Open access • Posted Content • DOI:10.1101/127902

Learning The High-Dimensional Immunogenomic Features That Predict Public And Private Antibody Repertoires — Source link

Victor Greiff, Cédric R. Weber, Johannes Palme, Ulrich Bodenhofer ...+3 more authors

Institutions: ETH Zurich, Austrian Institute of Technology, Johannes Kepler University of Linz

Published on: 18 Apr 2017 - bioRxiv (Cold Spring Harbor Laboratory)

The Journal of Immunology

Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires

Victor Greiff, Cédric R. Weber, Johannes Palme, Ulrich Bodenhofer, Enkelejda Miho, Ulrike Menzel and Sai T. Reddy

J Immunol 2017; 199:2985-2997; Prepublished online 18 September 2017;
doi: 10.4049/jimmunol.1700594
http://www.jimmunol.org/content/199/8/2985

The Journal of Immunology is published twice each month by
The American Association of Immunologists, Inc.,
1451 Rockville Pike, Suite 650, Rockville, MD 20852

Print ISSN: 0022-1767 Online ISSN: 1550-6606.

Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires

Victor Greiff,*[,1] Cédric R. Weber,*[,1] Johannes Palme,[†,‡] Ulrich Bodenhofer,[†] Enkelejda Miho,* Ulrike Menzel,* and Sai T. Reddy*

Recent studies have revealed that immune repertoires contain a substantial fraction of public clones, which may be defined as Ab or TCR clonal sequences shared across individuals. It has remained unclear whether public clones possess predictable sequence features that differentiate them from private clones, which are believed to be generated largely stochastically. This knowledge gap represents a lack of insight into the shaping of immune repertoire diversity. Leveraging a machine learning approach capable of capturing the high-dimensional compositional information of each clonal sequence (defined by CDR3), we detected predictive public clone and private clone–specific immunogenomic differences concentrated in CDR3's N1–D–N2 region, which allowed the prediction of public and private status with 80% accuracy in humans and mice. Our results unexpectedly demonstrate that public, as well as private, clones possess predictable high-dimensional immunogenomic features. Our support vector machine model could be trained effectively on large published datasets (3 million clonal sequences) and was sufficiently robust for public clone prediction across individuals and studies prepared with different library preparation and high-throughput sequencing protocols. In summary, we have uncovered the existence of high-dimensional immunogenomic rules that shape immune repertoire diversity in a predictable fashion. Our approach may pave the way for the construction of a comprehensive atlas of public mouse and human immune repertoires with potential applications in rational vaccine design and immunotherapeutics. ***The Journal of Immunology*****, 2017, 199: 2985–2997.**

The clonal identity, specificity, and diversity of adaptive immune receptors are largely defined by the CDR3 sequence of variable H and variable β chains of Abs and TCRs, respectively (1–6). CDR3 encompasses the junction region of recombined V, D, and J gene segments, as well as nontemplated nucleotide (N- and P-nucleotides) addition (7). As a result of the enormous theoretical diversity of Ab and TCR repertoires ($>10^{13}$) (8–11) and technological limitations (Sanger sequencing), it was long believed that clonal repertoires were, to an overwhelming extent, private to each individual (12, 13). However, as a result of recent advances in high-throughput immune repertoire sequencing, it has been observed that a considerable fraction ($>1\%$) of CDR3s is shared across individuals (1, 5, 14–27). Thus, these shared clones (hereafter referred to as "public clones") are refining our view of adaptive immune repertoire diversity. Therefore, a fundamental question emerges: Are there immunogenomic differences that predetermine whether a clone becomes part of the public or private immune repertoire?

In the context of Ab and TCR repertoires, the large theoretical clonal (CDR3) diversity renders the investigation of public and private repertoires computationally challenging (28). Previous studies using conventional low-dimensional analysis suggested that public clones are "germline-like" clones with few insertions, thereby having higher occurrence probabilities, whereas private clones contain more stochastic elements (i.e., N1, N2 insertions) (18, 24). To investigate the composition of large numbers of sequences with the appropriate dimensionality, sequence kernels are increasingly used (29, 30). Sequence kernels are high-dimensional functions that measure the similarity of pairs of sequences, for example, by comparing the occurrence of specific subsequences (k-mers) in a high-dimensional space (31, 32). Supervised machine learning (e.g., support vector machine [SVM] analysis) is an approach that takes low- or high-dimensional feature functions as input to find a classification rule that discriminates between two (or more) given classes on a single-clone level (e.g., public versus private clones) (33). In contrast to using conventional low-dimensional features to analyze immune repertoires, the coupling of high-dimensional sequence kernels to SVM analysis may offer greater insight into the immunogenomic structure of repertoire diversity, specifically the difference between public and private repertoires. As opposed to previous approaches (34), a key advantage of sequence kernel–based SVM analysis is the prediction profile–based identification of CDR3 subregions that are most predictive for a respective class (public or private class) (31, 32). This approach may lead to predictive immunological and

*Department of Biosystems Science and Engineering, Swiss Federal Institute of Technology Zurich, CH-4058 Basel, Switzerland; [†]Institute of Bioinformatics, Johannes Kepler University, 4040 Linz, Austria; and [‡]Health and Environment Department, Austrian Institute of Technology, 1220 Vienna, Austria

[1]V.G. and C.R.W. contributed equally to this work.

ORCIDs: 0000-0003-2622-5032 (V.G.); 0000-0003-4802-8996 (C.R.W.); 0000-0001-6859-8828 (U.B.); 0000-0001-6461-0519 (E.M.); 0000-0002-9956-3043 (U.M.).

Received for publication April 25, 2017. Accepted for publication August 16, 2017.

This work was supported by the Swiss National Science Foundation (Project 31003A_143869), the Swiss Initiative in Systems Biology AntibodyX Research, Technology, and Development Project, the Swiss Vaccine Research Institute, and a European Research Council Starting Grant (all to S.T.R.). The professorship of S.T.R. is made possible by the generous endowment of the S. Leslie Misrock Foundation.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Address correspondence and reprint requests to Prof. Sai T. Reddy, ETH Zurich, Department of Biosystems Science and Engineering, Mattenstrasse 26, 4058 Basel, Switzerland. E-mail address: sai.reddy@ethz.ch

The online version of this article contains supplemental material.

Abbreviations used in this article: AUC, area under the receiver operating characteristic curve; BACC, balanced accuracy; KeBABS, kernel-based analysis of biological sequences; nBC, naive B cell; PC, plasma cell; preBC, pre-B cell; SVM, support vector machine.

mechanistic insight into the immunogenomic elements that shape repertoire diversity.

To identify the immunogenomic differences between public and private Ab repertoires (Fig. 1), we applied SVM (Fig. 1B) to six large-scale immune repertoire (Ab and TCR) sequencing datasets from mice and humans (Fig. 1A). When using low-dimensional features (germline gene and amino acid usage, CDR3 subregion length) as the input for SVM analysis, prediction accuracy of private and public status reached a maximum of 67%, which only moderately improves on a random classifier (50%). However, when implementing a high-dimensional sequence kernel (sequence composition)–based SVM analysis, we were able to detect strong immunogenomic differences concentrated in the N1–D–N2 region in public and private clones with a high prediction accuracy (balanced accuracy [BACC] ∼ 79–83%, Fig. 1C). Our results unexpectedly signify that public and private Ab repertoires contain predictive high-dimensional features that enable their accurate classification. Our SVM approach was sufficiently robust to be applied across individuals and repertoire studies with different library preparations and high-throughput sequencing protocols, demonstrating its broad applicability.

Materials and Methods

Immune repertoire high-throughput sequencing datasets

We conducted our analysis on six high-throughput immune repertoire sequencing datasets, all of which are described below. Quality and read statistics may be found in the respective publications.

Dataset 1

Murine B cell origin (C57BL/6JRj; Janvier Labs): Sequencing data were generated by Greiff et al. (17). Briefly, B cells were isolated from four C57BL/6 cohorts (n = 4 or 5), including untreated mice and mice that received prime-boost immunization with hapten/protein Ags. Cells were sorted into pre-B cell (preBC), naive B cell (nBC), and plasma cell (PC) subsets by flow cytometry. Cell numbers per mouse were 750,000 for preBCs, 1,000,000 for nBCs, and 90,000 for PCs. RNA was isolated from cells, and Ab H-chain libraries were prepared by RT-PCR and sequenced using an Illumina MiSeq platform (2 × 300 bp paired end). The sequencing data have been submitted to the ArrayExpress Archive of Functional Genomics Data under accession number E-MTAB-5349 (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5349/) along with full experimental details, and they were preprocessed using MiXCR software for VDJ annotation, clonotype formation by CDR3, and error correction, as described previously (17, 35). For downstream analyses, functional clonotypes (clones, CDR3) were only retained if they were composed of at least four aa and had a minimal read count of two (36, 37). Public clones were defined as those clones that occurred in at least two different individuals within the same B cell population and cohort.

Dataset 2

Murine B cell origin (BALB/cJRj; Janvier Labs): Sequencing data were generated by Greiff et al. (17) and have been submitted to the ArrayExpress Archive of Functional Genomics Data under accession number E-MTAB-5349 (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5349/) with full experimental details. Briefly, nBCs (1,000,000 cells per mouse) from four unimmunized BALB/c mice were isolated using the sorting panel from Dataset 1, and Ab H-chain libraries were prepared and sequenced analogously to Dataset 1. Data preprocessing was performed analogously to Dataset 1. Public clones were defined as those clones that occurred at least twice across mice.

Dataset 3

Murine B cell origin (pet shop mice): Sequencing data were generated by Greiff et al. (17) and have been submitted to the ArrayExpress Archive of Functional Genomics Data under accession number E-MTAB-5349 (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5349/) with full experimental details. Briefly, nBCs (∼671,000 cells per mouse) from three pet shop mice were isolated, and library preparation, sequencing, and data preprocessing were performed analogously to Dataset 1. Public clones were defined as those clones that occurred at least twice across mice.

Dataset 4

Murine B cell origin (C57BL/6J): Sequencing data were published by Yang et al. (21). Mature B cells were extracted from C57BL/6J mice and sorted (1–2 × 10^4 per B cell population) into developmentally distinct subsets (splenic follicular B cells [n = 5], marginal zone B cells [n = 7], and peritoneal B2 B cells [n = 5] and B-1a B cells [n = 43]). Data preprocessing was performed analogously to Dataset 1. Public clones were defined as those clones of a given B cell population that occurred at least twice across mice.

Dataset 5

Human B cell origin: Sequencing data of nBCs and memory B cells from three healthy donors were published by DeWitt et al. (14) and downloaded already preprocessed from http://datadryad.org/resource/doi:10.5061/dryad.35ks2. Public clones were defined as those clones that occurred at least twice across individuals within a given B cell population (naive, memory). The numbers of nBCs and memory B cells were 2–4 × 10^7 and 1.5–2 × 10^7, respectively.

Dataset 6

Murine T cell origin: TCRβ-chain sequencing data were published by Madi et al. (18). CD4 T cells were isolated from 28 mice (three cohorts; untreated [n = 12], immunized with CFA [n = 7], or immunized with CFA and OVA [n = 9]). Data preprocessing was performed using MiXCR software for annotation and error correction, as described previously (17, 35). Public clones were defined as those clones that occurred at least twice across mice of a given cohort.

Determination of statistical significance

Significance was tested using the Wilcoxon rank-sum test unless indicated otherwise. Where applicable, the significance of correlation coefficients was tested using the cor.test() function in R with default parameters.

Statistical analysis and plots

Statistical analysis was performed using R (38) and Python (39). Graphics were generated using the R packages ggplot2 (40), RColorBrewer (41), and ComplexHeatmap (42). Parallel computing of SVM analyses was performed using the R packages BatchJobs (43) and doParallel (44).

Definition of a clone

For all analyses, clones were defined by 100% amino acid sequence identity of CDR3 (Ab H chain, TCRβ) regions (1, 17, 36). CDR3 regions were annotated and defined by MiXCR software (35) according to ImMunoGeneTics nomenclature (45).

Quantification of overlap and correlation

As defined previously (17), the percentage of clones shared between two repertoires X and Y: $overlap(X,Y) = \frac{|X \cap Y|}{\min(|X|,|Y|)} \times 100$, where $|X|$ and $|Y|$ are the clonal sizes (number of unique clones, species richness) of repertoires X and Y. A repertoire was mathematically defined as a set of unique clones. Correlation of germline gene/CDR3 subregion abundances between public and private repertoires was calculated using the Spearman correlation coefficient.

Determination of sequence similarity among clones within a repertoire

Sequence similarity among clones within a repertoire was calculated as previously described (17). Briefly, the Levenshtein distance between all pairwise CDR3 amino acid sequence combinations of identical CDR3 length and V/J genes was calculated, and each Levenshtein distance was subsequently normalized by the sequence length of the respective sequence combination. Finally, to correct for the baseline sequence similarity that is common to all clonal families, the similarity calculated among all CDR3s of identical length (but undefined V/J gene usage) was subtracted from that calculated exclusively among CDR3s of identical V/J gene usage and CDR3 length. Levenshtein distances were computed using the stringdist package in R (46).

Junction analysis

V, N1, D, N2, and J subregion annotation of sequences was performed using ImMunoGeneTics/HighV-QUEST (47) (after initial preprocessing by MiXCR software) (35). Deletions were determined by finding the longest common substring between the germline genes and the V, D, and J subregions identified in the CDR3 sequences.

Determination of Shannon evenness

Shannon evenness was calculated as previously described (48). Briefly, we calculated the Hill diversity for $\alpha = 1$ ${}^{\alpha}D = (\sum_{i=1}^{n} {f_i}^{\alpha})^{\frac{1}{1-\alpha}}$) for a given CDR3 subregion frequency distribution ($\vec{f}$, enumeration of the abundance of each of the *n* V, N1, D, N2, J subregions or combinations thereof). Subsequently, we obtained the Shannon evenness ${}^{\alpha=1}E$ by normalizing ${}^{\alpha=1}D$ by the respective total number of V, N1, D, N2, or J regions or combinations thereof (*n*) in the given repertoire. The Shannon evenness varies between ~0 and 1; higher values indicate an increasingly uniform frequency distribution.

SVM analysis

To classify clones into public and private classes, a supervised learning approach was chosen in the form of an SVM model. As input for all SVM analyses, CDR3-length equilibrated (normalized) and class-balanced datasets were built for each sample (Table I). Briefly, for each sample, all public clones were paired in equal numbers with private clones of the same sample, such that public and private repertoires followed identical CDR3 length distributions, with the following exceptions: for analyses of aggregated datasets (Fig. 4D), neither CDR3 length distributions nor public and private repertoire sizes were matched, and for Dataset 1, we also show predictive performance for CDR3-length nonequilibrated public and private repertoires (Supplemental Fig. 4C). Thus, within all datasets, be they CDR3 length normalized or not, public and private repertoires showed a wide range of CDR3 lengths (4–29 aa; see Table I legend).

SVM analysis was performed using kernel-based analysis of biological sequences (KeBABS) (31) and sklearn (49), both of which are described in more detail below. For all SVM analyses, each dataset was split into training (80%) and test (20%) sets. Cross-validation and SVM training were performed on the training set, and class prediction was performed on the test set. Prediction accuracy of class discrimination was quantified by calculating $BACC = \frac{1}{2} \times (spec + sens)$ (50), where specificity was defined as $spec = \frac{TN}{TN+FP}$ and sensitivity was defined as $sens = \frac{TP}{TP+FN}$ (*TP* = true positive, *TF* = true negative, *FP* = false positive, and *FN* = false negative). Additionally, the area under the receiver operating characteristic curve (AUC) was calculated using KeBABS R package (31). An AUC value of 1 means perfect prediction accuracy (BACC = 100%), whereas an AUC value of 0.5 (BACC = 50%) is equivalent to a random classifier.

KeBABS SVM analysis

To discriminate public and private clones based on the CDR3 sequence, we used the R package KeBABS (31), which implements KeBABS. For all datasets, unless mentioned otherwise, we used the position-independent gappy pair kernel (51, 52), which splits sequences into features of length *k* with a gap of maximal length *m* (Fig. 4A). Independently of species (mouse, human) or lymphocyte type (B/T cell), for the analysis of nucleotide sequences the parameters were set to $k = 3$, $m = 1$, and cost parameter (C) = 10, whereas the analysis of amino acid sequences was performed using parameters $k = 1$, $m = 1$, and $C = 100$. Optimal parameter combinations maximizing prediction accuracy were determined by cross-validation on the training set. *C* sets the cost for the misclassification of a sequence. The maximal numbers of possible features used in the gappy kernel are $4^{2\times k} \times (m + 1) = 8192$ for nucleotide sequences and $20^{2\times k} \times (m + 1) = 800$ for amino acid sequences.

Prediction profiles

Prediction profiles were computed from feature weights, as we described previously (31, 32, 52). Prediction profiles quantify the contribution of each sequence position to the decision value (public, private). Thus, as opposed to single feature weights, prediction profiles provide improved biological interpretability of learning results, because they render those individual positions or sequence stretches visible that substantially contribute to classification accuracy (31).

sklearn SVM analysis

For public versus private discrimination based on amino acid and V, N1, D, N2, and J composition (counts), the sklearn SVM implementation (49) for Python (39) was used with a linear kernel, and the cost parameter was set at $C = 10$, as determined by cross-validation.

Results

Public and private clone repertoires cannot be predicted by germline gene or amino acid usage

As the basis for elucidating the immunogenomic differences between public and private clones, we used a recently published large-scale high-throughput sequencing Ab repertoire dataset (17) (Dataset 1, *Materials and Methods*). This dataset contains 400 million full-length Ab variable H sequences derived from 19 mice and stratified into key stages of B cell differentiation: preBCs (IgM), nBCs (IgM), and PCs (IgG). Thus, this dataset provided the important advantages of high sequencing and biological depth (preBCs and nBCs represent Ag-inexperienced cells, whereas PCs are postclonal selection and expansion due to Ag exposure). Public clones, precisely defined here as CDR3 sequences (100% amino acid identity) occurring in at least two mice, were found to compose, on average, 15% (preBCs), 23% (nBCs), and 26% (PCs) of Ab repertoires across B cell stages (Figs. 1, 2A). As previously reported, we found that public clones are biased to higher frequencies and are enriched in sequences from natural Abs (Supplemental Fig. 1C, 1D) (18, 25, 53). Irrespective of public or private status, clones were, on average, 13–16% similar on the amino acid sequence level (Supplemental Fig. 1B). Thus, public clones were not substantially more similar to one another than private clones. Across B cell development, public and private clones used nearly identical V, D, J, VJ, and VDJ germline genes (overlap > 95%), with nearly identical frequency in preBCs and nBCs (Spearman $r \sim +1$) and with varying frequencies in PCs (Spearman $r > +0.5$–0.8) (Fig. 2B). Thus, overall, neither public nor private clones showed any preferential germline gene usage. On average, higher-frequency amino acids (e.g., A, C, and D) occurred more often in public clones, whereas lower-frequency amino acids (e.g., H, I, and K) could be found at higher frequency in private clones (Fig. 2C). This observation held true across all B cell stages ($r = +0.5$–0.76, $p < 0.05$, Supplemental Fig. 1A). Repertoire-wide absolute differences in amino acid usage between private and public clones were slight (0.2–1.4 percentage points, Fig. 2C). To test whether these repertoire-level differences were sufficient to predictively discriminate between public and private clones on a single-clone level, we used SVM analysis (Fig. 1B). For all SVM analyses in this study, we strove to minimize technological classification bias by constructing SVM datasets such that public and private repertoires had identical CDR3 length distributions and were class balanced (identical number of public and private clones; see Table I and *Materials and Methods*). Nevertheless, as a control, all SVM analyses were also validated without CDR3 length distribution matching (Supplemental Fig. 4C). Datasets constructed for SVM analyses were divided into 80% training sequences and 20% test sequences (Fig. 1B). We found that amino acid usage (dimensionality: 20) was a suboptimal predictor of clonal status, with a prediction accuracy ≤ 66% (Fig. 2D). Prediction accuracy (hereafter, the terms BACC, prediction accuracy, and classification accuracy are used interchangeably) is defined as the mean of specificity and sensitivity (31, 48, 50).

Public and private clones do not differ predictively in CDR3 subregion length

Because public and private clones did not differ in germline gene usage, we asked whether they differed with respect to length and diversity of CDR3 subregions (V, N1, D, N2, and J). The V, D, and J subregions are derived from germline gene segments (IGHV, IGHD, IGHJ), whereas N1 and N2 represent insertions (n- and p-nucleotides) introduced during the junctional somatic recombination process. Public clones in preBC and nBC repertoires possessed a relative V subregion length of 23–24% (Fig. 3A), whereas private clones had slightly shorter V subregions (~21%, $p < 0.05$, Supplemental Fig. 2A). The J subregion length behaved analogously (public 40%, private 36%), whereas the D subregion length did not differ between classes (public 25%,

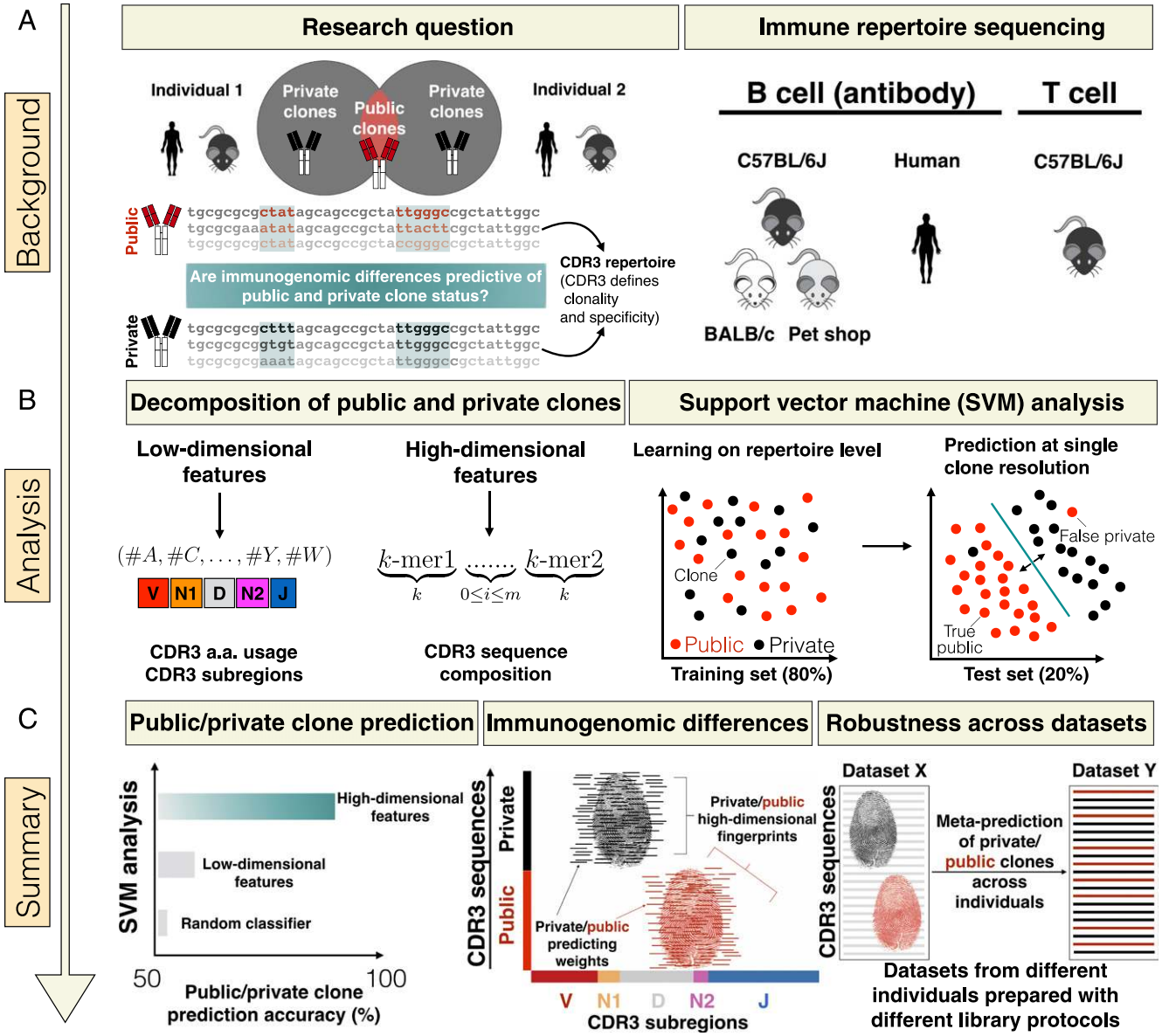

FIGURE 1. Immunogenomic analysis of public and private Ab repertoires. (**A**) We asked whether immunogenomic differences exist that predetermine a clonal sequence's (CDR3) public or private status within an immune repertoire. The public repertoire is composed of clones being shared among at least two individuals (we also explored an alternative public clone definition, Fig. 5C). Private clones are those unique to each individual. We defined Ab and T cell clones based on 100% H/β-chain CDR3 identity. For statistical power, we used six large-scale immune repertoire datasets (Table I) comprising different B cell populations, species (humans, mice), and immune AgRs (BCR/TCR). (**B**) To answer our question, we decomposed public and private immune repertoires into conventional low-dimensional features (e.g., CDR3 amino acid usage [Figs. 2, 3]) or novel high-dimensional features (CDR3 sequence decomposition into subsequences of length k (k-mers) separated by a gap of length m [Figs. 4–6]). Leveraging supervised machine learning (SVMs), we tested whether low- and high-dimensional features can identify immunogenomic differences between public and private repertoires and, consequently, can be used for prediction of public and private status at single-clone resolution. (**C**) We found that low-dimensional features are poor predictors of public and private clone status. In contrast, we detected strong predictive immunogenomic differences, concentrated in the N1–D–N2 CDR3 subregion, between public and private clones using high-dimensional features. Thus, public and private clones each possess a class-specific high-dimensional immunofingerprint composed of class-specific subsequences that enables their classification with high accuracy. Our SVM approach was generalizable across individuals and datasets produced in different laboratories with varying library-preparation and high-throughput sequencing protocols.

private 25%). We observed the largest difference between public and private clones in the relative length of N1 and N2 subregions with deviations of 36–46 percentage points from a 1:1 ratio (N1: public ~ 6.5%, private ~ 8.2%; N2: public ~ 4.3%, private ≈ 7.7%, $p < 0.05$, Fig. 3A, Supplemental Fig. 2A, 2B). Conversely, PC CDR3 subregion lengths did not differ between public and private clones (with the exception of N1, which was slightly longer in public clones, Fig. 3A, Supplemental Fig. 2B).

Regardless of public or private designation, nearly all CDR3s (>94%) had at least one nucleotide insertion (N1 or N2) and at least one deletion (Fig. 3B); thus, only a very small portion of clones were germline-like, having neither insertion nor deletion (≤4%, Supplemental Fig. 3C, 3D). Furthermore, across B cell populations, N1 and N2 insertions were present in >50 and >70% of public and private clones, respectively. Of note, N1 and N2 insertions showed no preferential selection of germline gene segments (IGHV, D, J) (Supplemental Fig. 2D).

Deletion length was highest in D subregions (mean of 5′ and 3′ D deletions ~ 7 nt), whereas it was lowest in V subregions (~0.8 nt, Supplemental Fig. 2C). Although private clones showed a higher

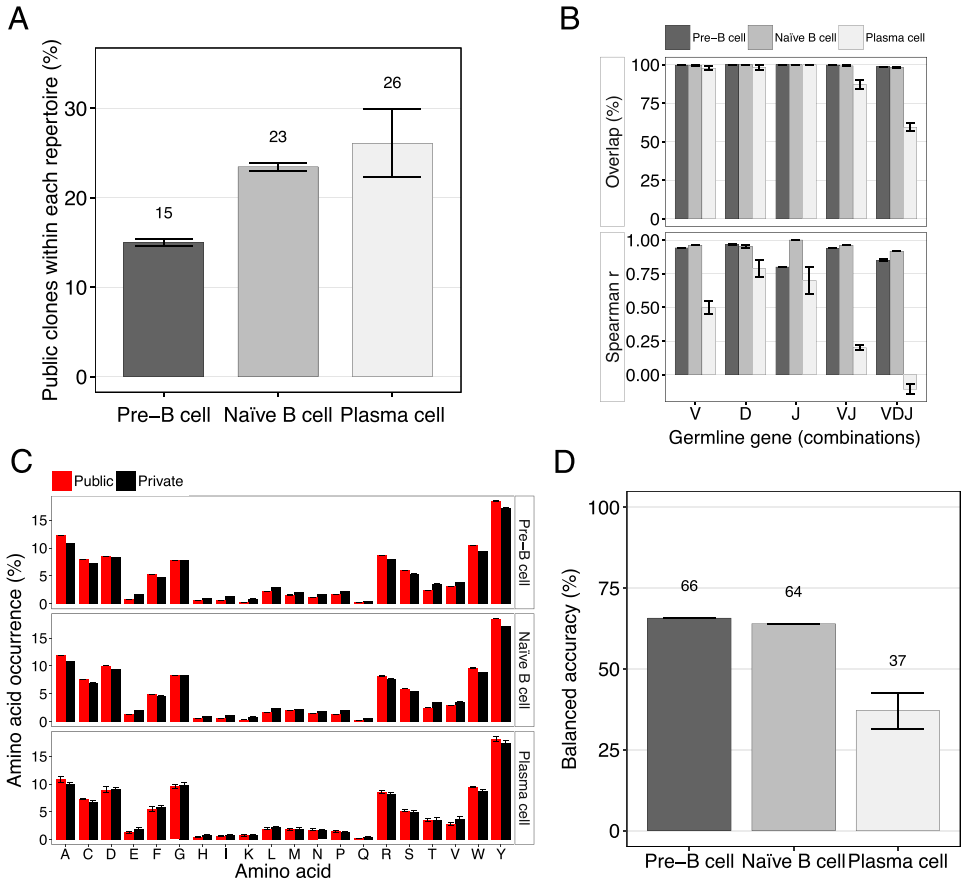

FIGURE 2. Public and private repertoires do not differ predictively in germline gene usage or amino acid composition. (**A**) Public clones represent 15–26% of murine Ab repertoires throughout B cell ontogeny. Public clones were defined as being shared in at least two mice (see *Materials and Methods*). (**B**) Overlap of V, D, and J germline genes (as well as respective combinations: V-J, and V-D-J) and the Spearman correlation of their frequencies between private and public clones by B cell population. (**C**) Relative amino acid composition of public and private clones. Differences between public and private clones were not significant ($p > 0.05$, Kolmogorov–Smirnov test). (**D**) SVM-based discrimination (dimensionality: 20, number of amino acids) of public and private clones based on CDR3 amino acid composition (linear SVM kernel). For SVM-based classification, a class-balanced dataset composed of equal numbers of public and private clones, as well as identical CDR3-length distributions, was assembled for each repertoire (Table I, see *Materials and Methods*). Balanced prediction accuracy was defined as the mean of specificity (detection rate of public clones) and sensitivity (detection rate of private clones). Bar graphs show mean ± SEM across samples. All data shown stem from Dataset 1.

number of deletions, differences between public and private clones were slight (maximum difference ~ 0.6 nt, Supplemental Fig. 2C).

Despite statistically significant differences in CDR3 subregion length and occurrence of insertions and deletions between public and private clones (Fig. 3A, Supplemental Figs. 2A–C, 3C, 3D, $p < 0.05$), training an SVM based on CDR3 subregion length led to a suboptimal prediction accuracy of public/private clone discrimination (BACC ≤ 67%, Fig. 3E). Therefore, CDR3 subregion length (dimensionality 5) is not a reliable predictor of public/private clone status.

Table I. Size of CDR3 length–equilibrated datasets used for SVM classification

Data Origin	Cell Type	No. of CDR3 Sequences
Dataset 1: mouse (C57BL/6J), B cell (17)	preBC (IgM) (19)	48,682 ± 10,493
	nBC (IgM) (19)	177,197 ± 28,393
	PC (IgG) (19)	51 ± 30
Dataset 2: mouse (BALB/c), B cell (17)	nBC (IgM) (4)	244,067 ± 11,293
Dataset 3: mouse (pet shop), B cell (17)	nBC (IgM) (3)	34,218 ± 1207
Dataset 4: mouse (C57BL/6J), B cell (21)	B-1a B cells (43)	2867 ± 1742
	Marginal zone B cells (7)	2987 ± 1151
	Follicular B cells (5)	2295 ± 406
	Peritoneal B2 B cells (5)	1519 ± 429
Dataset 5: human (healthy), B cell (14)	nBCs (IgM) (3)	289,598 ± 36,627
	Memory B cells (IgM, IgG) (3)	35,221 ± 5163
Dataset 6: mouse (C57BL/6J), T cell (18)	CD4 T cells (28)	2621 ± 1777

For each of the six datasets used in this study, a dataset of CDR3 length–equilibrated sequences was constructed consisting of 50% public and 50% private CDR3 sequences. Mean and SD of sequences used across all samples of a given dataset and B/T cell population are displayed. Numbers in parentheses indicate the number of samples for a given category. Amino acid CDR3 lengths are presented as mean ± SD with maximum in parentheses, for Dataset 1, 13 ± 2 (28); Dataset 2, 13 ± 2 (23); Dataset 3, 12 ± 1 (21); Dataset 4, 12 ± 3 (22); Dataset 5, 14 ± 3 (29); and Dataset 6, 10 ± 1 (13).

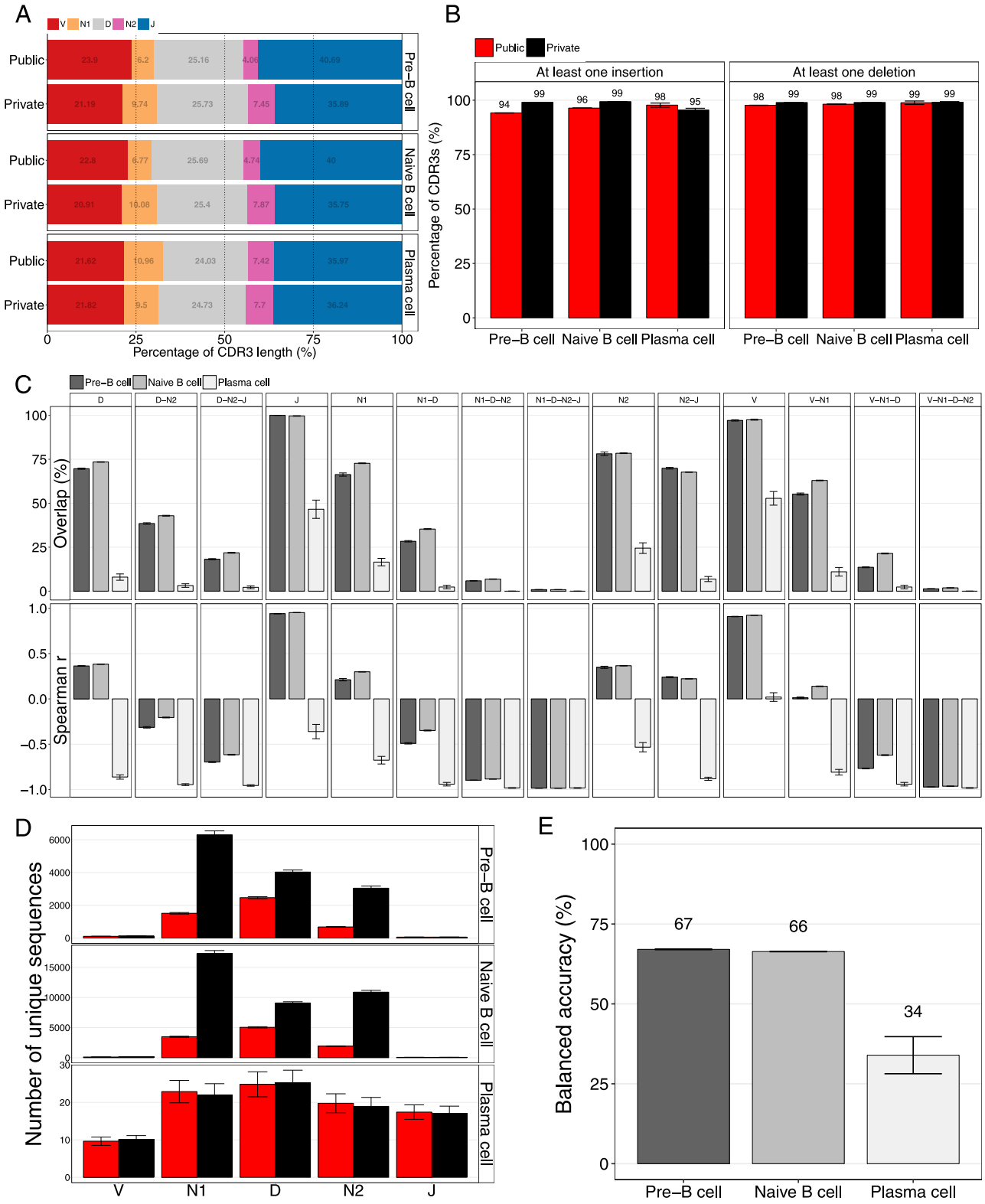

FIGURE 3. CDR3 subregion length does not predict clonal public/private status. (**A**) Normalized CDR3 subregion (V, N1, D, N2, J) lengths (median) of public and private clones by B cell population. (**B**) Percentage of clones (public, private) with at least one N1/N2 insertion or deletion occurrence by B cell population. (**C**) Overlap and Spearman correlation of CDR3 subregions and combinations thereof by B cell population. (**D**) Number of unique V, N1, D, N2, and J CDR3 subregions (species richness) of public and private clones. Species richness of private clone CDR3 subregions was obtained by accounting for private and public clone size differences (bootstrapping, see *Materials and Methods*). See Supplemental Fig. 3A for nonbootstrapped version. (**E**) SVM-based prediction (dimensionality: 5, number of CDR3 subregions, linear SVM kernel) of public and private clones based on relative V, N1, D, N2, and J subregion composition [(A), see *Materials and Methods*]. Class-balanced datasets, as described for Fig. 2, were used for SVM classification. Balanced (prediction) accuracy was defined as the mean of specificity (detection rate of public clones) and sensitivity (detection rate of private clones). Bar graphs show mean ± SEM across samples. All data shown stem from Dataset 1.

Public and private clones show differences in sequence composition

Because low-dimensional features (CDR3 amino acid and subregion length) only achieved ≤67% classification accuracy (Figs. 2D, 3E), we investigated whether CDR3 sequence composition (potential dimensionality $>10^{13}$ different CDR3 sequences) (8, 17) differed between public and private clones. In preBCs and nBCs, V and J subregions were almost completely overlapping with regard to sequence (>97%) and frequency (Spearman $r >$ +0.95, Fig. 3C). Consequently, we observed no difference in V or J subregion diversity (number of unique V and J subregions) between public and private clones (Fig. 3D). In contrast, we observed considerable differences between private and public repertoires with respect to the diversity of N1, D, and N2 subregions (Fig. 3C, 3D, Supplemental Fig. 2A). Specifically, combinations of CDR3 subregions showed low overlap between public and private repertoires, irrespective of B cell population (e.g., N1–D–N2 overlap in nBCs was ∼6%, Fig. 3C, Supplemental Fig. 3B). To summarize, in general, sequence composition differed substantially between public and private clone repertoires.

High-dimensional CDR3 sequence composition analysis predicts public and private clones with 80% accuracy

To test whether the detected differences in sequence composition were predictive, we used high-dimensional sequence kernels for SVM analysis (31). Specifically, we used the gappy-pair sequence kernel (31, 51, 52), which decomposes each CDR3 into subsequences (features) of length k (k-mers) separated by a gap of length m (Fig. 4A; see *Materials and Methods*). Applying this kernel function to all CDR3s of a given training dataset generates a feature matrix of dimension $n \times f$, which serves as input for the SVM analysis: n is the number of CDR3s in the training dataset, and f is the number of features. By cross-validation, we selected the parameter combinations that resulted in the highest prediction accuracy: $k = 3$, $m = 1$ at the nucleotide level (maximal feature diversity = 8192) and $k = 1$, $m = 1$ at the amino acid level (maximal feature diversity = 800). On the nucleotide and the amino acid levels, public and private clones in preBCs and nBCs could be classified with ∼80% accuracy, with very low variation across mice (Fig. 4A). We validated the robustness of this sequence-based SVM approach in three ways. We showed that the SVM was incapable of separating public from public and private from private clones of different individuals (BACC ∼ 50%, Fig. 5D). Furthermore, we showed that the high prediction accuracy was maintained for an alternative and more stringent definition for public clones (BACC = 83–84%, Fig. 5C). Finally, we confirmed that the prediction accuracy was close to random (50%) when shuffling CDR3 nucleotide and amino acid sequences (Fig. 5A) and when shuffling public and private labels across clones (Fig. 5B). In sum, we have performed a combination of simulation controls that exclude the possibility that factors other than class-specific repertoire sequence features substantially impact classification accuracy.

Furthermore, we confirmed that the differences in immunogenomic composition between public and private clones were not exclusively species specific (mouse) or mouse strain specific (C57BL/6) by replicating a classification accuracy ∼ 80% in BALB/c and pet shop mice (Datasets 2 and 3, Supplemental Fig. 4A). Also, public and private clones could be discriminated with >80% accuracy in human nBC and memory B cell repertoires (Fig. 4B, Dataset 5), thus indicating robust classification accuracy for hypermutated (memory) repertoires, as well. Finally, we showed that our approach demonstrated reasonable classification accuracy in mouse TCR variable β repertoires (BACC = 74%, Fig. 4B, Dataset 6).

Theoretically, successful classification of public/private clones within each individual (mouse or human) could be due to lineage-driven (clonal relatedness) effects and, thus, may not be generalizable. We excluded this possibility and showed cross-individual generalizability as follows: we aggregated public and private clones across individuals into datasets of up to 3×10^6 unique clonal sequences and showed that classification accuracy was maintained (maximum BACC = 83%, AUC = 0.90, Fig. 4C), and we used nBC and T cell repertoires of distinct mouse/human individuals as training set and predicted with ∼80% accuracy private/public status of clones in repertoires of unrelated individuals (but of identical species and lymphocyte population [B cell/T cell], Fig. 4C). These results signified that the same set of features used to predict public and private clones within one individual is sufficient for prediction across individuals of the same species. Thus, the public/private-specific features identified using sequence kernel-based SVM on the repertoire level were generalizable enough to gain a species-wide high-dimensional representation of public and private repertoires allowing the discrimination of public from private clones in humans and mice at single-clone resolution with high accuracy.

Prediction by CDR3 sequence composition is dependent on dataset size and is possible across studies

Our high-dimensional sequence composition–based SVM approach was unable to predict public and private clones in PCs (BACC = 50%, Fig. 4A, Dataset 1). With respect to unique CDR3s, the PC SVM dataset was three to four orders of magnitude smaller than that of preBCs and nBCs (Table I, Dataset 1); therefore, we tested whether the lower accuracy was due to sample size. We performed SVM analysis on datasets ranging in size from 100 to 230,000 unique CDR3 sequences (Fig. 5D) and found that prediction accuracy was indeed a function of sample size, increasing from 56% for 100 clonal sequences to 80% for 230,000 clonal sequences. Thus, small sample size may explain the lower prediction accuracies observed in the PC (IgG) dataset. In further support of this hypothesis, we found that, in a dataset of human memory B cells (mixed IgM, IgG; Dataset 5) that was three orders of magnitude larger than the PC dataset, we were able to achieve > 80% accuracy (Fig. 4B), suggesting that prediction of public clones may also be possible for Ag-experienced B cell populations (such as memory cells and PCs) and, thus, is not limited to Ag-inexperienced ones (such as preBCs and nBCs).

Because we observed that dataset size was important for attaining higher prediction accuracy (Fig. 5E), we asked whether large datasets could function as training sets for performing public and private clone prediction in other (smaller) datasets (obtained from studies with possibly different library preparation and high-throughput sequencing protocols). To answer this question, we investigated the prediction accuracy of the sequence composition–based SVM classifier trained on Dataset 1 (nBC B2 B cell population), applied to a test dataset 100 times smaller (177,197 versus 1,519 sequences), consisting of repertoires from various C57BL/6 B2 B cell populations (21) (Dataset 4, Table I). By using the SVM model computed on the larger dataset (Dataset 1), prediction accuracy could be improved by up to 7 percentage points (76–77 versus 69–73%, Fig. 4D), approaching the prediction accuracy within Dataset 1 (Fig. 4A). Thus, sequence kernel-based SVM models can be effectively trained on large, openly accessible datasets, enabling robust predictive performance for meta-analysis across studies of different laboratories using custom library-preparation methods and sequencing protocols.

Stereotypical immunogenomic differences between public and private clones are concentrated in the N1–D–N2 subregions

To identify the subregions that contributed most to classification accuracy, we performed sequence kernel-based SVM on each

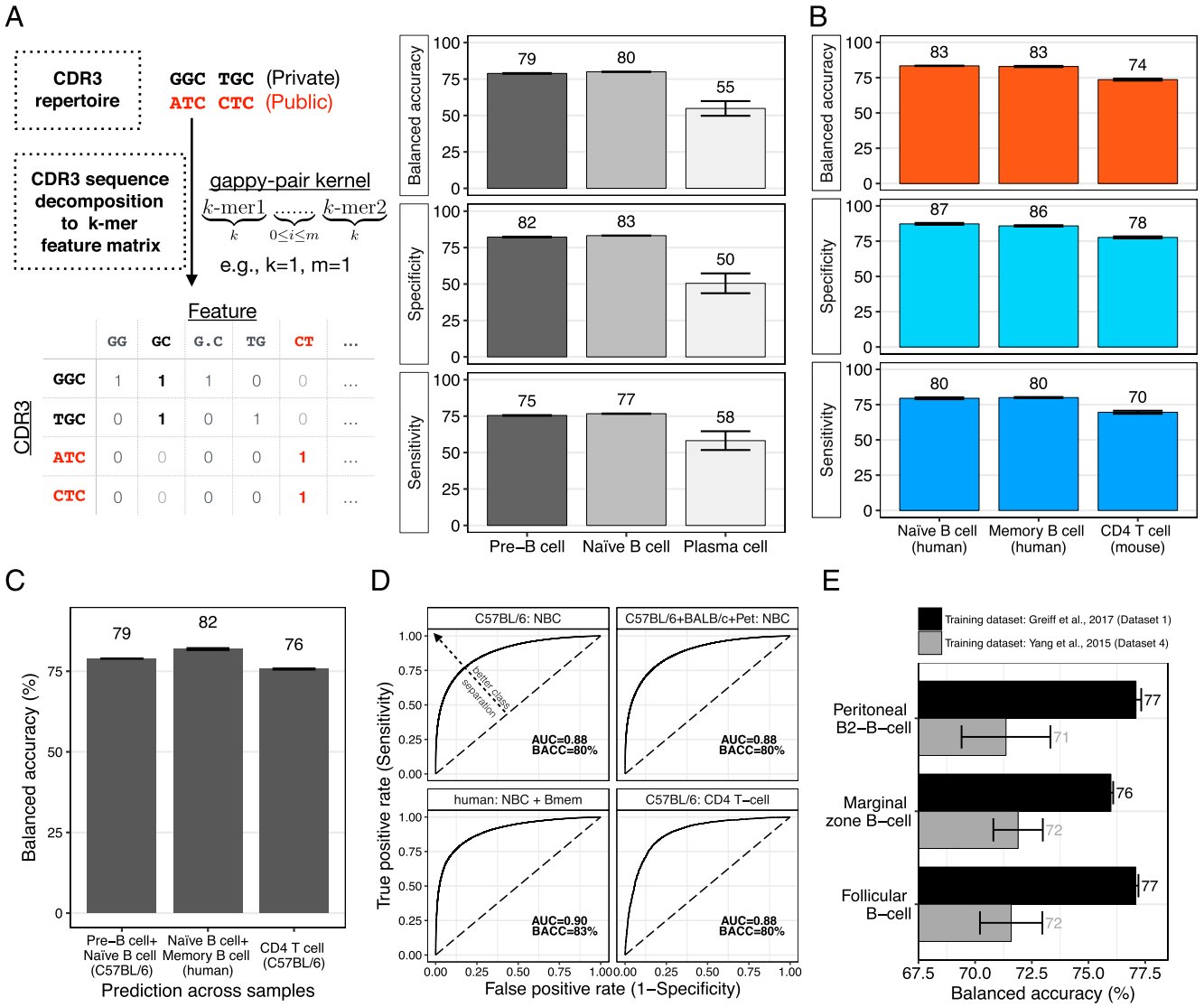

FIGURE 4. Public and private clones are predictable with 80% accuracy using high-dimensional CDR3 sequence decomposition. (**A**) Gapped k-mer–based SVM discrimination of murine public and private clones (nucleotide sequence, Dataset 1). For each repertoire, a dataset composed of equal numbers of public and private clones (nucleotide sequences, length equilibrated) was assembled (Table I) analogously to Figs. 2D and 3E. Subsequently, as displayed in the schematic diagram, the gappy pair kernel function decomposes each CDR3 sequence into features made of two k-mers separated by a gap of maximal length *m*. Parameters maximizing classification accuracy were determined via cross-validation on the training set [$k = 3$, $m = 1$, $4(2 \times k) \times (m + 1) = 8192$ possible features per dimensionality]. Based on the feature decomposition, a feature matrix of dimension #CDR3s × #Features is constructed. Thus, each row of the feature matrix corresponds to a feature vector for a CDR3 and contains counts of each feature as it occurs in the CDR3 sequence. These feature vectors serve as the input to the linear SVM analysis. Results for amino acid–based classification are displayed in Supplemental Fig. 4A, 4B. (**B**) SVM-based prediction of human B cell (Dataset 5) and murine CD4 T cell (Dataset 6) public and private clones. Dataset preparation and SVM method (gappy-pair kernel) per parameter were identical to those used in (A). (**C**) SVM-based prediction [SVM method identical to (A)] of public and private clones when training the classifier on one respective sample *i* of each of the three datasets (Datasets 1, 5, 6) to predict public/private status of sequences from all other respective $n - 1$ samples of Datasets 1, 5, and 6. Thus, training and test SVM sets stem from entirely distinct individuals (cross-sample prediction). (**D**) Public clones were aggregated across mice by B/T cell populations (nBCs, CD4), strain (nBCs: C57BL/6, BALB/c, pet) or across B cell populations (human nBCs and memory B cells [Bmem]) to subsequently perform SVM-based classification, as described in (A) (Datasets 1, 2, 3, 5, and 6). Sizes of aggregated SVM datasets ranged between ~5×10^4 (CD4 T cell) and 3×10^6 (nBCs: C57BL/6, BALB/c, pet) clones. Receiver operating characteristic curves show excellent classification results across unrelated datasets with identical library preparation (AUC ~ 0.90). (**E**) SVM-based prediction of public versus private clones across experimental studies with different library preparations. nBC repertoires of Dataset 1 (mean size ~ 180,000 clones) were used to predict public and private clones in the B2 B cell repertoires of Dataset 4 (mean size ~ 2400 clones, Table I). Bar graphs show mean ± SEM across samples.

CDR3 subregion separately, as well as all 10 relevant combinations thereof (Fig. 6A). Classification based on each single or paired CDR3 subregion resulted in a BACC < 70% (Fig. 6A). Among the partial combinations, the N1–D–N2 subregion combination achieved maximum prediction accuracy (74%, Fig. 6A, Supplemental Fig. 4D), approaching that of the full combination (V–N1–D–N2–J, ~80%), indicating that the sequence composition between public and private clones differed most within N1–D–N2 subregions. J subregions contributed least to prediction accuracy, because V–N1–D (BACC ~ 73%) and N1–D–N2 (BACC ~ 73%) surpassed D–N2–J (BACC ~ 70%, Fig. 6A). To confirm that subregion differences between public and private clones were largely dictated by the N1, D, and N2 subregions and not by the overhang regions linking N1, D, and N2, we showed

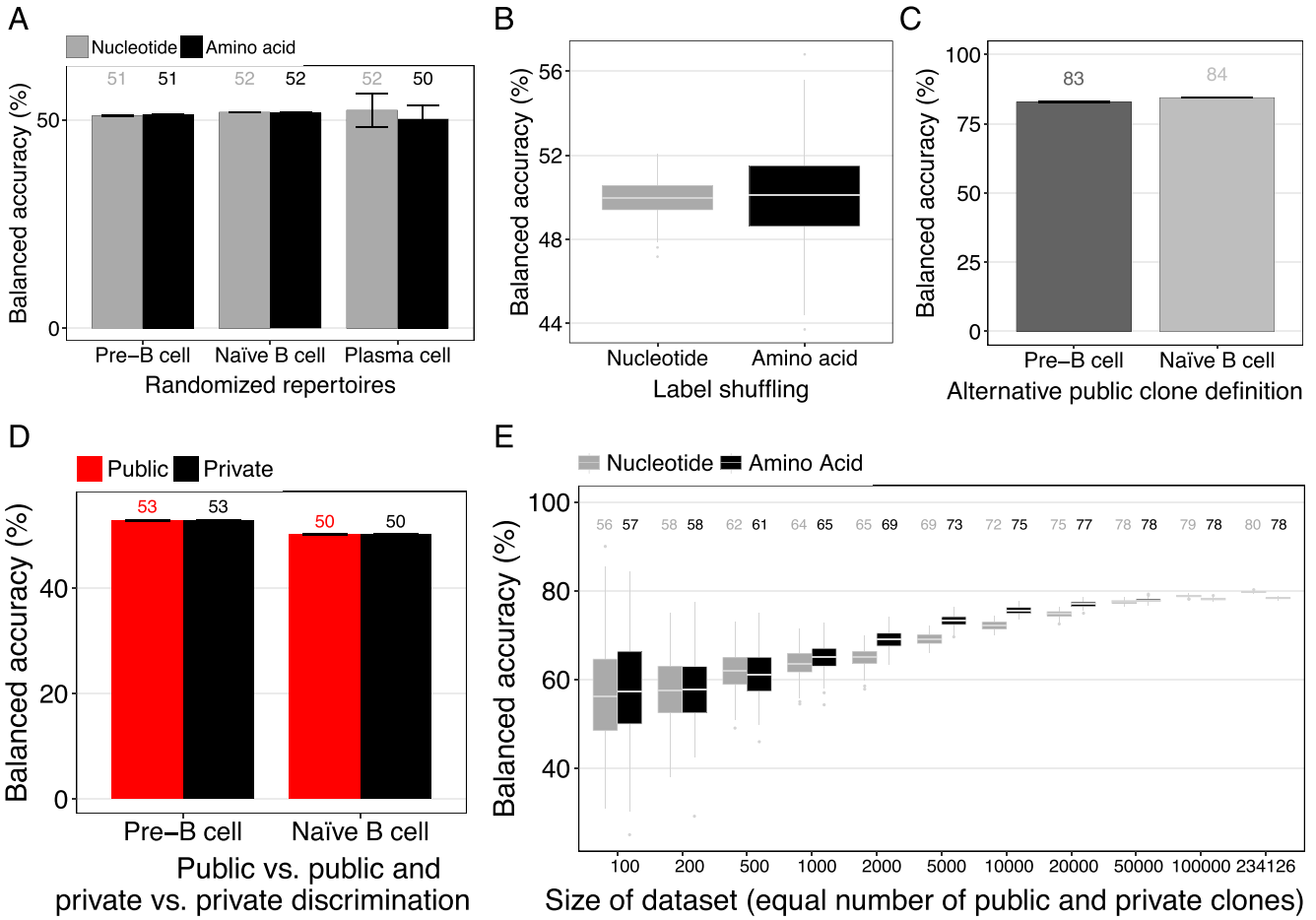

FIGURE 5. Validation of robustness of sequence-based SVM public/private clone classifier. (**A**) SVM-based discrimination of randomized public and private CDR3 sequences (Dataset 1). CDR3 sequences were randomized by nucleotide/amino acid shuffling. SVM was performed as described for Fig. 4A. (**B**) SVM-based discrimination performed on a nBC sample (murine, Dataset 1) of which the labels (public, private) were randomly shuffled (label shuffling). SVM was performed as described for Fig. 4A. (**C**) Validation that public/private clone BACC is independent of public clone definition (Dataset 1). In contrast to the default public clone definition (public: sharing among two individuals), public clones were defined as those clones that were shared among all mice of a given cohort and B cell population (size of CDR3 length–equilibrated SVM datasets: 4682 ± 657 [preBCs, mean ± SD], 28,249 ± 6,736 [nBCs]). Subsequently, SVM-based discrimination of public and private amino acid was carried out analogously to that described for Fig. 4A. (**D**) SVM-based discrimination of public versus public and private versus private clones of different samples (Dataset 1) to confirm that public and private clones of different individuals are indistinguishable from one another and, thus, possess a common sequence signature. SVM was performed as described for Fig. 4A. (**E**) Public and private clone prediction accuracy as a function of dataset size. From the largest murine nBC repertoire (Dataset 1), 100–234,126 CDR3 sequences were drawn randomly 100 times to subsequently perform SVM-based prediction of public and private clones (analogously to Fig. 4A). For each randomly drawn dataset, the number of public and private clones was kept equal. Bar graphs show mean ± SEM across samples.

that subregion shuffling impacted prediction accuracy only negligibly (Supplemental Fig. 4B). Furthermore, we confirmed that N1, D, and N2 subregions are the drivers of public and private clone discrimination by constructing prediction profiles, which quantify for each CDR3 sequence, in contrast to few selected features (30, 54), the contribution of each position to the decision value (public, private). Differences in contribution to the decision value were highest in the sequence positions belonging to the N1, D, and N2 subregions (Fig. 6B, 6D).

Next, we set out to answer the question whether both public and private repertoires possess class-specific (stereotypic; i.e., nonrandom) sequence signatures in the N1–D–N2 region because classification theoretically could be driven by dominant signals from one class. To this end, we posited that if both classes (public, private) contained class-specific sequence features, neither public nor private repertoires should be indistinguishable from their randomized counterparts (public versus public randomized, private versus private randomized; randomization was performed by nucleotide sequence shuffling of the N1–D–N2 region, V and J region were left nonrandomized). However, for all samples tested (nBCs, Dataset 1), public and private repertoires could be discriminated from randomized repertoires (prediction accuracy ≥ 90%, Fig. 6C). Thus, public and private repertoires contain nonrandom class-specific sequence features. In contrast, public-randomized versus private-randomized repertoires were nearly indistinguishable from one another (prediction accuracy = 63%, nonrandomized V and J still contain class-specific information leading to a prediction accuracy ≠ 50%).

To visualize prediction weight distribution from the above simulations, we again constructed prediction profiles (Fig. 6D). We found that classification involving randomized repertoires (columns 1–3, Fig. 6D) led to more even prediction weight distributions in the N1–D–N2 region of randomized repertoires, which, in addition, was mostly uncorrelated with the (non)randomized counterpart. In contrast, when classifying nonrandomized public and private repertoires (column 4, Fig. 6D), the distribution of prediction weights across subregions was skewed, correlated, and independent in magnitude with respect to the D region, which

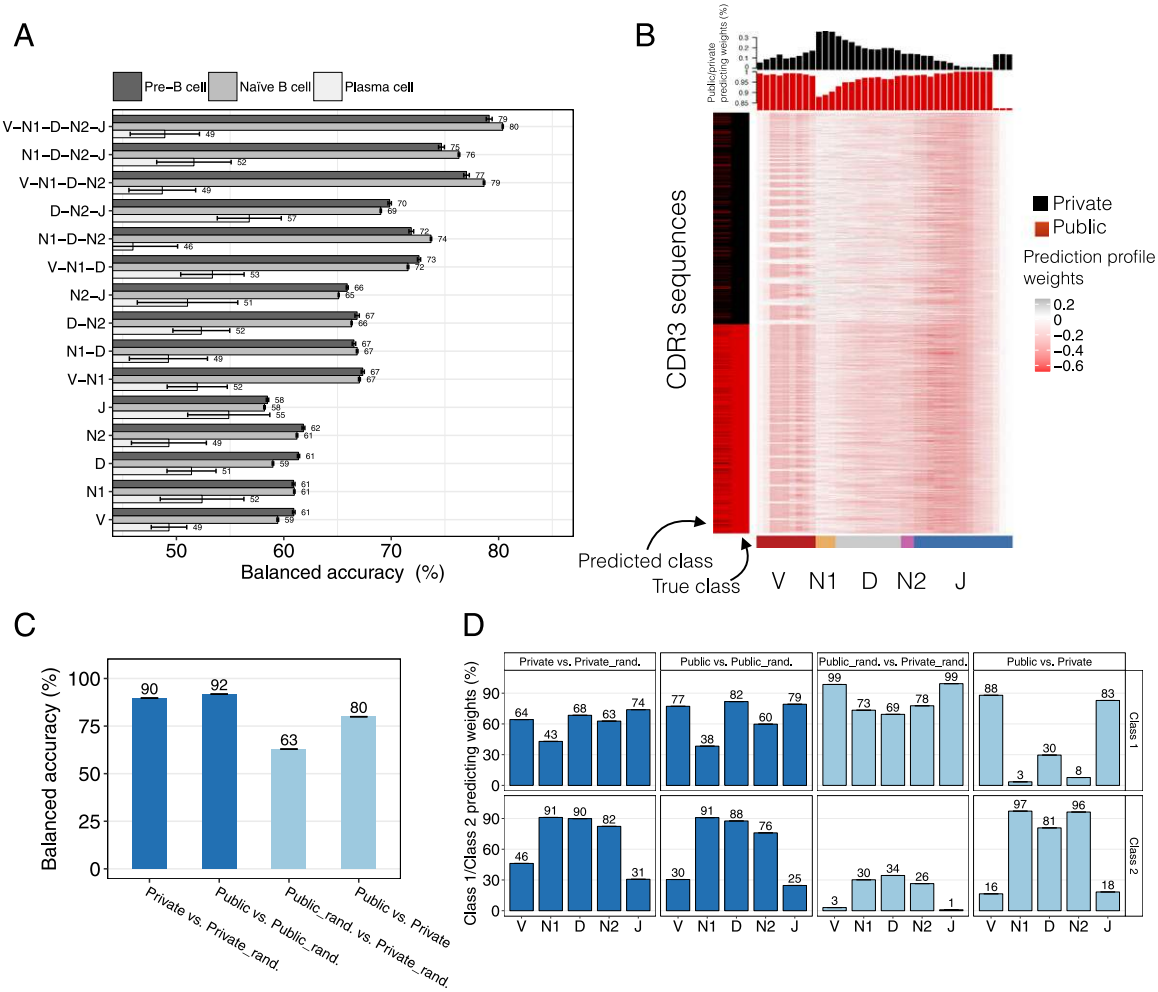

FIGURE 6. N1, D, and N2 CDR3 subregions dominate the public/private clone classification accuracy. (**A**) Public/private clone discrimination based on (combinations of) CDR3 subregions using sequence kernel-based SVM analysis of nucleotide sequences (Dataset 1). For each combination of CDR3 subregion, gappy pair kernel parameters (k, m, C) were determined by cross-validation. (**B**) Exemplary visualization of prediction profiles of one test dataset (nBC, Dataset 1) of CDR3s (rows) of length 39 nt. Prediction profiles were computed as means of feature weights at each CDR3 position (1–39 bp) and indicate the importance of each CDR3 sequence position/subregion for public/private clone classification (see *Materials and Methods*). Positions colored red (<0) count toward "public" prediction of the respective CDR3s, whereas black-colored ones (>0) bias prediction toward the "private" clone status. Bar graphs indicate the percentage of private (black) or public predicting weights at each of the 39 positions. Horizontal colored bars at the bottom indicate the median length of V (red), N1 (orange), D (gray), N2 (purple), and J (blue) subregions (see Fig. 3A). (**C**) SVM-based classification (nucleotide level) of various combinations of public and private repertoires and their randomized counterparts ("rand."). For all results shown, murine naive B cell repertoires (Dataset 1) were used. Nucleotide sequence randomization was performed as described for Fig. 5A (but only for N1, D, and N2; V and J CDR3 subregions were left nonrandomized), and SVM was performed as described for Fig. 4A. (**D**) For classification scenarios shown in (C), the percentage of class 1–predicting SVM weights for class 1 and the percentage of class 2–predicting weights for class 2 were determined by CDR3 subregion (V, N1, D, N2, J). For example, class 1 in columns 1 and 3 is Private and Public_rand., respectively. Bar graphs show mean ± SEM across samples.

makes up the largest part of the N1–D–N2 region (Fig. 3A). The higher N1, D, and N2 diversity of private repertoires (Fig. 3D) translates into a higher proportion of sequence positions required to define the private class (column 4, Fig. 6D). To summarize, our results indicate that the N1, D, and N2 subregions of public and private clone sequences contain class-specific stereotypic predictive signatures (accumulation of k-mers) that enable the prediction of their status (public, private).

Discussion

We have performed a comprehensive immunogenomic decomposition of public and private immune repertoires that led us to conclude that low-dimensional features (Figs. 2, 3), including CDR3 subregion length, germline gene usage, and amino acid usage, were insufficient in detecting the immunogenomic shift between public and private clonal repertoires. In contrast, a high-dimensional sequence decomposition (sequence kernel) approach could predict the public and private status of Ab clones with 80% accuracy. We excluded the possibility that high predictive performance was achieved as the result of trivially high sequence similarity among public clones by showing that public and private clones were of similar average similarity (Supplemental Figs. 1B, 3C–D). We validated the robustness of the sequence-based SVM approach across species and mouse strains, B and T cells, naive and Ag-experienced (somatically mutated) B cells, individuals, library-preparation methods, public clone definitions, and various simulation and randomization controls (Figs. 4–6, Supplemental Fig. 4).

The high computational scalability of our machine learning approach, tested with as many as 3×10^6 public and private clonal sequences (Fig. 4D), allowed us to establish that dataset size is decisive for achieving high prediction accuracy (34). In simula-

tions, prediction accuracy increased by ~25 percentage points when increasing the dataset size by four orders of magnitude from $\sim 10^{1-2}$ to $\sim 10^{5}$ clonal sequences (Fig. 5E). In experimental data, increasing training dataset size by one to two orders of magnitude (sequence data generated in a different laboratory using different experimental library-preparation methods) increased prediction accuracy by up to 7 percentage points, suggesting that large-scale cross-study detection of public clones is possible (Fig. 4E). Additionally, the higher prediction accuracy of human public and private memory B cell clones (Fig. 4B) suggested that the lower accuracy of PC (IgG) repertoires (Fig. 4A) may be due to small dataset size (Table I) rather than Ag-specific effects. In the future, it may be of interest to investigate the differences in naive and Ag-driven public clonal sequence signatures (17, 18, 55–62).

Because several definitions for public clones have recently been used, and a single accepted definition has not yet reached general agreement (5), we validated SVM analyses with two different definitions, which encompassed a definition range from lenient to stringent (Figs. 4A, 5C) (5, 23, 63). The fact that our SVM approach is robust to several public clone definitions suggests that the need for a consensus definition might be secondary. Nevertheless, once single-cell sequencing has reached the depth of bulk sequencing, we will be able to investigate to what extent paired chain information (H/C, α/β, γ/δ) influences public/private clone prediction (64–67).

Technologically, we speculate that the prediction accuracies reported in this article merely represent lower bounds; future studies that combine standardized (http://www.airr-community.org) advanced experimental and computational error correction methodologies (e.g., single-cell sequencing, unique molecular identifiers, replicate sequencing, construction of individual germline gene databases) (67–73), high sampling and sequencing depth (1), and novel sequence-based deep learning (neural networks) approaches accounting for long-range sequence interactions (74–77) may lead to even higher prediction accuracies.

Biologically, sequence kernel-based machine learning analysis revealed stereotypical and predictive high-dimensional immunogenomic composition biases (high-dimensional fingerprints) in the N1–D–N2 CDR3 subregions of public and private clones (Fig. 6). Although the relative size of the human CDR3 N1–D–N2 subregion is larger than that of mice [~65 (78) versus 42% in mice, Fig. 3A] (9, 26), identical feature space sizes (SVM parameters) for both species led to highly similar prediction accuracies (Fig. 4B). Thus, species-specific differences in clonal sequence length and diversity did not impact prediction accuracy. Not only is it remarkable that a feature space of dimension $< 10^{4}$ suffices for detecting subrepertoire clonal expansion, as well as Ag-driven changes in individuals of different immunological status as previously shown (30, 48, 54), it also provides ample flexibility for defining fingerprints that discriminate whole-repertoire properties (public, private) within a $>10^{13}$ dimensional space (9, 11). This may point to evolutionarily conserved traces in the immunogenome; indeed, we found that murine B cell public clones were enriched in natural Ab specificities (Supplemental Fig. 1D), which is in line with previous public T cell repertoire studies (18, 79).

Previous probabilistic work on modeling repertoire diversity revealed a broad range of clonal sequence-generation probabilities, with (B/T cell) public clones suggested to be biased toward higher generation probabilities (25). Corroborating these observations, we found that B cell public clones are more likely to have higher clonal abundance (Supplemental Fig. 1C). In general, however, public clones were distributed throughout the entire frequency spectrum from high to very low clonal frequency, suggesting that clonal frequency is not a reliable predictor of public status (53). Instead of attributing a generation probability to each clonal sequence, our work complements previous probabilistic work by leveraging a high-dimensional repertoire-level trained classifier for binary classification on a per-sequence basis. Thus, the unique advantage of sequence-based machine learning, as opposed to probabilistic approaches for inference of generation probabilities (11, 25, 26, 56), is the detection of predictive class-determining sequence signatures. Specifically, sequence composition–based machine learning led to the unexpected finding that private clones, which were thought to be mostly stochastically generated, also possess a high-dimensional fingerprint (predictive immunogenomic features [Figs. 1 and 6]).

To conclude, the existence of high-dimensional immunogenomic rules shaping immune repertoire diversity in a predictable fashion provides further insight into the hitherto insufficiently understood mechanisms of repertoire predetermination (17, 25, 26, 80, 81). Furthermore, we note that mouse and human trained SVM classifiers may be applied to experimental data, as well as to synthetic repertoire data (82), which could pave the way toward the construction of a comprehensive atlas of human and mouse public clones with possible applications in predictive immunotherapeutic targeting of clones that have desirable sequence features or occur often within a population (rational vaccine design) (17, 24, 53, 83, 84). Finally, we believe that our study represents a proof-of-principle for large-scale and high-dimensional machine learning on immune repertoire sequence data, serving as a guideline for important future studies, such as the dissection of public and private Ag-associated signatures (30, 60–62, 64, 85–88).

Acknowledgments

We thank Dr. Christian Beisel, Manuel Kohler, Ina Nissen, and Elodie Burcklen (Genomics Facility Basel, Eidgenössische Technische Hochschule Zürich) for their expert technical assistance with Illumina high-throughput sequencing and Sepp Hochreiter (Johannes Kepler University) for helpful discussions.

Disclosures

The authors have no financial conflicts of interest.

References

1. Greiff, V., E. Miho, U. Menzel, and S. T. Reddy. 2015. Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol.* 36: 738–749.
2. Hershberg, U., and E. T. Luning Prak. 2015. The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370: 20140239.
3. Xu, J. L., and M. M. Davis. 2000. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* 13: 37–45.
4. Kunik, V., B. Peters, and Y. Ofran. 2012. Structural consensus among antibodies defines the antigen binding site. *PLOS Comput. Biol.* 8: e1002388.
5. Castro, R., S. Navelsaker, A. Krasnov, L. Du Pasquier, and P. Boudinot. 2017. Describing the diversity of Ag specific receptors in vertebrates: contribution of repertoire deep sequencing. *Dev. Comp. Immunol.* 75: 28–37.
6. Davis, M. M., and P. J. Bjorkman. 1988. T-cell antigen receptor genes and T-cell recognition. [Published erratum appears in 1988 *Nature* 335: 744.] *Nature* 334: 395–402.
7. Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature* 302: 575–581.
8. Glanville, J., W. Zhai, J. Berka, D. Telman, G. Huerta, G. R. Mehta, I. Ni, L. Mei, P. D. Sundar, G. M. R. Day, 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.
9. Saada, R., M. Weinberger, G. Shahaf, and R. Mehr. 2007. Models for antigen receptor gene rearrangement: CDR3 length. *Immunol. Cell Biol.* 85: 323–332.
10. Warren, R. L., J. D. Freeman, T. Zeng, G. Choe, S. Munro, R. Moore, J. R. Webb, and R. A. Holt. 2011. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res.* 21: 790–797.
11. Murugan, A., T. Mora, A. M. Walczak, and C. G. Callan, Jr. 2012. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci. USA* 109: 16161–16166.

12. Arnaout, R., W. Lee, P. Cahill, T. Honan, T. Sparrow, M. Weiand, C. Nusbaum, K. Rajewsky, and S. B. Koralov. 2011. High-resolution description of antibody heavy-chain repertoires in humans. *PLoS One* 6: e22365.
13. Jiang, N., J. A. Weinstein, L. Penland, R. A. White, III, D. S. Fisher, and S. R. Quake. 2011. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proc. Natl. Acad. Sci. USA* 108: 5348–5353.
14. DeWitt, W. S., P. Lindau, T. M. Snyder, A. M. Sherwood, M. Vignali, C. S. Carlson, P. D. Greenberg, N. Duerkopp, R. O. Emerson, and H. S. Robins. 2016. A public database of memory and naive B-cell receptor sequences. *PLoS One* 11: e0160853.
15. Galson, J. D., J. Trück, A. Fowler, M. Münz, V. Cerundolo, A. J. Pollard, G. Lunter, and D. F. Kelly. 2015. In-depth assessment of within-individual and inter-individual variation in the B cell receptor repertoire. *Front. Immunol.* 6: 531.
16. Georgiou, G., G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, and S. R. Quake. 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* 32: 158–168.
17. Greiff, V., U. Menzel, E. Miho, C. Weber, R. Riedel, S. Cook, A. Valai, T. Lopes, A. Radbruch, T. H. Winkler, and S. T. Reddy. 2017. Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B cell development. *Cell Reports* 19: 1467–1478.
18. Madi, A., E. Shifrut, S. Reich-Zeliger, H. Gal, K. Best, W. Ndifon, B. Chain, I. R. Cohen, and N. Friedman. 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.
19. Robinson, W. H. 2015. Sequencing the functional antibody repertoire–diagnostic and therapeutic discovery. *Nat. Rev. Rheumatol.* 11: 171–182.
20. Yaari, G., and S. H. Kleinstein. 2015. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.* 7: 121.
21. Yang, Y., C. Wang, Q. Yang, A. B. Kantor, H. Chu, E. E. Ghosn, G. Qin, S. K. Mazmanian, J. Han, and L. A. Herzenberg. 2015. Distinct mechanisms define murine B cell lineage immunoglobulin heavy chain (IgH) repertoires. *eLife* 4: e09083.
22. Jackson, K. J., M. J. Kidd, Y. Wang, and A. M. Collins. 2013. The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor. *Front. Immunol.* 4: 263.
23. Covacu, R., H. Philip, M. Jaronen, J. Almeida, J. E. Kenison, S. Darko, C. C. Chao, G. Yaari, Y. Louzoun, L. Carmel, et al. 2016. System-wide analysis of the T cell response. *Cell Rep.* 14: 2733–2744.
24. Venturi, V., D. A. Price, D. C. Douek, and M. P. Davenport. 2008. The molecular basis for public T-cell responses? *Nat. Rev. Immunol.* 8: 231–238.
25. Elhanati, Y., A. Murugan, C. G. Callan, Jr., T. Mora, and A. M. Walczak. 2014. Quantifying selection in immune receptor repertoires. *Proc. Natl. Acad. Sci. USA* 111: 9875–9880.
26. Elhanati, Y., Z. Sethna, Q. Marcou, C. G. Callan, Jr., T. Mora, and A. M. Walczak. 2015. Inferring processes underlying B-cell repertoire diversity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370: 20140243.
27. Mora, T., A. M. Walczak, W. Bialek, and C. G. Callan, Jr. 2010. Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. USA* 107: 5405–5410.
28. Kidd, B. A., L. A. Peters, E. E. Schadt, and J. T. Dudley. 2014. Unifying immunology with informatics and multiscale biology. *Nat. Immunol.* 15: 118–127.
29. Lodhi, H., C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. 2002. Text classification using string kernels. *J. Mach. Learn. Res.* 2: 419–444.
30. Sun, Y., K. Best, M. Cinelli, J. M. Heather, S. Reich-Zeliger, E. Shifrut, N. Friedman, J. Shawe-Taylor, and B. Chain. 2017. Specificity, privacy, and degeneracy in the CD4 T cell receptor repertoire following immunization. *Front. Immunol.* 8: 430.
31. Palme, J., S. Hochreiter, and U. Bodenhofer. 2015. KeBABS: an R package for kernel-based analysis of biological sequences. *Bioinformatics* 31: 2574–2576.
32. Schwarzbauer, K., U. Bodenhofer, and S. Hochreiter. 2012. Genome-wide chromatin remodeling identified at GC-rich long nucleosome-free regions. *PLoS One* 7: e47924.
33. Bishop, C. M. 2007. *Pattern Recognition and Machine Learning.* Springer, Berlin.
34. Thomas, N., K. Best, M. Cinelli, S. Reich-Zeliger, H. Gal, E. Shifrut, A. Madi, N. Friedman, J. Shawe-Taylor, and B. Chain. 2014. Tracking global changes induced in the CD4 T cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics* 30: 3181–3188.
35. Bolotin, D. A., S. Poslavsky, I. Mitrophanov, M. Shugay, I. Z. Mamedov, E. V. Putintseva, and D. M. Chudakov. 2015. MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* 12: 380–381.
36. Greiff, V., U. Menzel, U. Haessler, S. C. Cook, S. Friedensohn, T. A. Khan, M. Pogson, I. Hellmann, and S. T. Reddy. 2014. Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC Immunol.* 15: 40.
37. Menzel, U., V. Greiff, T. A. Khan, U. Haessler, I. Hellmann, S. Friedensohn, S. C. Cook, M. Pogson, and S. T. Reddy. 2014. Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. *PLoS One* 9: e96727.
38. R. Development Core Team, 2009. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
39. Rossum, G. V., and F. L. Drake. 2011. *The Python Language Reference Manual.* Network Theory Ltd., Godalming, U.K.
40. Wickham, H. 2009. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag, New York.
41. Neuwirth, E. 2014. Package 'RColorBrewer'. Available at: https://cran.r-project.org/web/packages/RColorBrewer/RColorBrewer.pdf. Accessed: August 8, 2016.
42. Gu, Z. 2015. Making Complex Heatmaps. Available at: https://bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html. Accessed: February 27, 2016.
43. Bischl, B., M. Lang, O. Mersmann, J. Rahnenführer, and C. Weihs. 2015. BatchJobs and batchExperiments: abstraction mechanisms for using R in batch environments. *J. Stat. Softw.* 64: 1–25.
44. Revolution Analytics and S. Weston. 2014. doParallel: Foreach Parallel Adaptor for the "parallel" Package. Available at: http://CRAN.R-project.org/package=doParallel. Accessed: October 10, 2016.
45. Lefranc, M.-P., V. Giudicelli, C. Ginestoux, J. Bodmer, W. Müller, R. Bontrop, M. Lemaitre, A. Malik, V. Barbié, and D. Chaume. 1999. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.* 27: 209–212.
46. van der Loo, M. P. J. 2014. The stringdist package for approximate string matching. *R J.* 6: 111–122.
47. Li, S., M.-P. Lefranc, J. J. Miles, E. Alamyar, V. Giudicelli, P. Duroux, J. D. Freeman, V. D. Corbin, J.-P. Scheerlinck, M. A. Frohman, et al. 2013. IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat. Commun.* 4: 2333.
48. Greiff, V., P. Bhat, S. C. Cook, U. Menzel, W. Kang, and S. T. Reddy. 2015. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med.* 7: 49.
49. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12: 2825–2830.
50. Jiao, Y., and P. Du. 2016. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant. Biol.* 4: 320–330.
51. Leslie, C., and R. Kuang. 2004. Fast string kernels using inexact matching for protein sequences. *J. Mach. Learn. Res.* 5: 1435–1455.
52. Mahrenholz, C. C., I. G. Abfalter, U. Bodenhofer, R. Volkmer, and S. Hochreiter. 2011. Complex networks govern coiled-coil oligomerization–predicting and profiling by means of a machine learning approach. *Mol. Cell. Proteomics* 10: M110.004994.
53. Miho, E., V. Greiff, R. Roskar, and S. T. Reddy. 2017. The fundamental principles of antibody repertoire architecture revealed by large-scale network analysis. *bioRxiv.* DOI: 10.1101/124578.
54. Cinelli, M., Y. Sun, K. Best, J. M. Heather, S. Reich-Zeliger, E. Shifrut, N. Friedman, J. Shawe-Taylor, and B. Chain. 2017. Feature selection using a one dimensional naïve Bayes' classifier increases the accuracy of support vector machine classification of CDR3 repertoires. *Bioinformatics* 33: 951–955.
55. Callan, C. G. Jr., T. Mora, and A. M. Walczak. 2017. Repertoire sequencing and the statistical ensemble approach to adaptive immunity. *Curr. Opin. Syst. Biol.* 1: 44–47.
56. Marcou, Q., T. Mora, and A. M. Walczak. 2017. IGoR: a tool for high-throughput immune repertoire analysis. *arXiv.* 1705.08246. Available at: https://arxiv.org/abs/1705.08246. Accessed: May 25, 2017.
57. Calis, J. J., and B. R. Rosenberg. 2014. Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends Immunol.* 35: 581–590.
58. Strauli, N. B., and R. D. Hernandez. 2016. Statistical inference of a convergent antibody repertoire response to influenza vaccine. *Genome Med.* 8: 60.
59. Emerson, R. O., W. S. DeWitt, M. Vignali, J. Gravley, J. K. Hu, E. J. Osborne, C. Desmarais, M. Klinger, C. S. Carlson, J. A. Hansen, et al. 2017. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.* 49: 659–665.
60. Adaptive Immunity Group. 2017. *VDJdb: A Curated Database of T-Cell Receptors with Known Antigen Specificity.* Available at: https://zenodo.org/record/838663#.WZ7bxGPYmxo. Accessed: July 4, 2017.
61. Tickotsky, N., T. Sagiv, J. Prilusky, E. Shifrut, and N. Friedman. 2017. McPAS-TCR: A manually-curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*. DOI:10.1093/bioinformatics/btx286.
62. Parameswaran, P., Y. Liu, K. M. Roskin, K. K. Jackson, V. P. Dixit, J.Y. Lee, K. L. Artiles, S. Zompi, M. J. Vargas, B. B. Simen, et al. 2013. Convergent antibody signatures in human dengue. *Cell Host Microbe* 13: 691–700.
63. Li, H., C. Ye, G. Ji, X. Wu, Z. Xiang, Y. Li, Y. Cao, X. Liu, D. C. Douek, D. A. Price, and J. Han. 2012. Recombinatorial biases and convergent recombination determine interindividual TCRβ sharing in murine thymocytes. *J. Immunol.* 189: 2404–2413.
64. Dash, P., A. J. Fiore-Gartland, T. Hertz, G. C. Wang, S. Sharma, A. Souquette, J. C. Crawford, E. B. Clemens, T. H. Nguyen, K. Kedzierska, et al. 2017. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 547: 89–93.
65. Howie, B., A. M. Sherwood, A. D. Berkebile, J. Berka, R. O. Emerson, D. W. Williamson, I. Kirsch, M. Vignali, M. J. Rieder, C. S. Carlson, and H. S. Robins. 2015. High-throughput pairing of T cell receptor α and βαsequences. *Sci. Transl. Med.* 7: 301ra131.
66. DeKosky, B. J., G. C. Ippolito, R. P. Deschner, J. J. Lavinder, Y. Wine, B. M. Rawlings, N. Varadarajan, C. Giesecke, T. Dörner, S. F. Andrews, et al. 2013. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* 31: 166–169.
67. Friedensohn, S., T. A. Khan, and S. T. Reddy. 2017. Advanced methodologies in high-throughput sequencing of immune repertoires. *Trends Biotechnol.* 35: 203–214.
68. Khan, T. A., S. Friedensohn, A. R. Gorter de Vries, J. Straszewski, H.J. Ruscheweyh, and S. T. Reddy. 2016. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci. Adv.* 2: e1501371.

69. Vollmers, C., R. V. Sit, J. A. Weinstein, C. L. Dekker, and S. R. Quake. 2013. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc. Natl. Acad. Sci. USA* 110: 13463–13468.
70. Shugay, M., O. V. Britanova, E. M. Merzlyak, M. A. Turchaninova, I. Z. Mamedov, T. R. Tuganbaev, D. A. Bolotin, D. B. Staroverov, E. V. Putintseva, K. Plevova, et al. 2014. Towards error-free profiling of immune repertoires. *Nat. Methods* 11: 653–655.
71. Wardemann, H., and C. E. Busse. 2017. Novel approaches to analyze immunoglobulin repertoires. *Trends Immunol.* 38: 471–482.
72. Corcoran, M. M., G. E. Phad, N. Vázquez Bernat, C. Stahl-Hennig, N. Sumida, M. A. Persson, M. Martin, and G. B. Karlsson Hedestam. 2016. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat. Commun.* 7: 13642.
73. Watson, C. T., J. Glanville, and W. A. Marasco. 2017. The individual and population genetics of antibody immunity. *Trends Immunol.* 38: 459–470. doi.10.1016/j.it.2017.04.003.
74. Hochreiter, S., and J. Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9: 1735–1780.
75. Angermueller, C., T. Pärnamaa, L. Parts, and O. Stegle. 2016. Deep learning for computational biology. *Mol. Syst. Biol.* 12: 878.
76. Alipanahi, B., A. Delong, M. T. Weirauch, and B. J. Frey. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33: 831–838.
77. Ching, T., D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, W. Xie, G. L. Rosen, et al. 2017. Opportunities and obstacles for deep learning in biology and medicine. *bioRxiv* DOI: 10.1101/142760.
78. Mroczek, E. S., G. C. Ippolito, T. Rogosch, K. H. Hoi, T. A. Hwangpo, M. G. Brand, Y. Zhuang, C. R. Liu, D. A. Schneider, M. Zemlin, et al. 2014. Differences in the composition of the human antibody repertoire by B cell subsets in the blood. *Front. Immunol.* 5: 96.
79. Madi, A., A. Poran, E. Shifrut, S. Reich-Zeliger, E. Greenstein, I. Zaretsky, T. Arnon, F. V. Laethem, A. Singer, J. Lu, et al. 2017. T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. *eLife* 6: e22057.
80. Rubelt, F., C. R. Bolen, H. M. McGuire, J. A. Vander Heiden, D. Gadala-Maria, M. Levin, G. M. Euskirchen, M. R. Mamedov, G. E. Swan, C. L. Dekker, et al. 2016. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naive and antigen-experienced cells. *Nat. Commun.* 7: 11112.
81. Glanville, J., T. C. Kuo, H.-C. von Büdingen, L. Guey, J. Berka, P. D. Sundar, G. Huerta, G. R. Mehta, J. R. Oksenberg, S. L. Hauser, et al. 2011. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl. Acad. Sci. USA* 108: 20066–20071.
82. Safonova, Y., A. Lapidus, and J. Lill. 2015. IgSimulator: a versatile immunosequencing simulator. *Bioinformatics* 31: 3213–3215.
83. Miles, J. J., S. L. Silins, and S. R. Burrows. 2006. Engineered T cell receptors and their potential in molecular medicine. *Curr. Med. Chem.* 13: 2725–2736.
84. Jardine, J. G., D. W. Kulp, C. Havenar-Daughton, A. Sarkar, B. Briney, D. Sok, F. Sesterhenn, J. Ereño-Orbea, O. Kalyuzhniy, I. Deresa, et al. 2016. HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen. *Science* 351: 1458–1463.
85. Glanville, J., H. Huang, A. Nau, O. Hatton, L. E. Wagar, F. Rubelt, X. Ji, A. Han, S. M. Krams, C. Pettus, et al. 2017. Identifying specificity groups in the T cell receptor repertoire. *Nature* 547: 94–98.
86. Boyd, S. D., and J. E. Crowe, Jr. 2016. Deep sequencing and human antibody repertoire analysis. *Curr. Opin. Immunol.* 40: 103–109.
87. Buerckert, J.-P., A. R. Dubois, W. J. Faison, S. Farinelle, E. Charpentier, R. Sinner, A. Wienecke-Baldacchino, and C. P. Muller. 2017. Functionally convergent B cell receptor sequences in transgenic rats expressing a human B cell repertoire in response to tetanus toxoid and measles antigens. *bioRxiv*. DOI: 10.1101/159368.
88. Davis, M. M., C. M. Tato, and D. Furman. 2017. Systems immunology: just getting started. *Nat. Immunol.* 18: 725–732.