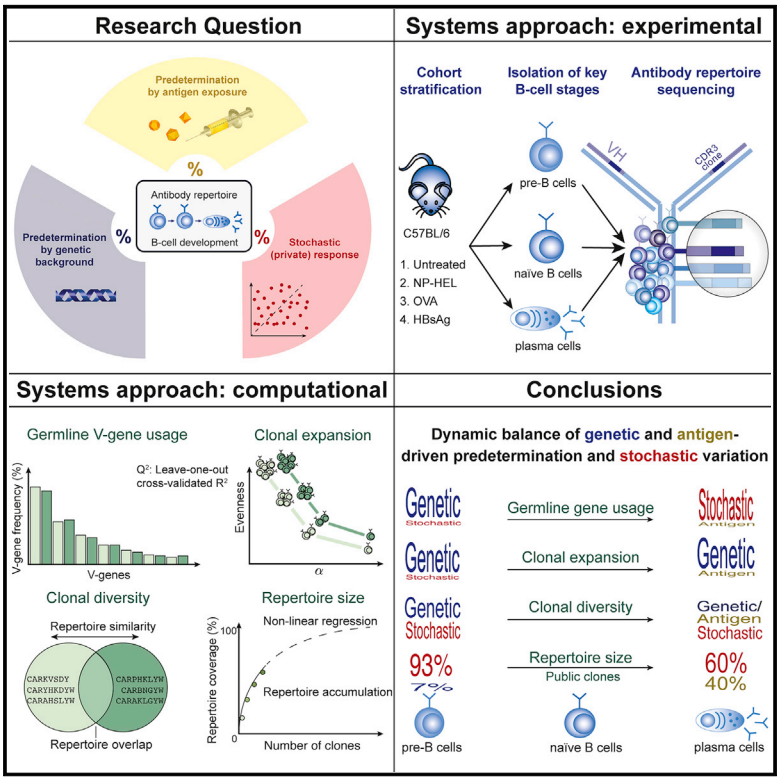
Resource

Cell Reports

Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development

Graphical Abstract



Highlights

- Systems approach allows quantification of antibody repertoire predetermination
- Antibody repertoire predetermination reaches a maximum of 99%
- Implications for the prediction and manipulation of humoral immunity

Authors

Victor Greiff, Ulrike Menzel, Enkelejda Miho, ..., Andreas Radbruch, Thomas H. Winkler, Sai T. Reddy

Correspondence

sai.reddy@ethz.ch

In Brief

Greiff et al. develop an integrated systems immunology approach for quantifying the extent of antibody repertoire predetermination. They find a dynamic balance of both high genetic (maximum: 99%) and antigen-driven (maximum: 40%) repertoire predetermination. The authors also uncover stochastic variation across B cell development, antigen exposure, and repertoire components (germline gene usage, clonal expansion, clonal diversity, repertoire size), which has implications for the prediction and manipulation of humoral immunity.

Accession Numbers

E-MTAB-5349
FR-FCM-ZY4N

Greiff et al., 2017, Cell Reports *19*, 1467–1478
May 16, 2017
http://dx.doi.org/10.1016/j.celrep.2017.04.054

Cell Reports
Resource

Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development

Victor Greiff,[1,5] Ulrike Menzel,[1,5] Enkelejda Miho,[1] Cédric Weber,[1] René Riedel,[2] Skylar Cook,[1] Atijeh Valai,[3] Telma Lopes,[1] Andreas Radbruch,[2] Thomas H. Winkler,[4] and Sai T. Reddy[1,6,*]
[1]Department of Biosystems Science and Engineering, ETH Zürich, Basel 4058, Switzerland
[2]German Rheumatism Research Center, a Leibniz Institute, Berlin 10117, Germany
[3]Shire, Research & Innovations, Research Immunology, Vienna 1221, Austria
[4]Nikolaus-Fiebiger-Zentrum für Molekulare Medizin, Department Biologie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen 91054, Germany
[5]These authors contributed equally
[6]Lead Contact
*Correspondence: sai.reddy@ethz.ch
http://dx.doi.org/10.1016/j.celrep.2017.04.054

SUMMARY

Antibody repertoire diversity and plasticity is crucial for broad protective immunity. Repertoires change in size and diversity across multiple B cell developmental stages and in response to antigen exposure. However, we still lack fundamental quantitative understanding of the extent to which repertoire diversity is predetermined. Therefore, we implemented a systems immunology framework for quantifying repertoire predetermination on three distinct levels: (1) B cell development (pre-B cell, naive B cell, plasma cell), (2) antigen exposure (three structurally different proteins), and (3) four antibody repertoire components (V-gene usage, clonal expansion, clonal diversity, repertoire size) extracted from antibody repertoire sequencing data (400 million reads). Across all three levels, we detected a dynamic balance of high genetic (e.g., >90% for V-gene usage and clonal expansion in naive B cells) and antigen-driven (e.g., 40% for clonal diversity in plasma cells) predetermination and stochastic variation. Our study has implications for the prediction and manipulation of humoral immunity.

INTRODUCTION

The humoral immune system ensures host protection and maintenance via highly diverse antibody repertoires, capable of responding to a plethora of antigens (Burnet, 1960; Landsteiner, 1947; Tonegawa, 1983). These repertoires are generated by somatic recombination of germline V, (D), and J segments in the B cell genomic locus (Tonegawa, 1983) and change along B cell development: after functional recombination, pre-B cells mature to naive B cells, which upon antigen encounter undergo clonal selection and expansion and terminally differentiate into memory plasma cells (Manz et al., 1997).

The diverse and dynamic structure of antibody repertoires has been an object of intense investigation for decades. However, only recently through the advent of high-throughput sequencing has the quantitative analysis of antibody repertoires become possible—as it is now routine to obtain millions of antibody sequences in a single study (DeWitt et al., 2016; Georgiou et al., 2014; Greiff et al., 2015a; Hoehn et al., 2016; Lindau and Robins, 2017; Robinson, 2015). Before high-throughput (and high-coverage) repertoire sequencing, quantitative analysis of repertoires was intractable (Benichou et al., 2012). Therefore, important questions related to the extent to which antibody repertoire diversity is predetermined have yet to be addressed. Gaining further insight here would allow for a greater understanding and potential for prediction in humoral immunity, which may eventually have implications for precision vaccine design and immunotherapeutics (Wang et al., 2015; Haynes et al., 2012).

In general, since Tonegawa's discovery of the genetic origin of antibody diversity more than 30 years ago (Tonegawa, 1983), textbook knowledge (Janeway and Murphy, 2011) held that antibody repertoire diversity is ruled by stochastic mechanisms irrespective of genetic background (genetically controlled mechanisms of VDJ recombination and selection) or antigen exposure. However, this view was largely due to insufficient technological and biological sampling depth; thus, it could not quantitatively represent antibody repertoire diversity. Ever since its first implementation in 2009 (Weinstein et al., 2009), high-throughput sequencing of immunoglobulin/antibody repertoires (Ig-seq) has enabled unprecedented quantitative insight into the diversity of humoral immunity. Several studies have reported deterministic convergence in multiple components of repertoire structure: germline gene usage (Avnir et al., 2016; Galson et al., 2015a; Glanville et al., 2011; Rubelt et al., 2016; Trück et al., 2014; Wang et al., 2015), clonal expansion (Greiff et al., 2015b; Mora et al., 2010; Weinstein et al., 2009), clonal sequence diversity (Henry Dunand and Wilson, 2015; Elhanati et al., 2015; Galson et al., 2015b; Jackson et al., 2013; McHeyzer-Williams et al., 1993; Mora et al., 2010; Parameswaran et al., 2013; Reddy et al.,

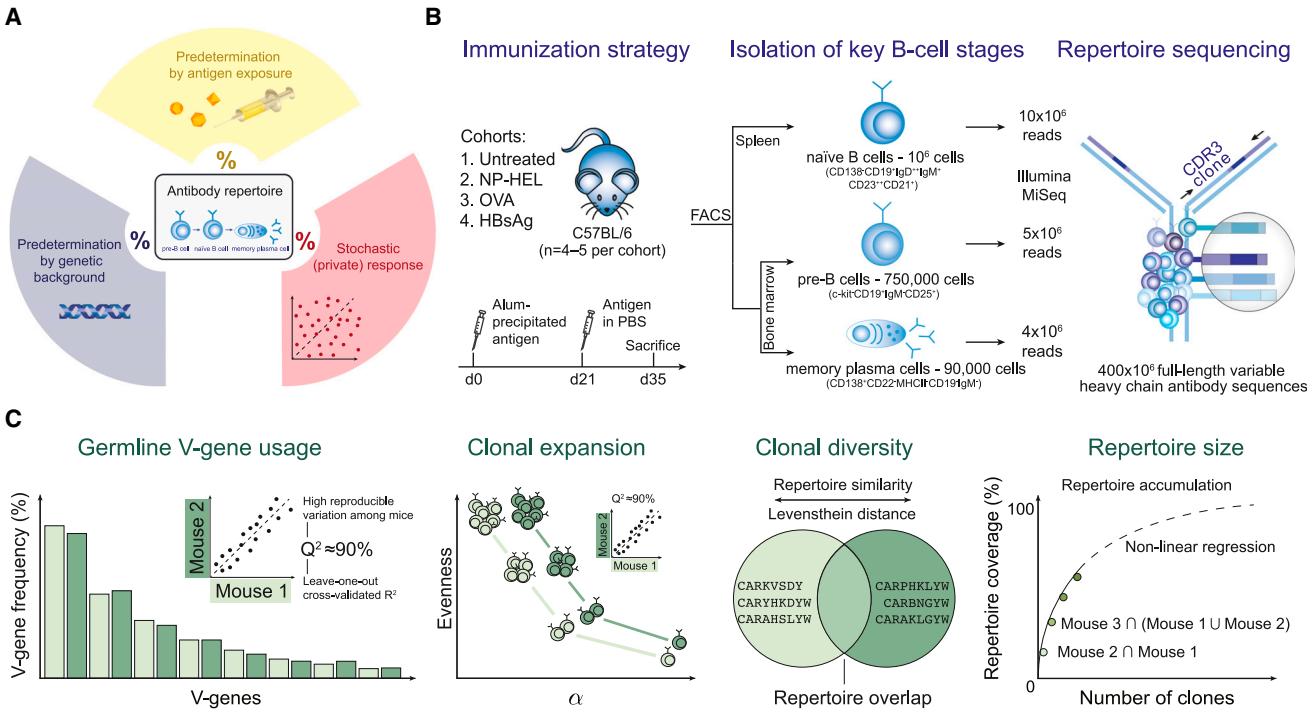


Figure 1. A Systems Approach for Quantifying the Balance of Antibody Repertoire Predetermination and Stochastic Variation

(A) We set out to quantify the extent of genetic and antigen-driven predetermination and stochastic variation in antibody repertoires throughout B cell development.

(B) Experimentally, we performed antibody heavy chain high-throughput sequencing of pre-B cells (preBC), naive B cells (nBC), and memory plasma cells (PC) from four C57BL/6 mouse cohorts.

(C) Computationally, we quantified the extent of predetermination and stochastic variation in four repertoire components (germline V-gene usage, clonal expansion, clonal diversity, and repertoire size).

See also Figures S1–S3 and Tables S1 and S2.

2010; Trück et al., 2014; Vollmers et al., 2013; Weinstein et al., 2009; Yang et al., 2015) and repertoire size (Elhanati et al., 2015; Glanville et al., 2009).

The majority of these Ig-seq studies have, however, been performed in humans, where the main components of antibody repertoires (germline gene usage, clonal expansion, clonal diversity, repertoire size) cannot be interrogated with sufficient quantitative and statistical rigor. For example, ethical and practical considerations restrict human repertoire studies primarily to the peripheral blood compartment, which does not contain many of the key stages of B cell development such as pre-B cells (bone marrow) or long-lived plasma cells (bone marrow). Second, insufficient technological and biological sampling depth due to the large size of human B cell compartments provides an incomplete picture of repertoire diversity (Greiff et al., 2015a). Finally, the pre-existing immunity acquired during the long lifetime of human donors from infections and vaccinations represents a major confounding factor in the study of environmental effects (e.g., single antigen challenge) on antibody repertoire evolution.

In this work, we devised an integrated systems immunology framework to comprehensively quantify antibody repertoire predetermination in mice across several B cell developmental stages, antigens, and repertoire components.

RESULTS

An Integrated Experimental and Computational Systems Immunology Approach

We defined genetic and antigen-driven predetermination and stochastic variation for a given B cell population as follows: (1) genetic predetermination represents the reproducible variation across individuals due to genetic background, (2) antigen-driven predetermination is the reproducible variation within cohorts imposed by antigen exposure, and (3) stochastic variation is the private variation not explained by either genetic or antigen-driven predetermination (Figure 1).

To probe the deterministic and stochastic influences that drive antibody repertoire formation, we designed a molecular, cellular, and in vivo experimental system. First, to quantify genetic predetermination, we leveraged an inbred mouse model using mainly C57BL/6 mice (n = 19), which have a fully sequenced genome and a completely annotated immunoglobulin genomic locus (Collins et al., 2015; Johnston et al., 2006). To exclude dependency of our findings on a single murine genetic background and thus ensure generalizability, we performed control experiments with two additional mouse models: (1) mice from another commonly used inbred mouse strain BALB/c (n = 4) and (2) outbred mice with unknown genetic background obtained from

a pet shop (n = 3) (Figures 1B, S1A, and S1B). Importantly, under the null hypothesis of complete repertoire stochasticity (Janeway and Murphy, 2011), even inbred mice should show highly diverging repertoires since the stochasticity hypothesis does not pertain only to the germline gene configuration, but also to the VDJ recombination *mechanism* (V(D)J-gene selection and generation of junctional diversity). Next, to quantify antigen-driven predetermination, mice from the C57BL/6 group were stratified into cohorts consisting of untreated (n = 5) or prime-boost immunized mice, using three structurally different antigens: ovalbumin (OVA, n = 5), 4-hydroxy-3-nitrophenylacetyl conjugated to hen egg lysozyme (NP-HEL, n = 5), and hepatitis B virus surface antigen (HBsAg, n = 4) (Figures 1B, S1A, and S1B).

We verified that our experimental framework provides both high biological and technological coverage of antibody repertoire immunobiology. Specifically, we performed Ig-seq of three key differentiation stages of the B cell lineage (C57BL/6 group) from two major lymphoid organs (spleen, bone marrow); pre-B cells (preBCs, bone marrow), naive B cells (nBCs, spleen), and long-lived memory plasma cells (PCs, bone marrow) (Figures 1, S1, and S2). Sufficient technological coverage was ensured by using total RNA leading to 400 million full-length variable heavy chain region (VH) sequences (Figures 1B, S2D, and S2E). The VH complementarity determining region 3 (CDR3) served as an accepted proxy for antibody clonality and specificity (Greiff et al., 2015a; Hershberg and Luning Prak, 2015; Xu and Davis, 2000). Ig-Seq reproducibility was confirmed by technical replicates (Figure S6). Error correction of sequencing data was performed by combining the MiXCR platform (Bolotin et al., 2015) with a previously published bioinformatics workflow for the reduction of artificial diversity (Greiff et al., 2014). We did not perform experimental error correction on all samples with unique molecular identifiers (UIDs), as this would require oversampling of the entirety of a sample's antibody RNA molecules for consensus read construction (multiple reads for each UID) (Greiff et al., 2015a; Khan et al., 2016; Shugay et al., 2014; Vollmers et al., 2013; Zhang et al., 2016) (Figure S7). Since the aim of our study was to capture maximum clonal diversity, the need to oversample UIDs made it prohibitive to perform error correction by consensus read building for all samples.

Computationally, we quantified the balance of predetermination and stochastic variation in four components that determine the biological diversity and function of antibody repertoires (Figure 1C): (1) "germline genes" pre-constrict the clonal diversity generated by V(D)J recombination, (2) "clonal diversity" (based on amino acid [aa] sequence of of VH-CDR3) defines B cell clonality and correlates with antigen binding potential (Greiff et al., 2015a; Xu and Davis, 2000), (3) "clonal expansion" characterizes the propensity to which B cell clones can or have encountered antigen (immunological status), and (4) "repertoire size" represents the naive repertoire's potential reactive breadth.

Genetic Background Dictates Germline V-Gene Usage in preBCs and nBCs, whereas PC V-Gene Selection Is Mostly Stochastic

In the C57BL/6 group, we examined the germline V-gene repertoire, defined here as the number of germline V-gene segments and their respective frequencies across clones for each individual. The total number of V-genes used by preBCs, nBCs, and PCs was 125, 125, and 86, respectively. Per mouse and independent of cohort, each preBC and nBC repertoire utilized ≈90% (112/125) of the entire V-gene diversity, whereas PC repertoires used ≈50% (40/86) (Figure 2A). Throughout B cell development, V-gene usage was non-uniform, as demonstrated by two to four orders of magnitude difference between minimum and maximum V-gene frequencies (mean ranges: preBCs: 0.0005%–7%, nBCs: 0.0005%–8%, PCs: 0.5%–12%, Figure 2A). Remarkably, the observed V-gene frequency distribution was consistent among preBCs and nBCs, both within and across cohorts (mean Pearson correlation $r > 0.95$, Figures 2A–2C) but varied widely among PC repertoires (range $r = -0.38 - 0.67$, $r_{PC,median} = 0.15$, Figures 2A–2C and S5C). Of note, the 4-hydroxy-3-nitrophenylacetyl (NP)-hen egg lysozyme (HEL) cohort showed the highest ($p < 0.05$) median PC Pearson correlation coefficient with $r_{PC,median} = 0.50$ (all other cohorts $-0.04 < r_{PC,median} < 0.21$, Figures 2C and S5C). Hierarchical clustering indicated that V-gene repertoires differed across B cell development (Figure 2C).

To determine the extent to which V-gene repertoires are predetermined, we developed a custom linear regression framework. Building on Fisher's formula of phenotypic variation (Fisher, 1918), V-gene repertoire genetic and antigen-driven predetermination was determined by computing the cross-validated regression (Q^2: leave-one-out cross-validated R^2) of one V-gene repertoire (from a given B cell stage and cohort of a given mouse) with that of another mouse's V-gene repertoire (Figure S3). In contrast to Pearson correlation, cross-validation enables the quantification of the predictive performance of each regression model thereby contributing to this study's main goal of quantifying repertoire predetermination (predictability). The extent of antigen-driven predetermination was determined only for PCs by a second linear regression model (Figure S3).

PreBC and nBC V-gene repertoires were nearly entirely genetically predetermined (genetic predetermination >99% in three out of four C57BL/6 cohorts, Figures 2D and S4A). This finding was confirmed for the nBC V-gene repertoires from BALB/c and pet mice (Figure S4C). This signifies that if the genetic composition and distribution is once determined, preBC and nBC V-gene repertoires can be accurately predicted a priori (without Ig-seq). In contrast, PC V-gene repertoires were largely stochastic (stochastic variation ≈80%–100%). The NP-HEL cohort (C57BL/6) showed the highest extent of antigen-driven predetermination (antigen-driven predetermination ≈20%, Figures 2D and 4A) due to significantly increased usage of V-genes (IGHV1-50/72/80) previously known to be implicated in the formation of NP-specific clones (Jacob et al., 1991a) (Figure 2D; Table S3).

Genetic Background Dictates Clonal Expansion throughout B Cell Development

The state of clonal expansion of a B cell population provides a static representation of its clonal dynamics. A repertoire's state of clonal expansion is defined as the set of its clonal frequencies (termed hereafter as clonal frequency distribution) (Greiff et al., 2015b), which can be converted—with minimal loss of information—to multidimensional evenness profiles $(\overrightarrow{{}^{\alpha}E})$ (Greiff et al.,

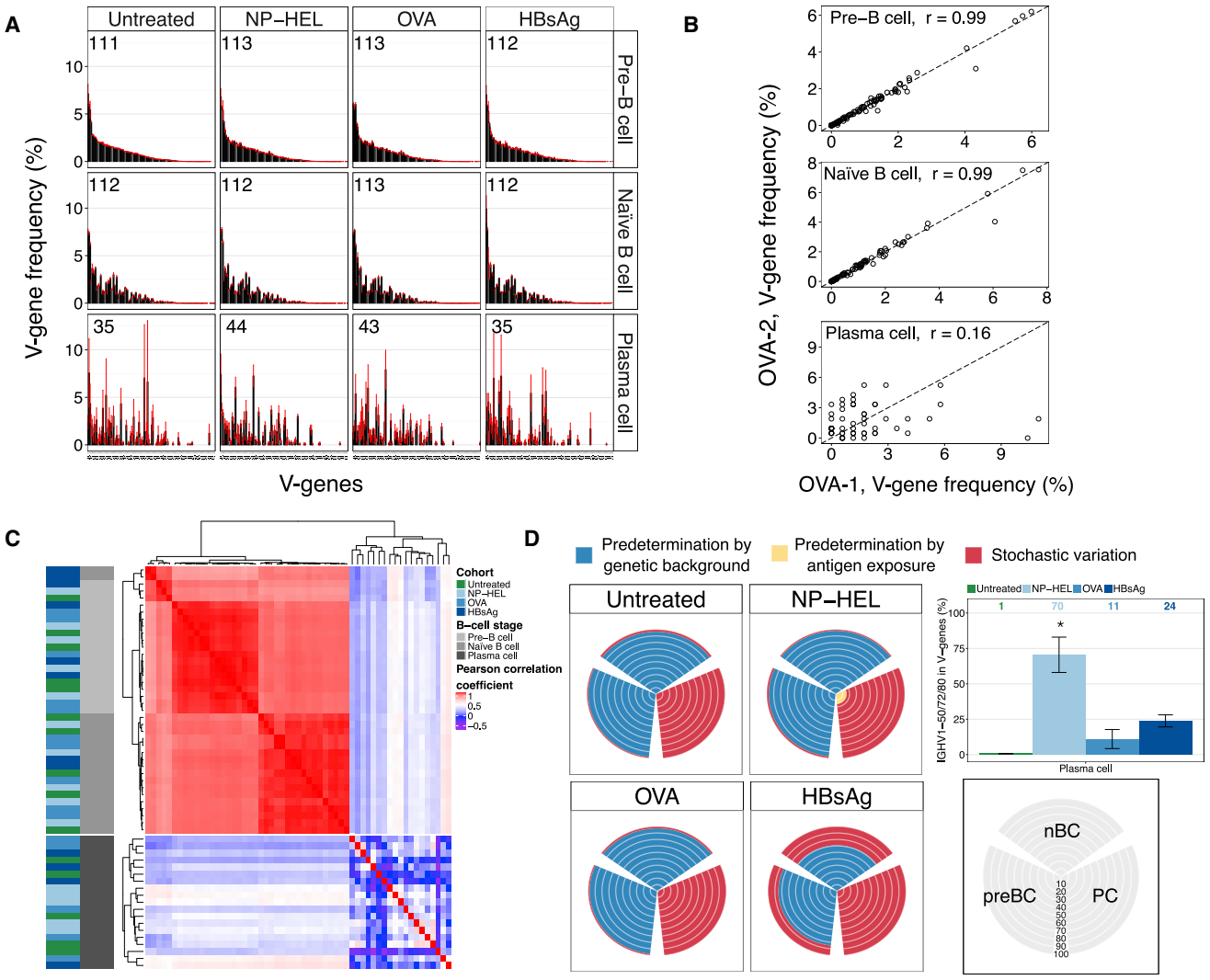


Figure 2. Genetic Background Dictates Germline V-Gene Usage of preBCs and nBCs, whereas that of PCs Is Mostly Stochastic

(A) Average V-gene frequency distributions by B cell stage and cohort (mean ± SEM). In each panel, the mean number of V genes used in the given B cell stage and cohort is indicated. V genes on the x axis were sorted by the preBC distribution of the untreated cohort.

(B) The Pearson correlation of V-gene repertoires is shown exemplarily across investigated B cell stages for two OVA-immunized mice.

(C) Hierarchical clustering of the Pearson correlation of V-gene repertoires. Each tile in the heatmap represents the pairwise Pearson correlation of two V-gene repertoires. Color legend indicates magnitude of correlation. Range of preBC, nBC, and PC correlation coefficients (within B cell stage, across cohorts): r_{preBC} = 0.86 – 0.99 (median: 0.99), r_{nBC} = 0.88 – 0.99 (median: 0.98), r_{PC} = −0.38 – 0.67 (median: 0.16). By cohort, nBC-PC median Pearson correlation coefficients were $r_{nBC–PC,Untreated}$ = 0.30, $r_{nBC–PC,NP-HEL}$ = 0.50, $r_{nBC–PC,OVA}$ = 0.34, $r_{nBC–PC,HBsAg}$ = 0.26 (differences between NP-HEL and all other cohorts are significant, $p < 0.05$).

(D) Genetic and antigen-driven predetermination and stochastic variation in V-gene repertoires by B cell stage. As subpanel, the frequency of NP-specific V genes IGHV1-50/72/80 among all V genes is shown for PCs (mean ± SEM) in order to explain the significantly higher antigen-driven predetermination detected in the NP-HEL cohort.

See also Figures S3–S5.

2015b). A repertoire with a uniform clonal frequency distribution shows a low difference between minimal and maximal clonal frequency, whereas only a few clones would dominate a polarized, clonally expanded repertoire.

While preBC and nBC repertoires were relatively uniformly distributed (Shannon Evenness: ${}^{\alpha=1}E \approx 0.7$, Figure 3A) (Hess et al., 2001), PC repertoires were extensively polarized (${}^{\alpha=1}E \approx 0.1$, Figure 3A). Applying a regression approach similar to V-gene repertoires (Figure S3), we found that clonal expansion was almost entirely genetically predetermined (independent of mouse strain in nBC, genetic background: 81%–99%, Figures 3C, S4A, and S4C). PC repertoires stored up to 16% (HBsAg cohort) of reproducible antigen-specific information (4%–16%, Figure 3C) in sequence-*associated* (clonal frequency) form. While this may appear to be a small amount of antigen-specific information, it was sufficient to cluster PC repertoires by

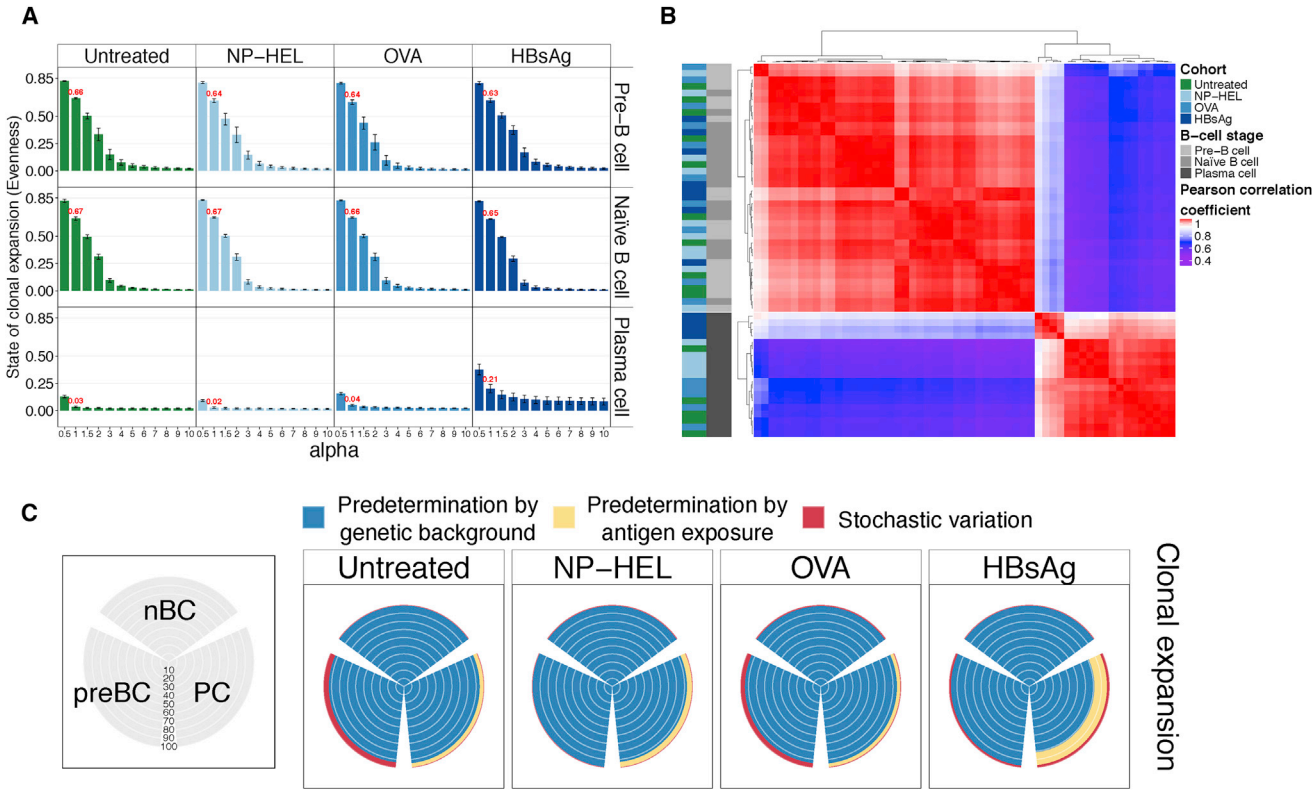


Figure 3. Genetic Background Dictates the State of Clonal Expansion throughout B Cell Development

(A) Average evenness profiles quantifying antibody repertoire clonal expansion (mean ± SEM). Shannon evenness (alpha = 1), indicated in red, differed significantly ($p < 0.05$) between preBCs/nBCs and PCs, indicating highly polarized PC repertoires.

(B) Hierarchical clustering of evenness profiles. The color legend indicates the magnitude of pairwise Pearson correlation. All correlation coefficients were significant ($p < 0.05$).

(C) Extent of genetic and antigen-driven predetermination and stochastic variation in clonal expansion.

See also Figures S3 and S4.

antigen-immunized cohort (Figure 3B), supporting previous findings with clonal expansion (Greiff et al., 2015b).

Antigen Exposure Increases Repertoire Sequence Similarity

The clonal sequence diversity determines the antigen binding potential of an antibody repertoire and can be represented by the pairwise similarity of all clones (defined here by their CDR3 aa sequences). We quantified the clonal sequence diversity of antibody repertoires by using a custom-developed Repertoire Similarity Index (RSI). The RSI determines the median aa CDR3 sequence similarity between all CDR3s of identical length with identical V and J genes *within* or *between* repertoires. When using the RSI to measure the CDR3 sequence similarity of antibody repertoires *between* mice (Figure S5), it can be used to estimate genetic and antigen-driven predetermination (Figure S3B). The RSI ranges between 0% and 100%, where an increasing percentage reflects greater clonal sequence similarity. We calculated the RSI among CDR3s of identical V and J genes and CDR3 length; this confined comparisons to only clones that possibly underwent similar routes of VDJ recombination, which share similar lineage dynamics (minimization of cross-lineage comparisons). It is widely accepted that sequences belonging to the same clonal family (lineage) have (1) identical V and J genes and (2) CDR3 length (Greiff et al., 2015a; Gupta et al., 2017; Hershberg and Luning Prak, 2015; Stern et al., 2014). In order to correct to a certain degree for lineage-unspecific conserved aa residues shared by all clonal families, we subtracted from each RSI value the baseline RSI calculated among all CDR3s of the same length irrespective of V- and J-gene usage (Figure S5B). For this V-J-excluded region composed mainly of non-templated insertions and deletions, we used the simulation of random sequence strings (hereafter referred to as "unbiased repertoire") as null distribution.

Both preBC and nBC repertoires displayed an RSI value of ≈13% (range: 11%–14%), which did not differ significantly from that of the unbiased repertoire (12.6%, Figure 4A). In contrast, we found that the clonal sequence diversity of the antigen-experienced PC compartment (51%) differed significantly from that of antigen-inexperienced preBCs and nBCs (Figure 4A), which suggests that antigen-induced clonal expansion leads to a high number of similar clonal variants. Similarly across mice, antigen-experienced PC repertoires showed increased clonal sequence convergence compared to antigen-inexperienced B cell stages,

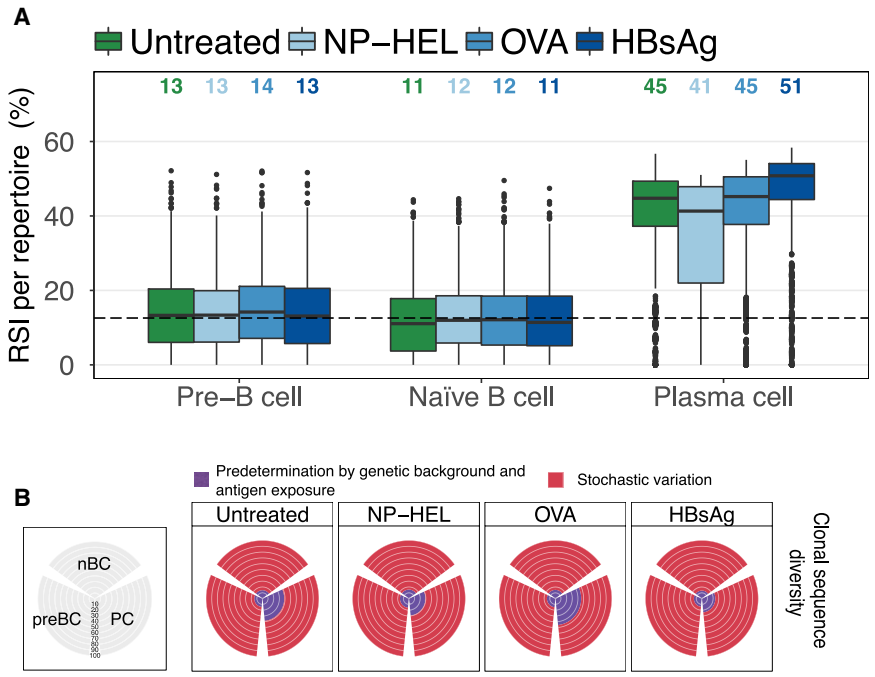


Figure 4. Antigen Exposure Increases Repertoire Sequence Similarity

(A) Within each repertoire, the baseline-corrected RSI was quantified by determining the median CDR3 sequence similarity (normalized Levenshtein distance) for those clones with identical V-J genes and CDR3 length. The difference between preBCs/nBCs and PCs is significant ($p < 0.05$). The median for each respective subset (B cell stage, cohort) is indicated. The black dashed line indicates the RSI of randomly constructed aa sequences (RSI: 12.6%, unbiased repertoire).

(B) Quantification of genetic and antigen-driven predetermination of clonal sequence diversity.

See also Figures S3–S5.

which did not differ from that of unbiased repertoires (preBCs: 15%, nBCs: 15%, PCs: 24%–46%, Figures 4, S4B, and S5). Thus, the fraction of reproducible sequence-*dependent* antigen-specific information (antigen-driven predetermination) encoded in antibody repertoires was maximally ≈31% [calculated based on RSI of (PC%) – (preBC% or nBC%) = 46% – 15%] (Figure 4B), which was double the amount of antigen-specific information contained in clonal expansion sequence-*associated* form (antigen-driven predetermination = 16%, Figure 3C).

Genetic Background and Antigen Challenge Predetermine the Formation of Public Clones

The theoretical diversity of antibody repertoires is immense—to the extent that mathematical estimations suggest the existence of clones shared among individuals (public clones) to be highly improbable (Saada et al., 2007). However, experimental studies have reported the existence of convergent clones in both humans and mice (DeWitt et al., 2016; Galson et al., 2015b; Jackson et al., 2013; Yang et al., 2015). Here, we asked to what extent genetic background and antigen challenge govern clonal convergence (presence of public clones). To quantify repertoire convergence, we determined the percentage of clones shared (public) between any two repertoires of a given B cell stage and cohort. Public clones were reproducibly observed at 7%–8% in preBCs and 12%–14% in nBCs in C57BL/6 mice (Figure 5A). A high amount of clonal overlap was mouse-strain independent, as nBC public clones were also observed at 7%–14% in BALB/c and pet mice (Figure 5C). The mean percentage of nBC clones shared *across* mouse strains was similarly high at ≈14% (Figure 5D), demonstrating that differences in genetic background do not result in different percentages of shared clones.

We asked whether the high extent of clonal convergence among antigen-inexperienced preBC and nBC repertoires differed significantly from that of unbiased repertoires. As null distribution for antibody repertoire clonal diversity, we used unbiased clonal repertoires from randomly constructed CDR3 sequences. The untreated cohort does not represent an appropriate null distribution, because even in the untreated cohort all B cell stages examined are post-VDJ recombination and thus already subject to any inherent genetic predetermination. The pairwise overlap of simulated unbiased repertoires (Figure 7B) with an underlying clonal diversity of 10^{13} (estimated size of naive repertoire, see Figure 7A) was five orders of magnitude lower (0.0002%, $p < 0.05$, Figure 7B) than that of preBC/nBC repertoires (Figure 5A). Therefore, the clonal convergence in these B cell populations is evidence of genetic predetermination. Leveraging UIDs and technical replicates, we unambiguously excluded the possibility that the presence of public clones was due to sample contamination (Figure S7).

For determining antigen-driven predetermination of public or convergent PC clones, the untreated cohort may serve as an appropriate control group since they have not been acutely challenged with antigen. In untreated mice, public PC clones were rarely observed ($p < 0.05$, Figure 5A); this was in contrast to the notable presence of PC public clones in antigen-challenged cohorts, which ranged between 14% and 37% ($p < 0.05$) and suggests that antigen challenge drives public clone formation. In some cases, public PC clones were nearly exclusive to a specific antigen, for example, in HBsAg-immunized mice (93% cohort-specific, Figure 6A). For HBsAg, the nBC-PC clonal overlap was also significantly higher compared to OVA and NP-HEL (Figure 5E).

NP-Specific Responses Are Private but Reproducibly Structured across Mice

It is well established that C57BL/6 mice mount a stereotypical antibody response against NP, with defined germline V-gene usage and common CDR3 aa motifs (Imanishi and Mäukelä, 1973; Jacob and Kelsoe, 1992; Jacob et al., 1991a, 1991b; Savelyeva

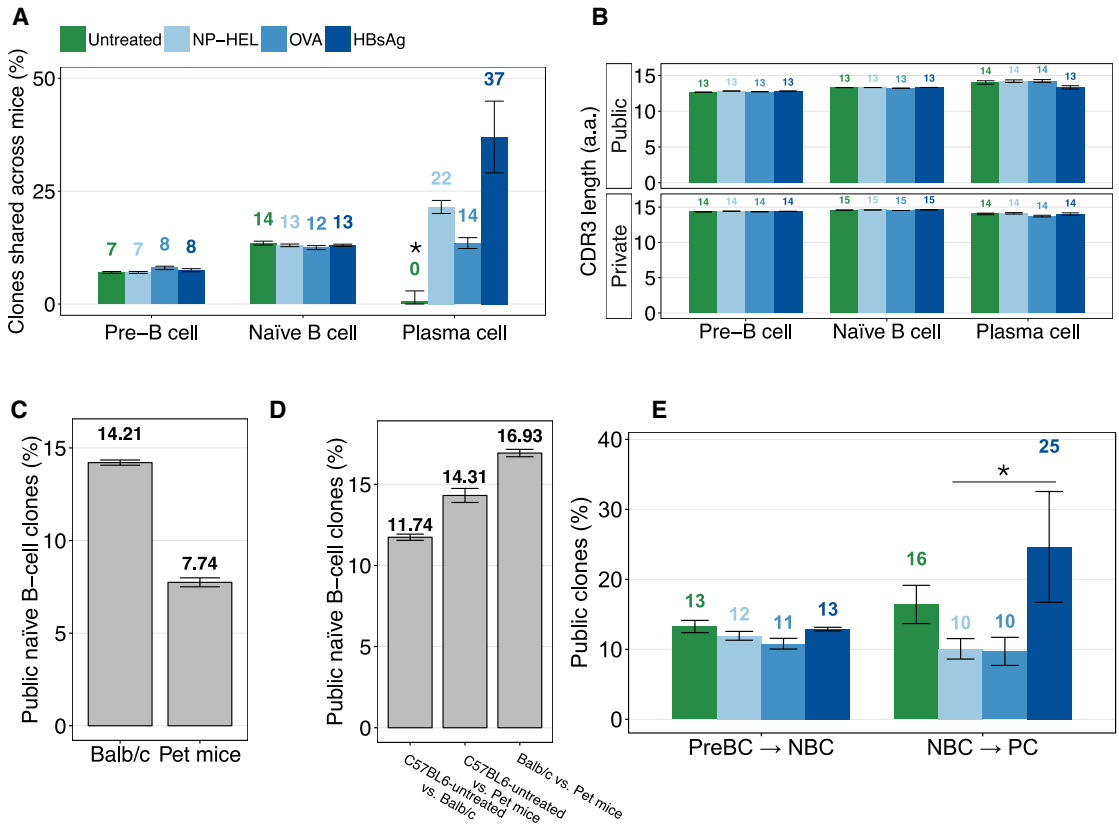


Figure 5. The Extent of Public Clones Is High throughout B Cell Development and Independent of Mouse Strain

(A) Clones shared between any two repertoires of a given B cell stage and cohort (referred to as public clones, y axis). Differences in repertoire convergence between untreated and immunized PC repertoires as well as among immunized ones are significant (*$p < 0.05$).

(B) The CDR3 length of preBC and nBC differs significantly (*$p < 0.05$) between public and private clones. The average CDR3 length of public and private clones is greater than 13 aa across B cell development.

(C) Public clones among BALB/c and pet mice, respectively.

(D) Clones shared among nBC originating from mice of different genetic background (C57BL/6, BALB/c, pet mice).

(E) Clones shared across B cell developmental stages. For nBC → PC, the difference between HBsAg and NP-HEL/OVA is significant (*$p < 0.05$).

All bar plots depict mean ± SEM. See also Figures S6 and S7.

et al., 2011). In each PC repertoire of NP-HEL-immunized mice, candidate NP-specific clones (identified by stereotypical V-genes and CDR3 sequence motifs, Table S3) were found both among the high- and low-frequency clones (Table S3), which is in line with theoretical predictions (Reshetova et al., 2017). While on average 22% of PC clones in the NP-HEL cohort were public (Figure 5A), only 3% (1/38) of the candidate NP-specific clones were public, and only 8% (3/38) of them were found in PC repertoires outside the NP-HEL cohort, suggesting a diverse NP-specific response across mice. However, within mice, clones were highly similar through their shared motifs (Figure 6B; Table S3). Indeed, the NP-specific repertoire was composed of four clusters (Figure 6B) and three of the four clusters were predominantly or entirely composed of clones from one NP-HEL immunized mouse (NP-HEL2/3/5). In general, the network organization of PC NP-specific repertoires was reproducible across each of the five NP-HEL-immunized mice, as the highest frequency clones were generally the most connected ones (Figure 6B).

The Size of the Available Naive Clonal Repertoire Is 10^{13}

Although the theoretical size of the clonal repertoire has been estimated to be $\sum_{i=4}^{20} 20^i = 10^{26}$ unique aa clones (including somatic hypermutations and accounting for aa CDR3 lengths that occur most commonly, e.g., 99.9% of all CDR3s in our dataset are of length four to 20 aa) (Saada et al., 2007), the size of the available *naive* repertoire (antigen-inexperienced = preBC + nBC) has yet to be determined. To this end, we first performed repertoire accumulation curves to quantify the extent to which the 38 combined antigen-inexperienced repertoires (preBC and nBC from 19 C57BL/6 mice) cover the clonal diversity of any naive repertoire. Strikingly, we found that 38 naive repertoires sufficed to cover already 42% of the clonal diversity of *any* naive repertoire (Figure 7A). In contrast, we found that, across all accumulation steps, unbiased repertoires showed five orders of magnitude lower coverage (median: 0.0002%, $p < 0.05$, Figures 7A and 7B), thus suggesting genetically predetermined preBC and nBC sequence coverage and repertoire size (Figure 7A).

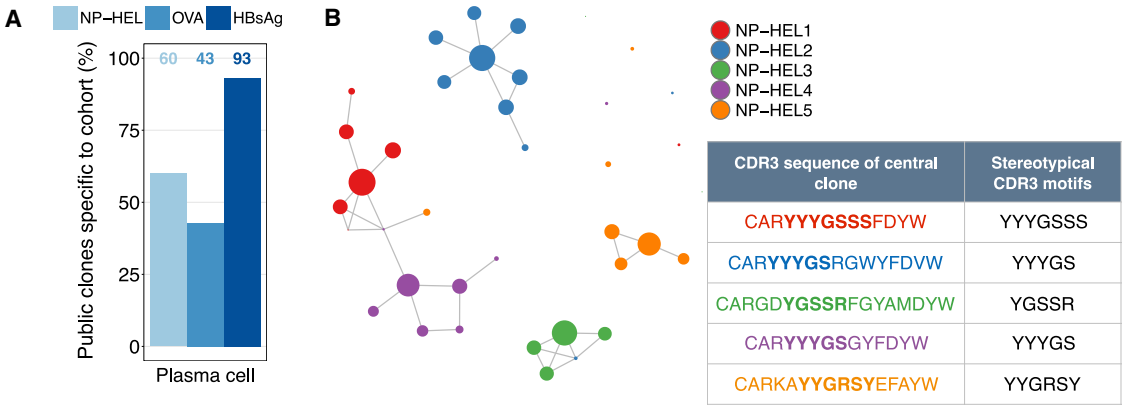


Figure 6. Public PC Clones Are Highly Cohort Specific, and Candidate NP-Specific Clones Are Predominantly Private but Highly Similar within Mice

(A) Public PC clones exclusive to each cohort (mean).

(B) Network representation of 38 candidate NP-specific clones. Nodes represent CDR3 aa clones and edges represent a Levenshtein (edit) distance of 1 between respective nodes. Nodes were scaled by relative clonal frequency and colored according the NP-HEL immunized mouse (NP-HEL1–5). Table inset shows sequence of central clones within each cluster.

See also Table S3.

To extrapolate the number of clones necessary to achieve full repertoire coverage (95%) of any naive repertoire, we performed non-linear regression on the repertoire accumulation curves (Figure 7A). Whereas 10^7 accumulated clones were sufficient to cover 42% of any naive repertoire, we estimated that 10^{13} unique aa CDR3s are required to encompass the *full clonal diversity* (≥95%) of any naive repertoire (Figure 7A). Since the cross-mouse overlap of BALB/c nBC was equally 14% (Figure 5A), it is likely that estimations performed using the 38 preBC/nBC C57BL/6 repertoire samples will be reflected across mouse strains. These results provide an empirical estimation of the available repertoire size in mice, which was calculated to be 10^{13} unique CDR3s (Figure 7B).

DISCUSSION

Leveraging a systems framework, we have discovered that the humoral immune response is more predictable than previously thought (Brodin et al., 2015; Glanville et al., 2011; Janeway and Murphy, 2011; Jiang et al., 2011; Honjo and Habu, 1985). Specifically, we show that genetic background and antigen exposure highly predetermine antibody repertoires throughout B cell development independently of mouse strain, thereby unifying previous notions on the extent of repertoire determinism and stochasticity (Galson et al., 2015b; Glanville et al., 2011; Janeway and Murphy, 2011; Rubelt et al., 2016; Honjo and Habu, 1985; Wang et al., 2015). This opens new avenues of research toward the elucidation of the selective advantage and mechanistic regulation of the balance of genetic predetermination and stochastic variation in the development of antibody repertoires (Cobey et al., 2015; Greiff et al., 2017). A high degree of repertoire predisposition was recently proposed as a necessary evolutionary adaptation for lymphocyte populations (Mayer et al., 2015). Public preBC and nBC clones thus may have a functional role, such as an evolutionarily selected first line of natural antibody defense against pathogens with relatively conserved epitopes (Baumgarth et al., 2015; Covacu et al., 2016; Greiff et al., 2017; Madi et al., 2014).

Our analyses revealed that the balance of predetermination and stochastic variation is highly dynamic across B cell development, repertoire components, and antigens. Specifically, while preBC V-gene repertoires differed only slightly from those of nBC, clonal overlap was substantially higher in nBCs (7% versus 14%, Figure 5A). PC behavior was not consistent across repertoire components: whereas PC V-gene repertoires and clonal sequence diversity varied within or among cohorts (Figures 2, 4, and S5), clonal expansion and convergence was cohort dependent (Figures 3 and 5).

Although we aimed to comprehensively reflect the complexity of antibody repertoires, future research exploiting improved experimental designs and statistical approaches may be able to draw an even more quantitative picture of antibody repertoire architecture (Bolen et al., 2017; Callan et al., 2017; Miho et al., 2017; Strauli and Hernandez, 2016). For example, the accurate determination of cross-individual lineage convergence as well as the benchmarking of repertoire simulation frameworks are matters of current debate and deserve further investigations (Greiff et al., 2015a; Gupta et al., 2017; Hershberg and Luning Prak, 2015; Miho et al., 2017; Safonova et al., 2015; Stern et al., 2014).

Our study, although performed in mice, provides a path toward improving the design of precision vaccines and targeted immunotherapies in humans (Bonsignori et al., 2016; Haynes et al., 2012; Jackson et al., 2014; Jardine et al., 2016). Specifically, a recent study describes the design of an HIV immunogen capable of eliciting broadly neutralizing antibody (bnAb) responses by selective germline-targeting of bnAb-precursor nBCs (Briney et al., 2016; Jardine et al., 2016). Within a conceptual framework of total repertoire stochasticity (Janeway and Murphy, 2011), the use of germline-targeting immunogens

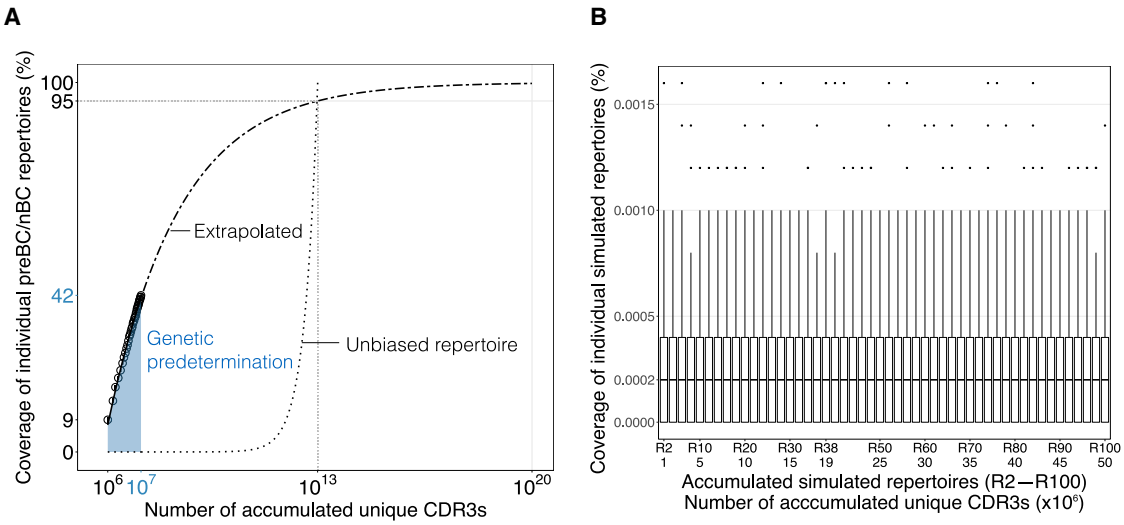


Figure 7. Repertoire Convergence Allows for Estimation of the Size of the Naive Repertoire

(A) The diversity coverage of individual preBC and nBC repertoires (y axis) was determined by their incremental accumulation (x axis) and measuring the amount of already discovered clones of a given repertoire not part of the accumulation (black points, solid line, mean and 95% confidence interval are displayed). The large amount of coverage achieved is a result of the high extent of genetic predetermination in clonal diversity (Figure 5A). The shaded area in blue indicates the difference between the diversity coverage observed by the accumulation of preBC and nBC repertoires and that of an unbiased repertoire. Repertoire accumulation was performed 100-fold for random repertoire orderings. Using non-linear regression (extrapolated), we estimated the size of the naive antibody repertoire (number of unique CDR3 aa sequences) to be 10^{13} (coverage at 95%). The repertoire accumulation curve of unbiased repertoires (dotted line) modeling the percentage of already discovered clones (k) in a repertoire by all previous repertoire accumulations was computed by leveraging the hypergeometric distribution $\left(p(k) = \left(\binom{K}{k}\binom{N-K}{n-k} \Big/ \binom{N}{n}\right)\right)$ assuming for each repertoire a size of 500,000 unique clones (n, ≈nBC repertoire size, Figure S2E) sampled from a theoretical diversity of 10^{13} (N) with an increasing number of already discovered clones by previous repertoire accumulations (K, all possible values on x axis from 10^6 until 10^{13}).

(B) Accumulation curve of 100 unbiased repertoires (R) shown in (A) using boxplots (computed analogously to that of experimental data; see A). R38 indicates equivalent size of experimental dataset: 38 preBC and nBC repertoires. R2 measures the number of shared clones among any two of the 100 unbiased repertoires, thus serving as null distribution to test for genetic predetermination in preBC and nBC clonal convergence (see Figure 5A).

would lead to highly variable results across individuals. In contrast, our results suggest that if human nBC also have a highly predetermined V-gene (DeWitt et al., 2016; Galson et al., 2015b; Glanville et al., 2011) and clonal repertoire, elicitation of neutralizing antibodies may be accomplished in a *large* proportion of the population by targeting highly represented germlines and public clones using computationally designed immunogens (Correia et al., 2014).

EXPERIMENTAL PROCEDURES

Experimental Model and Subject Details

All mouse experiments were performed under the guidelines and protocols approved by the Basel-Stadt cantonal veterinary office (Basel-Stadt Kantonales Veterinäramt Tierversuchsbewilligung #2582). Four cohorts of five female C57BL/6 J mice (Janvier Laboratories, France) 8–10 weeks old were housed under specific pathogen-free conditions. Primary immunizations of three cohorts consisted of intraperitoneal injections of alum-precipitated antigen (100 μg OVA, Invivogen), 100 μg 4-hydroxy-3-nitrophenylacetyl-conjugated hen egg lysozyme (NP-HEL, Biosearch Technologies), 4 μg HBsAg (Cell Sciences). Mice were boosted intraperitoneally after 3 weeks with identical amounts of antigen in PBS. Immunized mice were sacrificed 14 days post-secondary immunization; untreated control mice were sacrificed at corresponding age. Spleen, bone marrow, and blood were collected from all mice. Four untreated female BALB/c mice (Janvier Laboratories, France) were housed under the same conditions and spleens were isolated at 11 weeks of age. Three pet mice of unknown genetic background (white fur) were obtained as adult mice at pet shops in Berlin. All pet mice experiments were performed according to institutional guidelines and licensed under German animal protection regulations.

Fluorescence-Activated Cell Sorting

The following cellular populations were isolated (Figure S1B): preBCs from bone marrow (c-kit$^-$CD19$^+$IgM$^-$CD25$^+$PI$^-$) (Osmond et al., 1998), long-lived memory PCs from bone marrow (CD138$^+$CD22$^-$MHCII$^-$CD19$^-$IgM$^-$PI$^-$) (Shen et al., 2014), and naive follicular B cells from spleen (CD138$^-$CD19$^+$IgD^{2+}IgM$^+$CD23^{2+}CD21$^+$PI$^-$) (Srivastava et al., 2005). Naive splenic B cells of pet mice were sorted based on CD19$^+$CD38$^+$IgM$^+$IgD$^+$IgG2b$^-$CD93$^-$CD138$^-$F4/80$^-$CD11c$^-$GL7$^-$PI$^-$.

Ig-Seq Library Preparation

Antibody variable heavy chain (VH) libraries were prepared as previously described (Figures S2A–S2C) (Menzel et al., 2014). Amplification was performed using a forward mix of primers (specific to framework region 1) containing an extension 1 region (primer [Pr.]1–19, Table S1). PreBC and nBC were amplified with an extension-2-containing IgM-specific reverse primer (Pr. 20, Table S1), while PCs were amplified with mixed IgG- and IgM-specific reverse primers (Pr.21&22, Table S1).

Ig-Seq Library Preparation with UIDs

Two nBC samples (10^6 cells) from an NP-HEL and HBsAg immunized mouse were used for UID-based Ig-seq library as previously described (Khan et al., 2016) (Figure S7B).

Illumina-Based Antibody Repertoire Sequencing

Antibody library pools were sequenced on the Illumina MiSeq platform (2 × 300 cycles, paired-end, Figures S2D and S2E). Mean base call quality of all samples was in the range of Phred score 30.

Preprocessing of Antibody Repertoire Sequencing Data

Ig-seq data (excluding UID-tagged samples) were processed (VDJ alignment, clonotyping) using the MiXCR software package (clonotype formation by CDR3 region). For downstream analyses, functional clonotypes were only retained if: (1) they were composed of at least four aa and (2) had a minimal read count of 2 (Greiff et al., 2014; Menzel et al., 2014). UID-tagged data were preprocessed as described previously (Khan et al., 2016). Although, both IgM and IgG reads were obtained for the PC compartment, in this study we focused our analysis on IgG.

Quantification of Repertoire Convergence: Percentage of Public Clones

The percentage of public clones shared between two repertoires (sets of unique clones) X and *Y* of identical B cell stage and cohort was calculated as $overlap(X,\ Y) = (|X \cap Y|\ /\min(|X|, |Y|)) * 100$, where $|X|$ and $|Y|$ are the clonal sizes (number of unique clones) of repertoires X and *Y*.

Generation of Unbiased Antibody Repertoires

"Unbiased" repertoires, simulating the diversity of uniform combinatorial diversity with no bias in germline gene selection and N,P-nucleotide addition, were generated by drawing a random character from the aa alphabet (independently and identically distributed) for every sequential position according to a uniform distribution. The length of each random sequence string was chosen randomly between 4 and 20 (99% of all CDR3s lay within this length interval) equally, according to a uniform distribution.

Quantification of the Predictive Performance of Linear Regression Models

The predictive performance (Q^2) of each linear regression model ($\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$) was calculated using leave-one-out-cross validation (LOOCV): $Q^2 = (1 - (PRESS/TSS)) \cdot 100$, where PRESS is the predictive error sum of squares $(\sum_{j=1}^{n}(\boldsymbol{Y_j} - \hat{\boldsymbol{Y}}_{[j]})^2$ with $\hat{\boldsymbol{Y}}_{[j]}$ denoting the prediction of the model when the j^{th} case is deleted from the training set and TSS is the total sum of squares $(\sum_{i=1}^{n}(\boldsymbol{Y_j} - \overline{\boldsymbol{Y}})^2)$ (Greiff et al., 2012). $\boldsymbol{X}$ and $\boldsymbol{Y}$ are either V-gene repertoires or evenness profiles (vectors) of different mice.

Quantification of Genetic and Antigen-Driven Predetermination and Stochastic Variation for Germline Gene Repertoires and State of Clonal Expansion: Evenness Profiles

Building on Fisher's formula of phenotypic variation (Fisher, 1918) and assuming a superposition of genomic background-driven and antigen-driven effects, the genetic and antigen-driven predetermination of V-gene and clonal expansion profiles was calculated using a linear regression model (Figure S3).

Quantification of Repertoire Predetermination by Genetic Background, Antigen Exposure, and Stochastic Variation on the Clonal Sequence Level

We performed clonal sequence-based repertoire comparisons using the baseline-corrected RSI (Figure S3B, block 7). Genetic and antigen-driven predetermination was calculated by forming the median baseline-corrected RSI across mice by B cell stage and cohort (block 8). Stochastic variation was described as the complement of the baseline-corrected RSI (100%-baseline-corrected RSI, block 9).

Estimation of the Coverage and Size of the Theoretical Murine Naive Clonal Repertoire

In order to quantify the extent to which the entirety of the sequenced preBC and nBC clonal repertoires cover any preBC/nBC repertoire, species accumulation curves of repertoires were computed. We defined the repertoire coverage (C_i) of a given repertoire AB_i as the percentage overlap of its set of unique clones $\{CDR3\}_i$ with the set of clones contained in all previously accumulated repertoires $(\cup_{j=1}^{i-1} AB_j)$: $C_i = \left(\left|AB_i \cap \cup_{j=1}^{i-1} AB_j\right| \Big/ \left|AB_i\right|\right)$, where $i \in \{1, \ldots, R\}$ with R being the total number of preBC and nBC repertoires $(R = 38)$.

To infer the number of clones necessary for any given coverage, we used non-linear regression analysis using an exponential fit $(C_i \sim 100 - s * b^{\left(-\log\left(\left|\cup_{i=1}^{R} AB_i\right|\right)\right)}$ (Soberón and Llorente, 1993), where $\left|\cup_{i=1}^{R} AB_i\right|$ Uis the number of unique clones contained within the accumulated repertoires and s and b are the parameters to be inferred. For $\geq$ 95% coverage, this is the estimated size of the murine naive repertoire.

ACCESSION NUMBERS

The accession number for the antibody repertoire sequencing data reported in this paper is ArrayExpress: E-MTAB-5349. The accession number for the FACS data reported in this paper is FlowRepository: FR-FCM-ZY4N. The code for Q^2 and RSI determination is available on the github repository https://github.com/victorgreiff/repertoire_predetermination_project.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2017.04.054.

AUTHOR CONTRIBUTIONS

V.G., U.M., and S.T.R. conceived the project; V.G., U.M., and S.T.R. designed experiments; A.R. and T.H.W. designed supporting experiments; U.M. performed wet-lab experiments; R.R., A.V., and T.L. performed supporting wet-lab experiments; V.G. performed primary data analysis; E.M., C.W., and S.C. performed supporting data analysis; V.G., U.M., E.M., and S.T.R. wrote manuscript; all authors edited the manuscript.

ACKNOWLEDGMENTS

We thank Dr. Christian Beisel, Manuel Kohler, Ina Nissen, and Elodie Burcklen from the Genomics Facility Basel of ETH Zürich for their expert technical assistance with Illumina high-throughput sequencing. We thank Tarik Khan, Simon Friedensohn, Christoph Berger, and Brian Lang for scientific discussions and Gabriele Lillacci for critical reading of the manuscript. This work was funded by the Swiss National Science Foundation (Project #: 31003A_143869 and 31003A_170110 to S.T.R.), SystemsX.ch – AntibodyX RTD project (to S.T.R.), and Swiss Vaccine Research Institute (to S.T.R.). The professorship of S.T.R. is made possible by the generous endowment of the S. Leslie Misrock Foundation.

Received: July 7, 2016
Revised: March 21, 2017
Accepted: April 19, 2017
Published: May 16, 2017

REFERENCES

Avnir, Y., Watson, C.T., Glanville, J., Peterson, E.C., Tallarico, A.S., Bennett, A.S., Qin, K., Fu, Y., Huang, C.-Y., Beigel, J.H., et al. (2016). IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. Sci. Rep. *6*, 20842.

Baumgarth, N., Waffarn, E.E., and Nguyen, T.T.T. (2015). Natural and induced B-1 cell immunity to infections raises questions of nature versus nurture. Ann. N Y Acad. Sci. *1362*, 188–199.

Benichou, J., Ben-Hamo, R., Louzoun, Y., and Efroni, S. (2012). Rep-Seq: Uncovering the immunological repertoire through next-generation sequencing. Immunology *135*, 183–191.

Bolen, C.R., Rubelt, F., Vander Heiden, J.A., and Davis, M.M. (2017). The Repertoire Dissimilarity Index as a method to compare lymphocyte receptor repertoires. BMC Bioinformatics *18*, 155.

Bolotin, D.A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I.Z., Putintseva, E.V., and Chudakov, D.M. (2015). MiXCR: Software for comprehensive adaptive immunity profiling. Nat. Methods *12*, 380–381.

Bonsignori, M., Zhou, T., Sheng, Z., Chen, L., Gao, F., Joyce, M.G., Ozorowski, G., Chuang, G.-Y., Schramm, C.A., Wiehe, K., et al.; NISC Comparative Sequencing Program (2016). Maturation pathway from germline to broad HIV-1 neutralizer of a CD4-mimic antibody. Cell *165*, 449–463.

Briney, B., Sok, D., Jardine, J.G., Kulp, D.W., Skog, P., Menis, S., Jacak, R., Kalyuzhniy, O., de Val, N., Sesterhenn, F., et al. (2016). Tailored immunogens direct affinity maturation toward HIV neutralizing antibodies. Cell *166*, 1459–1470.e11.

Brodin, P., Jojic, V., Gao, T., Bhattacharya, S., Angel, C.J.L., Furman, D., Shen-Orr, S., Dekker, C.L., Swan, G.E., Butte, A.J., et al. (2015). Variation in the human immune system is largely driven by non-heritable influences. Cell *160*, 37–47.

Burnet, F.M. (1960). Theories of immunity. Perspect. Biol. Med. *3*, 447–458.

Callan, C.G., Jr., Mora, T., and Walczak, A.M. (2017). Repertoire sequencing and the statistical ensemble approach to adaptive immunity. Curr. Opin. Syst. Biol. *1*, 44–47.

Cobey, S., Wilson, P., and Matsen, F.A. (2015). The evolution within us. Philos. Trans. R. Soc. B Biol. Sci. *370*, Published online September 5, 2015. http://dx.doi.org/10.1098/rstb.2014.0235.

Collins, A.M., Wang, Y., Roskin, K.M., Marquis, C.P., and Jackson, K.J.L. (2015). The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140236.

Correia, B.E., Bates, J.T., Loomis, R.J., Baneyx, G., Carrico, C., Jardine, J.G., Rupert, P., Correnti, C., Kalyuzhniy, O., Vittal, V., et al. (2014). Proof of principle for epitope-focused vaccine design. Nature *507*, 201–206.

Covacu, R., Philip, H., Jaronen, M., Almeida, J., Kenison, J.E., Darko, S., Chao, C.-C., Yaari, G., Louzoun, Y., Carmel, L., et al. (2016). System-wide analysis of the T cell response. Cell Rep. *14*, 2733–2744.

DeWitt, W.S., Lindau, P., Snyder, T.M., Sherwood, A.M., Vignali, M., Carlson, C.S., Greenberg, P.D., Duerkopp, N., Emerson, R.O., and Robins, H.S. (2016). A public database of memory and naive B-cell receptor sequences. PLoS ONE *11*, e0160853.

Elhanati, Y., Sethna, Z., Marcou, Q., Callan, C.G., Jr., Mora, T., and Walczak, A.M. (2015). Inferring processes underlying B-cell repertoire diversity. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140243.

Fisher, R.A. (1918). The correlation between relatives on the supposition of mendelian inheritance. Trans. R. Soc. Edinb. *52*, 399–433.

Galson, J.D., Kelly, D.F., and Trück, J. (2015a). Identification of antigen-specific B cell receptor sequences from the total B cell repertoire. Crit. Rev. Immunol. *35*, 463–478.

Galson, J.D., Trück, J., Fowler, A., Münz, M., Cerundolo, V., Pollard, A.J., Lunter, G., and Kelly, D.F. (2015b). In-depth assessment of within-individual and inter-individual variation in the B cell receptor repertoire. Front. Immunol. *6*, 531.

Georgiou, G., Ippolito, G.C., Beausang, J., Busse, C.E., Wardemann, H., and Quake, S.R. (2014). The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat. Biotechnol. *32*, 158–168.

Glanville, J., Zhai, W., Berka, J., Telman, D., Huerta, G., Mehta, G.R., Ni, I., Mei, L., Sundar, P.D., Day, G.M.R., et al. (2009). Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc. Natl. Acad. Sci. USA *106*, 20216–20221.

Glanville, J., Kuo, T.C., von Büdingen, H.-C., Guey, L., Berka, J., Sundar, P.D., Huerta, G., Mehta, G.R., Oksenberg, J.R., Hauser, S.L., et al. (2011). Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proc. Natl. Acad. Sci. USA *108*, 20066–20071.

Greiff, V., Redestig, H., Lück, J., Bruni, N., Valai, A., Hartmann, S., Rausch, S., Schuchhardt, J., and Or-Guil, M. (2012). A minimal model of peptide binding predicts ensemble properties of serum antibodies. BMC Genomics *13*, 79.

Greiff, V., Menzel, U., Haessler, U., Cook, S.C., Friedensohn, S., Khan, T.A., Pogson, M., Hellmann, I., and Reddy, S.T. (2014). Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. BMC Immunol. *15*, 40.

Greiff, V., Miho, E., Menzel, U., and Reddy, S.T. (2015a). Bioinformatic and statistical analysis of adaptive immune repertoires. Trends Immunol. *36*, 738–749.

Greiff, V., Bhat, P., Cook, S.C., Menzel, U., Kang, W., and Reddy, S.T. (2015b). A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. Genome Med. *7*, 49.

Greiff, V., Weber, C., Miho, E., Menzel, U., Palme, J., Bodenhofer, U., and Reddy, S.T. (2017). Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. bioRxiv, Published online April 18, 2017. http://dx.doi.org/10.1101/127902.

Gupta, N.T., Adams, K.D., Briggs, A.W., Timberlake, S.C., Vigneault, F., and Kleinstein, S.H. (2017). Hierarchical clustering can identify B cell clones with high confidence in Ig repertoire sequencing data. J. Immunol. *198*, 2489–2499.

Haynes, B.F., Kelsoe, G., Harrison, S.C., and Kepler, T.B. (2012). B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study. Nat. Biotechnol. *30*, 423–433.

Henry Dunand, C.J., and Wilson, P.C. (2015). Restricted, canonical, stereotyped and convergent immunoglobulin responses. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140238.

Hershberg, U., and Luning Prak, E.T. (2015). The analysis of clonal expansions in normal and autoimmune B cell repertoires. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140239.

Hess, J., Werner, A., Wirth, T., Melchers, F., Jäck, H.-M., and Winkler, T.H. (2001). Induction of pre-B cell proliferation after de novo synthesis of the pre-B cell receptor. Proc. Natl. Acad. Sci. USA *98*, 1745–1750.

Hoehn, K.B., Fowler, A., Lunter, G., and Pybus, O.G. (2016). The diversity and molecular evolution of B-cell receptors during infection. Mol. Biol. Evol. *33*, 1147–1157.

Honjo, T., and Habu, S. (1985). Origin of immune diversity: Genetic variation and selection. Annu. Rev. Biochem. *54*, 803–830.

Imanishi, T., and Mäukelä, O. (1973). Strain differences in the fine specificity of mouse anti-hapten antibodies. Eur. J. Immunol. *3*, 323–330.

Jackson, K.J.L., Kidd, M.J., Wang, Y., and Collins, A.M. (2013). The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor. Front. Immunol. *4*, 263.

Jackson, K.J.L., Liu, Y., Roskin, K.M., Glanville, J., Hoh, R.A., Seo, K., Marshall, E.L., Gurley, T.C., Moody, M.A., Haynes, B.F., et al. (2014). Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. Cell Host Microbe *16*, 105–114.

Jacob, J., and Kelsoe, G. (1992). In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. II. A common clonal origin for periarteriolar lymphoid sheath-associated foci and germinal centers. J. Exp. Med. *176*, 679–687.

Jacob, J., Kassir, R., and Kelsoe, G. (1991a). In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. I. The architecture and dynamics of responding cell populations. J. Exp. Med. *173*, 1165–1175.

Jacob, J., Kelsoe, G., Rajewsky, K., and Weiss, U. (1991b). Intraclonal generation of antibody mutants in germinal centres. Nature *354*, 389–392.

Janeway, C.A., and Murphy, K. (2011). Janeway's Immunobiology (Taylor & Francis).

Jardine, J.G., Kulp, D.W., Havenar-Daughton, C., Sarkar, A., Briney, B., Sok, D., Sesterhenn, F., Ereño-Orbea, J., Kalyuzhniy, O., Deresa, I., et al. (2016). HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen. Science *351*, 1458–1463.

Jiang, N., Weinstein, J.A., Penland, L., White, R.A., 3rd, Fisher, D.S., and Quake, S.R. (2011). Determinism and stochasticity during maturation of the zebrafish antibody repertoire. Proc. Natl. Acad. Sci. USA *108*, 5348–5353.

Johnston, C.M., Wood, A.L., Bolland, D.J., and Corcoran, A.E. (2006). Complete sequence assembly and characterization of the C57BL/6 mouse Ig heavy chain V region. J. Immunol. *176*, 4221–4234.

Khan, T.A., Friedensohn, S., Gorter de Vries, A.R., Straszewski, J., Ruscheweyh, H.-J., and Reddy, S.T. (2016). Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. Sci. Adv. *2*, e1501371.

Landsteiner, K. (1947). The Specificity of Serological Reactions, Revised Edition (Harvard University Press).

Lindau, P., and Robins, H.S. (2017). Advances and applications of immune receptor sequencing in systems immunology. Curr. Opin. Syst. Biol. *1*, 62–68.

Madi, A., Shifrut, E., Reich-Zeliger, S., Gal, H., Best, K., Ndifon, W., Chain, B., Cohen, I.R., and Friedman, N. (2014). T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. Genome Res. *24*, 1603–1612.

Manz, R.A., Thiel, A., and Radbruch, A. (1997). Lifetime of plasma cells in the bone marrow. Nature *388*, 133–134.

Mayer, A., Balasubramanian, V., Mora, T., and Walczak, A.M. (2015). How a well-adapted immune system is organized. Proc. Natl. Acad. Sci. USA *112*, 5950–5955.

McHeyzer-Williams, M.G., McLean, M.J., Lalor, P.A., and Nossal, G.J. (1993). Antigen-driven B cell differentiation in vivo. J. Exp. Med. *178*, 295–307.

Menzel, U., Greiff, V., Khan, T.A., Haessler, U., Hellmann, I., Friedensohn, S., Cook, S.C., Pogson, M., and Reddy, S.T. (2014). Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. PLoS ONE *9*, e96727.

Miho, E., Greiff, V., Roskar, R., and Reddy, S.T. (2017). The fundamental principles of antibody repertoire architecture revealed by large-scale network analysis. bioRxiv, Published online April 5, 2017. http://dx.doi.org/10.1101/124578.

Mora, T., Walczak, A.M., Bialek, W., and Callan, C.G., Jr. (2010). Maximum entropy models for antibody diversity. Proc. Natl. Acad. Sci. USA *107*, 5405–5410.

Osmond, D.G., Rolink, A., and Melchers, F. (1998). Murine B lymphopoiesis: towards a unified model. Immunol. Today *19*, 65–68.

Parameswaran, P., Liu, Y., Roskin, K.M., Jackson, K.K.L., Dixit, V.P., Lee, J.-Y., Artiles, K.L., Zompi, S., Vargas, M.J., Simen, B.B., et al. (2013). Convergent antibody signatures in human dengue. Cell Host Microbe *13*, 691–700.

Reddy, S.T., Ge, X., Miklos, A.E., Hughes, R.A., Kang, S.H., Hoi, K.H., Chrysostomou, C., Hunicke-Smith, S.P., Iverson, B.L., Tucker, P.W., et al. (2010). Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. Nat. Biotechnol. *28*, 965–969.

Reshetova, P., van Schaik, B.D., Klarenbeek, P.L., Doorenspleet, M.E., Esveldt, R.E., Tak, P.P., Guikema, J.E., de Vries, N., and van Kampen, A.H. (2017). Computational model reveals limited correlation between germinal center B-cell subclone abundancy and affinity: Implications for repertoire sequencing. Front. Immunol. *8*, 221.

Robinson, W.H. (2015). Sequencing the functional antibody repertoire—diagnostic and therapeutic discovery. Nat. Rev. Rheumatol. *11*, 171–182.

Rubelt, F., Bolen, C.R., McGuire, H.M., Vander Heiden, J.A., Gadala-Maria, D., Levin, M., Euskirchen, G.M., Mamedov, M.R., Swan, G.E., Dekker, C.L., et al. (2016). Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. Nat. Commun. *7*, 11112.

Saada, R., Weinberger, M., Shahaf, G., and Mehr, R. (2007). Models for antigen receptor gene rearrangement: CDR3 length. Immunol. Cell Biol. *85*, 323–332.

Safonova, Y., Bonissone, S., Kurpilyansky, E., Starostina, E., Lapidus, A., Stinson, J., DePalatis, L., Sandoval, W., Lill, J., and Pevzner, P.A. (2015). IgRepertoireConstructor: a novel algorithm for antibody repertoire construction and immunoproteogenomics analysis. Bioinformatics *31*, i53–i61.

Savelyeva, N., Shipton, M., Suchacki, A., Babbage, G., and Stevenson, F.K. (2011). High-affinity memory B cells induced by conjugate vaccines against weak tumor antigens are vulnerable to nonconjugated antigen. Blood *118*, 650–659.

Shen, P., Roch, T., Lampropoulou, V., O'Connor, R.A., Stervbo, U., Hilgenberg, E., Ries, S., Dang, V.D., Jaimes, Y., Daridon, C., et al. (2014). IL-35-producing B cells are critical regulators of immunity during autoimmune and infectious diseases. Nature *507*, 366–370.

Shugay, M., Britanova, O.V., Merzlyak, E.M., Turchaninova, M.A., Mamedov, I.Z., Tuganbaev, T.R., Bolotin, D.A., Staroverov, D.B., Putintseva, E.V., Plevova, K., et al. (2014). Towards error-free profiling of immune repertoires. Nat. Methods *11*, 653–655.

Soberón, J., and Llorente, J. (1993). The use of species accumulation functions for the prediction of species richness. Conserv. Biol. *7*, 480–488.

Srivastava, B., Quinn, W.J., 3rd, Hazard, K., Erikson, J., and Allman, D. (2005). Characterization of marginal zone B cell precursors. J. Exp. Med. *202*, 1225–1234.

Stern, J.N., Yaari, G., Heiden, J.A.V., Church, G., Donahue, W.F., Hintzen, R.Q., Huttner, A.J., Laman, J.D., Nagra, R.M., Nylander, A., et al. (2014). B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. Sci. Transl. Med. *6*, 248ra107.

Strauli, N.B., and Hernandez, R.D. (2016). Statistical inference of a convergent antibody repertoire response to influenza vaccine. Genome Med. *8*, 60.

Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature *302*, 575–581.

Trück, J., Ramasamy, M.N., Galson, J.D., Rance, R., Parkhill, J., Lunter, G., Pollard, A.J., and Kelly, D.F. (2014). Identification of antigen-specific B cell receptor sequences using public repertoire analysis. J. Immunol. *194*, 252–261.

Vollmers, C., Sit, R.V., Weinstein, J.A., Dekker, C.L., and Quake, S.R. (2013). Genetic measurement of memory B-cell recall using antibody repertoire sequencing. Proc. Natl. Acad. Sci. USA *110*, 13463–13468.

Wang, C., Liu, Y., Cavanagh, M.M., Le Saux, S., Qi, Q., Roskin, K.M., Looney, T.J., Lee, J.-Y., Dixit, V., Dekker, C.L., et al. (2015). B-cell repertoire responses to varicella-zoster vaccination in human identical twins. Proc. Natl. Acad. Sci. USA *112*, 500–505.

Weinstein, J.A., Jiang, N., White, R.A., 3rd, Fisher, D.S., and Quake, S.R. (2009). High-throughput sequencing of the zebrafish antibody repertoire. Science *324*, 807–810.

Xu, J.L., and Davis, M.M. (2000). Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. Immunity *13*, 37–45.

Yang, Y., Wang, C., Yang, Q., Kantor, A.B., Chu, H., Ghosn, E.E., Qin, G., Mazmanian, S.K., Han, J., and Herzenberg, L.A. (2015). Distinct mechanisms define murine B cell lineage immunoglobulin heavy chain (IgH) repertoires. eLife *4*, e09083.

Zhang, T.-H., Wu, N.C., and Sun, R. (2016). A benchmark study on error-correction by read-pairing and tag-clustering in amplicon-based deep sequencing. BMC Genomics *17*, 108.