Memory persistence and differentiation into antibody-secreting cells accompanied by positive selection in longitudinal BCR repertoires

Artem I. Mikelov[1,2,3†], Evgeniia I. Alekseeva[1†], Ekaterina A. Komech[2,3], Dmitriy B. Staroverov[2], Maria A. Turchaninova[2], Mikhail Shugay[2,3], Dmitriy M. Chudakov[1,2,3], Georgii A. Bazykin[1,4], Ivan V. Zvyagin[2,3*]

[1] Skolkovo Institute of Science and Technology, Moscow, Russia;
[2] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia;
[3] Institute of Translational Medicine, Pirogov Russian National Research Medical University, Moscow, Russia;
[4] A.A. Kharkevich Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow, Russia.

*Corresponding author: izvyagin@gmail.com (I.V.Z.)

[†]These authors contributed equally to this work.

Key words: memory B cells, plasmablasts, plasma cells, BCR repertoire, somatic hypermutations, B-cell somatic evolution, natural selection, affinity maturation

Abstract

B-cell mediated immune memory holds both plasticity and conservatism to respond to new challenges and repeated infections. Here, we analyze the dynamics of immunoglobulin heavy chain (IGH) repertoires of memory B cells, plasmablasts and plasma cells sampled several times during one year from peripheral blood of volunteers without severe inflammatory diseases. We reveal a high degree of clonal persistence in individual memory B-cell subsets with inter-individual convergence in memory and antibody-secreting cells (ASCs). Clonotypes in ASCs demonstrate clonal relatedness to memory B cells and are transient in peripheral blood. Two clusters of expanded clonal lineages displayed different prevalence of memory B cells, isotypes, and persistence. Phylogenetic analysis revealed signs of reactivation of persisting memory B cell-enriched clonal lineages, accompanied by new rounds of affinity maturation during proliferation to ASCs. Negative selection contributes to both, persisting and reactivated lineages, saving functionality and specificity of BCRs to protect from the current and future pathogens.

Introduction

B cells, as a part of adaptive immunity, play a crucial role in protection from various pathogens and cancer cells or regulation of the immune response. The structural diversity of B-cell receptors (B-cell receptors (BCRs) is responsible for the potential of B-cell immunity to recognise a huge variety of different antigens via a large pool of clones of naive B cells, each having a unique BCR.

Antigenic challenge triggers the proliferation of a naive B cell whose progeny represents a number of cell subsets with different functions and lifespan. Initial structure of BCR can be changed by somatic hypermutations (SHMs) during affinity maturation, a process accompanying B-cell proliferation after antigen-specific activation. The cells bearing BCRs with higher affinity to the antigen are favoured in rounds of selection process by getting signals for further differentiation and expansion (De Silva and Klein 2015). Besides affinity maturation, another process called class-switch recombination further increases the dimensionality of the BCR space. The five main classes (isotypes) of antibodies have different functions in immune response (Stavnezer, Guikema, and Schrader 2008; Vidarsson, Dekkers, and Rispens 2014) and can be switched during clonal proliferation thus changing the functional abilities of the B cells and antibodies. Thus antigen challenge forms a population of clonally related cells with different BCRs and functional abilities.

Clonal relatedness of B cells in a lineage, the number and dynamics of B cell groups with distinct antigen specificities can be studied using BCR sequence homology. Rapidly developed immune repertoire sequencing technologies provide valuable insights into the development and structure of B-cell immunity with clonal level resolution (Chaudhary and Wesemann 2018). Numerous studies reported valuable data by analysing different characteristics of repertoires, such as clonal diversity and tissue distribution, the magnitude of clonal expansions and BCR somatic hypermutations, VDJ usage frequency and the distribution of CDR3 length, the degree of repertoire convergence and individuality (Briney et al. 2019; Soto et al. 2019; Shah et al. 2019; Mandric et al. 2020; Yang et al. 2021). Studies of BCR repertoires of patients with different diseases make a significant contribution to the understanding of mechanisms of pathology and B-cell immunity (Bashford-Rogers et al. 2019; S. C. A. Nielsen et al. 2020; Gaebler et al. 2021; Sakharkar et al. 2021).

Longitudinal analysis of repertoires in different timepoints allowed to study the dynamics of B cell response upon antigenic challenge or therapy (Laserson et al. 2014; Davydov et al. 2018; Horns et al. 2019; Nourmohammad et al. 2019; Hoehn et al. 2021). Reconstruction of BCR evolution in B-cell clonal lineages and phylogenetic analysis can reveal which evolutionary forces predominate at different stages of clonal lineage development. Age-related differences in the structure of clonal lineages and B-cell repertoire ability to generate novel specificities upon vaccination was recently reported (de Bourcy et al. 2017). Other studies described in detail the action of positive selection in the evolution of clonal lineages in vaccination and chronic HIV

infection (Bonsignori et al. 2017; Horns et al. 2019; Nourmohammad et al. 2019). Persisting clonal lineages which predominantly were represented by the cells with IgM/IgD isotypes and demonstrated signs of neutral evolution were recently reported (Horns et al. 2019). Concordantly, Hoehn and colleagues showed reactivation of clonal lineages after seasonal influenza vaccination (Hoehn et al. 2021).

Comparing BCR repertoires between different cell subsets allows investigating factors governing the functional assignment of B cells during proliferation to understand fundamental aspects of B-cell immunity. The difference in BCR repertoires of IgM and switched memory B cells as well as the complex interplay between CD27 high and low B-cell memory subsets was recently described, showing the complex nature of B cell immune memory (Wu et al. 2010; Grimsholm et al. 2020).

BCR repertoires of antigen-experienced B-cell subsets and their dynamics are usually studied in pathologic conditions and vaccination. But there is a lack of such knowledge in absence of acute or chronic immune response. We investigated immune repertoires of immunoglobulin heavy chains of memory B cells, plasmablasts and plasma cells from peripheral blood of volunteers without severe pathologies collected at three time points within a year. To get detailed and unbiased repertoire data we used the advanced PCR-bias sustainable IGH repertoire profiling technology providing the full-length IGH sequences with isotype annotation. With all the advantages of such study design and the power of phylogenetic analysis, we describe the structure and distinctive features, clonal relations, isotype distribution and temporal dynamics of the B-cell subsets repertoires, as well as phylogenetic history of large clonal lineages.

Results

IGH repertoire sequencing statistics and analysis depth

Memory B cells (Bmem: $CD19^{+}$ $CD20^{+}$ $CD27^{+}$), plasmablasts (PBL: $CD19^{low/+}$ $CD20^{-}$ $CD27^{high}$ $CD138^{-}$) and plasma cells (PL: $CD19^{low/+}$ $CD20^{-}$ $CD27^{high}$ $CD138^{+}$) were sorted (**Supplementary Fig. S1A**) from peripheral blood of 6 donors in 3 time points (**Supplementary Table S1**), achieving one month and one year distance between sample collection dates (**Fig. 1A**). Most of the cell samples were collected and processed in two independent replicates (**Supplementary Data SD1**). For each cell sample full-length IGH clonal repertoires were obtained using the 5'-RACE-based protocol, which allows unbiased amplification of full-length IGH variable domain cDNA preserving isotype information, with subsequent UMI-based sequencing data normalisation and error correction (Turchaninova et al. 2016; Shugay et al. 2014). For a total of 83 cell samples we obtained 1.06×10^{7} unique IGH cDNA molecules, each covered by at least 3 sequencing reads, resulting overall in 8.4×10^{5} unique IGH clonotypes (**Supplementary Data SD1**). An IGH clonotype was defined as a unique nucleotide sequence from the beginning of IGH V gene Framework 1 to the 5'-part of C-segment, that was sufficient to determine isotype.

Number of unique clonotypes (species richness) depended on the number of cells per sample (**Supplementary Fig. S2A**) even after data normalisation by sampling an equal number of unique IGH cDNA sequences. To characterise the number of distinct IGH clonotypes in cell subsets we selected the samples with the most common number of sorted cells for each sample set. The median number of clonotypes was 20072 (14572 - 32806, n=14) per 5 x 10^4 memory B cells, 628 (528 - 981, n=8) per 1 x 10^3 plasmablasts and 800 (623 - 1183, n=9) per 1 x 10^3 plasma cells. Rarefaction analysis in memory B-cell subpopulation (**Supplementary Fig. S2A** left) revealed an asymptotic increase of species richness not reaching plateau, indicating that the averaged species richness can only serve as a lower limit of sample diversity estimation. For all samples of PBL and PL subpopulations species richness curves plateaued, meaning that we reached sufficient sequencing depth to evaluate the clonal diversity of the sorted cell samples (**Supplementary Fig. S2A** center and right).

Subsets display divergent and similar characteristics of IGH repertoires

First, we aimed to characterise common and distinct features of IGH repertoire of the memory B cells, plasmablasts and plasma cells by several key properties of the IGH clonotypes: usage of the germ-line encoded IGHV segments, clonal distribution by isotypes, rate of somatic hypermutations in CDR1-2 and FWR1-3, and features of hypervariable CDR3 region.

The proportion of overall clonal diversity occupied by the five major IGH isotypes were strikingly different between memory B cells and antibody-secreting cells (ASCs, PBL and PL). IgM represented more than a half of the repertoire in Bmem, while IgA was dominant in PBL and PL (**Fig. 1B**, **Supplementary Data SD2**). The second prevalent isotype in ASCs was IgG, which was also less abundant in Bmem compared to IgA. As expected, IgD represented a substantial part of Bmem clonal repertoire, but < 1% clonotypes of ASCs expressed IgD. The proportion of each isotype varied between donors and time points; still, IgM and IgA or IgA and IgG remained the most abundant isotypes in memory B cells or ASCs, correspondingly (**Supplementary Fig. S1B, Supplementary Data SD2**). In all studied subsets, isotype distribution by number of unique clonotypes roughly followed the isotype distribution by the number of IGH cDNA molecules, indicating absence of large clonal expansions or differences in IGH expression level distorting abundance of isotypes. This can not be achieved by sequencing of bulk PBMC, as higher levels of IGH expression by ASCs can change the isotype proportions and, hence, bias the quantitation of a clonotype abundance (**Supplementary Fig. S2B**). The obtained IGH isotype distribution by unique clonotypes roughly corresponds to the distribution of IGH isotypes typically detected by flow cytometry of the same subsets (Perez-Andres et al. 2010).

The level of somatic hypermutation (SHM) was on average significantly higher in antibody-secreting cell subsets, reflecting that plasmablasts and plasma cells are enriched by the clones which have undergone affinity maturation (**Fig. 1C**). The averaged number of SHMs for IgE clonotypes didn't differ between cell subsets significantly, while it was higher compared

to the level detected for IgM and IgD clonotypes of B memory. Of note, the rate of SHM in plasmablasts was higher than that in plasma cells in clonotypes of the three most abundant isotypes (IgM, IgA, IgG).

Further, we compared the distributions of length of hypervariable CDR3 region between IGH clonotypes found in the different cell subsets. Plasmablasts had significantly longer CDR3 compared to memory B cells on average in four most represented isotypes (**Fig. 1D**). Of note, the averaged length of CDR3 of plasma cell clonotypes was higher compared to Bmem only for IgA, but not for all other isotypes.

IGHV gene segment usage was roughly similar between memory B cells, plasmablasts and plasma cells of all donors, indicating generally equal probabilities of memory-to-ASC conversion for B cells carrying BCRs encoded by distinct gene segments (**Fig. 1E**, **Supplementary Fig. S2C**). This distribution was significantly different between the studied cell subsets and naive B cells (data from Gidoni et al. 2019), while the repertoire of total B cells (Btot, $CD19^{+}$ $CD20^{+}$), containing large fraction of naive B cells, demonstrated similar IGHV gene segment usage to naive B-cell repertoire (**Supplementary Fig. S2C**): Pearson correlation of IGHV gene frequencies > 0.95 for all pairs between Bmem, PBL, PL (p-value < 0.01), naive vs Btot 0.79 (p-value < 0 .01), Btot vs Bmem, PBL, PL each < 0.24, naive vs Bmem, PBL, PL each < 0.45 (p-value > 0 .05).

We observed high level of concordance in under- or overrepresentation of IGHV gene segments in repertoires of all antigen-experienced B-cell subsets compared to naive B-cell repertoires: Pearson correlation coefficients on fold change of IGHV gene segment frequencies were 0.95 - Bmem/PBL, 0.96 - Bmem/PL, 0.98 - PB/PL (p.value < 0.01 for each pair). Moreover, whether the IGHV gene segment was under- or overrepresented clearly depended on the IGHV gene sequence. We clustered IGHV genes based on their sequence similarity and observed that most IGHV segments in each of the four major clusters behaved concordantly with other IGHV segments in that cluster (**Fig. 1E**). This effect was also observed at the level of individual repertoires (**Supplementary Fig. S2D**) with discrepancies which probably could be attributed to genetic polymorphism of IGH locus of particular donors.

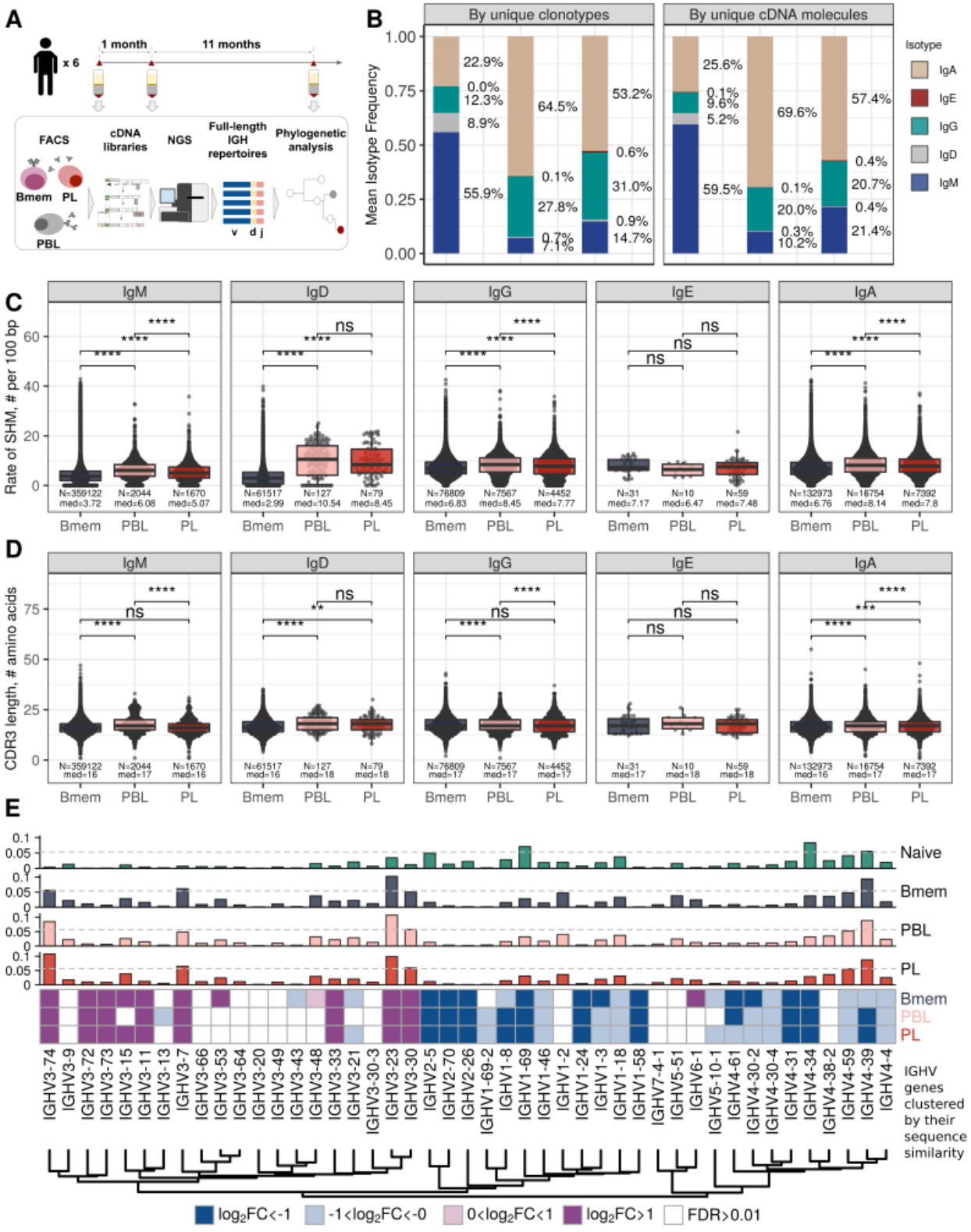

Figure 1. General characteristics of IGH repertoires in differentiated B-cell lineage subsets. A: Study design. Peripheral blood of 6 donors was sampled at three time points: T1 - initial time point, T2 - 1 month and T3 - 12 months later from the start of the study. At each time point we isolated PBMCs and sorted memory B cells (Bmem: $CD19^{+}$ $CD20^{+}$ $CD27^{+}$), plasmablasts (PBL: $CD19^{low/+}$ $CD20^{-}$ $CD27^{high}$ $CD138^{-}$) and plasma cells (PL: $CD19^{low/+}$ $CD20^{-}$ $CD27^{high}$ $CD138^{+}$) in two replicates using FACS. For each cell sample we obtained clonal repertoires of full-length IGH by sequencing of respective cDNA libraries; **B**: Proportion of isotypes in studied cell subsets averaged across all obtained repertoires. Left panel - frequency of unique IGH clonotypes with each particular isotype. Right panel - frequency of each isotype by IGH cDNA molecules detected in a sample; **C**: Distribution of the number of somatic hypermutations identified per 100 bp length of IGHV-segment for clonotypes with each particular isotype; **D**: Distribution of CDR3 length of clonotypes in studied cell subsets by isotype; **E**: Distributions of average IGHV gene frequencies by number of clonotypes of naive B-cell (Gidoni et al. 2019), memory B-cell, plasmablast and plasma cell repertoires are shown at the top. Colored squares on heatmap indicate significantly different (FDR < 0.01) IGHV-gene segments by their frequency in corresponding B-cell subsets compared to naive B-cell repertoires (data from Gidoni et al. 2019). Color intensity reflects the magnitude of the difference (FC = fold change). Only V-genes which were represented by more than 2 clonotypes on average are shown. IGHV-gene segments are clustered by the similarity of their amino acid sequence, as indicated by the dendrogram at the bottom. In C and D the number at the bottom of the plot represents a number of clonotypes in the corresponding group, pooled from all donors. Comparison between subsets was performed with two-sided Mann-Whitney U test, notation of the level of significance is the following: * - $p \leqslant 0.05$, ** - $p \leqslant 0.01$, *** - $p \leqslant 10^{-3}$, **** - $p \leqslant 10^{-4}$.

Memory B-cell repertoires are stable over time and contain a high number of public clonotypes

We further studied the similarity of IGH clonal repertoire of the B-cell subsets across time points and between individuals evaluating repertoire stability (distance between different time points) and degree of individuality (distance between repertoires from different donors). The repertoire similarity were studied on two levels of IGH sequence identity: by frequency of clonotypes with identical nucleotide sequence-defined variable regions (FR1-FR4) and by the number of clonotypes having identical amino acid sequences of CDR3 region, same IGHV-gene segments and isotype. Both metrics showed significantly higher inter-individual differences compared to the divergence of repertoires derived from the same donor, reflecting the fact that IGH repertoires of the Bmem, PBL and PL are private to a large degree (**Fig. 2A,B**). We observed identical clonotypes in the repertoires of PBL and PL collected at different time points, while the repertoire similarity was much lower compared to that between the replicate samples, reflecting transient nature of PBL and PL in peripheral blood. Notably, lower clonal overlap in PBL and PL was observed for more distant timepoints (11 or 12 months) than for closer ones (1 month) (**Supplementary Fig. S3A**). The dissimilarity between samples collected at the same day and 1 month or even 1 year later was much lower for Bmem, demonstrating a high level of stability of its' clonal repertoire and long time persistence of IGH clonotypes in memory B cells (**Fig. 2B**, **Supplementary Fig. S3A**).

To better describe the inter-individual IGH repertoire convergence we analysed the number of most expanded IGH amino acid clonotypes shared between different donors (public clonotypes), assuming that functional convergence could be detected among the most abundant clonotypes due to clonal expansions upon response to common pathogens. Indeed, in Bmem the average number of shared clonotypes was higher between fractions of the most abundant clonotypes compared to the randomly sampled clonotypes (**Fig. 2C**). Moreover, it was also significantly higher than the average number of shared clonotypes between two naive repertoires (data from Gidoni et al. 2019) or between pre-immune IGH repertoires obtained by *in silico* generation (**Fig. 2C**). We also highlight that no shared clonotypes defined by their full-length nucleotide sequence were detected (**Supplementary Fig. S3B**). Public clonotypes were also hypermutated, while slightly lower than clonotypes specific to one donor (private) (**Supplementary Fig. S3C**).These observations indicate the presence of functional convergence in memory B-cell repertoires presumably driven by exposure to common pathogens. Of note, clonal overlap between naive repertoires was significantly higher than that of synthetic repertoires assuming functional convergence even in pre-immune repertoires. Furthermore, the distance between V-segment usage distributions in the most abundant Bmem repertoires was not significantly different than that between the naive B-cell subset repertoires. That fact points out that the higher clonotype sharing in memory B cells can not be attributed to the lower diversity in IGHV germline usage (**Fig. 2D**).

The same analysis in plasmablast and plasma cell subpopulations for 600 and 200 most abundant clonotypes respectively yielded no shared clonotypes between repertoires of different donors demonstrating no detectable convergence at this sampling depth.

Finally, we show that public clonotypes were more likely to be detected in samples collected at different time points (**Fig. 2E**), compared to private ones, therefore assuming persistent memory to common antigens.

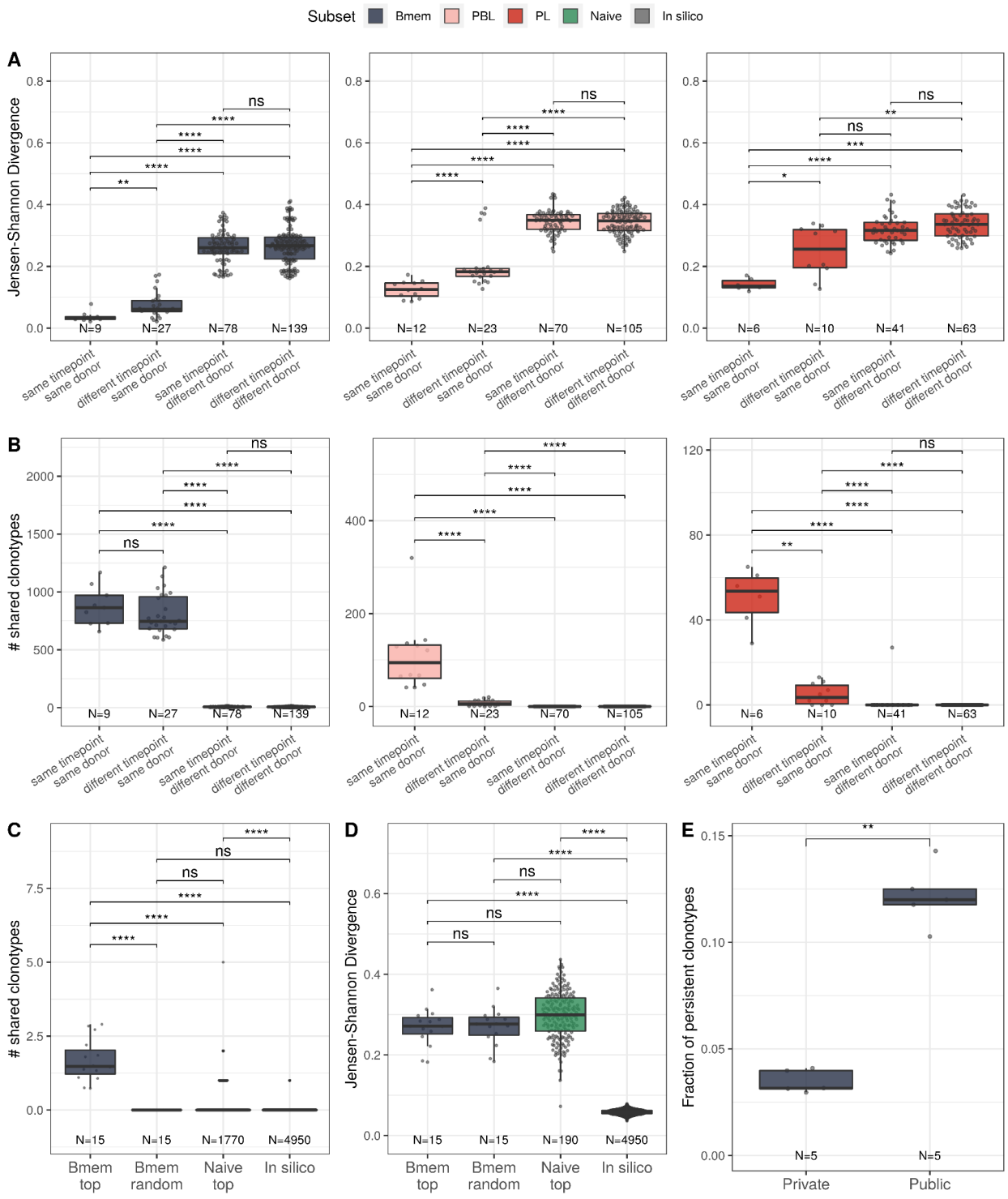

Figure 2. Memory B-cell, plasmablast and plasma cell IGH repertoire reproducibility in time and similarity between individuals. A: Distance between repertoires obtained at different time points from same or different donor calculated as Jensen-Shannon divergence index for IGHV-gene frequency distribution; **B**: Number of shared clonotypes between pairs of repertoires from the same or different donor and time point. For data normalization 14 000 Bmem, 600 PBL and 300 PL most abundant clonotypes were considered; **C:** Number of shared clonotypes between pairs of repertoires from unrelated donors of the following type from left to right: Most abundant clonotypes from Bmem repertoires (Bmem top), randomly selected clonotypes from Bmem repertoires (Bmem random), most abundant clonotypes from naive repertoires of unrelated donors (Naive top, from Gidoni et al. 2019) or from synthetic repertoires, generated with OLGA software (In silico). Number of shared clonotypes between repertoires of the corresponding type, each containing 5000 clonotypes, was calculated. Each dot represents an averaged number of shared clonotypes for a pair of donors; **D**: Inter-individual distance between distributions of V-genes in repertoires calculated as Jensen-Shannon divergence index for pairs of repertoires of the same types as on C; **E**: Fraction of clonotypes, detected in more than one time point (persistent) among clonotypes detected in repertoires from only one donor (private) or in at least two donors (public). Each dot represents a fraction of persistent clonotypes from one donor; In all plots clonotypes are defined as having identical CDR3 amino acid sequence, same IGHV gene segment and isotype. Each dot in A, B, C and D represents a pair of repertoires of corresponding type, numbers below each box indicate the number of pairs of repertoires in the group. Comparisons in all panels were performed with two-sided Mann-Whitney U test, notation of the level of significance is the following: * - $p \leqslant 0.05$, ** - $p \leqslant 0.01$, *** - $p \leqslant 10^{-3}$, **** - $p \leqslant 10^{-4}$.

Temporal dynamics of clonal lineages is associated with cell subset composition

Somatic hypermutation process during affinity maturation of BCR leads to formation of clonal lineages, i.e. BCR clonotypes evolved from a single ancestor after B-cell activation. To study the structure and dynamics of clonal lineages originating from a single BCR ancestor, we grouped clonotypes of each individual based on their sequence similarity (see Material and Methods for details). In what follows, we focused on the larger clonal lineages that consisted of at least 20 unique clonotypes of the corresponding donor. On average, these clonal lineages covered 3.4% of the donor's repertoire, and there were 190 such lineages across the four donors for whom samples were collected at each of the three time points (**Supplementary Fig. S4A**).

We found that these clonal lineages could be divided into two large clusters (further referred as HBmem and LBmem) according to the proportions of cell subsets and BCR isotypes (**Fig. 3A**, **Supplementary Fig. S4B**). The more abundant HBmem cluster includes 138 clonal lineages, and is mostly composed of memory B-cell clonotypes with the non-switched isotype IgM. Conversely, the smaller LBmem cluster (52 clonal lineages) is more diverse and largely composed of ASCs, and enriched in IgG and IgA clonotypes detected in antibody secreting cells (PBL and PL) (**Fig. 3B**). The average size of clonal lineages, i.e. the number of unique clonotypes per lineage, did not differ between the two clusters (**Supplementary Fig. S4C**).

Clonal lineages belonging to both clusters were observed in repertoires of all donors, and the HBmem cluster was more prevalent in each donor (**Supplementary Fig. S4D**).

To better understand the differences between HBmem and LBmem clonal lineages, we tracked the abundance of each clonal lineage in the repertoire across time points. HBmem and LBmem lineages demonstrated different temporal behavior: while HBmem groups were quite stable through time, LBmem lineages had a burst in frequency at one of the time points (**Fig. 3C**).

To compare the temporal stability of clonal lineages, we defined lineage persistence metric, which equals 1 when the clonal lineage was equally frequent at all three time points and is close to 0 when it was detected at just one time point (**Fig. 3D**). Persistence of a clonal lineage was strongly associated with its composition (**Fig. 3E, F**). Clonal lineages enriched with clonotypes of memory B cells or with the IgM isotype, including all HBmem lineages, were more likely to persist through time. Conversely, lineages with larger proportions of antibody-producing cells or IgG/IgA isotypes, including most LBmem lineages, tended to have lower persistence, i.e. had a burst of frequency at some time point. The persistence of a clonal lineage was not associated with its size and the frequencies of clonal lineages were highly correlated among replicate samples (**Supplementary Fig. S4E, F**), indicating that the difference in persistence can not be attributed to clonotype sampling noise .

Besides their higher persistence, the HBmem lineages were enriched in clonotypes detected at multiple time points (**Supplementary Fig. S4G**), indicating that persistent clonal lineages are supported by persistent clonotypes. Furthermore, 29.7% of the HBmem cluster was represented by public clonal lineages (i.e., those shared between at least two donors), compared to 3.8% for the LBmem cluster; the sole two shared LBmem lineages had untypically high persistence, which made them similar to HBmem (**Fig. 3G**).

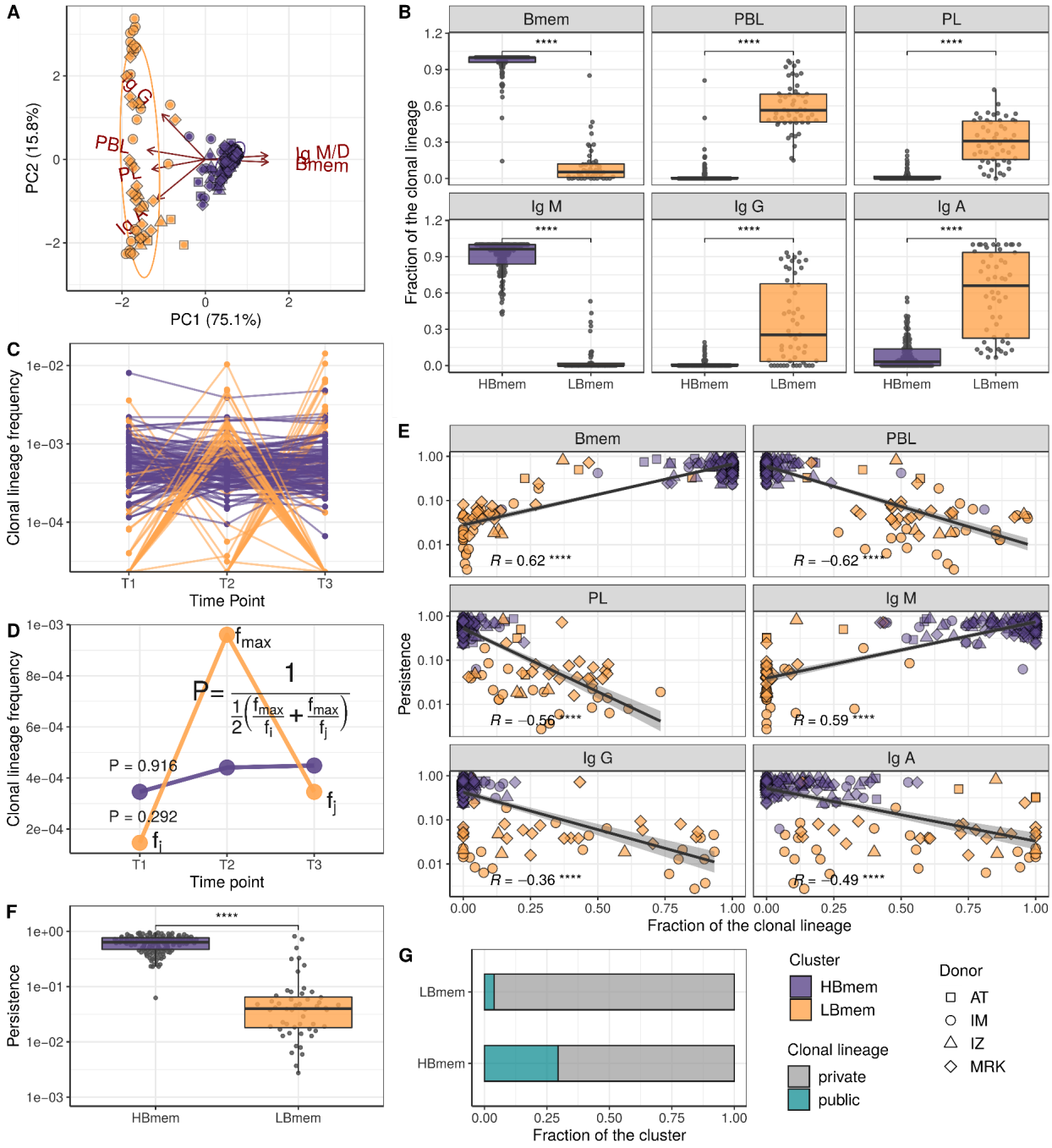

Figure 3. Temporal dynamics and composition of clonal lineages. A: Principal component analysis (PCA) of clonal lineage composition: proportions of B memory cells (Bmem), plasmablasts (PBL) and plasma cells (PL) as well as proportions of isotypes. The arrows represent the projections of the corresponding variables onto the two dimensional PCA plane, with lengths reflecting how well the variable explains the variance of the data. The two principal components (PC1 and PC2) cumulatively explain 90.9% of the variance; **B**: Proportion of clonotypes of a certain cell subset or isotype for clonal lineages falling into HBmem or LBmem clusters; **C**: Dynamics of clonal lineage frequency, defined as the number of clonotypes in a lineage divided by the total number of clonotypes detected at this time point, for HBmem and LBmem clonal lineages. Each line connecting the points represents a unique clonal lineage (N=190); **D**: A schematic representation of calculation of clonal lineage persistence. f_{max} is the maximum clonal lineage frequency among the three time points, and $f_{i,j}$ are the frequencies at the remaining two timepoints; **E**: Spearman's correlation between persistence of a clonal lineage and fractions of its clonotypes attributed to the B-cell subset or isotype; **F**: Comparison of persistence between HBmem and LBmem clonal lineages; **G:** Fraction of public clonal lineages in the cluster, i.e., those shared between at least two donors. Statistical significance for B, F and G is calculated by the two-sided Mann-Whitney test, notation is the following: * - $p \leqslant 0.05$, ** - $p \leqslant 0.01$, *** - $p \leqslant 10^{-3}$, **** - $p \leqslant 10^{-4}$.

LBmem clonal lineages could arise from HBmem clonal lineages

The evolutionary past of a clonal lineage can be described by inferring the history of accumulation of SHMs leading to individual clonotypes, i.e., by reconstructing the phylogenetic tree of the clonal lineage. The initial germline sequence of each clonal lineage partially matches the germline VDJ segments and can be reconstructed, corresponding to the root of the phylogenetic tree of this lineage (see Methods). However, the first node of the phylogenetic tree (green diamond in **Fig. 4A**), the most recent common ancestor (MRCA) of the sampled part of the lineage, can be different from the inferred germline sequence. These differences, referred to as the G-MRCA distance, correspond to the SHMs that were accumulated during the evolution of the clonal lineage prior to divergence of the observed clonotypes. The G-MRCA distance depends on how clonotypes of the tree were sampled. Sampling of clonotypes regardless of their position on the tree results in a low G-MRCA distance (**Fig. 4A**, top panel), while sampling just those clonotypes belonging to some particular clade can hide the early part of lineage evolution from observation and thus result in a large G-MRCA distance (**Fig. 4A**, bottom panel).

The G-MRCA distance was on average 5-fold higher in the LBmem clonal lineages, compared to the HBmem clonal lineages (median = 0.044 vs. 0.008, **Fig. 4B**). This means that, while nearly all the evolution of an HBmem clonal lineage leaves a trace in the observed diversity of such a lineage, the sequence variants of an LBmem lineage typically results from divergence of an already hypermutated clonotype. In most (38 out of 52) LBmem lineages, some B memory clonotypes were observed at the time point preceding the time point of their expansion. Moreover, clonotypes of LBmem lineages are typically characterized by lower pairwise divergence, compared to HBmem lineages (median = 0.11 vs 0.13, **Fig. 4C-E**). Together with the burst-like dynamics characteristic of LBmem lineages (**Fig. 3F**), this implies that LBmem lineages may represent recent rapid clonal expansion of preexisting memory.

Based on these results, together with the compositional features of the two clusters of clonal lineages, we further hypothesized that LBmem clonal lineages may arise from reactivation of pre-existing memory cells belonging to the HBmem cluster. In search of examples of such a transition, we examined all clonal lineages that were persistent but included ASC clonotypes. We found one clear example of a transition from HBmem to LBmem state in the evolutionary history of a clonal lineage (**Fig. 4F**). While the MRCA of this lineage nearly matched the germline sequence, all ASC clonotypes grouped in a single monophyletic clade (sublineage) such that its ancestral node was remote from the MRCA. The ASC sublineage demonstrated all features characteristic for LBmem lineages, namely, predominance of IgG and IgA isotypes, low persistence and low clonotype divergence. Conversely, the remainder of the clonal lineage had features of an HBmem cluster: predominance of IgM, high persistence, and high level of clonotype divergence.

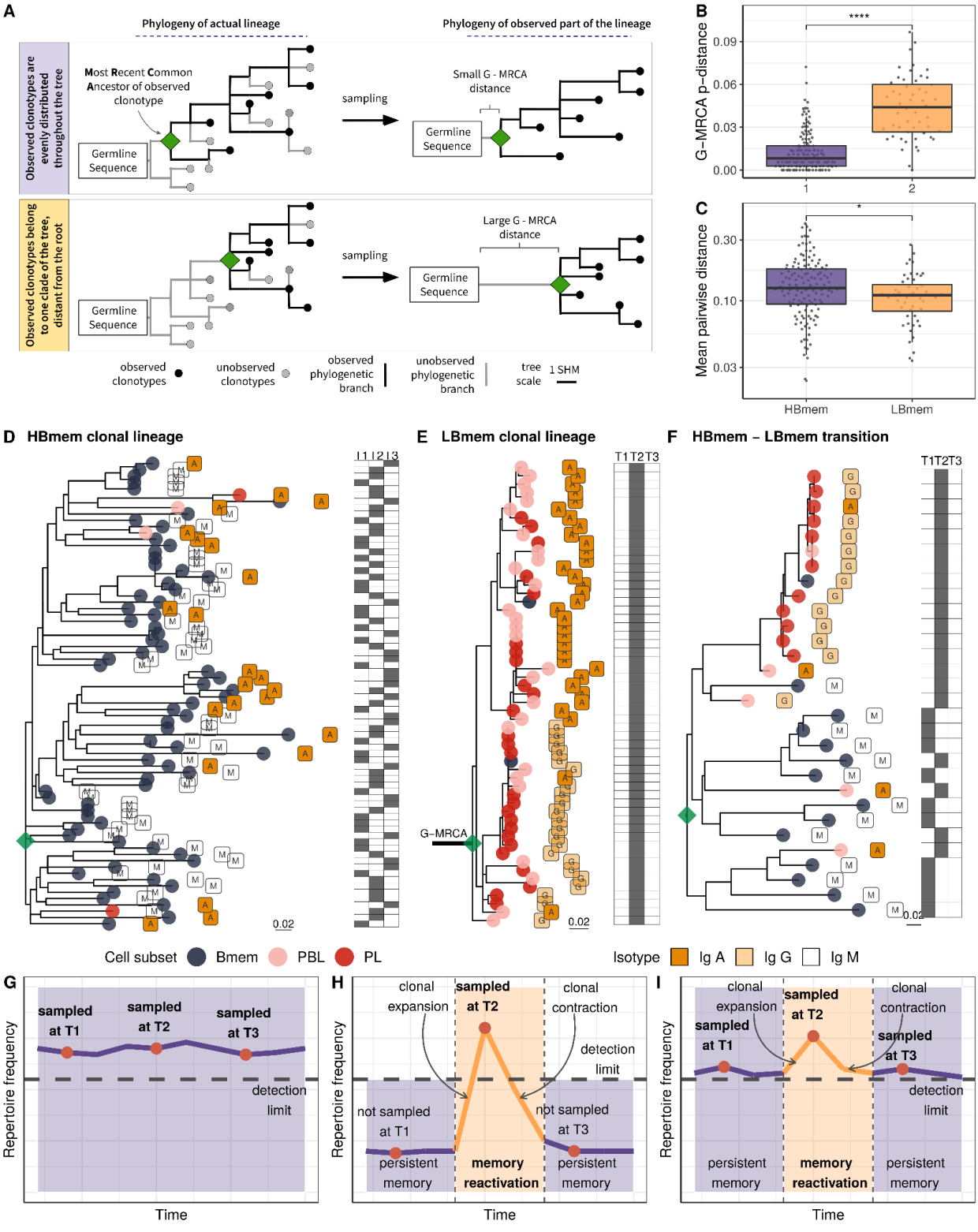

Figure 4. Phylogenetic history of HBmem and LBmem clonal lineages. A: A schematic illustration of how the distances between the germline sequence and the MRCA of a clonal lineage (G-MRCA distance) varies depending on which subset of clonotypes is sampled: a sample uniform with regard to the position on the tree (top panel) or only those belonging to a particular clade of the tree (bottom panel). **B**: Comparison of G-MRCA p-distance (i.e., the fraction of differing nucleotides) for lineages of HBmem and LBmem clusters; **C**: Mean pairwise phylogenetic distance (i.e., the distance along the tree) between the clonotypes of the same lineage for lineages of HBmem and LBmem clusters; **D,E,F**: Representative phylogenetic trees for clonal lineages belonging to HBmem (**D**) and LBmem (**E**) clusters and the case of HBmem-LBmem transition (**F**). LBmem sublineage in **F** is nested deep in the phylogeny of the memory clonotypes and is not characterized by a particularly long ancestral branch, reflecting that it is not an artefact of clonal lineage assignment. Circles correspond to individual clonotypes, with the cellular subset indicated by color, and the isotype, by label. The table at the right of each tree indicates the presence or absence of the corresponding clonotype at each time point. The G-MRCA distance is indicated with a thick line. **G,H,I:** Schematic representation of the hypothetical dynamics of relative size for clonal lineages represented in D, E and F respectively. Level of significance for B and C is obtained by the two-sided Mann-Whitney test: * - $p \leqslant 0.05$, ** - $p \leqslant 0.01$, *** - $p \leqslant 10^{-3}$, **** - $p \leqslant 10^{-4}$.

Reactivation of LBmem clonal lineages is driven by positive selection

Having shown that the LBmem lineages likely originate from clonal expansion of pre-existing memory, we further compared the contribution of positive (favoring new beneficial SHMs) and negative (preserving the current variant) selection between lineages of LBmem and HBmem clusters.

Since we observed only one clear example of an HBmem-LBmem transition in our data (**Fig. 4F**), we could not claim with certainty that LBmem lineages always emerge from preexisting HBmem lineages rather than from some other memory type. Still, we were able to study the LBmem reactivation by comparing the differences in substitution patterns at the origin of HBmem and LBmem clusters. We reasoned that the G-MRCA distance of an HBmem lineage contains the mutations fixed by primary affinity maturation after the first lineage activation; while the G-MRCA distance of an LBmem lineage contains both mutations arose during primary affinity maturation, and subsequent changes that could have happened later in evolution of the lineage. The differences in the characteristics of the G-MRCA mutations between clusters therefore are informative of the process prior to observed expansion of LBmem lineages.

To assess selection at the origin of HBmem and LBmem groups, we measured the divergence of nonsynonymous sites relative to synonymous sites (i.e. dN/dS ratio). In the classical dN/dS test, dN/dS > 1 is interpreted as evidence for positive selection. However, dN/dS > 1 is rare, because the signal of positive selection is usually swamped by that of negative selection. In the McDonald-Kreitman (MK) framework, positive selection is instead revealed from the excess of nonsynonymous divergence relative to nonsynonymous polymorphism (dN/dS > pN/pS, see Methods and **Supplementary Table S2** for examples), under the logic that advantageous

changes contribute more to divergence than to polymorphism (McDonald and Kreitman 1991). The fraction of adaptive nonsynonymous substitutions (α) can then be estimated from this excess. We designed an MK-like analysis, comparing the relative frequencies of nonsynonymous and synonymous SHMs at the G-MRCA branch (equivalent to divergence in the MK test) to those in subsequent evolution of clonal lineages (equivalent to polymorphism in the MK test; **Fig. 5A**, see Methods).

Both in HBmem and LBmem clonal lineages, a higher ratio of nonsynonymous to synonymous SHMs was observed in the G-MRCA branches compared to that found in the subsequent tree branches, meaning that a fraction of SHMs acquired by MRCA was further fixed by positive selection. However, this fraction was higher in the LBmem clonal lineages (Fisher's exact test: α = 0.58 and 0.65 with p-value < 10^{-6} an < 10^{-15} in HBmem and LBmem clusters respectively). α of distinct clonal lineages was also generally higher in the LBmem cluster than in the HBmem cluster (median α = 0.57 vs α = 0.18, **Fig. 5B**), showing that positive selection more frequently preceded the expansion of LBmem than HBmem lineages. The observation of excess α in the LBmem cluster, compared to the HBmem cluster, was robust to the peculiarities of the MK analysis (**Supplementary Table S3**).

The higher α for LBmem compared to HBmem lineages implies that a fraction of SHMs was positively selected in LBmem clonal lineages already after their primary affinity maturation.

Subsequent evolution of LBmem clonal lineages is affected by negative and positive selection

Next, we considered selection which has acted on the HBmem and LBmem clusters since their divergence from their MRCAs, i.e., in the subsequent evolution of a clonal lineage leading to the diversity of the observed clonotypes. For this, we calculated the per-site ratio of nonsynonymous and synonymous SHMs among those that originated after the MRCA (πNπS). πNπS of both clusters was lower than 1. This deficit of nonsynonymous SHMs indicates negative selection in the observed part of clonal lineage evolution. The πNπS ratio was lower in the LBmem cluster, indicating stronger negative selection (**Fig. 5C**).

To examine the selection affecting these post-MRCA SHMs in more detail, we studied the frequency distribution of SHMs in individual lineages, or their site frequency spectra (SFS) (R. Nielsen 2005; Neher and Hallatschek 2013; Nei and Kumar 2000; Horns et al. 2019; Nourmohammad et al. 2019) (**Fig. 5A**). SFS reflects the effect on selection on these SHMs. Deleterious SHMs are held back by negative selection, so that their frequency in the lineage remains low. By contrast, positive selection favors the spread of adaptive SHMs, increasing their frequency. Therefore, negative selection biases the SFS towards low frequencies, and positive selection, towards high frequencies.

For each clonal lineage, we reconstructed the SFS of the SHMs accumulated since its divergence from MRCA (**Fig. 5A**), and then averaged these SFSs within HBmem and LBmem

clusters. A larger proportion of the LBmem SFS corresponds to high frequencies, compared to the HBmem SFS (**Fig. 5D**), indicating weaker negative and/or stronger positive selection in LBmem SFS.

To distinguish between these selection types, we calculated, for each frequency bin, the proportion of the SFS distribution falling into this bin for nonsynonymous SHMs, and divided it by the same value for synonymous SHMs (normalised $\pi N \pi S$, see Methods, **Fig. 5E**). The inter-cluster differences in the normalised $\pi N \pi S$ in low-frequency bins are generally reflective of negative selection, while the differences in the high-frequency bins are reflective of positive selection. The normalised $\pi N \pi S$ was significantly higher in the high-frequency (>60%) bins of SHMs in LBmem clonal lineages. This indicates that in the LBmem cluster, those nonsynonymous changes that were not removed by negative selection reached high frequencies more often than in the HBmem cluster. In total, these data indicate that a fraction of nonsynonymous mutations accumulated by LBmem lineages were adaptive.

Taken together, we observe that reactivation of LBmem lineages is coupled with strengthening of both types of selection: positive on the G-MRCA branch, and both positive and negative during subsequent clonal lineage expansion. Most likely, this pattern is a footprint of new rounds of affinity maturation, which result in acquisition of new advantageous changes and preserve the resulting BCRs from deleterious ones. HBmem instead evolved more neutrally under weaker negative selection, suggesting absence of antigen challenge during observation period (**Fig. 5F**).

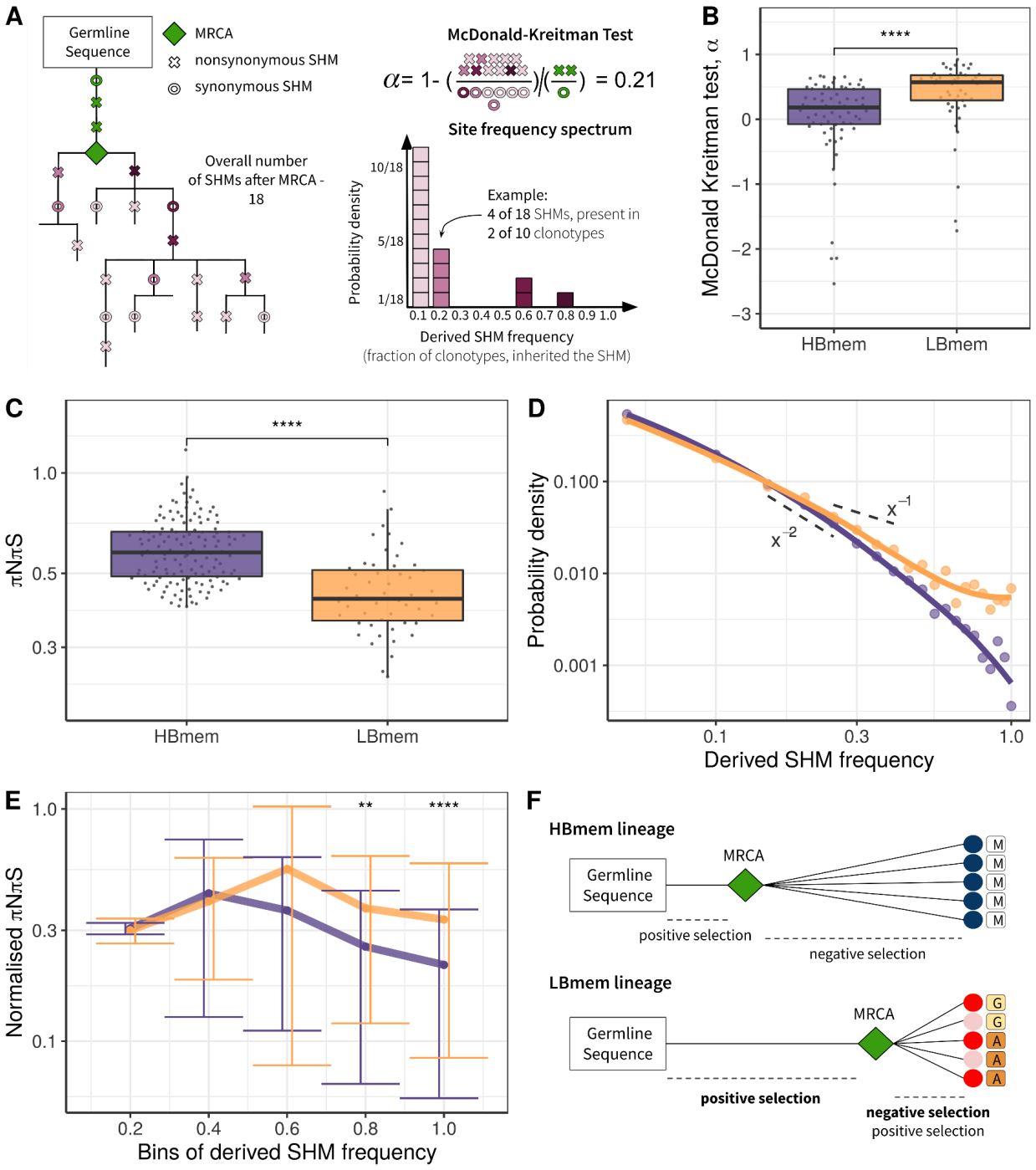

Figure 5. Signatures of positive and negative selection in HBmem and LBmem clusters. A: The schematic representation of McDonald-Kreitman test and site frequency spectrum (SFS) concept; **B**: McDonald-Kreitman estimate of the fraction of adaptive non-synonymous changes α between germline and MRCA in HBmem and LBmem clonal lineages (only those lineages having nonzero G-MRCA distance included, n=68 for HBmem and n=49 for LBmem, see **Supplementary Table S3**); **C**: Comparison of mean pairwise πNπS of HBmem and LBmem lineages; **D**: Averaged site frequency spectrum for HBmem and LBmem clonal lineages. The two dashed lines correspond to $f(x)=x^{-1}$, the expected neutral SFS under Kingman's coalescent model (Kingman 1982), and $f(x)=x^{-2}$; **E**: Comparison of normalised πNπS HBmem and LBmem clonal lineages in bins of SHM frequencies. The number of polymorphisms in each bin is normalised by the overall number of polymorphisms in a corresponding clonal lineage; **F**: Concluding scheme, highlighting features of HBmem and LBmem clonal lineages. Comparisons on B, C and E were performed by two-sided Mann-Whitney test with Bonferroni-Holm multiple testing correction in E, notation of the level of significance is the following: * - $p \leqslant 0.05$, ** - $p \leqslant 0.01$, *** - $p \leqslant 10^{-3}$, **** - $p \leqslant 10^{-4}$.

Discussion

Using advanced library preparation technology, we performed a longitudinal study of full-length BCR repertoires of the three main antigen-experienced B-cell subsets (memory B cells, plasmablasts and plasma cells (together ASC)) from peripheral blood of 6 donors, sampled three times within a year. We analysed repertoires from two conceptually different but complementary points of view. First, we compared various repertoire features between the cell subsets, including clonotype stability in time and convergence between individuals. Second, we tracked the most abundant B-cell clonal lineages in time and analysed their cell subset and isotype composition, phylogenetic history and the mode of selection.

Comparative analysis of the cell subsets revealed significant differences in IGH isotype distribution, rate of SHMs and CDR3 length. IgM clonotypes predominate in the memory subset, when in ASCs switched isotypes, IgA and IgG, together represent more than 80% of repertoire diversity on average. As expected, classical switched isotypes have higher rate of SHMs, and the rate of SHMs in ASCs is in general higher than in memory B-cell subset. Still remaining enigmatic, the IgD isotype in memory B cells shows similarity with IgM: most of IgD clonotypes have low number of SHMs, but there was a fraction of heavily mutated clonotypes. Detected in PB IgD-switched plasma cells and plasmablasts had on average a comparable number of SHMs with IgG- and IgA-expressing clonotypes of ASCs. Notably, the level of SHMs and CDR3 length in plasmablasts on average exceeds thereof in plasma cells in IgM, IgA and IgG isotypes. We hypothesize that such plasmablasts with heavily hypemutated BCRs could be the subset of B-cell progeny that continued to acquire mutations after the optimal affinity have been achieved, while another part of clonal progeny committed to long-lived plasma cell fate and acquired CD138 marker, characteristic for this cell subset (Garimilla et al. 2019).

While different in many aspects, immune-experienced B-cell subsets are similar in terms of IGHV gene segment usage and are concordantly distinct in that from naive B cells. Moreover, we observed that the correlated enrichment/depletion in V segment usage frequency in general coincides with the level of sequence similarity of the V segments: most of IGHV-3 family members are more frequent in antigen-experienced B cells compared to naive subset, in all donors and time points, while most of other well-represented in naive subset V genes decreased. The differences in V usage frequencies between naive and antigen-experienced B-cell subsets were also reported in several previous studies, despite different FACS gating strategies were used (Mitsunaga and Snyder 2020; Ghraichy et al. 2021). Our findings further support the idea that initial recruitment of the B cells to the immune response in many cases is determined by the germline-encoded parts of the B-cell receptors, presumably CDR1 and CDR2. High level of convergence in IGHV usage between B-cell clonotypes specific for particular pathogens or self-antigens was shown in previous studies (Peng et al. 2019; Galson et al. 2015; Bashford-Rogers et al. 2019).

We further analysed the level of repertoire similarity of cell subsets in time and between individuals. Intuitively, the memory B-cell subset is the most stable in time, showing less repertoire divergence and higher number of shared clonotypes between sampling time points of the same individuals. Comparing this with between-individual sharing, we detected the very low number of common clonotypes in memory B cells. Those clonotypes have a comparable number of SHM as private ones, assuming germinal center dependent origin. Two recent studies on extra-deep repertoires of bulk peripheral blood B cells reported 1-6% (Soto et al. 2019) or ~1% (Briney et al. 2019) of shared V-CDR3aa-J clonotypes between a pair of unrelated donors with lower repertoire convergence for class-switched clonotypes shown in the latter study. Similarly, using the same method we measured 0,06% of repertoire overlap in the Bmem subset (**Supplementary Figure S3D**). Complementing the model proposed by Briney with colleagues, - that initial IGH repertoires are dissimilar, then they homogenize during B-cell development and finally become highly individual after immunological exposure, we found significantly higher number of shared clonotypes between IGH repertoires among most abundant memory B-cell clonotypes, indicating functional convergence presumably due to exposure to common environmental antigens. The latter is further supported by the higher number of persisting memory B-cell clonotypes observed among public clonotypes compared to private ones.

Next, we focused on the most abundant B-cell clonal lineages, which are large enough to study the interconnection between cell subsets and phylogenetic features of the lineages. In all individuals the observed clonal lineages clearly fall into two clusters by their features. The first HBmem cluster represents persistent memory with the predominance of IgM isotype. Such clonal lineages were equally sampled from all time points and rarely included clonotypes from ASCs. Most recent common ancestor (MRCA) of observed clonotypes in HBmem lineages almost matches with the predicted germline sequence (in 14.5% of the lineages the match is complete), indicating that probability to observe clonotype from these lineages has no

association with the position on lineage's phylogeny. Horns and colleagues observed lineages with very similar features to our HBmem cluster. They also possessed persistent dynamics on the background of vaccine-responsive lineages and were predominantly composed of IgM isotype. However their study was performed on bulk B cells, so there was no possibility to track their relatedness to B-cell memory subset (Horns et al. 2019).

The second LBmem cluster demonstrates completely different features: LBmem lineages are mostly composed of ASC clonotypes with switched IgA or IgG isotypes, showing active involvement in ongoing immune response. MRCA of LBmem lineages stood out from the germline sequence by some number of SHMs, and only 1.9% of LBmem lineages had complete match between MRCA and the germline sequence. Large G-MRCA distance implies that observed clonotypes originated from the already hypermutated ancestor and we thus sampled clonotypes from a single clade of lineage phylogeny. Such an effect can be caused by both rapid expansion of the clade and migration of clade's clonotypes in a peripheral blood. We also observe that most LBmem lineages expanded at time point 2 or 3 (38 out of 45, > 80%) had at least one clonotype detected in Bmem subset at the previous time point (T1 or T2 respectively), leading as to the conclusion, that LBmems represent progeny of reactivated persistent memory B cells. We found one lineage, which possesses all features of HBmem cluster except one monophyletic clade, typical for LBmem lineage, thus at least some of LBmem clonal lineages represent progeny of reactivated persistent memory B cells, represented among HBmem lineages. This example of HBmem-LBmem transition is very similar to reactivated persistent memory, observed by Hoehn et al. in response to seasonal flu vaccination (Hoehn et al. 2021).

Analysis of selection mode in HBmem and LBmem lineages supported our assumptions. We showed that both types of lineages experienced positive selection from the germline sequence to MRCA of observed clonotypes. Such observation is expected, assuming that primary B-cell activation is followed by affinity maturation associated with clonal lineage expansion. However the pressure of positive selection is stronger in LBmem lineages than in HBmem. In addition we detected an excess of sites under positive selection in LBmem lineages in evolution after the MRCA as well. Listed above lead us to the conclusion that LBmems underwent additional rounds of affinity maturation after their reactivation. The mode of selection in reactivated lineages, observed by Hoehn et al., was not studied, however some clonotypes were sampled from germinal centers, assuming involvement in affinity maturation.

In subsequent evolution after the MRCA we detected negative selection in both groups of lineages and again it was stronger in LBmem. We consider this excess of negative selection in LBmem as an additional footprint of affinity maturation, purifying the lineage from deleterious BCR variants.

To conclude, we performed a detailed longitudinal analysis of BCR repertoires of immune experienced B-cell subsets from donors without severe pathologies, and provided a framework for comprehensive analysis of selection in BCR clonal lineages.

Our results demonstrate the interconnection of B-cell subsets on clonal level, B-cell memory convergence in unrelated donors and long-term persistence of the memory-enriched clonal lineages in peripheral blood. Signs of positive selection were detected in both memory- and ASC-dominated B-cell lineages. Together the results of evolutionary analysis of B cell clonal lineages coupled with B-cell subset annotation suggest that the reactivation of pre-existing memory B cells is accompanied by new rounds of affinity maturation.

Methods

Donors, cells, timepoints

Blood samples from 6 young and middle-aged (**Supplementary Table S1**) donors without severe inflammatory diseases, chronic and recent acute infectious diseases or vaccinations were collected at three time points: T1 - 0, T2 - 1 month, T3 - 12 months **(Fig. 1A)**. 4 of the donors suffered with allergic rhinitis to pollen, 2 of which also suffered from food allergy. Informed consent was obtained from each donor. The study was approved by the Ethical Committee of Pirogov Russian National Research Medical University, Moscow, Russia.

At each time point 18-22 ml of peripheral blood was collected in BD Vacuette tubes with EDTA. Peripheral blood mononuclear cells were isolated using ficoll gradient density centrifugation.To isolate sub populations of interest cells were stained with anti-CD19-APC, anti-CD20-VioBlue, anti-CD27-VioBright FITC, anti-CD138-PE-Vio770 (all Miltenyi Biotec) in presence of FcR Blocking Reagent (Miltenyi Biotec) according to the manufacturer protocol and sorted using fluorescence activated cell sorting (BD FacsAria III, BD Biosciences) into the following populations: memory B cells ($CD19^{+}$ $CD20^{+}$ $CD27^{+}$ $CD138^{-}$, plasmablasts ($CD20^{-}$ $CD19^{Low/+}$ $CD27^{++}$ $CD138^{-}$), plasma cells ($CD20^{-}$ $CD19^{Low/+}$ $CD27^{++}$ $CD138^{+}$). For each donor in time point T1 one replicate sample of each cell subpopulation was collected, for time points T2 and T3 two replicate samples were collected (50×10^3 to 100×10^3 memory B cells, 1×10^3 to 2×10^3 plasmablasts, 0.5×10^3 to 1×10^3 plasma cells per sample).

Full-length IGH cDNA libraries and sequencing

Immunoglobulin heavy chain (IGH) cDNA libraries were prepared according to the protocol described earlier (Turchaninova et al. 2016) with several modifications. Briefly, we used RACE (Rapid Amplification of cDNA Ends) approach with a template-switch effect to introduce 5' adaptors during cDNA synthesis. These adaptors contained both unique molecular identifiers (UMIs), allowing error-correction, and sample barcodes (described in Zvyagin et al. 2017) allowing to rule out all potential cross-sample contaminations. Besides, a universal sequence for annealing forward PCR primer, also introduced within a 5' adaptor during reverse transcription (RT) reaction, allows avoiding usage of multiplex forward primers specific for V segments to reduce PCR amplification biases.

Multiplex C-segment specific primers used for RT and PCR allowed us to preserve isotype information. Prepared libraries were then sequenced with Illumina HiSeq 2000/2500, paired-end reading (2x310 bp).

Sequencing data pre-processing and repertoire reconstruction

Sample demultiplexing by sample-barcodes introduced in 5' adapter and UMI-based error-correction were performed using MIGEC v1.2.7 software (Shugay et al. 2014). For further analysis we used sequences covered by at least two sequencing reads. Alignment of sequences, V-,D-,J-, C-segment annotation and reconstruction of clonal repertoires was accomplished using MiXCR v3.0.10 (Bolotin et al. 2015) with prior removal of the primer-originated part in C-segment. We defined clonotypes as a unique IGH nucleotide sequence, starting from Framework 1 region of V-segment to the end of J-segment and taking into account isotype.

Using TIgGER (Gadala-Maria et al. 2015) software we derived an individual database of V gene alleles for each donor and realigned all sequences for precise detection of hypermutations. For general repertoire characteristics analysis (Isotype frequencies, somatic hypermutation levels, CDR3 length, IGHV gene usage and repertoire similarity metrics) we used samples covered by at least 0.1 cDNA molecules per cell for Bmem, and at least 5 for PBL and PL.

Repertoire characteristics analysis

Isotype frequencies, rate of somatic hypermutations and CDR3 lengths were determined using MiXCR v3.0.10 (Bolotin et al. 2015). For calculation of background IGHV gene segment usage and number of shared clonotypes we utilized data derived from Gidoni et al. 2019 (European Nucleotide Archive accession number ERP108501) that represents naive B-cell IGH repertoires. We used repertoires containing more than 5000 clonotypes and processed them in the same way with our data. IGHV gene frequencies were calculated as a number of unique clonotypes to which a particular IGHV gene was annotated by MiXCR divided by total number of clonotypes identified in this sample. To assess IGHV gene segments over- and under-represented in studied subsets we utilized edgeR package v0.4.4 (Robinson et al., 2010) with 'trended' dispersion model. To evaluate pairwise similarity between repertoires based on IGHV gene segment frequency distributions we utilized Jensen-Shannon divergence, which was calculated using formula:

$$JS(P,Q) = \frac{1}{2}\sum_i p_i log_2 p_i + \frac{1}{2}\sum_i q_i log_2 q_i - \sum_i \left(\frac{p_i+q_i}{2} log_2 \left(p_i + q_i \right) \right)$$

where P and Q represent distributions of IGHV gene segment in two repertoires, p_i and q_i represent frequencies of individual member *i* of the population (IGHV gene segment).

In-silico repertoires used for calculation of background clonal overlap were generated with OLGA software v1.0.2 (Sethna et al. 2019) under standard settings utilizing the built-in model.

For clonal overlap calculation we downsized repertoires to a fixed number of clonotypes. For **Fig. 1B** 14 000 most abundant clonotypes were considered in Bmem, 600 in PBL and 300 in PL. For **Fig. 1C** we considered 5000 clonotypes for all cell subsets. Clonotypes with identical CDR3 amino acid sequence and same IGHV gene segment detected in both analyzed samples were considered shared. Clonotypes shared between repertoires of at least two individuals were termed as public.

Assignment of clonal lineages

Change-O v0.4.4 (Gupta et al. 2015) was utilized to assign clonal groups, defined as groups of clonotypes with the same V-segment, CDR3 length and at least 85% similarity in CDR3 nucleotide sequence. Before clonal group assignment we excluded all clonotypes with counts equal to 1. Clonal groups represent observed subsets of clonal lineages originating from a single BCR ancestor, so for simplicity we use the term of clonal lineages instead. To study evolutionary dynamics of clonal lineages we joined all replicas, three time points (T1, T2, T3) and cell subsets for each patient in a single dataset and excluded clonotypes that were presented by a single UMI. Phylogenetic analysis was performed on four patients, for whom we had samples in all considered time points, and on clonal lineages, containing at least 20 unique clonotypes as in (Nourmohammad et al. 2019).

Clusterization of clonal lineages on HBmem and LBmem clusters

We performed principal component analysis on six scaled variables of clonal lineage composition: fractions of memory B cells, plasmablasts, plasma cells and fractions of IgM, IgG and IgA. IgE isotype was not detected in clonal lineages involved in phylogenetic analysis, so we did not include its fraction as a variable. HBmem and LBmem clusters were defined using the K-means clustering algorithm.

Metric of persistence of clonal lineages

We estimated the frequency of the clonal lineage in the repertoire of a given time point as a ratio between the number of unique clonotypes in the clonal lineage, detected at this time point, to the overall number of unique clonotypes, detected in this time point. If the clonal lineage was not detected at some time point, we assigned its frequency to pseudocount, as it would be a single clonotype detected from this time point. To estimate persistence of clonal lineage frequency in the repertoire through time we introduced a metric of the same name:

$$P = \frac{1}{\frac{1}{2}\left(\frac{f_{max}}{f_i} + \frac{f_{max}}{f_j}\right)},$$

where f_{max} is a maximum frequency of the clonal lineage among three time points and $f_{i,j}$ are frequencies in the other two (**Fig. 3D**). Persistence is equal to 1, if the clonal lineage has equal repertoire frequencies at all three time points. If the clonal lineage was detected just once in the experiment and frequencies at other two time points were assigned to pseudocounts, persistence becomes close to zero.

Reconstruction of clonal lineage germline sequence

We used MiXCR-derived reference V-, D- and J-segment sequences to reconstruct IGH germline sequences for each clonal lineage, concatenating only those sequence fragments which were present at CDR3 junctions of original MiXCR-defined clonotypes. Thus, random nucleotide insertions were disregarded, making them appear as gaps in the alignment of lineage clonotypes with the germline sequence. We excluded them from all parts of the phylogenetic analysis, where germline sequence was required.

Reconstruction of clonal lineage phylogeny and MRCA

For phylogenetic analysis of clonal lineages first we aligned clonotypes of clonal lineages together with reconstructed germline sequence using MUSCLE version 3.8.31 with 400 gap open penalty (Edgar 2004). Next we reconstructed the clonal lineage's phylogeny by RAxML version 8.2.11 using GTRGAMMA evolutionary model and germline sequence as an outgroup, and computed marginal ancestral states (Stamatakis 2014). The ancestral sequence of the node closest to the root of the tree, represented by germline sequence, is a most recent common ancestor of the sampled clonotypes (MRCA). It can match with the germline sequence or stand out from it by some amount of SHMs, reflecting the starting point of subsequent evolution of observed clonotypes. It allowed us to distinguish between SHMs, fixed in the clonal lineage on the way from the germline sequence to the MRCA (G-MRCA SHMs), and polymorphisms within the observed part of lineage. G-MRCA p-distance on **Fig. 4B** was measured as a fraction of diverged positions between germline and MRCA sequences.

McDonald-Kreitman test

McDonald-Kreitman (MKT) test is aimed to detect positive or negative selection on the way of population divergence from another species or its ancestral state (McDonald and Kreitman 1991). It is based on comparison of ratios of nonsynonymous to synonymous substitutions, observed in diverged and polymorphic sites and estimates the fraction of diverged amino acid substitutions fixed by positive selection:

$$\alpha = 1 - \frac{P_n}{P_s} \cdot \frac{D_s}{D_n},$$

where P reflects the number of polymorphisms, nonsynonymous (P_n) and synonymous (P_s) and D - the number of divergences, fixed in the population, nonsynonymous (D_n) and synonymous (D_s) as well.

Under neutral evolution nonsynonymous and synonymous changes are equally likely to be fixed or appear in the population as polymorphisms, so $\frac{D_n}{D_s} = \frac{P_n}{P_s}$ and α = 0. Positive selection favors adaptive nonsynonymous changes to be fixed and increases $\frac{D_n}{D_s}$ ratio over $\frac{P_n}{P_s}$, that results in $\alpha > 0$. Negative selection has the opposite effect and produces $\alpha < 0$.

To detect selection in the origin of clonal lineages, we considered G-MRCA SHMs as divergent changes, and the remaining SHMs in a clonal lineage after the MRCA as polymorphic ones (**Fig. 5A**). If we observed different nucleotides in the germline sequence and MRCA at the site, which was also polymorphic, we considered it as divergent only if the germline variant was not among polymorphisms (**Supplementary Table S2**, examples of codons q and r). Codons with unknown germline state were excluded from the MKT test (**Supplementary Table S2**, example of codon j).

To perform MKT test on joined HBmem or LBmem cluster variation we summed variation of all clonal lineages of the same cluster in each category (D_n, D_s, P_n, P_s). Calculations of α of distinct clonal lineages for comparison of its distributions between two clusters were complicated by zero G-MRCA distance in some clonal lineages, mostly belonging to HBmem cluster. To deal with it we used three approaches, presented in **Supplementary Table S3**. In the first one we added pseudocounts to D_n and D_s in each clonal lineage, so clonal lineages with zero G-MRCA distance $\frac{D_n}{D_s} = 1$. In the second one we excluded clonal lineages with zero G-MRCA distance from the analysis, still adding pseudocounts to D_n and D_s in each clonal lineage in the case if G-MRCA distance consists of just one nonsynonymous or synonymous substitution. And in the third one we compared only those clonal lineages that had at least one nonsynonymous and at least one synonymous substitution on the G-MRCA branch. We also calculated the MKT test on joined variation for all types of exclusion criteria to check its robustness, however there is no need to exclude clonal lineages in the case of joined test (**Supplementary Table S3**).

In the first approach clonal lineages with zero G-MRCA distance always produced negative α and biased median α to negative values as well. Medians of α in the second and the third approaches were more consistent with results of the test on joined variation. However in the third approach the filter excluded the most part of HBmem cluster, so in the main test we presented results of the second one (**Fig. 5B**). To check the significance of deviation of α from

neutral expectations we used an exact Fisher test as in original MKT pipeline (McDonald and Kreitman 1991).

πNπS

To calculate πNπS we identified SHMs in each clonal lineage relative to the reconstructed MRCA sequence. In multiallelic sites (sites with multiple SHMs observed, see codon i in **Supplementary Table S2** as an example) we considered each variant as an independent SHM event. πN and πS were calculated as a number of nonsynonymous and synonymous SHMs in a clonal lineage, normalised by the number of nonsynonymous and synonymous sites in MRCA sequence respectively. Resulting πNπS value is a ratio between πN and πS:

$\pi N \pi S = \frac{N}{N_S} : \frac{S}{S_S}$, where N (S) - number of nonsynonymous (synonymous) SHMs, observed in the clonal lineage and N_S (S_S)- number of nonsynonymous (synonymous) sites in the MRCA sequence of the clonal lineage, calculated as in (Gojobori 1986).

Site frequency spectrum

Site frequency spectrum (SFS) reflects the distribution of SHMs frequencies in the clonal lineage. We calculated the frequency of each SHM as a number of unique clonotypes, carrying the SHM, relative to the overall number of unique clonotypes in the lineage. To visualise SFS we binned SHM frequencies by 20 equal intervals with the step 0.05 (0; 0.05; 0.1; 0.15; 0.2; …. 0.9; 0.95; 1) and counted SHM density in each bin as the number of SHMs of bin's frequencies normalised by the overall number of SHMs detected in the lineage. To obtain the cluster average SFSs we took a mean of clonal lineages of the same cluster in each frequency bin.

Normalised πNπS in bins of SHM frequencies

To compare ratios of nonsynonymous and synonymous SHMs of different frequencies between two clusters we calculated normalised πNπS in bins of SHM frequency. For this purpose we used a lower number of frequency bins (0; 0.2; 0.4; 0.6; 0.8; 1) to reduce the probability of bins without observed SHMs. To deal with remaining empty bins we added pseudocounts to nonsynonymous and synonymous SHMs in each frequency bin. Thus, normalised πNπS in *i*-th SHM frequency bin was calculated as following:

normalised $\pi N \pi S = \frac{(N_i+1)/(\sum_{i=1}^{5} N_i+5)/N_S}{(S_i+1)/(\sum_{i=1}^{5} S_i+5)/S_S}$, where $N_i(S_i)$ is the number of nonsynonymous (synonymous) SHMs in *i*-th frequency bin, $\sum_{i=1}^{5} N_i(S_i)$ is the overall number of nonsynonymous (synonymous) SHMs observed in the clonal lineage (sum of SHMs in all frequency bins); N_S (S_S)

- number of nonsynonymous (synonymous) sites in the MRCA sequence of the clonal lineage calculated as in (Gojobori 1986). To compare distributions of normalised πNπS between two clusters of clonal lineages in 5 frequency bins we used Mann-Whitney test with Bonferroni-Holm multiple testing correction.

Data analysis and visualisation

All analysis was performed using R language (R Core Team 2018) and visualized with the ggplot2 package (Ginestet 2011). Ggtree package was used to visualise phylogenetic trees of clonal lineages (Yu et al. 2017). The code for repertoire analysis is available at https://github.com/amikelov/igh_subsets; the code for clonal lineage analysis is available at https://github.com/EvgeniiaAlekseeva/Clonal_group_analysis.

Data availability

Sequencing data have been deposited in the ArrayExpress database (www.ebi.ac.uk/arrayexpress, acc. num. E-MTAB-11193).

Acknowledgements

We are grateful to our donors. We are grateful to Alexey Neverov for helpful discussion of inference of selection. The study was supported by grant of the Ministry of Science and Higher Education of the Russian Federation grant #075-15-2020-807 (to D.M.C), and by grant of the RBFR, project number 20-34-90153 (to E.I.A).

Author contribution

A.I.M. and I.V.Z. designed the study, collected samples and prepared cDNA libraries; M.A.T. participated in optimisation of cDNA library preparation; E.A.K. and D.B.S. performed FACS sorting; A.I.M. performed repertoire data processing and analysis; E.I.A. and G.A.B. designed evolutionary analysis of clonal lineages; E.I.A. performed evolutionary analysis of clonal lineages; M.A.S. and D.M.C. participated in study design and discussions; A.I.M., E.I.A., I.V.Z. and G.A.B. wrote the manuscript with the contribution of all authors.

Competing Interests Statement

Authors declare no competing interests.

References

Bashford-Rogers, R. J. M., L. Bergamaschi, E. F. McKinney, D. C. Pombal, F. Mescia, J. C. Lee, D. C. Thomas, et al. 2019. “Analysis of the B Cell Receptor Repertoire in Six Immune-Mediated Diseases.” *Nature* 574 (7776): 122–26. https://doi.org/10.1038/s41586-019-1595-3.

Bolotin, Dmitriy A, Stanislav Poslavsky, Igor Mitrophanov, Mikhail Shugay, Ilgar Z Mamedov, Ekaterina V Putintseva, and Dmitriy M Chudakov. 2015. “MiXCR: Software for Comprehensive Adaptive Immunity Profiling.” *Nature Methods* 12 (5): 380–81. https://doi.org/10.1038/nmeth.3364.

Bonsignori, Mattia, Hua-Xin Liao, Feng Gao, Wilton B. Williams, S. Munir Alam, David C. Montefiori, and Barton F. Haynes. 2017. “Antibody-Virus Co-Evolution in HIV Infection: Paths for HIV Vaccine Development.” *Immunological Reviews* 275 (1): 145–60. https://doi.org/10.1111/imr.12509.

Bourcy, Charles F. A. de, Cesar J. Lopez Angel, Christopher Vollmers, Cornelia L. Dekker, Mark M. Davis, and Stephen R. Quake. 2017. “Phylogenetic Analysis of the Human Antibody Repertoire Reveals Quantitative Signatures of Immune Senescence and Aging.” *Proceedings of the National Academy of Sciences* 114 (5): 1105–10. https://doi.org/10.1073/pnas.1617959114.

Briney, Bryan, Anne Inderbitzin, Collin Joyce, and Dennis R. Burton. 2019. “Commonality despite Exceptional Diversity in the Baseline Human Antibody Repertoire.” *Nature* 566 (7744): 393–97. https://doi.org/10.1038/s41586-019-0879-y.

Chaudhary, Neha, and Duane R. Wesemann. 2018. “Analyzing Immunoglobulin Repertoires.” *Frontiers in Immunology* 9 (March): 462. https://doi.org/10.3389/fimmu.2018.00462.

Davydov, Alexey N., Anna S. Obraztsova, Mikhail Y. Lebedin, Maria A. Turchaninova, Dmitriy B. Staroverov, Ekaterina M. Merzlyak, George V. Sharonov, et al. 2018. “Comparative Analysis of B-Cell Receptor Repertoires Induced by Live Yellow Fever Vaccine in Young and Middle-Age Donors.” *Frontiers in Immunology* 9 (October): 2309. https://doi.org/10.3389/fimmu.2018.02309.

De Silva, Nilushi S., and Ulf Klein. 2015. “Dynamics of B Cells in Germinal Centres.” *Nature Reviews Immunology* 15 (3): 137–48. https://doi.org/10.1038/nri3804.

Edgar, R. C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97. https://doi.org/10.1093/nar/gkh340.

Gadala-Maria, Daniel, Gur Yaari, Mohamed Uduman, and Steven H. Kleinstein. 2015. “Automated Analysis of High-Throughput B-Cell Sequencing Data Reveals a High Frequency of Novel Immunoglobulin V Gene Segment Alleles.” *Proceedings of the National Academy of Sciences* 112 (8): E862–70. https://doi.org/10.1073/pnas.1417683112.

Gaebler, Christian, Zijun Wang, Julio C. C. Lorenzi, Frauke Muecksch, Shlomo Finkin, Minami Tokuyama, Alice Cho, et al. 2021. “Evolution of Antibody Immunity to SARS-CoV-2.” *Nature* 591 (7851): 639–44. https://doi.org/10.1038/s41586-021-03207-w.

Galson, Jacob D, Elizabeth A Clutterbuck, Johannes Trück, Maheshi N Ramasamy, Márton Münz, Anna Fowler, Vincenzo Cerundolo, Andrew J Pollard, Gerton Lunter, and Dominic F Kelly. 2015. “BCR Repertoire Sequencing: Different Patterns of B-cell Activation after Two Meningococcal Vaccines.” *Immunology & Cell Biology* 93 (10): 885–95.

https://doi.org/10.1038/icb.2015.57.

Garimilla, Swetha, Doan C. Nguyen, Jessica L. Halliley, Christopher Tipton, Alexander F. Rosenberg, Christopher F. Fucile, Celia L. Saney, et al. 2019. “Differential Transcriptome and Development of Human Peripheral Plasma Cell Subsets.” *JCI Insight* 4 (9): e126732. https://doi.org/10.1172/jci.insight.126732.

Ghraichy, Marie, Valentin von Niederhäusern, Aleksandr Kovaltsuk, Jacob D Galson, Charlotte M Deane, and Johannes Trück. 2021. “Different B Cell Subpopulations Show Distinct Patterns in Their IgH Repertoire Metrics.” *ELife* 10 (October): e73111. https://doi.org/10.7554/eLife.73111.

Gidoni, Moriah, Omri Snir, Ayelet Peres, Pazit Polak, Ida Lindeman, Ivana Mikocziova, Vikas Kumar Sarna, et al. 2019. “Mosaic Deletion Patterns of the Human Antibody Heavy Chain Gene Locus Shown by Bayesian Haplotyping.” *Nature Communications* 10 (1): 628. https://doi.org/10.1038/s41467-019-08489-3.

Ginestet, Cedric. 2011. “Ggplot2: Elegant Graphics for Data Analysis: Book Reviews.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174 (1): 245–46. https://doi.org/10.1111/j.1467-985X.2010.00676_9.x.

Gojobori, Nei. 1986. “Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions.” *Molecular Biology and Evolution*, September. https://doi.org/10.1093/oxfordjournals.molbev.a040410.

Grimsholm, Ola, Eva Piano Mortari, Alexey N. Davydov, Mikhail Shugay, Anna S. Obraztsova, Chiara Bocci, Emiliano Marasco, et al. 2020. “The Interplay between CD27dull and CD27bright B Cells Ensures the Flexibility, Stability, and Resilience of Human B Cell Memory.” *Cell Reports* 30 (9): 2963-2977.e6. https://doi.org/10.1016/j.celrep.2020.02.022.

Gupta, Namita T., Jason A. Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Gur Yaari, and Steven H. Kleinstein. 2015. “Change-O: A Toolkit for Analyzing Large-Scale B Cell Immunoglobulin Repertoire Sequencing Data: Table 1.” *Bioinformatics* 31 (20): 3356–58. https://doi.org/10.1093/bioinformatics/btv359.

Hoehn, Kenneth B., Jackson S. Turner, Frederick I. Miller, Ruoyi Jiang, Oliver G. Pybus, Ali H. Ellebedy, and Steven H. Kleinstein. 2021. “Human B Cell Lineages Engaged by Germinal Centers Following Influenza Vaccination Are Measurably Evolving.” Preprint. Immunology. https://doi.org/10.1101/2021.01.06.425648.

Horns, Felix, Christopher Vollmers, Cornelia L. Dekker, and Stephen R. Quake. 2019. “Signatures of Selection in the Human Antibody Repertoire: Selective Sweeps, Competing Subclones, and Neutral Drift.” *Proceedings of the National Academy of Sciences* 116 (4): 1261–66. https://doi.org/10.1073/pnas.1814213116.

Laserson, Uri, Francois Vigneault, Daniel Gadala-Maria, Gur Yaari, Mohamed Uduman, Jason A. Vander Heiden, William Kelton, et al. 2014. “High-Resolution Antibody Dynamics of Vaccine-Induced Immune Responses.” *Proceedings of the National Academy of Sciences* 111 (13): 4928–33. https://doi.org/10.1073/pnas.1323862111.

Mandric, Igor, Jeremy Rotman, Harry Taegyun Yang, Nicolas Strauli, Dennis J. Montoya, William Van Der Wey, Jiem R. Ronas, et al. 2020. “Profiling Immunoglobulin Repertoires across Multiple Human Tissues Using RNA Sequencing.” *Nature Communications* 11 (1): 3126. https://doi.org/10.1038/s41467-020-16857-7.

McDonald, John H., and Martin Kreitman. 1991. “Adaptive Protein Evolution at the Adh Locus in Drosophila.” *Nature* 351 (6328): 652–54. https://doi.org/10.1038/351652a0.

Mitsunaga, Erin M., and Michael P. Snyder. 2020. “Deep Characterization of the Human Antibody Response to Natural Infection Using Longitudinal Immune Repertoire

Sequencing." *Molecular & Cellular Proteomics* 19 (2): 278–93. https://doi.org/10.1074/mcp.RA119.001633.

Neher, R. A., and O. Hallatschek. 2013. "Genealogies of Rapidly Adapting Populations." *Proceedings of the National Academy of Sciences* 110 (2): 437–42. https://doi.org/10.1073/pnas.1213113110.

Nei, Masatoshi, and Sudhir Kumar. 2000. *Molecular Evolution and Phylogenetics*. Oxford ; New York: Oxford University Press.

Nielsen, Rasmus. 2005. "Molecular Signatures of Natural Selection." *Annual Review of Genetics* 39 (1): 197–218. https://doi.org/10.1146/annurev.genet.39.073003.112420.

Nielsen, Sandra C.A., Fan Yang, Katherine J.L. Jackson, Ramona A. Hoh, Katharina Röltgen, Grace H. Jean, Bryan A. Stevens, et al. 2020. "Human B Cell Clonal Expansion and Convergent Antibody Responses to SARS-CoV-2." *Cell Host & Microbe* 28 (4): 516-525.e5. https://doi.org/10.1016/j.chom.2020.09.002.

Nourmohammad, Armita, Jakub Otwinowski, Marta Łuksza, Thierry Mora, and Aleksandra M Walczak. 2019. "Fierce Selection and Interference in B-Cell Repertoire Response to Chronic HIV-1." Edited by Thomas Leitner. *Molecular Biology and Evolution* 36 (10): 2184–94. https://doi.org/10.1093/molbev/msz143.

Peng, Wujian, Song Liu, Jingye Meng, Jiali Huang, Jianrong Huang, Donge Tang, and Yong Dai. 2019. "Profiling the TRB and IGH Repertoire of Patients with H5N6 Avian Influenza Virus Infection by High-Throughput Sequencing." *Scientific Reports* 9 (1): 7429. https://doi.org/10.1038/s41598-019-43648-y.

Perez-Andres, M., B. Paiva, W. G. Nieto, A. Caraux, A. Schmitz, J. Almeida, R. F. Vogt, et al. 2010. "Human Peripheral Blood B-Cell Compartments: A Crossroad in B-Cell Traffic." *Cytometry Part B: Clinical Cytometry* 78B (S1): S47–60. https://doi.org/10.1002/cyto.b.20547.

R Core Team. 2018. "R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing,." https://www.R-project.org/.

Sakharkar, Mrunal, C. Garrett Rappazzo, Wendy F. Wieland-Alter, Ching-Lin Hsieh, Daniel Wrapp, Emma S. Esterman, Chengzi I. Kaku, et al. 2021. "Prolonged Evolution of the Human B Cell Response to SARS-CoV-2 Infection." *Science Immunology* 6 (56): eabg6916. https://doi.org/10.1126/sciimmunol.abg6916.

Sethna, Zachary, Yuval Elhanati, Curtis G Callan, Aleksandra M Walczak, and Thierry Mora. 2019. "OLGA: Fast Computation of Generation Probabilities of B- and T-Cell Receptor Amino Acid Sequences and Motifs." Edited by Bonnie Berger. *Bioinformatics* 35 (17): 2974–81. https://doi.org/10.1093/bioinformatics/btz035.

Shah, Hemangi B., Kenneth Smith, Jonathan D. Wren, Carol F. Webb, Jimmy D. Ballard, Rebecka L. Bourn, Judith A. James, and Mark L. Lang. 2019. "Insights From Analysis of Human Antigen-Specific Memory B Cell Repertoires." *Frontiers in Immunology* 9 (January): 3064. https://doi.org/10.3389/fimmu.2018.03064.

Shugay, Mikhail, Olga V Britanova, Ekaterina M Merzlyak, Maria A Turchaninova, Ilgar Z Mamedov, Timur R Tuganbaev, Dmitriy A Bolotin, et al. 2014. "Towards Error-Free Profiling of Immune Repertoires." *Nature Methods* 11 (6): 653–55. https://doi.org/10.1038/nmeth.2960.

Soto, Cinque, Robin G. Bombardi, Andre Branchizio, Nurgun Kose, Pranathi Matta, Alexander M. Sevy, Robert S. Sinkovits, Pavlo Gilchuk, Jessica A. Finn, and James E. Crowe. 2019. "High Frequency of Shared Clonotypes in Human B Cell Receptor Repertoires." *Nature* 566 (7744): 398–402. https://doi.org/10.1038/s41586-019-0934-8.

Stamatakis, Alexandros. 2014. "RAxML Version 8: A Tool for Phylogenetic Analysis and

Post-Analysis of Large Phylogenies." *Bioinformatics* 30 (9): 1312–13. https://doi.org/10.1093/bioinformatics/btu033.

Stavnezer, Janet, Jeroen E.J. Guikema, and Carol E. Schrader. 2008. "Mechanism and Regulation of Class Switch Recombination." *Annual Review of Immunology* 26 (1): 261–92. https://doi.org/10.1146/annurev.immunol.26.021607.090248.

Turchaninova, M A, A Davydov, O V Britanova, M Shugay, V Bikos, E S Egorov, V I Kirgizova, et al. 2016. "High-Quality Full-Length Immunoglobulin Profiling with Unique Molecular Barcoding." *Nature Protocols* 11 (9): 1599–1616. https://doi.org/10.1038/nprot.2016.093.

Vidarsson, Gestur, Gillian Dekkers, and Theo Rispens. 2014. "IgG Subclasses and Allotypes: From Structure to Effector Functions." *Frontiers in Immunology* 5 (October). https://doi.org/10.3389/fimmu.2014.00520.

Wu, Yu-Chang, David Kipling, Hui Sun Leong, Victoria Martin, Alexander A. Ademokun, and Deborah K. Dunn-Walters. 2010. "High-Throughput Immunoglobulin Repertoire Analysis Distinguishes between Human IgM Memory and Switched Memory B-Cell Populations." *Blood* 116 (7): 1070–78. https://doi.org/10.1182/blood-2010-03-275859.

Yang, Xiujia, Minhui Wang, Jiaqi Wu, Dianchun Shi, Yanfang Zhang, Huikun Zeng, Yan Zhu, et al. 2021. "Large-Scale Analysis of 2,152 Ig-Seq Datasets Reveals Key Features of B Cell Biology and the Antibody Repertoire." *Cell Reports* 35 (6): 109110. https://doi.org/10.1016/j.celrep.2021.109110.

Yu, Guangchuang, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. 2017. "Ggtree: Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data." Edited by Greg McInerny. *Methods in Ecology and Evolution* 8 (1): 28–36. https://doi.org/10.1111/2041-210X.12628.

Zvyagin, I V, I Z Mamedov, O V Tatarinova, E A Komech, E E Kurnikova, E V Boyakova, V Brilliantova, et al. 2017. "Tracking T-Cell Immune Reconstitution after TCRαβ/CD19-Depleted Hematopoietic Cells Transplantation in Children." *Leukemia* 31 (5): 1145–53. https://doi.org/10.1038/leu.2016.321.

Supplementary Information

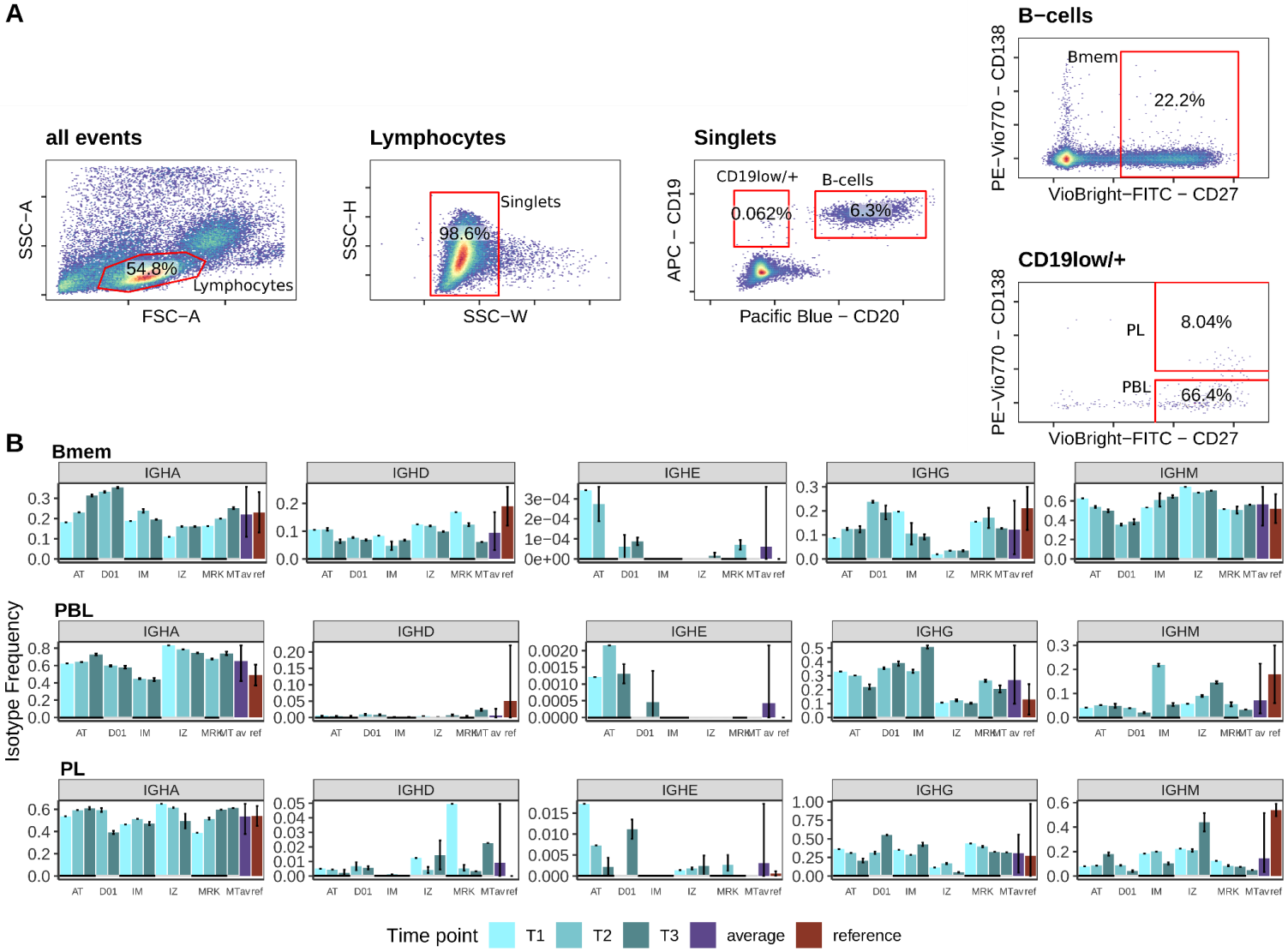

Figure S1. A: FACS gating strategy and the frequencies of studied cell subsets for representative peripheral blood sample (donor IZ time point T3): Memory B-cells (Bmem: $CD19^{+}$ $CD20^{+}$ $CD27^{+}$), plasmablasts (PBL: $CD19^{low/+}$ $CD20^{-}$ $CD27^{high}$ $CD138^{-}$) and plasma cells (PL: $CD19^{low/+}$ $CD20^{-}$ $CD27^{high}$ $CD138^{+}$); **B**: Isotype frequencies for individual samples by unique clonotypes. Whiskers illustrate minimal and maximal isotype frequencies for the group. Black and grey lines at the bottom of the plot indicate groups of bars corresponding to a particular donor.

Supplementary Table S1. Donor demographics and cell samples size. Several values in table cells separated by a semicolon represent replicates collected for corresponding donor, time point and cellular subset. AR - allergic rhinitis; FA - food allergy; HD - healthy donor.

				Number of cells per sample								
		Time point			T1			T2			T3	
Donor ID	Age	Sex	Status	Bmem	PBL	PL	Bmem	PBL	PL	Bmem	PBL	PL
D01	27	F	AR	n/a	n/a	n/a	50,300; 55,400	2,100; 2,100	1,020; 1,010	50,000; 50,000	1,000; 1,000	500; 500
IM	39	M	AR,FA	186,572	2,200	129	69,900; 68,400	2,000; 2,486	920	50,000; 50,000	2,000; 2,000	1,000; 1,000
MRK	27	M	AR	143,162	5,336	251	51,700; 50,600	2,130; 2,020	1,000; 1,035	50,000; 50,000	1,000; 1,000	400; 200
AT	23	M	AR,FA	101,400	7,200	1,800	50,600; 57,400	2,520	800	50000; 40800	1,000; 1,000	400; 200
IZ	33	M	HD	101 800	3,900	850	50,500; 56,300	1,140; 1,840	1,050; 625	50,000; 50,000	2,000; 2,000	200; 200
MT	33	F	HD	n/a	n/a	n/a	n/a	n/a	n/a	50,000; 50,000	1,000; 1,000	400

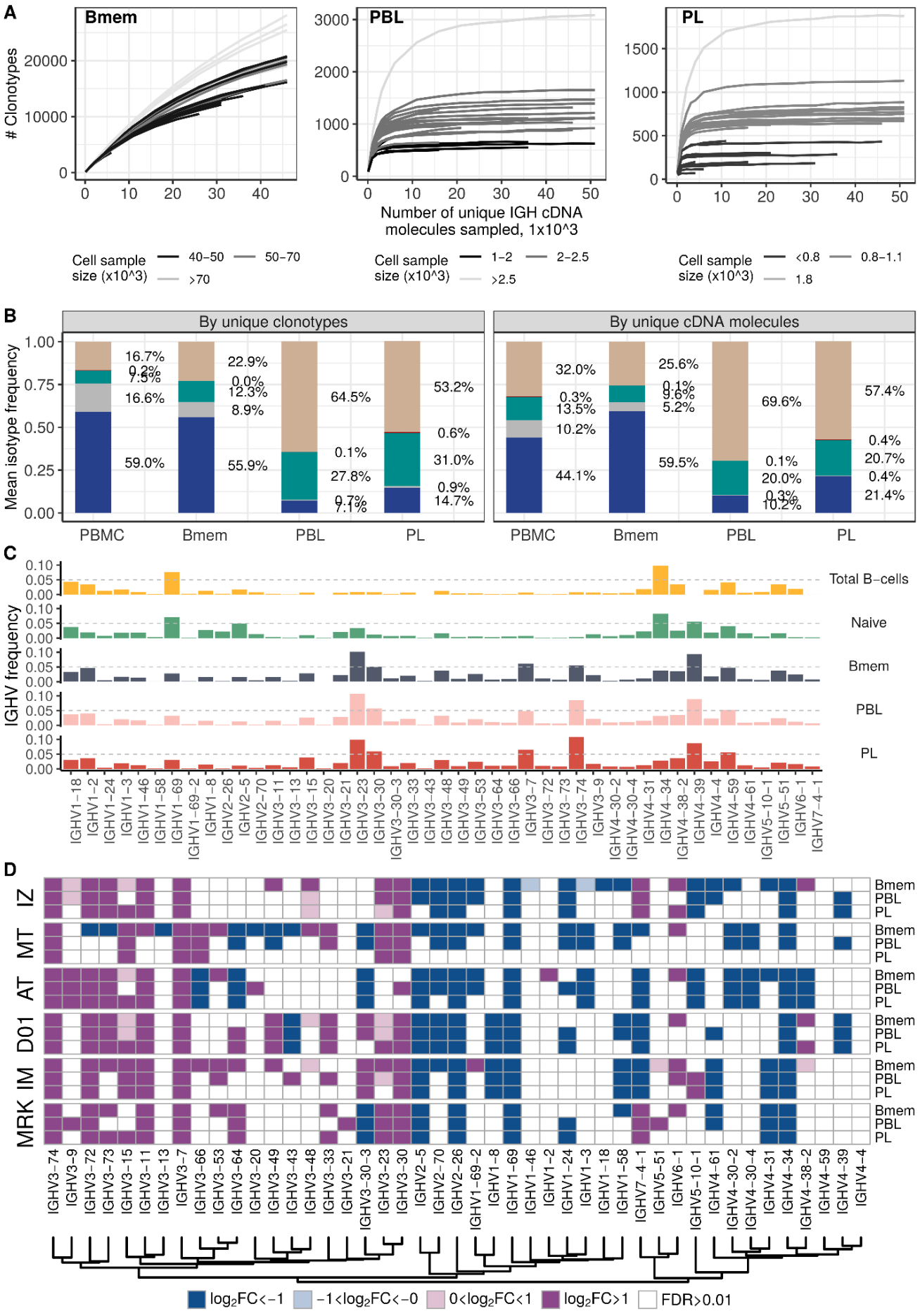

Figure S2. General characteristics of IGH repertoires in differentiated B-cell subsets. A: Rarefaction curves by IGH cDNA molecules. From each repertoire a defined number of unique IGH cDNA molecules was sampled and the number of unique IGH clonotypes was determined. Each line represents a single sample. Samples with representative cell number ($5x10^4$ Bmem, $1x10^3$ PBL, $1x10^3$ PL) are shown in black, samples of other sizes - in grey. **B**: Frequencies of isotypes in studied cell subsets as well as in bulk PBMCs averaged across all obtained samples. Left panel - isotype frequencies calculated as a number of IGH clonotypes (full-length unique nucleotide sequence) with specific isotype divided by total number of clonotypes. Right panel - isotype frequencies calculated as a number of cDNA molecules in isotype divided by total number of cDNA molecules. **C**: Distributions of average IGHV gene frequencies in repertoires of total B-cells, naive B-cell (from Gidoni et al. 2019), memory B-cells, plasmablasts and plasma cells. **D**: Heatmap of IGHV frequencies for individual donors. Colored squares on heatmap indicate significantly different (false discovery rate less than 0.01) IGHV-gene segments by their frequency in corresponding B-cell subsets than in publicly available naive B-cell repertoires (Gidoni et al. 2019). Color intensity reflects magnitude difference (FC=fold change). Only V-genes which were represented by more than 2 clonotypes on average are shown. IGHV-gene segments are ordered by the similarity of their amino acid sequence, as indicated by the amino acid similarity dendrogram at the bottom, and colored by four major clusters on the dendrogram.

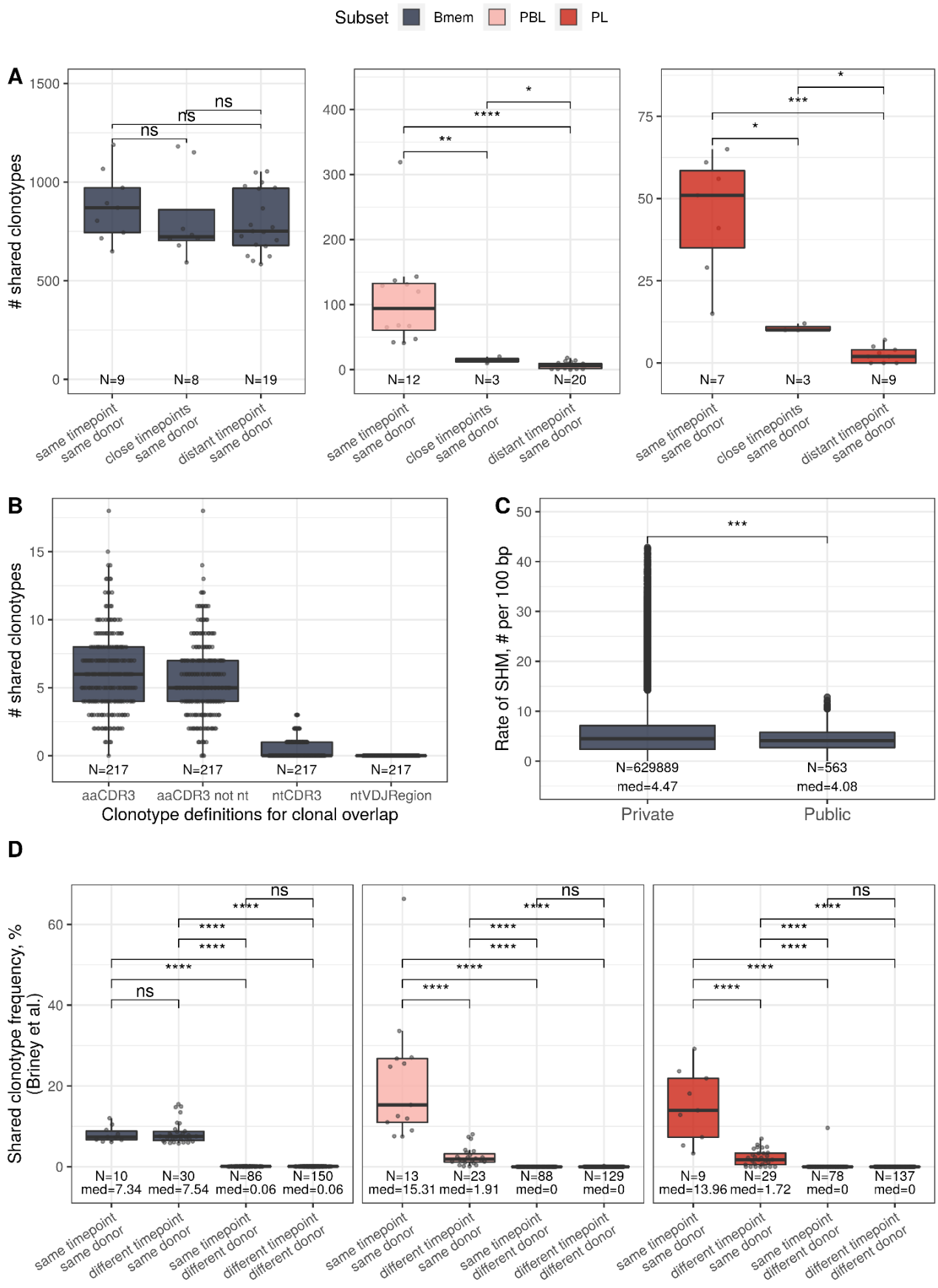

Figure S3. IGH repertoire similarity within subpopulations of B-cell lineage. A: Number of shared clonotypes between pairs of repertoires from the same donor and same or different time points. “Same time point” represents replicate samples derived from the same blood draw, “close time points” - samples were collected with approx. 1 month interval, and “distant time points” - samples were collected with approx. 1 year interval. **B**: Number of shared clonotypes between pairs of repertoires from different donors with different clonotype definitions used for overlap calculation: aaCDR3 - amino acid CDR3 sequence and V-gene label; aaCDR3 not nt - amino acid CDR3 sequence and V-gene label except clonotypes with the identical CDR3 nucleotide sequence; ntCDR3 - nucleotide CDR3 sequence and V-gene label; ntVDJRegion - full nucleotide sequence from the beginning of IGH Framework 1 region to the end of IGH Framework 4 region. **C**: Distribution of the number of somatic hypermutations identified per 100 bp length of IGHV-segment for clonotypes detected either in repertoires from only one donor (private) or in at least two donors (public). **D**: Shared clonotype frequency between pairs of repertoires calculated as in Briney et al. 2019. (N - number of pairs, med - median frequency in the group). In each plot for normalization in Bmem repertoires of 14 000 most abundant clonotypes were considered, in PBL - 600, in PL - 300. Each dot in each plot represents a pair of repertoires of corresponding type, numbers below each box indicate the number of pairs of repertoires in the group. Comparisons in all panels were performed with Mann-Whitney test, notation of the level of significance is the following: * - $p \leqslant 0.05$, ** - $p \leqslant 0.01$, *** - $p \leqslant 10^{-3}$, **** - $p \leqslant 10^{-4}$.

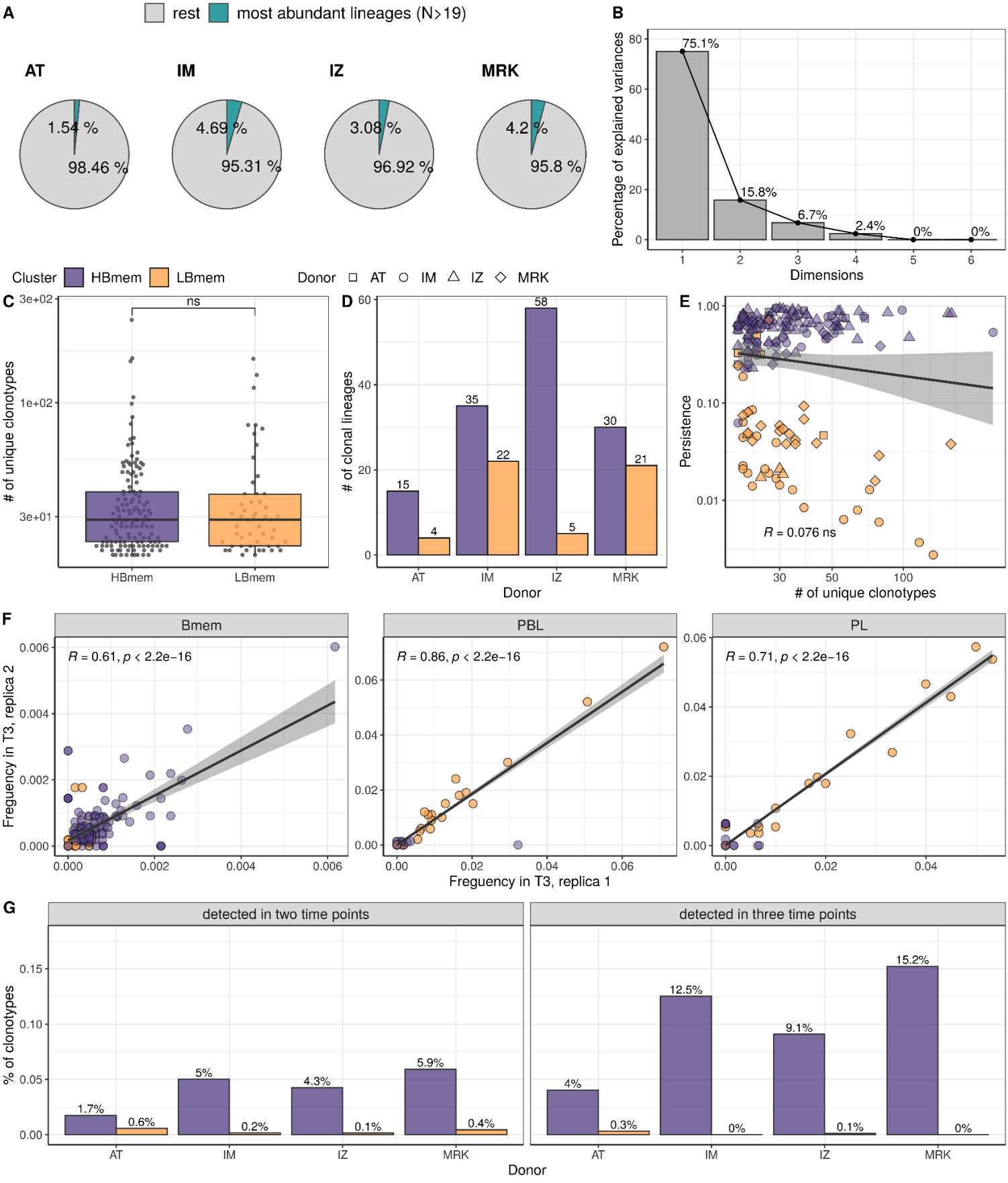

Figure S4 A: Proportion of IGH clonotype diversity, occupied by the most abundant clonal lineages (> 19 unique clonotypes); **B**: Scree plot for principal component analysis from Fig. 3A of composition of clonal lineages, where fractions of memory B cell, plasmablasts, plasma cells and fractions of Ig M, Ig G and Ig A were used as variables; **C**: Distribution of sizes, i.e. the number of unique clonotypes in a lineage, for HBmem and LBmem clonal lineages; **D**: The number of clonal lineages belonging to HBmem or LBmem clusters in each donor; **E**: Spearman's correlation between the size of the clonal lineage and its persistence; **F**: Spearman's correlation between frequencies of clonal lineages in two replicates of the time point 3 (T3) samples. Only clonal lineages which were sampled at least in one replica at this time point were included in the analysis; **G**: Fraction of clonotypes in HBmem or LBmem clonal lineages, detected in two or three time points.

Supplementary Table S2. Examples of divergent (*D*) and polymorphic (*P*) sites as they are calculated for MacDonald-Kreitman test. Synonymous and nonsynonymous substitutions from germ-line sequence are shown as underlined and in bold, correspondingly.

Codon №	**i**			**j**			**q**			**r**		
Germline	c	g	c	g	-	-	c	t	a	a	a	t
MRCA	c	g	t	g	t	a	c	t	c	a	**g**	t
Clonotypes in the clonal lineage	c c c c	g **t** g **a**	t t t t	g g g g	t t t t	a a a a	c c c c	t t t t	c a c c	a a a a	**g** **g** **c** **c**	t t t t
Dn	0	0	0	-	-	-	0	0	0	0	**1**	0
Ds	0	0	1	-	-	-	0	0	0	0	0	0
Pn	0	**2**	0	-	-	-	0	0	0	0	**1**	0
Ps	0	0	0	-	-	-	0	0	1	0	0	0
Comment	example of the codon, represented by multiple variants in clonal lineage (i.e. with the multiallelic site)			example of the codon, excluded from the analysis because of unknown germline sequence for the site			example of the codon, where divergence is not counted because of presence of the germline variant among sequence variants in the lineage			example of the codon, when the divergence is counted because there is no clonotypes identical to the germline sequence		

Supplementary Table S3. McDonald-Kreitman (MKT) test results under different inclusion criterion of clonal lineages of HBmem and LBmem clusters, which allows to deal with zero values in G-MRCA nonsynonymous or synonymous divergence. LBmem cluster demonstrated consistent results of MKT test under all types of inclusion criterion and α of joined inside cluster diversion corresponds well to the median α among clonal lineages. HBmem cluster is more sensible for the type of the filter, since in general it has much lower G-MRCA distance, and some clonal lineages have no divergence in MRCA from reconstructed part of the germline sequence. Estimated α on joined cluster divergence in HBmem cluster varies depending on the type of the filter, however it is always lower than α of LBmem cluster. Also consideration of all clonal lineages with addition of pseudocounts to *Dn* and *Ds* produces negative median α there, because α of a clonal lineage with zero G-MRCA distance will always produce negative α.

Inclusion criterion of clonal groups in MKT test	All clonal lineages. Pseudocounts are added to *Dn* and *Ds* to deal with zero values in the MKT test of distinct clonal lineages.		Clonal lineages with nonzero G-MRCA distance (at least one nonsynonymous or synonymous substitution). Pseudocounts are added to *Dn* and *Ds* to deal with zero values in the MKT test of distinct clonal lineages.		Clonal lineages with at least one nonsynonymous and synonymous substitution. No pseudocounts in *Dn* and *Ds* are required.	
Cluster	HBmem	LBmem	HBmem	LBmem	HBmem	LBmem
# of clonal lineages, passed the filter	138	52	68	49	18	29
Median α	-0.46	0.55	0.18	0.57	- 0.07	0.54
Mann-Whitney test	$p = 2.9 \cdot 10^{-11}$		$p = 4.8 \cdot 10^{-6}$		$p = 0.0028$	
MKT test on joined diversity of the cluster	$\alpha = 0.58$ $p = 4.97 \cdot 10^{-7}$	$\alpha = 0.65$ $p < 2.2 \cdot 10^{-16}$	$\alpha = 0.61$ $p = 6.05 \cdot 10^{-8}$	$\alpha = 0.66$ $p < 2.2 \cdot 10^{-16}$	$\alpha = 0.26$ $p = 0.1004$	$\alpha = 0.56$ $p = 2.05 \cdot 10^{-10}$