Different B cell subpopulations show distinct patterns in their IgH repertoire metrics

Marie Ghraichy[1], Valentin von Niederhäusern[1], Aleksandr Kovaltsuk[2], Jacob D. Galson[1,3], Charlotte M. Deane[2], Johannes Trück[1]

[1] Division of Immunology, University Children's Hospital and Children's Research Center, University of Zurich (UZH), Switzerland
[2] Department of Statistics, University of Oxford, United Kingdom
[3] Alchemab Therapeutics Ltd, London, United Kingdom

Abstract

Background:

Several human B-cell subpopulations are recognized in the peripheral blood, which play distinct roles in the humoral immune response. These cells undergo developmental and maturational changes involving VDJ recombination, somatic hypermutation and class switch recombination, altogether shaping their immunoglobulin heavy chain (IgH) repertoire.

Methods:

Here, we sequenced the IgH repertoire of naïve, marginal zone, switched and plasma cells from 10 healthy adults along with matched unsorted and *in silico* separated CD19$^+$ bulk B cells. We used advanced bioinformatic analysis and machine learning to thoroughly examine and compare these repertoires.

Results:

We show that sorted B cell subpopulations are characterised by distinct repertoire characteristics on both the individual sequence and the repertoire level. Sorted subpopulations shared similar repertoire characteristics with their corresponding *in silico* separated subsets. Furthermore, certain IgH repertoire characteristics correlated with the position of the constant region on the IgH locus.

Conclusion:

Overall, this study provides unprecedented insight over mechanisms of B cell repertoire control in peripherally circulating B cell subpopulations.

Introduction

B-cell development starts in the bone marrow where immature B cells must assemble and express on their surface a functional but non-self-reactive B cell antigen receptor (BCR).[1] The generation of the heavy and light chain of the BCR is mediated by the random and imprecise process of V(D)J recombination.[2] Further development of B cells occurs in the periphery in response to stimulation with the process of somatic hypermutation (SHM) through which point mutations are introduced in the genes coding for the V(D)J part of the immunoglobulin heavy (IgH) and light chain.[3] Subsequently, B cells with a mutated BCR providing increased antigen affinity are selected and show increased survival and proliferation capacity.[4]

Furthermore, class-switch recombination (CSR) modifies the IgH constant region resulting in the generation of B cells with nine different immunoglobulin isotypes or isotype subclasses, namely IgD, IgM, IgG1-4, IgA1/2 and IgE.[5] This process involves the replacement of the proximal heavy chain constant gene by a more distal gene. Class switching is an essential mechanism during humoral immune responses as the constant region of an antibody determines its effector function.[6] Both direct switching and sequential switching upon a second round of antigen exposure have been reported.[7–9]

Through developmental mechanisms and further differentiation in the periphery, several phenotypically distinct circulating B cell subpopulations are generated.[10] They include naïve, marginal zone (MZ), switched memory B cells and plasma cells (PC), which are mainly characterized by their differential expression of surface markers and by playing distinct roles in the adaptive immune response.[11] High-throughput sequencing of the IgH repertoire (AIRR-seq) has made it possible to improve our understanding of the different components of the adaptive immune system in health and disease, and following vaccine challenge.[12–16] Previous studies using both high- and low-throughput sequencing techniques have already reported important differences between B-cell subpopulations affecting their IgH repertoire composition, VDJ gene usage, mutations and clonality.[17–20]

Recent AIRR-seq workflows allow coverage of a sufficient part of the IgH constant region in addition to the VDJ region, making it possible to assign antibody classes and subclasses on an individual sequence level. It is common practice to use unsorted bulk B cells from peripheral blood as a starting material and use the constant region information combined with the degree of SHM to group transcripts *in silico* into different B cell populations.[21,22] Using isotype-resolved IgH sequencing of bulk B cells, isotype subclasses have been found to show differences in their repertoire characteristics.[23,24] However, it remains unknown how the IgH repertoire of bioinformatically separated transcripts originating from bulk-sequenced B cells compares to the repertoire of their corresponding circulating B cell subpopulations. It is also unknown how IgH sequences with the same constant region originating from different cell types compare.

Here, we used an established AIRR-seq workflow that captures the diversity of the variable IgH genes together with the isotype subclass usage to study in detail the repertoire of CD19$^+$ bulk B cells as well as flow cytometry sorted naïve, MZ, switched and plasma cells from 10 healthy adults. We applied advanced statistical methods and machine learning algorithms to combine several repertoire metrics and characterize the different B cell subpopulations. We show that transcripts from physically sorted B cell subpopulations share similar characteristics with their corresponding subsets in the bulk that were grouped *in silico* using isotype subclass information and number of mutations. We further demonstrate that sequences with the same isotype subclass originating from different cell types are closely related, suggesting the presence of isotype-specific rather than cell-type specific signatures in the IgH repertoire. We finally correlate these signatures to the isotype subclass positioning on the locus and find that downstream subclasses exhibit enhanced signs of maturity, overall providing new insights into the selection and the peripheral differentiation of distinct B cell subpopulations.

Results

Physically sorted B cell subpopulations and their corresponding subsets in the bulk share similar repertoire characteristics

We compared IgH repertoire characteristics between the following B cell subpopulations: B_{naive}, B_{MZ}, B_{PC_MD}, B_{PC_AG}, and $B_{switched}$ and their corresponding subsets that we obtained *in silico* from B_{bulk}: $B_{bulk\_naïve}$, B_{bulk_MD}, and $B_{bulk\_switched}$. We identified three separate clusters: one made of predominantly B_{MZ}, B_{bulk_MD} and B_{PC_MD}; another with only B_{naive} and $B_{bulk\_naïve}$; and a third cluster with predominantly $B_{bulk\_switched}$, B_{PC_AG} and $B_{switched}$ (*Figure 1A)* by combining all repertoire characteristics in a PCA and applying k-means clustering. To test whether this clustering pattern was driven by VJ gene usage, CDR3 physiochemical properties or the general repertoire metrics, we analysed these variables separately. Using V family and J gene usage, there was a clear separation between naïve and memory cells mostly driven by differences in V1/3 and J4/6 usage (*Supplementary figure 1*). However, no separation between B_{MZ}/B_{PC_MD}/ B_{bulk_MD} and $B_{switched}$/B_{PC_AG}/$B_{bulk\_switched}$ was observed (*Figure 1B)*. The CDR3 physiochemical properties alone created similar clusters as when combined together with the other metrics (*Figure 1C*). This separation was mostly driven by a lower basic and a higher aromatic content in addition to a higher gravy index and a lower polarity in B_{naive}/$B_{bulk\_naïve}$ compared to memory subpopulations (*Supplementary figure 2*). Global repertoire metrics also created a clear separation between B_{naive}/$B_{bulk\_naïve}$, $B_{switched}$/B_{PC_AG}/$B_{bulk\_switched}$ and B_{MZ}/B_{PC_MD}/B_{bulk_MD} subpopulations mostly driven by higher mutation counts, NP length and selection pressure in the CDR and lower junction length and diversity in $B_{switched}$ compared to B_{naive} (*Supplementary figure 3*).

In summary, we found that V family and J gene usage, the physiochemical properties of the CDR3, and global repertoire metrics similarly distinguish between B cell subpopulations: B_{naive}, B_{MZ}/B_{PC_MD} and $B_{switched}$/B_{PC_AG} were divergent but shared properties with their relative corresponding subsets in the bulk.

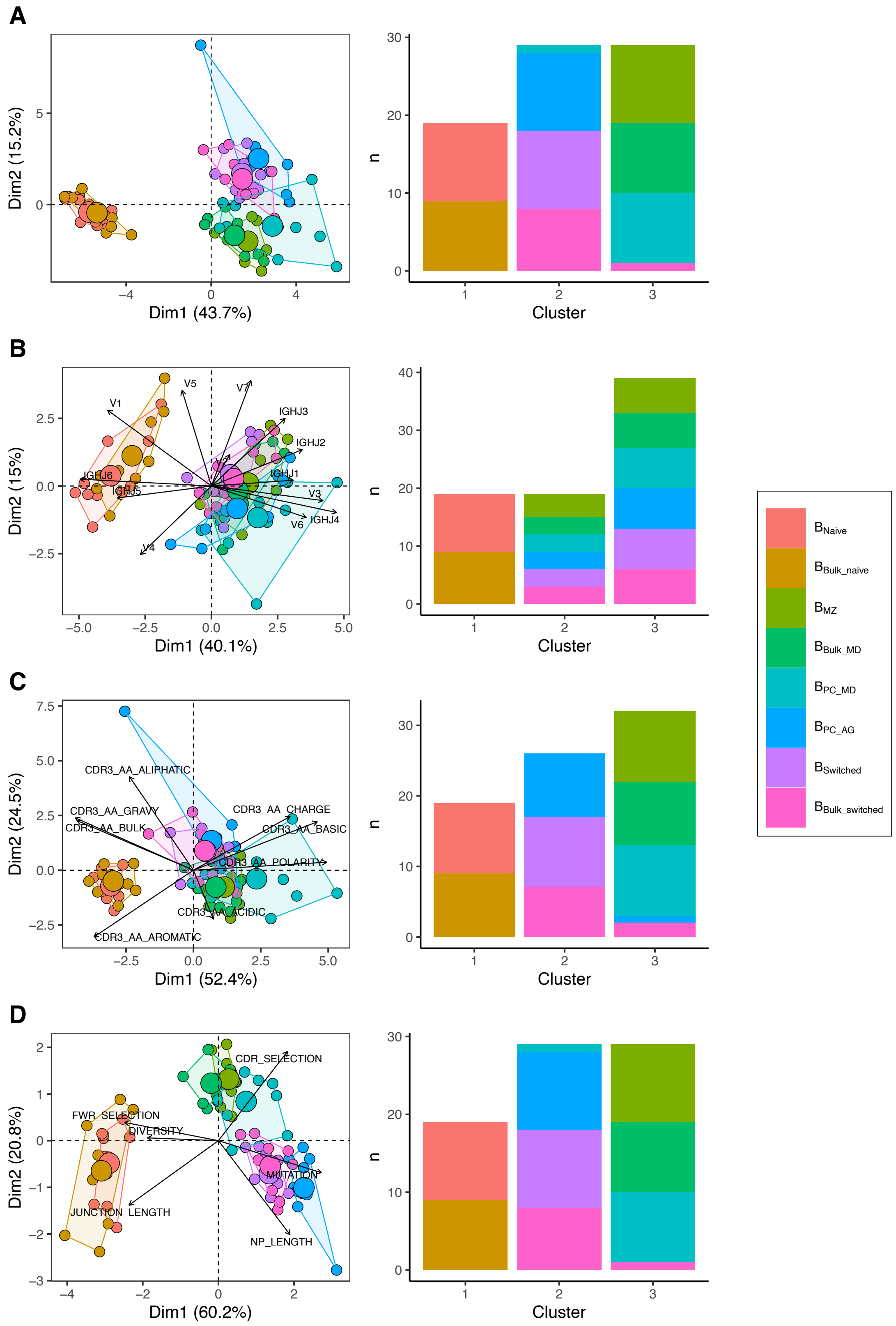


Figure 1: Different repertoire characteristics similarly separate between B cells subpopulations. PCA (left) and composition of the clusters formed using k-means clustering with k=3 (right) applied on A) all repertoire characteristics, B) V family and J gene usage, C) physiochemical properties of CDR3 junction, D) global repertoire metrics. The percentage of all variation in the data that is explained by PC1 and PC2 is shown on the x and y axis respectively between brackets. In the PCA plots, areas are the convex hulls of the subsets and the largest point of one color represents the center of that hull.

Accurate prediction of cell type based on repertoire features on a single-cell level

We constructed a sequence classifier that predicts the cell type of a sequence using sequence attributes and different repertoire metrics. Since we subsampled our data making our datasets perfectly balanced, we used only accuracy as a performance metric. Logistic regression, decision tree and random forest classifiers all performed satisfactorily (*Figure 2A*). However, logistic regression performed poorly on correctly classifying $B_{switched}$ and B_{PC_AG}, for which accuracy was almost equal to chance. The performance of all three classifiers was highest in distinguishing between B_{naive} and other cell types.

The random forest classifier was the most successful compared to the other two and the most accurate in predicting the cell type of a sequence. We assessed the relevance of specific predictors in properly classifying cell types by calculating feature importance scores for each cell pair (*Figure 2B*). The number of mutations was the highest scoring feature for all cell pairs except for distinguishing between $B_{switched}$ and B_{PC_AG} and between B_{MZ} and B_{PC_MD} for which CDR3 amino acid characteristics had higher scores. Within the CDR3 physiochemical properties, average bulkiness, average polarity and the gravy hydrophobicity index were the most differentiating between cell types whereas the basic and acidic content of the CDR3 chain seemed to be less important. R/S ratio in CDR and FWR and the junction length appeared to have similar scores and were more important in cases where B_{naive} were not one of the two cell types. V family and J gene appeared to have low importance in distinguishing between all cell pairs.

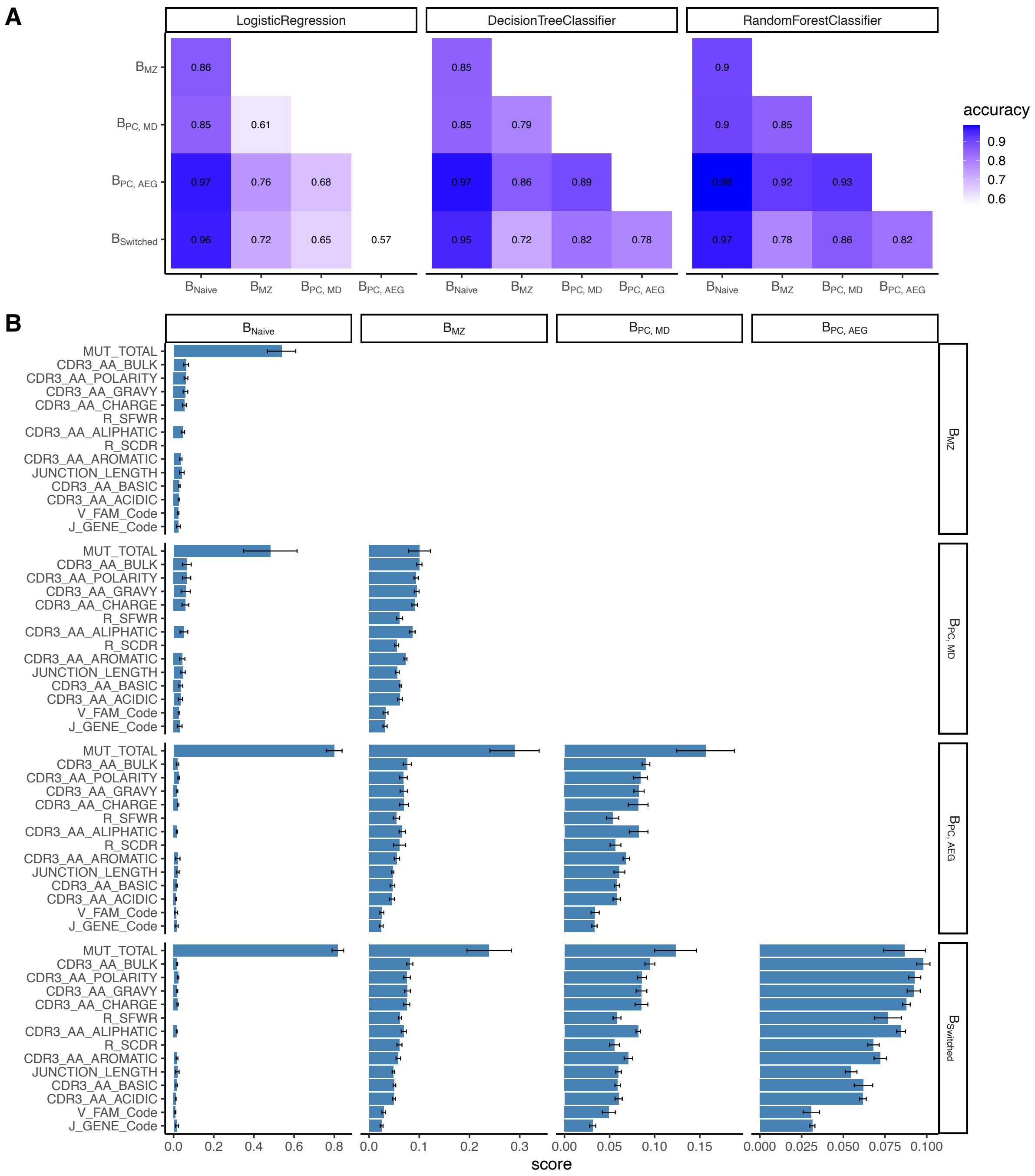


Figure 2: Classification accuracies and feature scores on a single-sequence level. A) Heatmap showing pairwise classification accuracy results using logistic regression, decision tree and random forest classifier. B) Random forest feature scores by cell pair.

Within class switched subsets, sequences with same constant region from different cell types show similar features.

When comparing class-switched transcripts originating from $B_{bulk\_switched}$, $B_{switched}$, and B_{PC_AG}, isotype subclasses were similarly distributed: IgA1 was the dominant subclass in IgA transcripts whereas IgA2 was less frequently used. All cells showed a dominant use of IgG1 and IgG2 with little IgG3 and negligible IgG4 (*Figure 3A*). Usage of IgA1 in B_{PC_AG} was similar to $B_{switched}$ and $B_{bulk\_switched}$ (p=0.28 and p=0.25, Kruskal-Wallis). IgG3 usage was significantly lower in B_{PC_AG} compared to $B_{bulk\_switched}$ and $B_{switched}$ (p=0.01, p=0.01, Kruskal-Wallis) while IgG1 usage tended to be lower (p=0.13 and p=0.11, Kruskal-Wallis) and IgG2 usage higher in B_{PC_AG} compared to the other two B cell subpopulations (p=0.11 and p=0.11, Kruskal-Wallis).

When combining repertoire characteristics by isotype subclass and cell type for class-switched transcripts resulting from $B_{bulk\_switched}$, $B_{switched}$ and B_{PC_AG}, we found that samples with the same constant region originating from different cell types overlapped. (*Figure 3B*) We identified two clusters: one mainly composed of IgG1 and IgG3 samples from all cell types and another with IgA1, IgA2 and IgG2 samples by applying k-means clustering with k=2 (*Figure 3C*). By further dividing the data and with increasing k, we observed that newly formed clusters were mainly composed of distinct isotype subclasses, while the cell type itself was not a defining factor for cluster formation. Interestingly, we couldn't see a clear separation between IgG2 and IgA2 samples with increasing number of clusters.

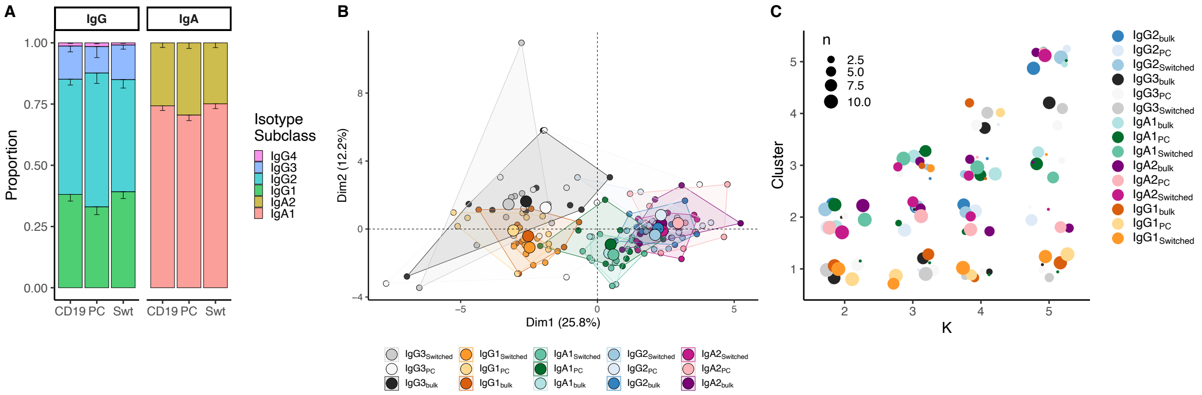


Figure 3: Analysis of isotype subclasses in IgG and IgA transcripts. A) Isotype subclass distribution by cell type. Error bars represent the standard error of the mean. B) PCA on all repertoire properties combined by cell type and isotype subclass. Areas are the convex hulls of a group and the largest point of one color represents the center of that hull. C) Composition of the clusters formed by applying the k-means clustering algorithm on all data with increasing k from k=2 to k=5

B cell repertoire metrics correlate with constant region positioning on the IgH locus.

The IgH locus contains 9 constant genes: the genes encoding for IgM and IgD are the closest to the V-D-J recombination sites while those for IgG3, IgG1 and IgA1 are further downstream but still close to IgM/IgD whereas more distant on the locus are the genes that encode for IgG2, IgG4, IgE and IgA2 (*Figure 4A*). We determined and compared B cell repertoire metrics between different subclasses in B_{PC} and $B_{switched}$ and compared those to B_{naive} and B_{MZ}. B_{naive} showed the lowest number of mutations and R/S ratio and longest CDR3 junction. Memory subsets had a high number of mutations, with B_{MZ} and B_{PC_MD} having fewer mutations than class switched transcripts (*Figure 4B*). IgM-distal subclasses IgG2 and IgA2 in both $B_{switched}$ and B_{PC_AG} showed the highest R/S ratio indicating high selection pressure (*Figure 4C*). All antigen-experienced subsets had a lower junction length compared to B_{naive} except for IgM-proximal transcripts IgG3 and IgG1 (*Figure 4E*). The proportion of IGHV4-34, the gene associated with self-reactivity[33], was lower in memory subsets compared to B_{naive} except for IgG3 from $B_{switched}$ for which the proportion of IGHV4-34 was similar to naïve subsets (*Figure 4F*). Within IgG and IgA sequences, genomic distance from IgM correlated with a higher R/S ratio, shorter junction and lower usage of IGHV4-34. B_{PC} had a significantly lower diversity compared to all other cell types (*Figure 4G*). Interestingly, transcripts from $B_{switched}$ showed a similar diversity to B_{naive} whereas B_{MZ} were less diverse. Within B_{PC_AG}, IgM-distal subclasses showed a lower diversity.

IGHV family and IGHJ gene usage also showed a discrepancy between different subsets: IGHV family usage in IgM-proximal subclasses IgG3 and IgG1 was similar to B_{Naive}. B_{MZ} and IgM-distal subclasses were enriched in IGHJ4 at the expense of IGHJ6 compared to naïve cells and

IgG1-3 B-cell subsets (*Supplementary figure 4*). To reduce the dimensionality of all data points into a single one-dimensional axis, we performed LDA fitted on the relative gene frequencies (*Figure 4H*). This showed a clear distinction between B_{naive}, IgG1-3 and B_{MZ}, IgG2 and IgA1-2. An LDA fitted on the physiochemical properties of the CDR3 junction also showed a clear distinction between naïve and memory subsets, with IgG3 and IgG1 being closest to B_{naive} and IgG2 and IgA2 overlapping and furthest away (*Figure 4I*).

In summary, we found that different B cell repertoire metrics correlate with the positioning of their respective subclass genes on the IgH locus, namely with the increasing genomic distance from IgM, with the proximal IgH subclasses being more similar to naïve.

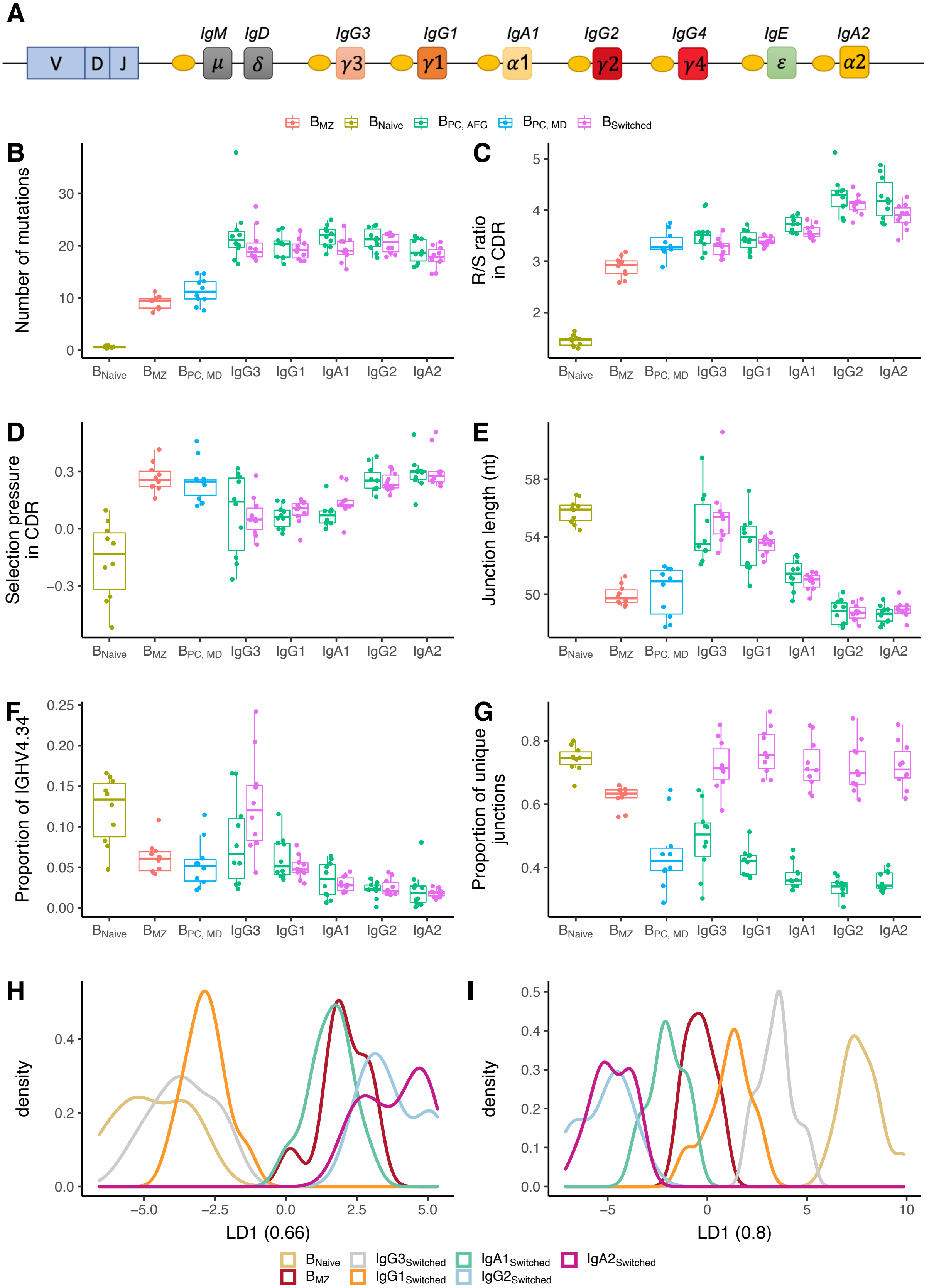


Figure 4: Analysis of repertoire metrics by isotype subclass and cell type. A) Overview of the IgH constant region locus. Comparison of A) mutation counts, B) R/S ratio, C) selection pressure, D) junction length, F) proportion of IGHV4-34 and G) diversity between different B cell subpopulations. LDA fitted on H) V family and J gene usage and I) CDR3 amino acid physiochemical properties.

Discussion

Here, we used AIRR-seq to characterize similarities and differences in the IgH repertoire of bulk B cells and different sorted naïve and memory B cell populations. This allowed for an in-depth understanding of the mechanisms underlying B-cell responses. We report differences in V family and J gene usage, CDR3 physiochemical properties and global repertoire characteristics that similarly distinguish between naïve, IgM/IgD memory and class switched subsets both at the repertoire and at the sequence level. Furthermore, we show differences in the repertoire characteristics at the isotype subclass level unrelated to cell type that correlate with the position of the constant gene on the IgH locus. This study provides powerful insight on biological mechanisms underlying the B cell response as well as novel understanding of AIRR-seq methodologies to be taken into account in future studies.

Previous work involving human naïve and antigen-experienced B cell repertoires have shown naïve B cells to have shorter junctions and higher usage of IGHJ6 and IGHV3, and lower usage of IGHJ4 and IGHV1 compared with IgM memory and switched B cells.[34–37] Differences in gene usage and CDR3 properties between IgM memory and switched B cells have also been reported.[20] IgM memory and switched B cells have been found to use more negatively charged residues and to have less hydrophobic junctions compared with naïve B cells.[18,20] Here, we focused on a more detailed examination of the repertoires by combining multiple characteristics using dimensionality reduction methods. Results of a previous study revealed that combining only a few repertoire characteristics is sufficient to discriminate between B cell subpopulations.[19] In addition, an LDA combining V gene family proportions has been found to successfully distinguish between IgM and IgG repertoires.[38] We extend these findings by showing that using V family and J gene usage, CDR3 physiochemical properties or global repertoire characteristics similarly allow to separate between naïve and memory subpopulations. This suggests that distinct B cell subpopulations derive from different developmental mechanisms and are subject to selective processes that lead to similar variable gene identity. This can also reflect that different types of B cells are stimulated by different types of antigens and therefore have distinctive junction compositions and properties.

Previous research has demonstrated that same B cell subpopulations from different donors are more similar in their repertoire characteristics than different B cell subpopulations within an individual.[39,40] This has led to the understanding that differences between naïve and memory cells are conserved across unrelated individuals. Our findings are in agreement with these observations, and we extend on those by showing that the main defining factor in repertoire similarity is the constant region type, namely the isotype subclass, and that differences between subclasses are conserved across both cell type and individual. This finding suggests the existence of an isotype-based mechanism for repertoire control that is constant across cell types and individuals.

In addition to the comparative analysis of the different peripheral B cell subsets, our study represents, to our knowledge, the first comparison of bulk B cell sequencing with sorted B cell subpopulations. We showed that sequencing unsorted B cells from peripheral blood and combining the constant region information with the degree of SHM to bioinformatically group transcripts yields accurate results comparable to physical sorting, especially when analysing global repertoire characteristics. We acknowledge that this might be limiting in tasks sensitive to potential biases from different RNA levels per cell such as identifying antigen-specific sequences from plasma cells.

Recent IgH repertoire studies have moved towards using machine learning and artificial intelligence in contrast to traditional statistical approaches for goals including vaccine design, immunodiagnostics and antibody discovery.[41–44] Previous work has focused on representing repertoires as sequence or subsequence-based features, i.e. overlapping amino acid k-mers and their Atchley biophysiochemical properties.[41,42] Here, we report a simple pairwise classifier that successfully predicts the cell type of a sequence based on only the commonly used sequence attributes such as number of mutations and junction length. Random forest and decision tree classifiers outperformed the logistic regression algorithm suggesting a non-linear separation between cell types. A common concern when applying machine learning is the possibility of over-fitting. To prevent this, we trained the algorithm on 80% of the data and tested its performance on the remaining unseen 20%. We also subsampled every pair of classes to equal number of sequences in order to balance the dataset. The model presented here is applied only within an individual and is thereby confined by repertoire signals that might be individual-specific. More work improving the generalisability of the model across individuals would be revolutionizing in terms of its potential practical applications. Unsurprisingly, the number of mutations was the most important feature in distinguishing between cell types. These results along with previous work are promising and suggest that increasing the predictive potential of machine learning methods could help in finding sequence characteristics that distinguish between groups, such as disease state and healthy.

Studies indicate that both direct and sequential CSR to IgM distal isotype subclasses can occur.[45,46] Several studies have provided evidence for sequential CSR. IgM was found to commonly switch to proximal subclasses (IgG1, IgA1, and IgG2), but direct switches from IgM to more downstream subclasses (IgG4, IgE, or IgA2) were rare.[7] It has also been reported that a deficiency in IgG3, the most IgM-proximal subclass, frequently results in a decrease in other IgG subclasses.[47] Although it is challenging to determine whether sequential CSR occurs during a primary response, by re-entry into the germinal center, or during a secondary response to the same antigen, we and others have shown that IgM-distal subclasses accumulate with age, likely due to secondary encounter with antigen.[22,48] Studies comparing the mean mutation number between isotype subclasses have shown contradicting results: in one study, mutations varied in relation to the constant region position on the IgH locus, with the closest to IgM (IgG3) having the lowest mutations,[23] while in another study, no such difference was observed.[24] We didn't find a difference in number of mutations among IgG subclasses. Our findings rather suggest that mutation is more efficient in more downstream subclasses as we found that these exhibit higher R/S ratios and selection pressure in the CDR, consistent with previous studies.[49] Generally, IgM distal subclasses showed signs of maturity (shorter junctions, lower IGHV4-34 usage) while transcripts from IgM proximal subclasses were more similar to those of naïve B cells. These results suggest that sequential CSR subjects B cells to selective forces leading to more mature variable gene properties without necessarily accumulating more mutations.

In summary, in this study we took an extensive look at the IgH repertoire of different flow cytometry sorted as well as bioinformatically grouped cell types and isotype subclasses of healthy individuals. Using advanced bioinformatic tools, statistical analysis and machine learning, this analysis provides deep insight into the different mechanisms of B cell development and boosts our understanding of the B cell system components in health.

Material and methods

1. Sample collection and cell sorting

Buffy coat samples were obtained from 10 anonymous healthy adults, hence no approval from the local ethics committee was necessary. B cells were first isolated by magnetic cell sorting using the human CD19 MicroBeads (Miltenyi Biotec, San Diego, CA) and the AutoMACS magnetic cell separator. From 9 out of the 10 samples, $3x10^6$ bulk $CD19^+$ B cells (B_{bulk}) were lysed and stored at -80C. The remaining cells were sorted by flow cytometry into 4 subpopulations using cell surface markers characteristic for naïve (B_{naive}), marginal zone (B_{MZ}), plasma cells (B_{PC}), and switched memory B cells ($B_{switched}$). Cells were then lysed and stored at -80C. Surface markers, demographics, number of cells and purity of each sample are outlined in *supplementary table 1*.

2. RNA extraction and library preparation

RNA extraction was performed on the lysate using the RNeasy Mini Kit (Qiagen, Hilden, Germany). Libraries were prepared as previously described.[22] Briefly, two reverse transcription (RT) reactions were carried out for each RNA sample resulting from B_{bulk} or B_{PC}: one with equal concentrations of IgM and IgD specific primers and another with IgA, IgG, and IgE specific primers. Only one RT reaction with IgM and IgD specific primers was performed on B_{naive} and B_{MZ} samples; similarly, we applied one RT reaction with IgA, IgG and IgE primers on samples obtained from $B_{switched}$. IgH cDNA rearrangements were then amplified in a two-round multiplex PCR using a mix of IGHV region forward primers and Illumina adapter primers, followed by gel extraction for purification and size selection. The final concentration of PCR products was measured using Qubit prior to library preparation and combined with a total of 12 equally concentrated samples. Final libraries barcoded with individual i7 and i5 adapters were sequenced in each run on the Illumina MiSeq platform (2x300bp protocol).

3. Data preprocessing

Preprocessing of raw sequences was carried out using the Immcantation toolkit and as per Ghraichy et al 2020.[22,25,26] Briefly, samples were demultiplexed based on their Illumina tags. A quality filter was applied, paired reads were joined and then collapsed according to their unique molecular identifier (UMI). Identical reads with different UMI were further collapsed resulting in a dataset of unique sequences. VDJ gene assignment was carried out using IgBlast.[27] Isotype subclass annotation was carried out by mapping constant regions to germline sequences using stampy.[28] The number and type of V gene mutations was determined as the number of mismatches with the germline sequence using the R package shazam.[26] The R package alakazam was also used to calculate the physicochemical properties of the CDR3 amino acid sequences.[26] Selection pressure was calculated using BASELINe and the statistical framework used to test for selection was CDR_R/(CDR_R + CDR_S)[29].

4. In silico grouping of sequences

For B_{bulk} samples, we used the constant region information combined with the mutation counts to classify individual sequences into different subsets: IgD and IgM sequences with up to 2 nt mutations across the entire V gene were considered "unmutated" ($B_{bulk\_naïve}$) to account for remaining PCR and sequencing bias. The remaining mutated IgD and IgM sequences were labelled as IgD/IgM memory (B_{bulk_MD}). All class-switched sequences were defined as antigen-experienced regardless of their V gene mutation count ($B_{bulk\_switched}$). We split the sequences originating from B_{PC} into two categories: IgM/IgD B_{PC} (B_{PC_MD}) and switched IgG/IgA PCs (B_{PC_AG}) according to the constant region of the sequences.

5. Summarising repertoire characteristics

V family and J gene usage was calculated in proportions for each individual and cell type. We summarised the mean of the following CDR3 physiochemical characteristics: hydrophobicity, bulkiness, polarity, normalized aliphatic index, normalized net charge, acidic side chain residue content, basic side chain residue content, aromatic side chain content by individual and cell type.

Mean junction length, number of mutations, and numbers of non-template (N) and palindromic (P) nucleotide added at the junction were calculated by individual and cell type. Selection pressure was summarised separately in complementarity-determining region (CDR) and framework region (FWR). Diversity was calculated as the proportion of unique junctions out of total transcripts. The preceding characteristics are referred to as global repertoire metrics.

6. Dimensionality reduction and clustering

Principal component analysis (PCA) and k-means clustering were applied to the different repertoire characteristics to explore and find associations in the data. They were applied using the internal R functions prcomp() and kmeans().[30] Linear discriminant analysis (LDA) was performed using the R function lda() from the package MASS[31].

7. Sequence classifier

We constructed the sequence classifier using the sklearn package in python[32]. Because we have the constant region information and to avoid error accumulation, we performed a pairwise classification thereby transforming the multiclass problem into a binary classification. Within every participant and for every pair of cells, we subsampled to the lower sequence number to avoid bias and dataset imbalance. We used the number of mutations, the physiochemical properties, and the junction length as numerical input features. The V gene family and J gene were one-hot encoded. In the case where the naïve cells were not one of the two classes, the replacement/silent (R/S) mutation ratios in CDR and FWR were included as features. We split the data into training and testing set using the default test size of 0.2. We used logistic regression, decision tree, and random forest classifiers for prediction. The accuracy was recorded to judge the overall performance of the models. For every pair of classes, the mean accuracy of the 10 samples was calculated.

8. Data Availability

Raw data used in this study are available at the NCBI Sequencing Read Archive (www.ncbi.nlm.nih.gov/sra) under BioProject number PRJNA748239 including metadata meeting MiAIRR standards (32). The processed dataset is available in Zenodo (https://doi.org/10.5281/zenodo.3585046) along with the protocol describing the exact processing steps with the software tools and version numbers.

Competing interests

None of the authors have declared any conflict of interest related to this work.

References

1. Lefranc, M.-P. & Lefranc, G. *The immunoglobulin factsbook*. *Academic Press* (2001).
2. Tonegawa, S. Somatic generation of antibody diversity. *Nature* **302**, 575–581 (1983).
3. Jolly, C. J. *et al.* The targeting of somatic hypermutation. *Semin. Immunol.* (1996). doi:10.1006/smim.1996.0020
4. Zheng, N. Y., Wilson, K., Jared, M. & Wilson, P. C. Intricate targeting of immunoglobulin somatic hypermutation maximizes the efficiency of affinity maturation. *J. Exp. Med.* (2005). doi:10.1084/jem.20042483
5. Stavnezer, J., Guikema, J. E. J. & Schrader, C. E. Mechanism and Regulation of Class Switch Recombination. *Annu. Rev. Immunol.* (2008). doi:10.1146/annurev.immunol.26.021607.090248
6. Vidarsson, G., Dekkers, G. & Rispens, T. IgG subclasses and allotypes: From structure to effector functions. *Front. Immunol.* **5**, 520 (2014).
7. Horns, F. *et al.* Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching. *Elife* (2016). doi:10.7554/elife.16578
8. Cameron, L. *et al.* SεSμ and SεSγ Switch Circles in Human Nasal Mucosa Following Ex Vivo Allergen Challenge: Evidence for Direct as Well as Sequential Class Switch Recombination. *J. Immunol.* (2003). doi:10.4049/jimmunol.171.7.3816
9. Zhang, K., Mills, F. C. & Saxon, A. Switch circles from IL-4-directed ε class switching from human B lymphocytes: Evidence for direct, sequential, and multiple step sequential switch from μ to ε Ig heavy chain gene. *J. Immunol.* (1994).
10. Allman, D. & Pillai, S. Peripheral B cell subsets. *Current Opinion in Immunology* (2008). doi:10.1016/j.coi.2008.03.014
11. Leandro, M. J. B-cell subpopulations in humans and their differential susceptibility to depletion with anti-CD20 monoclonal antibodies. *Arthritis Research and Therapy* (2013). doi:10.1186/ar3908
12. Mandric, I. *et al.* Profiling immunoglobulin repertoires across multiple human tissues using RNA sequencing. *Nat. Commun.* (2020). doi:10.1038/s41467-020-16857-7
13. Ghraichy, M., Galson, J. D., Kelly, D. F. & Trück, J. B-cell receptor repertoire sequencing in patients with primary immunodeficiency: a review. 1–16 (2017). doi:10.1111/imm.12865
14. Galson, J. D., Pollard, A. J., Trück, J. & Kelly, D. F. Studying the antibody repertoire after vaccination: Practical applications. *Trends in Immunology* **35**, 319–331 (2014).
15. Lindau, P. & Robins, H. S. Advances and applications of immune receptor sequencing in systems immunology. *Curr. Opin. Syst. Biol.* **1**, 62–68 (2017).
16. Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnology* **32**, 158–168 (2014).
17. Berkowska, M. A. *et al.* Human memory B cells originate from three distinct germinal center-dependent and -independent maturation pathways. *Blood* (2011). doi:10.1182/blood-2011-04-345579
18. Mroczek, E. S. *et al.* Differences in the composition of the human antibody repertoire by b cell subsets in the blood. *Front. Immunol.* (2014). doi:10.3389/fimmu.2014.00096
19. Galson, J. D. *et al.* BCR repertoire sequencing: Different patterns of B-cell activation after two Meningococcal vaccines. *Immunol. Cell Biol.* **93**, 885–895 (2015).
20. Wu, Y. C. *et al.* High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* (2010). doi:10.1182/blood-2010-03-275859
21. 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).

22. Ghraichy, M. *et al.* Maturation of the Human Immunoglobulin Heavy Chain Repertoire With Age. *Front. Immunol.* (2020). doi:10.3389/fimmu.2020.01734
23. Jackson, K. J. L., Wang, Y. & Collins, A. M. Human immunoglobulin classes and subclasses show variability in VDJ gene mutation levels. *Immunol. Cell Biol.* (2014). doi:10.1038/icb.2014.44
24. Kitaura, K. *et al.* Different somatic hypermutation levels among antibody subclasses disclosed by a new next-generation sequencing-based antibody repertoire analysis. *Front. Immunol.* (2017). doi:10.3389/fimmu.2017.00389
25. Vander Heiden, J. A. *et al.* PRESTO: A toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* **30**, 1930–1932 (2014).
26. Gupta, N. T. *et al.* Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31**, 3356–3358 (2015).
27. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34–W40 (2013).
28. Lunter, G. & Goodson, M. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011).
29. Yaari, G., Uduman, M. & Kleinstein, S. H. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Res.* **40**, e134–e134 (2012).
30. 3.5.1., R. D. C. T. A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* **2**, https://www.R-project.org (2018).
31. Venables, W. N. & Ripley, B. D. Modern Applied Statistics with S (fourth.). New York: Springer. *Retrieved from http//www.stats.ox.ac.uk/pub/MASS4* (2002).
32. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* (2015).
33. Bashford-Rogers, R. J. M., Smith, K. G. C. & Thomas, D. C. Antibody repertoire analysis in polygenic autoimmune diseases. *Immunology* **155**, 3–17 (2018).
34. Briney, B. S., Willis, J. R., Hicar, M. D., Thomas, J. W. & Crowe, J. E. Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire. *Immunology* (2012). doi:10.1111/j.1365-2567.2012.03605.x
35. Larimore, K., McCormick, M. W., Robins, H. S. & Greenberg, P. D. Shaping of Human Germline IgH Repertoires Revealed by Deep Sequencing. *J. Immunol.* **189**, 3221–3230 (2012).
36. DeWitt, W. S. *et al.* A public database of memory and naive B-cell receptor sequences. *PLoS One* (2016). doi:10.1371/journal.pone.0160853
37. Bautista, D. *et al.* Differential Expression of IgM and IgD Discriminates Two Subpopulations of Human Circulating IgM+IgD+CD27+ B Cells That Differ Phenotypically, Functionally, and Genetically. *Front. Immunol.* (2020). doi:10.3389/fimmu.2020.00736
38. Friedensohn, S. *et al.* Synthetic standards combined with error and bias correction improve the accuracy and quantitative resolution of antibody repertoire sequencing in human naïve and memory B cells. *Front. Immunol.* (2018). doi:10.3389/fimmu.2018.01401
39. Briney, B. S., Willis, J. R., McKinney, B. A. & Crowe, J. E. High-throughput antibody sequencing reveals genetic evidence of global regulation of the nave and memory repertoires that extends across individuals. *Genes Immun.* (2012).

doi:10.1038/gene.2012.20

40. Rubelt, F. *et al.* Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat. Commun.* (2016). doi:10.1038/ncomms11112

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

42. Ostmeyer, J. *et al.* Statistical classifiers for diagnosing disease from immune repertoires: A case study using multiple sclerosis. *BMC Bioinformatics* (2017). doi:10.1186/s12859-017-1814-6

43. Konishi, H. *et al.* Capturing the differences between humoral immunity in the normal and tumor environments from repertoire-seq of B-cell receptors using supervised machine learning. *BMC Bioinformatics* (2019). doi:10.1186/s12859-019-2853-y

44. Shemesh, O., Polak, P., Lundin, K. E. A., Sollid, L. M. & Yaari, G. Machine Learning Analysis of Naïve B-Cell Receptor Repertoires Stratifies Celiac Disease Patients and Controls. *Front. Immunol.* (2021). doi:10.3389/fimmu.2021.627813

45. Wesemann, D. R. *et al.* Immature B cells preferentially switch to IgE with increased direct Sμ to Sε recombination. *J. Exp. Med.* (2011). doi:10.1084/jem.20111155

46. Looney, T. J. *et al.* Human B-cell isotype switching origins of IgE. *J. Allergy Clin. Immunol.* (2016). doi:10.1016/j.jaci.2015.07.014

47. Meyts, I., Bossuyt, X., Proesmans, M. & De, B. Isolated IgG3 deficiency in children: To treat or not to treat? Case presentation and review of the literature. *Pediatric Allergy and Immunology* (2006). doi:10.1111/j.1399-3038.2006.00454.x

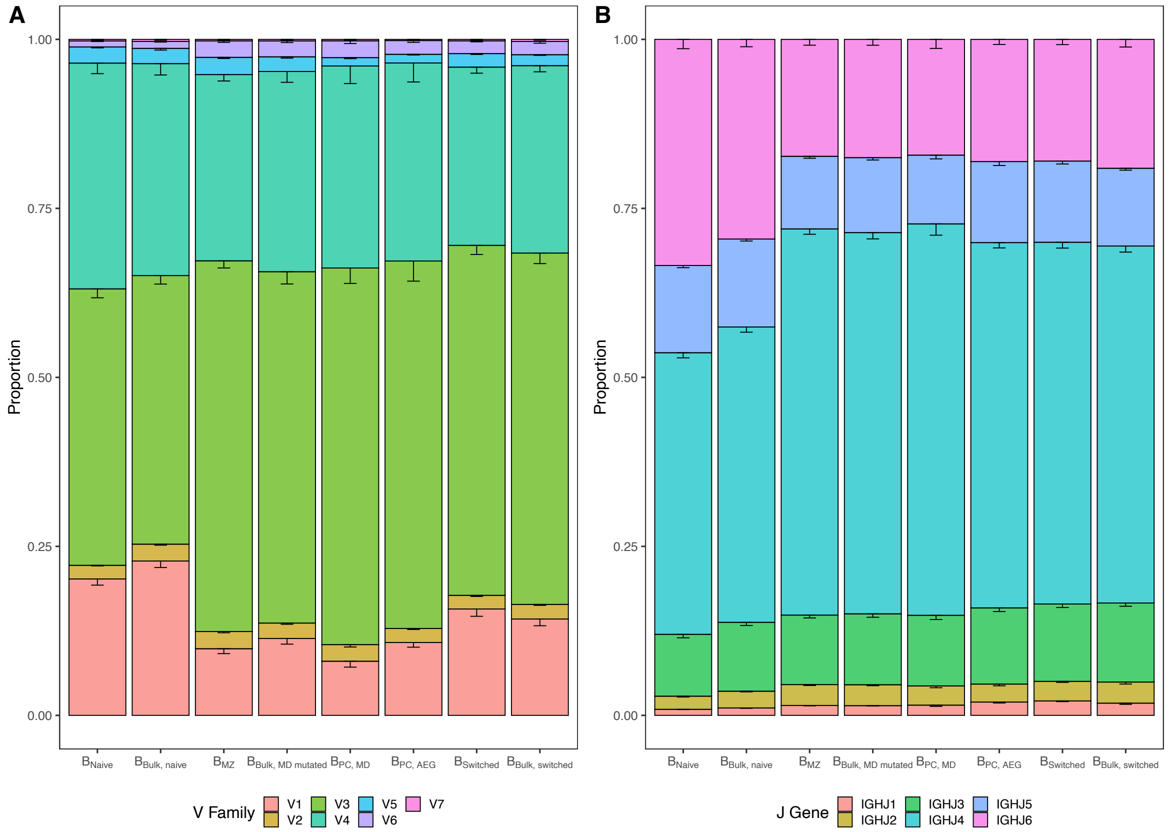
48. IJspeert, H. *et al.* Evaluation of the Antigen-Experienced B-Cell Receptor Repertoire in Healthy Children and Adults. *Front. Immunol.* **7**, 410 (2016).

49. De Jong, B. G. *et al.* Human IgG2- and IgG4-expressing memory B cells display enhanced molecular and phenotypic signs of maturity and accumulate with age. *Immunol. Cell Biol.* (2017). doi:10.1038/icb.2017.43

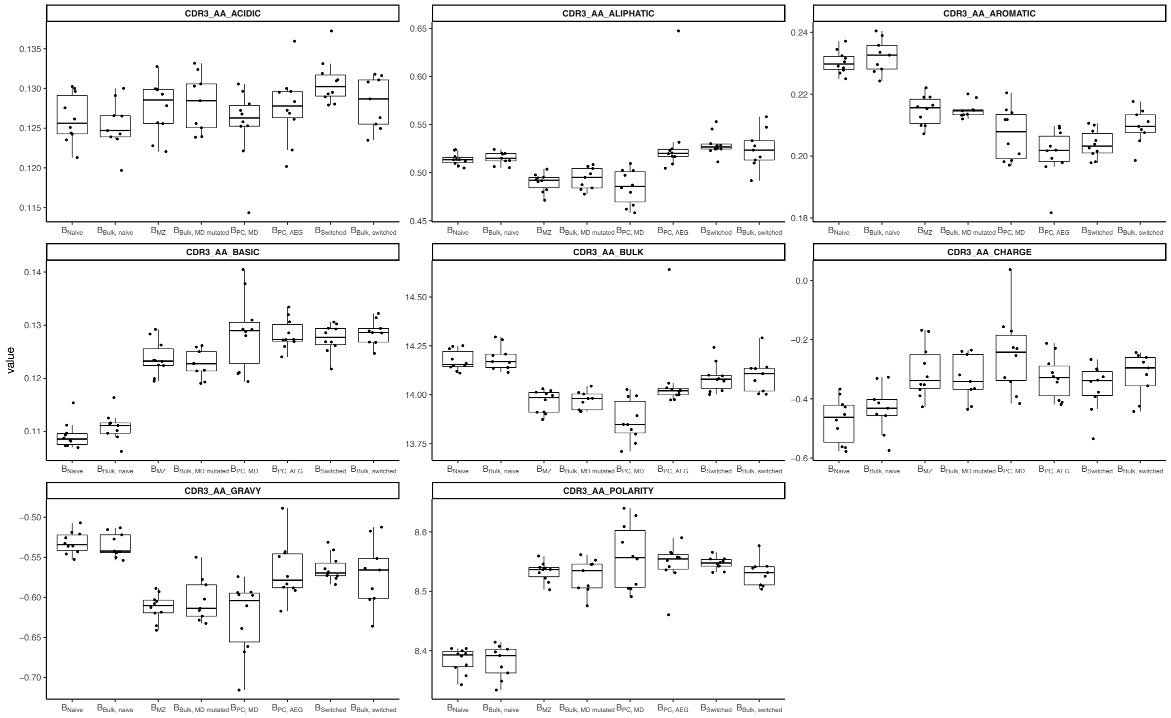
Supplementary material

Supplementary table 1:

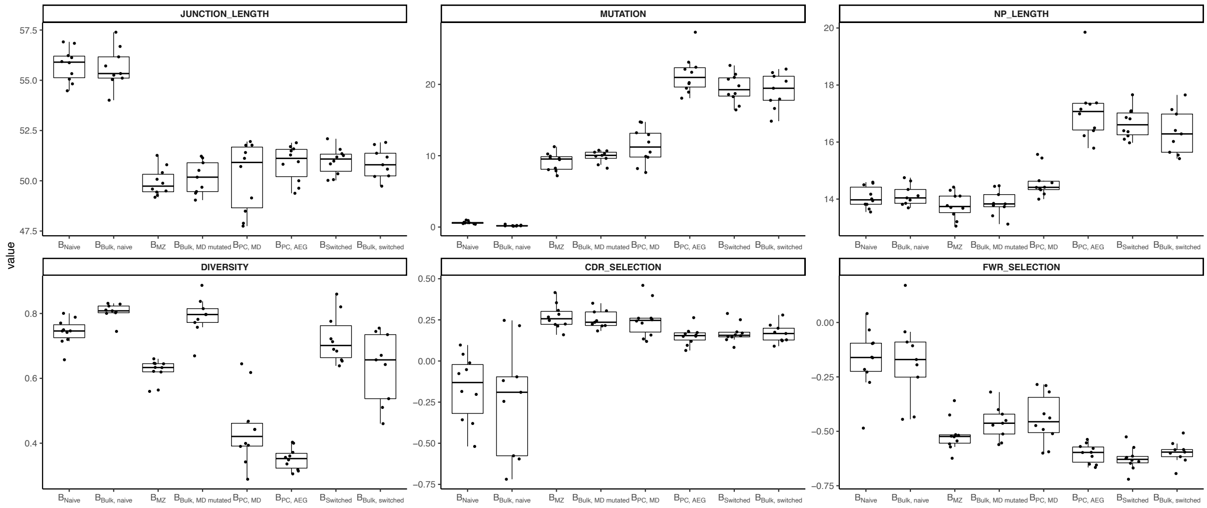
Participant ID	Cells	Age	Sex	B cell number	Purity
Co_C.081.1_BC	MZB	50	M	250000	93.6
Co_C.081.1_BC	Naive	50	M	250000	96.7
Co_C.080.1_BC	MZB	NA	M	250000	NA
Co_C.080.1_BC	Naive	NA	M	250000	NA
Co_C.082.1_BC	Naive	40	M	250000	97.6
Co_C.083.1_BC	MZB	18	M	250000	82.9
Co_C.083.1_BC	Naive	18	M	250000	98.2
Co_C.084.1_BC	MZB	36	F	250000	86.4
Co_C.084.1_BC	Naive	36	F	250000	96.7
Co_C.081.1_BC	Swt	50	M	250000	98.8
Co_C.080.1_BC	Swt	NA		250000	NA
Co_C.080.1_BC	PC	NA		55000	NA
Co_C.081.1_BC	PC	50	M	1.00E+05	45.8
Co_C.082.1_BC	MZB	40	M	250000	81.5
Co_C.082.1_BC	Swt	40	M	250000	98.7
Co_C.082.1_BC	PC	40	M	19000	67.8
Co_C.083.1_BC	Swt	18	M	250000	99.2
Co_C.083.1_BC	PC	18	M	21000	32.8
Co_C.081.1_BC	CD19	50	M	5.00E+05	NA
Co_C.084.1_BC	Swt	36	F	250000	98.2
Co_C.084.1_BC	PC	36	F	15000	30.3
Co_C.084.1_BC	CD19	36	F	5.00E+05	NA
Co_C.085.1_BC	CD19	41	M	5.00E+05	NA
Co_C.085.1_BC	PC	41	M	21000	32.2
Co_C.085.1_BC	Swt	41	M	250000	97.5
Co_C.085.1_BC	Naive	41	M	250000	98.9
Co_C.085.1_BC	MZB	41	M	250000	92.1
Co_BC7_BC	Naive	49	F	250000	93.6
Co_BC8_BC	MZB	59	F	250000	90.5
Co_BC8_BC	Naive	59	F	250000	95.2
Co_BC9_BC	MZB	44	F	250000	91.8
Co_BC9_BC	Naive	44	F	250000	99.2
Co_BC10_BC	MZB	51	F	250000	94.2
Co_BC10_BC	Naive	51	F	250000	96.2
Co_BC7_BC	CD19	49	F	5.00E+05	NA
Co_BC7_BC	Swt	49	F	250000	95.6
Co_BC7_BC	PC	49	F	14000	37
Co_BC8_BC	CD19	59	F	5.00E+05	NA
Co_BC8_BC	Swt	59	F	250000	97.3
Co_BC8_BC	PC	59	F	24000	68.2
Co_BC9_BC	CD19	44	F	5.00E+05	NA
Co_BC9_BC	Swt	44	F	250000	94.5
Co_BC9_BC	PC	44	F	22000	82.8
Co_BC10_BC	CD19	51	F	5.00E+05	NA
Co_BC10_BC	Swt	51	F	250000	99.1
Co_BC10_BC	PC	51	F	19000	60
Co_C.082.1_BC	CD19	40	M	5.00E+05	NA
Co_BC7_BC	MZB	49	F	250000	86.7
Co_C.083.1_BC	CD19	18	M	5.00E+05	NA



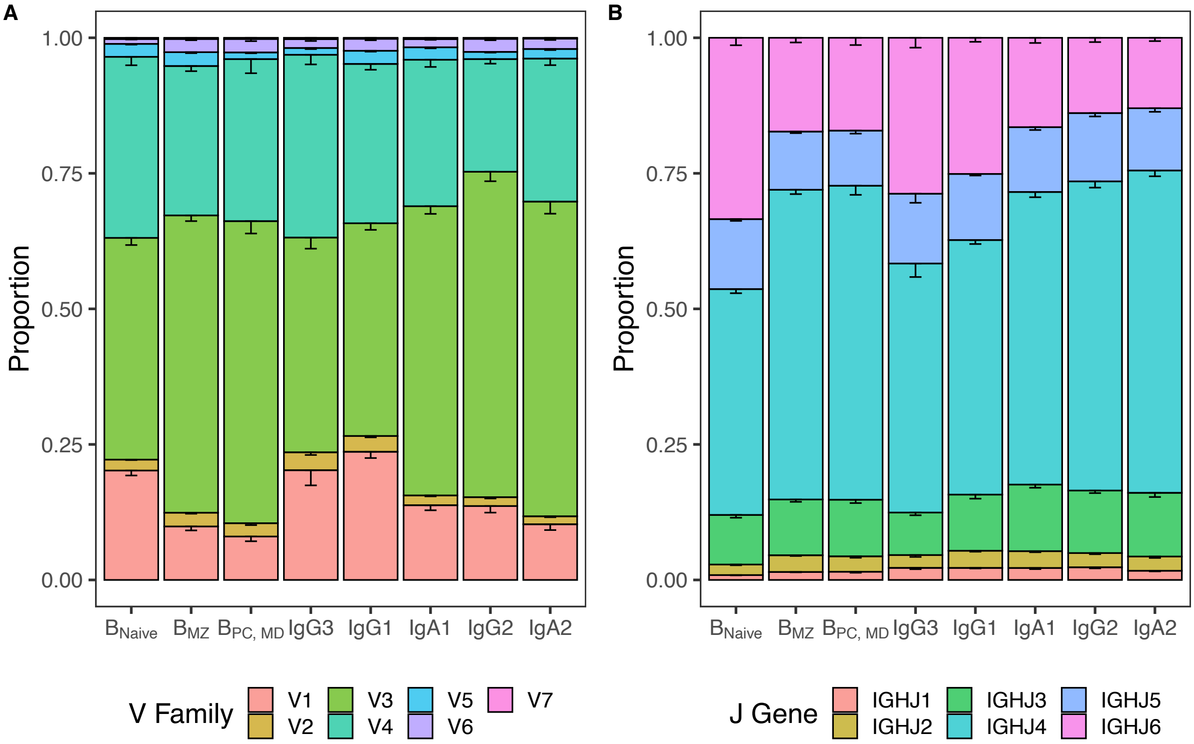
Supplementary figure 1 A) V family and B) J gene usage by B cell subpopulation. Bar plots indicate the proportion of sequences with a certain gene. Error bars represent the standard error of the mean.



Supplementary figure 2 : Comparison of CDR3 amino acid physiochemical properties in different B cell subpopulations.



Supplementary figure 3 : Comparison of global repertoire metrics in different B cell subpopulations.



Supplementary figure 4: A) V family and B) J gene usage in different B cell subpopulations and isotype subclasses. Bar plots indicate the proportion of sequences with a certain gene. Error bars represent the standard error of the mean.