic hypermutation within a repertoire belonged to the 20 most dominant predictors, repertoire richness seemed to play a minor role in the discrimination between NLPBL, DLBCL, CLL and HD.

Moreover, the model predictions relied on features such as biochemical properties (Kidera factors) and length of the CDR3 sequence of the most dominant clonotypes. Interestingly, we found that Kidera Factor 7, which is associated with flat and extended conformations, and Kidera Factor 9, which represents the partial positive charge of the side chain, served as predictors for DLBCL. In contrast, Kidera factor 8, which reflects the alpha-helical secondary structure, was associated with CLL. The length of the CDR3 of the most dominant clonotype was associated with NLPBL.

Finally, we found specific gene usages of the most dominant clonotypes to add to the discriminative power of the model. For instance, the expression of IGHV4/OR15-8 and IGH4-34 by the most dominant clonotype was predictive of DLBCL and NLPBL, while IGHJ3 was associated with NLPBL and CLL.

Discussion

We set out to develop a machine-learning tool capable of distinguishing between different types of lymphomas by analyzing the B-cell architecture within lymphoma-infiltrated tissue using BCR-repertoire NGS data. Interestingly, our data revealed that a simple logistic regression with the right set of predictors performed exceptionally well in achieving this goal. It demonstrated a high degree of accuracy when discriminating between the three lymphoma types in our test dataset.

The factors contributing to this successful discrimination were, to some extent, as we anticipated. For example, we found that broad repertoire metrics carried significant discriminatory value. This result aligns with our expectations, especially in diseases characterized by varying levels of immune bystander cells, where such metrics naturally play a role. For this proof-of-concept study, we intentionally selected lymphoma entities characterized by distinct microenvironments. It is conceivable that fitting the algorithm on lymphoma entities with greater similarities could potentially diminish the discriminatory capacity of the basic repertoire metrics, while highlighting the significance of more specific features.

What made our findings particularly intriguing was the discovery that while we saw some discriminative features, which might have been selected based on prior knowledge of the malignant clonotype, some were not as much prioritized by this regression model. For instance, the classical VDJ-rearrangement known to be present in the malignant NLPBL clonotype (V3D3J6) [29, 30] coupled with an unusually long CDR3 sequence length was not prioritized by the algorithm to the extent we would have expected. While one might have anticipated a clear separation based on these features, the discrimination between NLPBL and other lymphomas, such as DLBCL, relied instead quite strongly on other immune repertoire metrics. Overall, most of the discriminatory power came from features of the most dominant clonotypes in the repertoire, yet there was quite some weight on global repertoire metrics underscoring the complexity of lymphoid malignancies. The analysis of patterns of non-malignant bystander cells—unrecognized by traditional bioinformatic analysis—may perspectively shed more light on the immunobiology of lymphoma and may therefore even have broader implications beyond diagnostics.

From a technical perspective, it needs to be noted that several key predictors in our model exhibited correlation with each other, notably observed in the broad repertoire metrics of diversity and clonality. Generally, higher diversity tends to correlate with lower clonality, and vice versa, within biological systems. This relationship arises from the fact that higher diversity suggests a wider range of distinct clones within a population, naturally resulting in a decreased dominance of any single clone (lower clonality). Conversely, lower diversity implies a greater proportion of cells originating from a limited number of ancestral clones (higher clonality).

However, it's crucial to recognize that the association between diversity and clonality can vary depending on the specific context and biological system under investigation. For instance, in certain disease states or immune responses, an increase in diversity might coincide with an increase in clonality if certain clones are selectively expanded in response to specific stimuli. Thus, although correlated, these metrics provide distinct and non-overlapping information. To ensure a comprehensive understanding of the data, we included all relevant metrics in our analysis. While focusing on the comparison between logistic Regression and Random Forest models for fixed numbers of clonotypes, we performed hyperparameter tuning using a grid search approach. We opted for a predefined comprehensive feature set over other feature selection methods to manage computational resource effectively. While acknowledging the potential risk of overfitting, we mitigated this concern by employing a cross-validation regime.

In terms of model complexity, the robust performance of logistic regression and random forest models suggests that the problem at hand may initially appear linearly separable. However, it is noteworthy that logistic regression starts to show limitations when the number of clonotypes exceeds 20, rendering the problem non-linearly separable. This trend prompts the exploration of deep learning models, which are well-suited for handling complex, non-linear relationships within the data. Deep models, such as neural networks, have the capacity to capture intricate patterns and relationships within the data, making them a promising avenue for further exploration. Yet, larger sequencing datasets would be needed to be able to apply deep learning models. In this context, the role of data quantity in optimizing predictive power cannot be overstated. Accumulating more data, particularly diverse and representative samples, is instrumental in bolstering the performance of machine learning models. Additionally, careful consideration of sampling strategies is crucial. Properly balanced and stratified sampling can mitigate issues related to class imbalances and ensure that the model is trained on a comprehensive spectrum of cases. By addressing these aspects, we could refine our machine-learning approach to achieve higher diagnostic accuracy in lymphoma subtyping and potentially even uncover biologically valuable insights into lymphoma clonotypes and their microenvironment.

Our dataset should be considered within the broader framework of evolving research on the application of machine-learning algorithms for lymphoma diagnosis and subtyping. Numerous studies are emerging in the field of digital hematopathology [31–35], and an increasing body of data underscores the pivotal role of machine-learning in harmonizing sequencing data related to lymphoma driver-genes [21–25]. Our approach introduces a novel dimension by incorporating immune architecture as an additional layer of analysis. In this respect, the potential of AI may be particularly significant when dealing with challenging scenarios such as suboptimal samples characterized by their small or squeezed nature. In such cases, the BCR repertoire analysis may prove to be less susceptible to interpretation errors compared to traditional morphological assessments. Moreover, specific situations may stand to gain significant benefits, e.g. Richter's transformation that can sometimes be quite complex to diagnose, while in other instances discerning CLL from DLBCL poses fewer challenges. As an added diagnostic tool in complex cases, it also holds the potential to address the widely recognized shortage of personnel in the field of pathology, which has garnered significant attention within the medical community [36].

Together, our study represents a significant advancement in the field as we demonstrate a compelling proof-of-concept: the ability to differentiate between distinct lymphoma entities by leveraging BCR repertoire NGS on lymphoma-infiltrated tissues through the application of a trained machine learning model. This paves the way for further research and potential clinical applications. In the future, our approach may become a component of digital lymphoma diagnostics as an efficient resource to complement conventional techniques.

Methods

Ethics statement

New CLL BCR repertoire data has been acquired in a study reviewed by the Ethics Committee of the Martin-Luther-University Halle-Wittenberg after informed written consent (project number 2014–75). The BCR repertoire data of DLBCL samples derives from anonymized paraffin-embedded tissue sections. All studies were conducted in accordance with the ethical principles stated by the Declaration of Helsinki.

Study design

Most of the BCR repertoires have already been deposited in the context of different projects [29,37–49] listed in S1 Table. The so far unpublished BCR repertoires are deposited together with the previously published ones along with this manuscript to ensure replicability of our results for the comprehensive cohort in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB66357 (https://www.ebi.ac.uk/ena/browser/view/PRJEB66357).

Sample collection, DNA preparation and NGS of BCR repertoires

Peripheral mononuclear cells (PBMC) were isolated from blood of CLL patients or HD by standard density-gradient centrifugation using Ficoll. Genomic DNA was extracted from PBMCs using the GenElute Mammalian Genomic DNA Miniprep Kit (Sigma-Aldrich, St. Louis, USA). In DLBCL and NLPBL patients, DNA was extracted from paraffin-embedded lymphoma-infiltrated tissue as previously described [29].

We used a multiplex PCR based on BIOMED-FR1 primer pool to amplify VDJ rearranged immunoglobulin heavy chain (IGH) loci from 250 ng of genomic DNA. Purified amplicons were pooled at 4 nM, quality-assessed on a 2100 Bioanalyzer (Agilent Technologies) and sequenced on an Illumina MiSeq (paired-end, 2 x 301-cycles, v3 chemistry). Sequence reads were mapped to genomic V, D, J reference sequences using the MiXCR framework [50]. As reference for sequence alignment, the IMGT library v3 was used. For analysis, we defined each unique complementarity-determining region 3 (CDR3) nucleotide sequence as a clonotype. Non-productive reads and sequences with less than 2 read counts were discarded. IGHV genes that showed $\leq$ 98% identity to the germline sequence were considered somatically hypermutated.

Immune repertoire metrics

We determined the clonality of the sequenced IGH repertoires using the formula “1- Pielou’s evenness” [51,52]. In our context, evenness quantifies the relative prevalence of distinct B cell types within each repertoire. It is calculated according to the formula $J = H'/\log2(S)$ with H’ being the Shannon diversity index [53] and S the total clonotype number (richness) [54] in a distinct sample. A clonality index of 1 indicates that the analyzed sample comprises a single clonotype, while 0 indicates complete clonal diversity. Since richness, clonality, and Shannon diversity showed negligible to no correlation with total read count, we decided to utilize raw sequencing data without applying any read normalization techniques ($r_{richness} = 0.18$, $p < 0.001$; $r_{clonality} = 0.04$, $p = 0.22$; $r_{Shannon} = 0.08$, $p = 0.01$).

Machine learning

Data preprocessing. For each clonotype in all repertoires, we extracted the clonotype fraction, lengths of CDR3 sequences and the VDJ arrangement. Since VDJ genes are categorical features, we applied one-hot-encoding. Representing the theoretical physico-chemical properties, we calculated the ten Kidera Factors from the individual amino acid sequence of each CDR3 region and augmented the dataset with the clonality, Shannon diversity index and richness for each repertoire as described. In order to deal with the variable number of clonotypes we set a fixed number of clonotypes per repertoire ranging from n = 1 to 10 and subsequently extending to 20, 50 or 100 clonotypes. Based on their clonotype fractions, we extracted the n most dominant clonotypes within each repertoire while discarding the remaining clonotypes. For repertoires with fewer than n clonotypes we applied zero-padding to ensure consistent input dimensions across all samples. We concatenated the features of the n clonotypes to a single vector representing a single repertoire within 111 (n = 1) up to 10151 (n = 100 dimensions. Depending on a freely varying hyperparameter we scaled the numerical feature across all vectors to have zero mean and unit variance. We partitioned the data into a training (80%) and test set (20%) ensuring the proportions of classes were maintained in both subsets.

Training

Within the training process we fitted two different model types in a supervised manner. The first model was a random forest consisting of an ensemble of single decision trees. The second model consists of a logistic regression minimizing a multinomial loss. While focusing on the comparison between both models we performed hyperparameter tuning using a grid search approach. We opted for a predefined comprehensive feature set over feature selection methods to manage computational resource effectively. While acknowledging the potential risk of overfitting due to feature correlation (S2 Fig), we incorporated all relevant features available to ensure the most comprehensive information about the data. We tried to mitigated potential pitfalls by using a stratified k-fold cross-validation approach with k set to 3 folds fitting each model with a different set of hyperparameters (S2 Table) to a subset of the training data. We calculated the performance in form of the F1 score on the remaining part of the training set and averaged the obtained scores over all folds. The best performing settings of hyperparameters were chosen and used to train the model on the entire training set.

Testing

The best performing models with respect to the training phase were used to predict unseen data from the separate test set. We compared the predictions of each model to the known labels and calculated a final F1 score.

All analyses and data plotting were conducted using RStudio (version 1.1.456) or Python (version 3.11.5) within a Conda environment. Model fitting and evaluation were executed using scikit-learn 1.3.0. All computations were carried out on a MacBook Pro featuring an M1 processor, with Kernel Version Darwin 22.6.0 and macOS 13.5.2.

All code is available on github.com/paulovic96.

Supporting information

S1 Fig. Somatic hypermutation rate of top 10 clonotypes. Somatic hypermutation rate calculated based on the 10 most frequent clonotypes per repertoire.
(TIFF)

S2 Fig. Feature correlation Matrix. Correlation of major immune repertoire metrics along with strongest predictors from the best performing model for discrimination of HD vs. NLPBL vs. DLBCL vs. CLL.
(TIFF)

S1 Table. Comprehensive table of data samples.
(XLS)

S2 Table. Hyperparameter settings.
(DOCX)

Acknowledgments

We thank the lymphoma patients and healthy donors for their great support. Moreover, we thank Adeline Stiefvater, Christoph Wosiek, Aline Patzschke, Jenny Wehde, Yiqing Du and Katrin Nerger for excellent technical assistance. We also thank Roland Mertelsmann for his valuable advice and the CRIION for core AI support.

Author Contributions

Conceptualization: Mascha Binder, Maria Kalweit.

Data curation: Paul Schmidt-Barbo, Edith Willscher.

Formal analysis: Paul Schmidt-Barbo, Gabriel Kalweit, Stefan Dirnhofer, Alexandar Tzankov, Mascha Binder, Maria Kalweit.

Investigation: Paul Schmidt-Barbo, Maria Kalweit.

Methodology: Paul Schmidt-Barbo, Gabriel Kalweit, Mehdi Naouar, Maria Kalweit.

Project administration: Mascha Binder, Maria Kalweit.

Resources: Lisa Paschold, Edith Willscher, Christoph Schultheiß, Bruno Märkl, Stefan Dirnhofer, Alexandar Tzankov, Mascha Binder.

Software: Paul Schmidt-Barbo, Gabriel Kalweit, Mehdi Naouar, Maria Kalweit.

Supervision: Mascha Binder, Maria Kalweit.

Validation: Paul Schmidt-Barbo, Gabriel Kalweit, Christoph Schultheiß, Mascha Binder, Maria Kalweit.

Visualization: Paul Schmidt-Barbo.

Writing – original draft: Paul Schmidt-Barbo, Mascha Binder, Maria Kalweit.

Writing – review & editing: Paul Schmidt-Barbo, Gabriel Kalweit, Mehdi Naouar, Lisa Paschold, Edith Willscher, Christoph Schultheiß, Bruno Märkl, Stefan Dirnhofer, Alexandar Tzankov, Mascha Binder, Maria Kalweit.

References

1. Bonilla FA, Oettgen HC. Adaptive immunity. J Allergy Clin Immunol. 2010; 125(2 Suppl 2):S33–40. https://doi.org/10.1016/j.jaci.2009.09.017 PMID: 20061006
2. Cooper MD. The early history of B cells. Nat Rev Immunol. 2015; 15(3):191–7. https://doi.org/10.1038/nri3801 PMID: 25656707
3. Wang Y, Liu J, Burrows PD, Wang JY. B Cell Development and Maturation. Adv Exp Med Biol. 2020; 1254:1–22. https://doi.org/10.1007/978-981-15-3532-1_1 PMID: 32323265
4. Pieper K, Grimbacher B, Eibel H. B-cell biology and development. J Allergy Clin Immunol. 2013; 131 (4):959–71. https://doi.org/10.1016/j.jaci.2013.01.046 PMID: 23465663
5. Tanaka S, Baba Y. B Cell Receptor Signaling. Adv Exp Med Biol. 2020; 1254:23–36. https://doi.org/10.1007/978-981-15-3532-1_2 PMID: 32323266
6. Cyster JG, Allen CDC. B Cell Responses: Cell Interaction Dynamics and Decisions. Cell. 2019; 177 (3):524–40. https://doi.org/10.1016/j.cell.2019.03.016 PMID: 31002794

7. De Silva NS, Klein U. Dynamics of B cells in germinal centres. Nat Rev Immunol. 2015; 15(3):137–48. https://doi.org/10.1038/nri3804 PMID: 25656706

8. Chi X, Li Y, Qiu X. V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. Immunology. 2020; 160(3):233–47. https://doi.org/10.1111/imm.13176 PMID: 32031242

9. Zheng B, Yang Y, Chen L, Wu M, Zhou S. B-cell receptor repertoire sequencing: Deeper digging into the mechanisms and clinical aspects of immune-mediated diseases. iScience. 2022; 25(10):105002. https://doi.org/10.1016/j.isci.2022.105002 PMID: 36157582

10. Kovaltsuk A, Krawczyk K, Galson JD, Kelly DF, Deane CM, Truck J. How B-Cell Receptor Repertoire Sequencing Can Be Enriched with Structural Antibody Data. Front Immunol. 2017; 8:1753. https://doi.org/10.3389/fimmu.2017.01753 PMID: 29276518

11. Young C, Brink R. The unique biology of germinal center B cells. Immunity. 2021; 54(8):1652–64. https://doi.org/10.1016/j.immuni.2021.07.015 PMID: 34380063

12. Bispo JAB, Pinheiro PS, Kobetz EK. Epidemiology and Etiology of Leukemia and Lymphoma. Cold Spring Harb Perspect Med. 2020; 10(6). https://doi.org/10.1101/cshperspect.a034819 PMID: 31727680

13. Meng X, Min Q, Wang JY. B Cell Lymphoma. Adv Exp Med Biol. 2020; 1254:161–81.

14. Seifert M, Scholtysik R, Kuppers R. Origin and Pathogenesis of B Cell Lymphomas. Methods Mol Biol. 2019; 1956:1–33. https://doi.org/10.1007/978-1-4939-9151-8_1 PMID: 30779028

15. Kuppers R. Mechanisms of B-cell lymphoma pathogenesis. Nat Rev Cancer. 2005; 5(4):251–62. https://doi.org/10.1038/nrc1589 PMID: 15803153

16. Agathangelidis A, Psomopoulos F, Stamatopoulos K. Stereotyped B Cell Receptor Immunoglobulins in B Cell Lymphomas. Methods Mol Biol. 2019; 1956:139–55. https://doi.org/10.1007/978-1-4939-9151-8_7 PMID: 30779034

17. Agathangelidis A, Vardi A, Baliakas P, Stamatopoulos K. Stereotyped B-cell receptors in chronic lymphocytic leukemia. Leuk Lymphoma. 2014; 55(10):2252–61. https://doi.org/10.3109/10428194.2013.879715 PMID: 24397617

18. Schroers-Martin JG, Alig S, Garofalo A, Tessoulin B, Sugio T, Alizadeh AA. Molecular Monitoring of Lymphomas. Annu Rev Pathol. 2023; 18:149–80. https://doi.org/10.1146/annurev-pathol-050520-044652 PMID: 36130071

19. Hopken UE, Rehm A. Targeting the Tumor Microenvironment of Leukemia and Lymphoma. Trends Cancer. 2019; 5(6):351–64. https://doi.org/10.1016/j.trecan.2019.05.001 PMID: 31208697

20. Menter T, Tzankov A. Lymphomas and Their Microenvironment: A Multifaceted Relationship. Pathobiology. 2019; 86(5–6):225–36. https://doi.org/10.1159/000502912 PMID: 31574515

21. Albitar M, Xu-Monette ZY, Shahbaba B, De Dios I, Wang Y, Manman D, et al. Cell of Origin Classification of DLBCL Using Targeted NGS Expression Profiling and Deep Learning. Blood. 2019; 134 (Supplement_1):2891–.

22. Kanduri C, Pavlović M, Scheffer L, Motwani K, Chernigovskaya M, Greiff V, et al. Profiling the baseline performance and limits of machine learning models for adaptive immune receptor repertoire classification. bioRxiv. 2021:2021.05.23.445346.

23. Widrich M, Schäfl B, Pavlović M, Ramsauer H, Gruber L, Holzleitner M, et al. Modern Hopfield Networks and Attention for Immune Repertoire Classification. bioRxiv. 2020:2020.04.12.038158.

24. Bobée V, Drieux F, Marchand V, Sater V, Veresezan L, Picquenot J-M, et al. Combining gene expression profiling and machine learning to diagnose B-cell non-Hodgkin lymphoma. Blood Cancer Journal. 2020; 10(5):59. https://doi.org/10.1038/s41408-020-0322-5 PMID: 32444689

25. Xu-Monette ZY, Zhang H, Zhu F, Tzankov A, Bhagat G, Visco C, et al. A refined cell-of-origin classifier with targeted NGS and artificial intelligence shows robust predictive value in DLBCL. Blood Advances. 2020; 4(14):3391–404. https://doi.org/10.1182/bloodadvances.2020001949 PMID: 32722783

26. Tousseyn TA, King RL, Fend F, Feldman AL, Brousset P, Jaffe ES. Evolution in the definition and diagnosis of the Hodgkin lymphomas and related entities. Virchows Arch. 2023; 482(1):207–26. https://doi.org/10.1007/s00428-022-03427-z PMID: 36274093

27. Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. Journal of Protein Chemistry. 1985; 4:23–55.

28. Stone M. Cross-validatory choice and assessment of statistical predictions. Journal of the royal statistical society: Series B (Methodological). 1974; 36(2):111–33.

29. Paschold L, Willscher E, Bein J, Vornanen M, Eichenauer DA, Simnica D, et al. Evolutionary clonal trajectories in nodular lymphocyte-predominant Hodgkin lymphoma with high risk of transformation. haematologica. 2021; 106(10):2654. https://doi.org/10.3324/haematol.2021.278427 PMID: 33882641

30. Thurner L, Hartmann S, Fadle N, Regitz E, Kemele M, Kim Y-J, et al. Lymphocyte predominant cells detect Moraxella catarrhalis-derived antigens in nodular lymphocyte-predominant Hodgkin lymphoma. Nature communications. 2020; 11(1):2465. https://doi.org/10.1038/s41467-020-16375-6 PMID: 32424289

31. Li D, Bledsoe JR, Zeng Y, Liu W, Hu Y, Bi K, et al. A deep learning diagnostic platform for diffuse large B-cell lymphoma with high accuracy across multiple hospitals. Nature Communications. 2020; 11 (1):6004. https://doi.org/10.1038/s41467-020-19817-3 PMID: 33244018

32. El Achi H, Belousova T, Chen L, Wahed A, Wang I, Hu Z, et al. Automated diagnosis of lymphoma with digital pathology images using deep learning. Annals of Clinical & Laboratory Science. 2019; 49 (2):153–60.

33. Miyoshi H, Sato K, Kabeya Y, Yonezawa S, Nakano H, Takeuchi Y, et al. Deep learning shows the capability of high-level computer-aided diagnosis in malignant lymphoma. Laboratory Investigation. 2020; 100(10):1300–10. https://doi.org/10.1038/s41374-020-0442-3 PMID: 32472096

34. Steinbuss G, Kriegsmann M, Zgorzelski C, Brobeil A, Goeppert B, Dietrich S, et al. Deep learning for the classification of non-Hodgkin lymphoma on histopathological images. Cancers. 2021; 13(10):2419. https://doi.org/10.3390/cancers13102419 PMID: 34067726

35. Syrykh C, Abreu A, Amara N, Siegfried A, Maisongrosse V, Frenois FX, et al. Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning. NPJ digital medicine. 2020; 3(1):63. https://doi.org/10.1038/s41746-020-0272-0 PMID: 32377574

36. Märkl B, Füzesi L, Huss R, Bauer S, Schaller T. Number of pathologists in Germany: comparison with European countries, USA, and Canada. Virchows Archiv. 2021; 478:335–41. https://doi.org/10.1007/s00428-020-02894-6 PMID: 32719890

37. Paschold L, Gottschick C, Langer S, Klee B, Diexer S, Aksentijevich I, et al. T cell repertoire breadth is associated with the number of acute respiratory infections in the LoewenKIDS birth cohort. Scientific Reports. 2023; 13(1):9516. https://doi.org/10.1038/s41598-023-36144-x PMID: 37308563

38. Paschold L, Klee B, Gottschick C, Willscher E, Diexer S, Schultheiß C, et al. Rapid Hypermutation B Cell Trajectory Recruits Previously Primed B Cells Upon Third SARS-Cov-2 mRNA Vaccination. Frontiers in Immunology. 2022; 13. https://doi.org/10.3389/fimmu.2022.876306 PMID: 35615365

39. Paschold L, Simnica D, Brito RB, Zhang T, Schultheiß C, Dierks C, et al. Subclonal heterogeneity sheds light on the transformation trajectory in IGLV3-21R110 chronic lymphocytic leukemia. Blood Cancer Journal. 2022; 12(3):49. https://doi.org/10.1038/s41408-022-00650-4 PMID: 35354800

40. Paschold L, Simnica D, Willscher E, Vehreschild MJ, Dutzmann J, Sedding DG, et al. SARS-CoV-2–specific antibody rearrangements in prepandemic immune repertoires of risk cohorts and patients with COVID-19. The Journal of Clinical Investigation. 2021; 131(1). https://doi.org/10.1172/JCI142966 PMID: 33064671

41. Schultheiß C, Paschold L, Simnica D, Mohme M, Willscher E, von Wenserski L, et al. Next-generation sequencing of T and B cell receptor repertoires from COVID-19 patients showed signatures associated with severity of disease. Immunity. 2020; 53(2):442–55. e4. https://doi.org/10.1016/j.immuni.2020.06.024 PMID: 32668194

42. Schultheiß C, Paschold L, Willscher E, Simnica D, Wöstemeier A, Muscate F, et al. Maturation trajectories and transcriptional landscape of plasmablasts and autoreactive B cells in COVID-19. Iscience. 2021; 24(11). https://doi.org/10.1016/j.isci.2021.103325 PMID: 34723157

43. von Wenserski L, Schultheiß C, Bolz S, Schliffke S, Simnica D, Willscher E, et al. SLAMF receptors negatively regulate B cell receptor signaling in chronic lymphocytic leukemia via recruitment of prohibitin-2. Leukemia. 2021; 35(4):1073–86. https://doi.org/10.1038/s41375-020-01025-z PMID: 32826957

44. Schultheiß C, Simnica D, Willscher E, Oberle A, Fanchi L, Bonzanni N, et al. Next-generation Immunosequencing reveals pathological T-cell architecture in autoimmune hepatitis. Hepatology. 2021; 73 (4):1436–48. https://doi.org/10.1002/hep.31473 PMID: 32692457

45. Simnica D, Schliffke S, Schultheiß C, Bonzanni N, Fanchi LF, Akyüz N, et al. High-throughput immunogenetics reveals a lack of physiological T cell clusters in patients with autoimmune cytopenias. Frontiers in Immunology. 2019; 10:1897. https://doi.org/10.3389/fimmu.2019.01897 PMID: 31497012

46. Gomes de Castro MA, Wildhagen H, Sograte-Idrissi S, Hitzing C, Binder M, Trepel M, et al. Differential organization of tonic and chronic B cell antigen receptors in the plasma membrane. Nature communications. 2019; 10(1):820. https://doi.org/10.1038/s41467-019-08677-1 PMID: 30778055

47. Schliffke S, Buhs S, Bolz S, Gerull H, von Wenserski L, Riecken K, et al. The phosphotyrosine phosphatase SHP2 promotes anergy in chronic lymphocytic leukemia. Blood, The Journal of the American Society of Hematology. 2018; 131(15):1755–8. https://doi.org/10.1182/blood-2017-06-788166 PMID: 29475960

48. Schliffke S, Sivina M, Kim E, Von Wenserski L, Thiele B, Akyüz N, et al. Dynamic changes of the normal B lymphocyte repertoire in CLL in response to ibrutinib or FCR chemo-immunotherapy. Oncoimmunology. 2018; 7(4):e1417720. https://doi.org/10.1080/2162402X.2017.1417720 PMID: 29632735

49. Schliffke S, Akyüz N, Ford C, Mährle T, Thenhausen T, Krohn-Grimberghe A, et al. Clinical response to ibrutinib is accompanied by normalization of the T-cell environment in CLL-related autoimmune cytopenia. Leukemia. 2016; 30(11):2232–4. https://doi.org/10.1038/leu.2016.157 PMID: 27220665

50. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. Nature methods. 2015; 12(5):380–1. https://doi.org/10.1038/nmeth.3364 PMID: 25924071

51. Greiff V, Miho E, Menzel U, Reddy ST. Bioinformatic and statistical analysis of adaptive immune repertoires. Trends in immunology. 2015; 36(11):738–49. https://doi.org/10.1016/j.it.2015.09.006 PMID: 26508293

52. Pielou EC. The measurement of diversity in different types of biological collections. Journal of theoretical biology. 1966; 13:131–44.

53. Shannon CE. A mathematical theory of communication. The Bell system technical journal. 1948; 27 (3):379–423.

54. Postow MA, Manuel M, Wong P, Yuan J, Dong Z, Liu C, et al. Peripheral T cell receptor diversity is associated with clinical outcomes following ipilimumab treatment in metastatic melanoma. Journal for immunotherapy of cancer. 2015; 3(1):1–5.