Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires

Authors:

Victor Greiff[#1], Cédric R. Weber[#1], Johannes Palme[2,3], Ulrich Bodenhofer[2], Enkelejda Miho[1], Ulrike Menzel[1], Sai T. Reddy*[1]

Affiliations:

[1] Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland

[2] Institute of Bioinformatics, Johannes Kepler University, Linz, Austria

[3] Health & Environment Department, Molecular Diagnostics, AIT – Austrian Institute of Technology, Vienna, Austria

[#]These authors contributed equally to this study

*Correspondence to: sai.reddy@ethz.ch

Abstract

Recent studies have revealed that immune repertoires contain a substantial fraction of public clones, which are defined as antibody or T-cell receptor (TCR) clonal sequences shared across individuals. As of yet, it has remained unclear whether public clones possess predictable sequence features that separate 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 the complementarity determining region 3, CDR3), we detected predictive public- and private-clone-specific immunogenomic differences concentrated in the CDR3's N1-D-N2 region, which allowed the prediction of public and private status with 80% accuracy in both humans and mice. Our results unexpectedly demonstrate that not only public but also 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 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 towards the construction of a comprehensive atlas of public clones in immune repertoires, which may have applications in rational vaccine design and immunotherapeutics.

Introduction

The clonal identity, specificity, and diversity of adaptive immune receptors is largely defined by the sequence of complementarity determining region 3 (CDR3) of variable heavy (V_H) and variable beta (V_β) chains of antibodies and TCRs, respectively [1–5]. The CDR3 encompasses the junction region of recombined V-, D-, J-gene segments as well as non-templated nucleotide (n, p) addition [6]. Due to the enormous theoretical diversity of antibody and TCR repertoires ($>10^{13}$) [7–10] and technological limitations (Sanger sequencing), it was long believed that clonal repertoires were to an overwhelming extent private to each individual [11,12]. However, due to recent advances in high-throughput immune repertoire sequencing, it has been observed that a considerable fraction (>1%) of CDR3s are shared across individuals [1,5,13–26]. 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 antibody and TCR repertoires, the large theoretical clonal (CDR3) diversity renders the investigation of public and private repertoires computationally challenging [27]. 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) [17,23]. In order to investigate the composition of large numbers of sequences with the appropriate dimensionality, sequence kernels are increasingly used [28,29]. Sequence kernels are high-dimensional functions which measure the similarity of pairs of sequences, for example, by comparing the occurrence of specific subsequences (k-mers) in a high-dimensional space [30,31]. Supervised machine learning (e.g., support vector machine analysis) is an approach, which takes low and 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 vs. private clones) [32]. In contrast to using conventional low-dimensional features to analyze immune repertoires, the coupling of high-dimensional sequence kernels to support vector machine (SVM) analysis may lead to greater insight into the immunogenomic structure of repertoire diversity; specifically the difference between public and private repertoires. As opposed to previous approaches [33], a key advantage of sequence-kernel based SVM analysis is the prediction-profile-based identification of CDR3 subregions

68   that are most predictive for a respective class (public or private class) [30,31]. This approach may lead to

69   predictive immunological and mechanistic insight into the immunogenomic elements that shape repertoire

70   diversity.

71   In order to identify the immunogenomic differences between public and private antibody repertoires

72   (Figure 1), we applied support vector machine analysis (Figure 1B) to six large-scale immune repertoire

73   (antibody and TCR) sequencing datasets from mice and humans (Figure 1A). When using low-

74   dimensional features (germline gene and amino acid usage, CDR3 subregion length) as the input for

75   SVM analysis, prediction accuracy of private and public status reached maximally 66%, which only

76   slightly improves on a random classifier (50%). However, when implementing a high-dimensional

77   sequence-kernel (sequence composition) based support vector machine analysis, we were able to detect

78   strong immunogenomic differences concentrated in the N1-D-N2 region in public and private clones, with

79   a high prediction accuracy (balanced accuracy≈79–83%, Figure 1C). Our results unexpectedly signify that

80   both public *and* private antibody repertoires contain predictive high-dimensional features that enable their

81   accurate classification. Our SVM approach was sufficiently robust to be applied across repertoire studies

82   with different library preparations and high-throughput sequencing protocols demonstrating their

83   widespread applicability.

84   **Results**

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

86   As the basis for elucidating the immunogenomic differences between public and private clones, we used

87   a recently published high-throughput sequencing antibody repertoire dataset [16] (Dataset 1, *Methods*).

88   This dataset contains ~200 million full-length antibody V_H sequences derived from 19 different mice,

89   stratified into key stages of B-cell differentiation: pre-B cells (preBC, IgM), naïve B cells (nBC, IgM), and

90   plasma cells (PC, IgG). This dataset thus provided the important advantages of both high sequencing and

91   biological depth (preBC and nBC represent antigen-inexperienced cells, while PC are post-clonal

92   selection and expansion due to antigen exposure). Public clones, precisely defined here as CDR3

93   sequences (100% amino acid identity) occurring in at least two mice, were found to compose on average

94   15% (preBC), 23% (nBC), and 26% (PC) of antibody repertoires across B-cell stages (Figure 2A). As

95   previously reported, we found that public clones are both biased to higher frequencies and are enriched in

96   sequences from natural antibodies (Supplementary Figure 10) [17,24,34]. Throughout B-cell

3

97   development, public and private clones used nearly identical V, D, J, VJ and VDJ germline genes

98   (overlap >95%), which were at nearly identical frequencies in preBC and nBC (Spearman r≈1) and at

99   varied frequencies in PC (Spearman r>0.5–0.8) (Figure 2B). Thus, neither public nor private clones

100 showed any preferential germline gene usage. On average as well as at each CDR3 sequence position,

101 higher frequency amino acids occurred more often in public clones (e.g.: A, C, D), whereas lower

102 frequency amino acids could be found at higher frequency in private clones (e.g.: H, I, K) (Figure 2C,

103 Supplementary Figure 2B). This observation held true across all B-cell stages (r = 0.5–0.76; p<0.05,

104 Supplementary Figure 1A). Repertoire-wide absolute differences in amino acid usage between private

105 and public clones were slight (0.2–1.4 percentage points, Figure 2C). To test whether these repertoire-

106 level differences were sufficient to predictively discriminate between public and private clones on a single

107 clone level, we employed supervised support vector machine learning (SVM) analysis (*Methods*, Figure

108 1B). For all SVM analyses in this study, in order to minimize classification bias, a dataset was constructed

109 for each repertoire, which consisted of all public clones and an equal number of private clones from the

110 repertoire (Supplementary Table 1) such that both public and private clones had identical CDR3 length

111 distributions. Subsequently, the dataset constructed for SVM analyses was divided into 80% training

112 sequences and 20% test sequences (Figure 1B, *Methods*). We found that amino acid usage was a

113 suboptimal predictor of clonal status with a prediction accuracy ≤ 65% (Figure 2D) where prediction

114 accuracy is defined as the mean (balanced accuracy) of specificity and sensitivity (see *Methods*), as

115 described previously [30,35].

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

117 Since public and private clones did not differ in germline gene usage, we asked whether they differed with

118 respect to length and diversity of CDR3 subregions (V, N1, D, N2, J). The V, D and J subregions are

119 derived from germline genes (IGHV, IGHD, IGHJ), while N1 and N2 represent the insertions introduced

120 during the junctional recombination process (n- and p-nucleotides). Public clones in preBC and nBC

121 repertoires possessed a relative V subregion length of 23–24% (Figure 3A), whereas private clones had

122 slightly shorter V subregions (≈21%, p<0.05, Supplementary Figure 3A). The J subregion length behaved

123 analogously (public: 40%, private: 36%) while the D subregion length did not differ between groups

124 (public: 25%, private: 25%). We observed the largest difference between public and private clones in the

125 relative length of N1 and N2 subregions with deviations of 36–46 percentage points from a 1:1 ratio (N1:

4

126   public ≈6.5%, private ≈8.2%; N2: public ≈4.3%, private: ≈7.7%, p<0.05, Figure 3A, Supplementary

127   Figures 3A, B). Conversely, PC CDR3 subregion lengths did not differ between public and private clones

128   (with the exception of N1, which was slightly longer in public clones, Figure 3A, Supplementary Figure

129   3B).

130   Regardless of public or private designation, nearly all CDR3s (>94%) had at least one nucleotide insertion

131   (N1 or N2) and at least one deletion (Figure 3B), thus only a very small portion of clones were "germline-

132   like" having neither insertion nor deletion (≤4%, Supplementary Figure 4C,D). Furthermore, across B-cell

133   populations both N1 and N2 insertions were present in >50% and >70% of public and private clones,

134   respectively. Of note, N1 and N2 insertions showed no preferential selection of germline gene segments

135   (IGHV, D, J) (Supplementary Figure 3D) and the mean length of the sum of insertions (N1+N2) did not

136   correlate with V-D-J frequencies (Supplementary Figure 5A, Pearson r=0).

137   Deletion length was highest in D subregions (mean of 5' and 3' D-deletions: ≈7 nt, Supplementary Figure

138   3C) whereas it was lowest in V subregions (≈0.8 nt, Supplementary Figure 3C). Although private clones

139   showed a higher number of deletions, differences between public and private clones were slight (max

140   difference≈0.6 nt, Supplementary Figure 3C). Of interest, we were unable to detect an association

141   between the lengths of insertions and deletions (Supplementary Figure 5B).

142   Although differences in CDR3 subregion length and occurrence of insertions and deletions were

143   significant in preBC and nBC (Figure 3A, Supplementary Figures 3A–C, 4C, D, p<0.05), training a SVM

144   based on CDR3 subregion length, led to low prediction accuracy of public/private clone discrimination

145   (balanced accuracy ≤ 68%, Figure 3E). This indicates that the slight differences observed in CDR3

146   subregion length on the repertoire level are not reliable for class prediction.

147   **Public and private clones show differences in sequence composition**

148   Since *low-dimensional* features (CDR3 a.a. and subregion properties) did not achieve high discrimination

149   accuracy between public and private clones (Figures 2D, 3E), we investigated whether CDR3 sequence

150   composition (*potential dimensionality*: >10^{13} different CDR3 sequences) differed between public and

151   private clones. In preBC and nBC, V and J subregions neither differed in public and private clones with

152   regard to unique sequences (>97%) nor frequency thereof (Spearman r>0.95, Figure 3C). Consequently,

153   we observed no differences in V and J subregion diversity (number of unique V and J subregions)

154   between public and private clones (Figure 3D). Although there was a major difference in diversity of N1,

155   D, and N2 subregions between private and public repertoires, as the number of private preBC and nBC

156   clones surpassed that of public clones by 1.6–4.5-fold (Figure 3D, p<0.05, size adjusted, see also

157   Supplementary Figure 3A), N1, D, N2 subregion overlap between public and private clones was >66%

158   (Figure 3C). In PC repertoires, diversity differences between public and private repertoires were minimal

159   but overlap of subregions reached maximally 46% and Spearman correlation was consistently negative.

160   In contrast to single subregions, combinations of subregions showed low overlap between public and

161   private repertoires irrespective of B-cell population (e.g., N1-D-N2 overlap in nBC was ≈6%, Figure 3C),

162   which is explained by a large combinatorial diversity (Supplementary Figure 4B, Supplementary Table 2)

163   of CDR3 subregions. Thus, sequence composition differed substantially between public and private

164   clones.

165   **High-dimensional CDR3 sequence composition analysis predicts public and private clones with**

166   **high accuracy**

167   In order to test, whether the detected differences in sequence composition were predictive, we utilized

168   high-dimensional sequence kernels for SVM analysis [30]. We used the gappy-pair sequence kernel

169   [30,36,37], which decomposes each CDR3 into subsequences of length k (k-mers) separated by a gap of

170   length m (Figure 4A, see *Methods*). Applying this kernel function to all CDR3s of a given training dataset

171   generates a feature matrix of dimension $n*f$, which serves as input for the SVM analysis: here, n is the

172   number of CDR3s in the training dataset and f the number of features. By cross-validation, we selected

173   the parameter combinations that resulted in the highest prediction accuracy: k=3, m=1 at the nucleotide

174   level (potential feature diversity: 8192, *Methods*) and k=1, m=1 at the amino acid level (potential feature

175   diversity: 800, *Methods*). On both the nucleotide and the amino acid level, public and private clones in

176   preBC and nBC could be classified with ≈80% accuracy, with very low variation across mice (Figure 4A,

177   Supplementary Figure 6A, E). In order to validate the robustness of the chosen public clone definition, we

178   showed that the SVM was *incapable* of separating public from public and private from private clones

179   across individuals (balanced accuracy < 50%, Supplementary Figure 6D). In addition, we validated that

180   the high prediction accuracy was maintained for an alternative and more stringent definition for public

181   clones (balanced accuracy = 83–84%, Supplementary Figure 6F). In order to quantify the statistical

182   significance of our high-dimensional SVM approach, we confirmed that the balanced accuracy was close

183   to random (50%) when shuffling CDR3 nucleotide and amino acid sequences (Supplementary Figure 6B)

184   and when shuffling public and private labels across clones (Supplementary Figure 6C).

185   Furthermore, we confirmed that the differences in immunogenomic composition between public and

186   private clones were not exclusively mouse-strain-specific (C57BL/6); we replicated a balanced accuracy

187   of ≈80% with repertoires from BALB/c and pet shop mice (Datasets 2, 3, Supplementary Figure 6A).

188   Analogously, public and private clones could be discriminated with >80% accuracy in human B-cell

189   repertoires (Figure 4B, Dataset 5). Finally, we showed that our approach also demonstrated high

190   classification accuracy between public and private clones of mouse TCR V_β repertoires (balanced

191   accuracy = 74%, Figure 4B, Dataset 6).

192   Successful classification within each individual (mouse or human) proved that fundamental and

193   stereotypical differences between public and private classes do indeed exist. However, theoretically,

194   these differences could be specific to each individual and not generalizable. In order to exclude this

195   possibility, we accumulated public and private clones across individuals into datasets of up to 3×10^6

196   unique clonal sequences and showed that classification accuracy was maintained (Figure 4C), reaching a

197   maximum in human naïve and memory B cells (balanced accuracy = 83%, AUC [area under the ROC

198   curve] = 0.90). These results signified that the same set of features used to predict public and private

199   clones within one individual is sufficient for prediction across individuals of the same species. Thus, the

200   high-dimensional features provided by sequence kernels (800 for amino acid and 8192 for nucleotide)

201   and learned on the repertoire level, were sufficient and generalizable to discriminate public from private

202   clones in both humans and mice on a per clonal sequence basis (single clone resolution).

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

205   Our high-dimensional sequence-composition-based SVM approach was unable to predict public and

206   private clones in PCs (balanced accuracy = 50%, Figure 4A, Dataset 1). With respect to unique CDR3s,

207   the PC SVM-dataset was 3 to 4 orders of magnitude smaller than that of preBC and nBC (Supplementary

208   Table 1, Dataset 1); therefore we tested whether the low accuracy was due to sample size. We performed

209   SVM analysis on datasets ranging in size from 100 to 230'000 unique CDR3 sequences (Supplementary

210   Figure 7B) and found that prediction accuracy was indeed a function of sample size, increasing from 56%

211   (100 clonal sequences) to 80% (230'000 clonal sequences). Thus, small sample size may explain the

7

212 lower prediction accuracies observed in the PC (IgG) dataset. In further support of this hypothesis, we

213 found that in a dataset of human memory B-cells (mixed IgM, IgG) (Dataset 5) that was 3 orders of

214 magnitudes larger than the PC dataset, we were able to achieve >80% accuracy (Figure 4C), suggesting

215 that prediction of public clones may also be possible for antigen experienced B-cell populations and is

216 thus not limited to antigen-inexperienced ones.

217 Since we observed that dataset size was important for reaching high prediction accuracy (Supplementary

218 Figure 7B), we asked whether cross-dataset meta-analysis, which leverages large datasets as training

219 datasets for performing public and private clone prediction in other (smaller) datasets obtained from

220 studies using slightly different library preparation and high-throughput sequencing protocols. To answer

221 this question, we investigated the prediction accuracy of the sequence-composition-based SVM classifier

222 trained on Dataset 1 (nBC B2-B-cell population), applied to a test dataset 100 times smaller (177'197 vs

223 1519 sequences), consisting of repertoires from various C57BL/6 B2-B-cell populations [20] (Dataset 4,

224 Supplementary Table 1). By using the model based on the larger dataset (Dataset 1), prediction accuracy

225 could be improved by up to 7 percentage points (76%–77% vs. 69%–73%, Figure 4D), which neared the

226 prediction accuracy *within* Dataset 1 (Figure 4A). Thus, sequence-kernel-based SVM models can be

227 effectively trained on large datasets (openly accessible) enabling robust predictive performance for meta-

228 analysis *across* studies.

229 **Stereotypical immunogenomic differences between public and private clones are concentrated in**

230 **the N1-D-N2 subregions**

231 To identify the subregions that contributed most to classification accuracy, we performed sequence-

232 kernel-based SVM on each CDR3 subregion separately as well as all ten relevant combinations thereof

233 (Figure 5A). Classification based on each single or paired CDR3 subregions did not result in high

234 prediction accuracy (balanced accuracy ≤ 67%, Figure 5A). Among the partial combinations, it was the

235 N1-D-N2 subregion combination that achieved maximum prediction accuracy (74%, Figure 5A,

236 Supplementary Figure 7A) approaching that of the full combination (V-N1-D-N2-J, ≈80%), indicating that

237 the sequence composition between public and private clones differed most within N1-D-N2 subregions. J

238 subregions contributed least to prediction accuracy as V-N1-D (balanced accuracy≈73%) and N1-D-N2

239 (balanced accuracy≈73%) surpassed D-N2-J (balanced accuracy≈70%, Figure 5A). In order to confirm

240 that subregion differences between public and private clones were largely dictated by the N1, D and N2

241 subregions and not within the overhang regions linking N1, D, and N2, we showed that subregion

242 shuffling impacted prediction accuracy only negligibly (Supplementary Figure 6E). Visually and

243 numerically, we confirmed the N1, D, and N2 subregions to be the drivers of public and private clone

244 discrimination by constructing prediction profiles, which quantify for each sequence the contribution of

245 each position to the decision value (public, private). Differences in contribution to the decision value were

246 highest in the sequence positions belonging to the N1, D, and N2 subregions (Figure 5B, Supplementary

247 Figure 9). To summarize, our results indicate that the N1, D, N2 subregions of both public and private

248 clone sequences contain class-specific predictive subsequences (k-mers) that enable the prediction of

249 their status (public, private) with high accuracy.

Discussion

251 We have performed a comprehensive immunogenomic decomposition of immune repertoires, which led

252 us to conclude that *low-dimensional* features (Figures 2, 3, S1, S3–5) – CDR3 subregion length, germline

253 gene usage, amino acid usage (Figures 2D, 3E) – were insufficient in detecting the immunogenomic shift

254 between public and private clonal repertoires. In contrast, a *high-dimensional* sequence composition

255 (sequence-kernel) approach could predict the public and private status of antibody clones within any

256 individual with 80% accuracy. This CDR3 sequence-composition-based approach was generalizable

257 across individuals, B-cell populations, mouse strains, species (mouse, human), immune cell types (B-cell,

258 T-cell), and datasets produced in different laboratories (Figures 4B–D). While the appropriate definition of

259 "public" clones is subject to current debate [5], the public clone definition adopted in this study has been

260 used previously [5,22,38], and is the most lenient one possible. In fact, we showed that prediction

261 accuracy only increased when increasing the stringency of the public clone definition (Supplementary

262 Figure 6F). The fact that our SVM approach is robust to several public clone definitions, suggests there

263 may not be the need for a consensus definition.

264 Sequence-kernel based machine learning analysis revealed stereotypical and predictive high-dimensional

265 immunogenomic CDR3 subregion (N1, D, N2) composition biases (high-dimensional fingerprints) specific

266 to both public and private clones, respectively (Figure 5). Those fingerprints achieved up to 100%

267 prediction accuracy when isolated from V and J regions (Supplementary Figure 7A). Shuffling CDR3

268 subregions (V, N1, D, N2, J) impacted prediction accuracy only negligibly (Figure 5B, Supplementary

269 Figures 6E, 9), confirming that N1, D, and N2 held the highest amount of class-specific information

270 [25,39]. Of note, although the relative size of the human CDR3 N1-D-N2 subregion is larger than that of

271 mice (≈65% [40] vs. 42% in mice, Figure 3A) with the N1-D-N2 subregion being the main amplifier of

272 sequence diversity (Supplementary Table 2) [8,25], identical feature space sizes led to identical prediction

273 accuracies for both species (Figure 4B). Thus, potential species-specific differences in sequence length

274 and diversity did not impact the prediction accuracy of our approach. More generally, it is remarkable that

275 feature spaces of dimension $<10^4$ do not only suffice for detecting *sub*-repertoire clonal expansion-driven

276 changes in individuals of different immunological status [29,35] but also provide ample combinatorial

277 flexibility in defining fingerprints that discriminate *whole*-repertoire properties (public, private) within a

278 $>10^{13}$-dimensional space (Supplementary Table 2, [8,10]). This may point towards evolutionarily

279 conserved traces in the immunogenome; for example, we found that murine public clones were enriched

280 in natural antibody specificities (Supplementary Figure 10B).

281 Our results indicate that statistical significance does not necessarily translate into predictive performance:

282 although CDR3 subregion length differed significantly between public and private clones (Figure 3A,

283 Supplementary Figures 3A–C, 4C, D), the prediction accuracy of the low-dimensional SVM model based

284 on CDR3 subregion length (Figure 3D) remained inferior to the high-dimensional one based on the actual

285 sequence composition (Figure 4A). Furthermore, previous probabilistic work on modeling repertoire

286 diversity indicated a broad range in clonal sequence generation probabilities – with (T cell) public clones

287 suggested to be biased towards higher generation probabilities [24]. Corroborating these observations,

288 we found that B cell public clones are more likely to have higher clonal abundance (Supplementary Figure

289 10A) – in general, however, public clones were distributed throughout the entire frequency spectrum from

290 high to very low clonal frequency (Supplementary Figure 10A) [34]. Instead of attributing to each clonal

291 sequence a generation probability, our work complements previous *probabilistic* work by leveraging a

292 high-dimensional repertoire-level trained classifier for *binary classification* on a *per sequence basis*. It is

293 this sequence-composition-based machine learning approach that led to the unexpected finding that also

294 *private* clones – which were thought to be mostly stochastically generated – possess a high-dimensional

295 fingerprint (predictive immunogenomic features).

296 Our SVM-driven approach enables rapid and accurate separation of large repertoire datasets into public

297 and private repertoires. We note that mouse and human trained SVM-classifiers may not only be applied

298 to experimental but also to synthetic repertoire data [41], which could pave the way towards the

299 construction of a comprehensive atlas of human and mouse public clones. The high computational

300 scalability of our machine learning approach – tested with as many as 3×10^6 public and private sequences

301 (Figure 4C) – allowed us to establish that the dataset size is a deciding factor for high prediction accuracy

302 [33]: (i) in simulations, prediction accuracy increased by ≈30 percentage points when increasing the

303 dataset size by 4 orders of magnitude from ≈$10^{1–2}$ to ≈10^5 clonal sequences (Supplementary Figure 7B).

304 (ii) In experimental data, increasing training dataset size by 1–2 orders of magnitude (sequence data

305 generated in a different lab using different experimental library preparation methods) increased prediction

306 accuracy by up to 7 percentage points, suggesting large-scale cross-study detection of public clones is

307 possible. (iii) The high prediction accuracy of human (antigen-selected) public and private memory B-cell

308 clones (Figure 4B) suggested that the low accuracy of (antigen-selected) PC (IgG) repertoires (Figure 4A)

309 may be due to small dataset size (Supplementary Table 1). More generally, we speculate that the

310 prediction accuracies reported here merely represent lower bounds; future studies, which combine (i)

311 advanced experimental and computational error correction methodologies (e.g., unique molecular

312 identifiers) [42–44], (ii) high sampling and sequencing depth [1] and (iii) novel sequence-based deep

313 learning approaches [45–47] may lead to even higher prediction accuracies.

314 To conclude, the existence of high-dimensional immunogenomic rules shaping immune repertoire

315 diversity in a predictable fashion, leading to clones with higher occurrence probability within a population,

316 highlights the potential of public clones to be a promising target for rational vaccine design and targeted

317 immunotherapies [23,34,48,49].

318

319

320

321

322

323

324

11

Methods

Immune repertoire high-throughput sequencing datasets

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

Dataset 1

Murine B-cell origin (C57BL/6J): Sequencing data were generated by Greiff and colleagues [16]. B-cells were isolated from four C57BL/6 cohorts (n=4–5) including untreated and prime-boost immunized with protein antigens. Cells were sorted into the subsets pre-B cells (preBC), naïve B cell (nBC) and plasma cells (PC) by flow cytometry. Cell numbers per mouse were: 750'000 (preBC), 1'000'000 (nBC) and 90'000 (PC). RNA was isolated from cells, antibody libraries were prepared by RT-PCR and sequenced using Illumina MiSeq platform (2x300bp paired-end). The sequencing data has been deposited online (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5349/) along with full experimental details and were preprocessed using MiXCR for VDJ-annotation, clonotype formation by CDR3 and error correction as described previously [16,50]. Briefly, for downstream analyses, functional clonotypes were only retained if: (i) they were composed of at least 4 amino acids, and (ii) had a minimal read count of 2 [51,52]. 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/c): Sequencing data were generated by Greiff and colleagues [16] and have been deposited online (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5349/) with full experimental details. Briefly, naïve B-cells (1,000,000 cells per mouse) from 4 unimmunized BALB/c mice were isolated using the sorting panel from Dataset 1 and antibody 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 and colleagues [16] and have been deposited online (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5349/) with full experimental details. Briefly, naïve B-cells (≈671'000 cells per mouse) from three pet shop mice were

12

353   isolated and library preparation, sequencing, and data preprocessing was performed analogously to

354   Dataset 1. Public clones were defined as those clones that occurred at least twice across mice.

355   **Dataset 4**

356   Murine B-cell origin (C57BL/6J): Sequencing data were published by Yang and colleagues [20]: Mature B

357   cells were extracted C57BL/6J-mice and sorted ($1-2x10^4$ per cell population) into developmentally distinct

358   subsets (splenic follicular B-cells (FOB, n=5), marginal zone B-cells (MZB, n=7), peritoneal B2-B-cells

359   (n=5) and B-1a B-cells (n=43)). Data preprocessing was performed analogously to Dataset 1. Public

360   clones were defined as those clones that occurred at least twice across mice of a given B-cell population.

361   **Dataset 5**

362   Human B-cell origin: Sequencing data of naïve and memory B-cells from three healthy donors were

363   published   by   DeWitt   and   colleagues   [13]   and   downloaded   already   preprocessed   from

364   http://datadryad.org/resource/doi:10.5061/dryad.35ks2. Public clones were defined as those clones that

365   occurred at least twice across individuals within a given B-cell population. Cell numbers of naïve and

366   memory B cells were $2-4x10^7$ and $1.5-2x10^7$, respectively.

367   **Dataset 6**

368   Murine T-cell origin: Sequencing data were published by Madi and colleagues [17]. CD4 T cells were

369   isolated from 28 mice (three cohorts; untreated (n=12), immunized with complete Freud's adjuvant (CFA,

370   n=7) or immunized with CFA and ovalbumin (n=9). Data preprocessing was performed using MiXCR for

371   annotation and error correction as described previously [16,50]. Public clones were defined as those

372   clones that occurred at least twice across mice of a given cohort.

373   **Determination of statistical significance**

374   Significance was tested using the Wilcoxon rank-sum test if not indicated otherwise. Where applicable,

375   significance of correlation coefficients was tested using the R function cor.test() with default parameters.

376

377   **Statistical analysis and plots**

378   Statistical analysis was performed using R [53] and Python [54]. Graphics were generated using the R

379   packages ggplot2 [55], RColorBrewer [56], and Complex Heatmap [57]. Parallel computing of SVM

380   analyses was performed using the R packages RBatchJobs [58] and doParallel [59].

13

381 **Definition of a clone**

382 For all analyses, clones were defined by 100% amino acid sequence identity of CDR3 regions [1,16,51].

383 CDR3 regions were annotated and defined by MiXCR software [50] according to the nomenclature of the

384 Immunogenetics database (IMGT) [60].

385 **Quantification of overlap**

386 As defined previously [16], the percentage of clones shared between two repertoires X and Y:

387 overlap $(X,Y) = \frac{|X \cap Y|}{\min(|X|,|Y|)} \times 100$, where $|X|$ and $|Y|$ are the clonal sizes (number of unique clones) of

388 repertoires X and Y. A repertoire was mathematically defined as a set of unique clones.

389 **Junction Analysis**

390 V, N1, D, N2 and J subregion annotation of sequences was performed using IMGT/HighV-Quest [61]

391 (after initial preprocessing by MiXCR) [50]. Deletions were determined by finding the longest common

392 substring between the germline genes and the V, D and J subregions identified in the CDR3 sequences.

393 **Estimation of the technological coverage of V, N1, D, N2, J regions**

394 To estimate the technological coverage of each region (V, N1, D, N2, J), bootstrapping was conducted

395 (Supplementary Figure 2). Briefly, 5, 25, 50, 75 and 100% of the full diversity of each region was

396 sampled. Subsequently, the number of unique sequences per region present in the sample was

397 compared to the total number of unique sequences.

398 **Determination of Shannon Evenness**

399 The Shannon Evenness was calculated as previously described [35]. Briefly, we calculated the Hill-

400 diversity for alpha = 1 ($^{\alpha}D = (\sum_{i=1}^{n} f_i^{\alpha})^{\frac{1}{1-\alpha}}$) for a given frequency distribution ($\vec{f}$, enumeration of the

401 abundance of each subregion (combination)) of V, N1, D, N2, J subregions or combinations thereof.

402 Subsequently, we obtained the Shannon Evenness $^{\alpha=1} = E$ by normalizing $^{\alpha=1}D$ by the respective total

403 number of V, N1, D, N2, J regions or combinations thereof (n) in the given repertoire.

404 **Estimation of the theoretical nucleotide diversity of the murine naïve clonal repertoire**

405 The extent to which the entirety of the subregions V, N1, D, N2, J discovered in preBC and nBC of

406 Dataset 1 covered any preBC/nBC repertoire was quantified by species accumulation curves as

14

407   previously described [16]. Briefly, we defined the repertoire coverage (C_i) of a given CDR3 subregion (R_i)

408   as the percentage overlap of its set of unique regions $\{R\}_i$ with the set of regions contained in all

409   previously accumulated repertoires ($\cup_{j=1}^{i-1} R_j$): $C_i = \dfrac{\left|R_i \cap \cup_{j=1}^{i-1} R_j\right|}{|R_i|}$, where $i \in \{1, \ldots, m\}$ with m being the total

410   number of preBC and nBC repertoires ($m = 38$). To infer the number of subregions necessary for any

411   given   coverage,   we   used   non-linear   regression   analysis   using   an   exponential   fit

412   ($C_i \sim 100 - s * b^{(-\log(|\cup_{i=1}^{m} R_i|))}$) [62], where $|\cup_{i=1}^{m} R_i|$ is the number of unique subregions contained within

413   the accumulated repertoires and s and b are the parameters to be inferred. For ≥95% coverage, this is the

414   estimated size of each murine naïve V, N1, D, N2, J subregion repertoire. We opted to report the

415   coverage at 95% (Supplementary Table 2, column 2) to counter the effect of rare clones as described

416   previously [16]. The product of the extrapolated coverage at 95% of each region (Supplementary Table 2)

417   is the theoretical nucleotide diversity of the murine naïve clonal repertoire.

Determination of private clones with high similarity to public clones

419   For each public clone, the number of private clones within 1 amino acid edit distance was enumerated

420   (Figure 6B). Edit distance was determined using the stringdist() function (distance metric: Levenshtein

421   distance) from the stringdist R package [63] as well as igraph [64]

Support Vector Machine (SVM) analysis

423   In order to classify clones into public and private classes, a supervised learning approach was chosen in

424   the form of a support vector machine (SVM) model. As input for all SVM analyses, CDR3-length

425   equilibrated datasets were built for each sample (Supplementary Table 1). Briefly, for each sample, all

426   public clones were paired in equal numbers with private clones of the same sample such that both public

427   and private clones followed identical CDR3 length distributions. SVM analysis was performed using

428   kernel-based analysis of biological sequences (KeBABS) [30] and sklearn [65], both of which are

429   described in more detail below. For all SVM analyses, each dataset was split into training (80%) and test

430   subset (20%). Cross-validation and SVM training was performed on the training dataset and class

431   prediction on the test dataset. Prediction accuracy of class discrimination was quantified by calculating

432   the balanced accuracy $\mathrm{BACC} = \frac{1}{2} \times (\mathrm{spec} + \mathrm{sens})$, where specificity was defined as $\mathrm{spec} = \frac{TN}{TN+FP}$, and

433   sensitivity defined as $\mathrm{sens} = \frac{TP}{TP+FN}$ ($TP = \mathrm{True\ Positive}, TN = \mathrm{True\ Negative}, \quad FP = \mathrm{False\ Positive},$

434   FN = False Negative). Additionally, AUC (area under the curve, ROC curve) was calculated using the

435   KeBABS R package [30]. An AUC value of 1 means perfect prediction accuracy (BACC = 100%), while an

436   AUC value of 0.5 (BACC = 50%) is equivalent to random guessing.

437   **KeBABS support vector machine analysis**

438   To discriminate public and private clones based on CDR3 sequence, we used the KeBABS R package

439   [30], which implements kernel-based analysis of biological sequences. For all datasets, we used the

440   position-independent gappy pair kernel [36,37], which divides all sequences into features of length k with

441   gaps of maximal length m (Figure 4A). For the analysis of nucleotide sequences the parameters were set

442   to k=3, m=1, C = 10, whereas the analysis of amino acid sequences was performed using parameters

443   k=1, m=1, C = 100 (as determined by cross-validation). The cost parameter C sets the cost for the

444   misclassification of a sequence. The maximal number of possible features used in the gappy kernel is

445   determined by $4^{2 \times k} \times (m + 1) = 8'192$ for nucleotide sequences and $20^{2 \times k} \times (m + 1) = 800$ for amino acid

446   sequences.

447   **Prediction Profiles**

448   Prediction profiles were computed from feature weights as described by Palme and colleagues
449   [30,31,37]. Prediction profiles quantify the contribution of each sequence position to the decision value
450   (public, private). Thus, prediction profiles provide improved biological interpretability of the learning results
451   compared to single feature weights because those individual positions or sequence stretches that drive
452   classification accuracy most become visible [30].

453   **Sklearn support vector machine analysis**

454   For public vs. private discrimination based on amino acid and V, N1, D, N2, J composition, the sklearn

455   implementation of SVM [65] for Python [54] was employed with the cost parameter set at C=10 as

456   determined by cross-validation.

457

458

459

460

16

References

461

462   1.   Greiff V, Miho E, Menzel U, Reddy ST. Bioinformatic and Statistical Analysis of Adaptive Immune
463   Repertoires. Trends Immunol. 2015;36: 738–749. doi:10.1016/j.it.2015.09.006

464   2.   Hershberg U, Prak ETL. The analysis of clonal expansions in normal and autoimmune B cell repertoires. Phil
465   Trans R Soc B. 2015;370: 20140239. doi:10.1098/rstb.2014.0239

466   3.   Xu JL, Davis MM. Diversity in the CDR3 Region of VH Is Sufficient for Most Antibody Specificities.
467   Immunity. 2000;13: 37–45. doi:10.1016/S1074-7613(00)00006-6

468   4.   Kunik V, Peters B, Ofran Y. Structural Consensus among Antibodies Defines the Antigen Binding Site. PLoS
469   Comput Biol. 2012;8. doi:10.1371/journal.pcbi.1002388

470   5.   Castro R, Navelsaker S, Krasnov A, Du Pasquier L, Boudinot P. Describing the diversity of Ag specific
471   receptors in vertebrates: Contribution of repertoire deep sequencing. Dev Comp Immunol. 2017;
472   doi:10.1016/j.dci.2017.02.018

473   6.   Tonegawa S. Somatic generation of antibody diversity. Nature. 1983;302: 575–581. doi:10.1038/302575a0

474   7.   Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise determination of the diversity of a
475   combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci.
476   2009;106: 20216–20221. doi:10.1073/pnas.0909775106

477   8.   Saada R, Weinberger M, Shahaf G, Mehr R. Models for antigen receptor gene rearrangement: CDR3 length.
478   Immunol Cell Biol. 2007;85: 323–332. doi:10.1038/sj.icb.7100055

479   9.   Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, et al. Exhaustive T-cell repertoire sequencing of
480   human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size
481   of at least 1 million clonotypes. Genome Res. 2011;21: 790–797. doi:10.1101/gr.115428.110

482   10. Murugan A, Mora T, Walczak AM, Callan CG. Statistical inference of the generation probability of T-cell
483   receptors from sequence repertoires. Proc Natl Acad Sci. 2012;109: 16161–16166.
484   doi:10.1073/pnas.1212755109

485   11. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, et al. High-Resolution Description of Antibody
486   Heavy-Chain Repertoires in Humans. PLoS ONE. 2011;6: e22365. doi:10.1371/journal.pone.0022365

487   12. Jiang N, Weinstein JA, Penland L, White RA, Fisher DS, Quake SR. Determinism and stochasticity during
488   maturation of the zebrafish antibody repertoire. Proc Natl Acad Sci. 2011;108: 5348–5353.
489   doi:10.1073/pnas.1014277108

490   13. DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, Carlson CS, et al. A Public Database of
491   Memory and Naive B-Cell Receptor Sequences. PLOS ONE. 2016;11: e0160853.
492   doi:10.1371/journal.pone.0160853

493   14. Galson JD, Trück J, Fowler A, Münz M, Cerundolo V, Pollard AJ, et al. In-depth assessment of within-
494   individual and inter-individual variation in the B cell receptor repertoire. Front Immunol. 2015; 531.
495   doi:10.3389/fimmu.2015.00531

496   15. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of
497   high-throughput sequencing of the antibody repertoire. Nat Biotechnol. 2014;32: 158–168.
498   doi:10.1038/nbt.2782

499   16. Greiff V, Menzel U, Miho E, Weber C, Riedel R, Cook SC, et al. Systems analysis reveals high genetic and
500   antigen-driven predetermination of antibody repertoires throughout B-cell development. Cell Rep., "accepted
501   in principle", 2017.

17

502   17. Madi A, Shifrut E, Reich-Zeliger S, Gal H, Best K, Ndifon W, et al. T-cell receptor repertoires share a
503   restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. Genome
504   Res. 2014;24: 1603–1612. doi:10.1101/gr.170753.113

505   18. Robinson WH. Sequencing the functional antibody repertoire—diagnostic and therapeutic discovery. Nat Rev
506   Rheumatol. 2014;11: 171–182. doi:10.1038/nrrheum.2014.220

507   19. Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. Genome Med.
508   2015;7: 121. doi:10.1186/s13073-015-0243-2

509   20. Yang Y, Wang C, Yang Q, Kantor AB, Chu H, Ghosn EE, et al. Distinct mechanisms define murine B cell
510   lineage immunoglobulin heavy chain (IgH) repertoires. eLife. 2015; e09083. doi:10.7554/eLife.09083

511   21. Jackson KJL, Kidd MJ, Wang Y, Collins AM. The shape of the lymphocyte receptor repertoire: lessons from
512   the B cell receptor. Front B Cell Biol. 2013;4: 263. doi:10.3389/fimmu.2013.00263

513   22. Covacu R, Philip H, Jaronen M, Almeida J, Kenison JE, Darko S, et al. System-wide Analysis of the T Cell
514   Response. Cell Rep. 2016;14: 2733–2744. doi:10.1016/j.celrep.2016.02.056

515   23. Venturi V, Price DA, Douek DC, Davenport MP. The molecular basis for public T-cell responses? Nat Rev
516   Immunol. 2008;8: 231–238. doi:10.1038/nri2260

517   24. Elhanati Y, Murugan A, Callan CG, Mora T, Walczak AM. Quantifying selection in immune receptor
518   repertoires. Proc Natl Acad Sci. 2014;111: 9875–9880.

519   25. Elhanati Y, Sethna Z, Marcou Q, Callan CG, Mora T, Walczak AM. Inferring processes underlying B-cell
520   repertoire diversity. Phil Trans R Soc B. 2015;370: 20140243. doi:10.1098/rstb.2014.0243

521   26. Mora T, Walczak AM, Bialek W, Callan CG. Maximum entropy models for antibody diversity. Proc Natl Acad
522   Sci. 2010;107: 5405–5410. doi:10.1073/pnas.1001705107

523   27. Kidd BA, Peters LA, Schadt EE, Dudley JT. Unifying immunology with informatics and multiscale biology.
524   Nat Immunol. 2014;15: 118–127. doi:10.1038/ni.2787

525   28. Lodhi H, Saunders C, Shawe-Taylor J, Cristianini N, Watkins C. Text classification using string kernels. J
526   Mach Learn Res. 2002;2: 419–444.

527   29. Sun Y, Best K, Cinelli M, Heather JM, Reich-Zeliger S, Shifrut E, et al. Specificity, Privacy, and Degeneracy
528   in the CD4 T Cell Receptor Repertoire Following Immunization. Front Immunol. 2017;8.
529   doi:10.3389/fimmu.2017.00430

530   30. Palme J, Hochreiter S, Bodenhofer U. KeBABS: an R package for kernel-based analysis of biological
531   sequences. Bioinformatics. 2015; btv176. doi:10.1093/bioinformatics/btv176

532   31. Schwarzbauer K, Bodenhofer U, Hochreiter S. Genome-Wide Chromatin Remodeling Identified at GC-Rich
533   Long Nucleosome-Free Regions. PLOS ONE. 2012;7: e47924. doi:10.1371/journal.pone.0047924

534   32. Bishop CM. Pattern Recognition and Machine Learning. New edition. Springer, Berlin; 2007.

535   33. Thomas N, Best K, Cinelli M, Reich-Zeliger S, Gal H, Shifrut E, et al. Tracking global changes induced in the
536   CD4 T cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein
537   sequence. Bioinforma Oxf Engl. 2014; doi:10.1093/bioinformatics/btu523

538   34. Miho E, Greiff V, Roskar R, Reddy ST. The fundamental principles of antibody repertoire architecture revealed
539   by large-scale network analysis. bioRxiv. 2017; 124578. doi:10.1101/124578

35. Greiff V, Bhat P, Cook SC, Menzel U, Kang W, Reddy ST. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. Genome Med. 2015;7: 49. doi:10.1186/s13073-015-0169-8

36. Leslie C, Kuang R. Fast String Kernels Using Inexact Matching for Protein Sequences. J Mach Learn Res. 2004;5: 1435–1455.

37. Mahrenholz CC, Abfalter IG, Bodenhofer U, Volkmer R, Hochreiter S. Complex networks govern coiled-coil oligomerization--predicting and profiling by means of a machine learning approach. Mol Cell Proteomics MCP. 2011;10: M110.004994. doi:10.1074/mcp.M110.004994

38. Li H, Ye C, Ji G, Wu X, Xiang Z, Li Y, et al. Recombinatorial Biases and Convergent Recombination Determine Interindividual TCRβ Sharing in Murine Thymocytes. J Immunol. 2012;189: 2404–2413. doi:10.4049/jimmunol.1102087

39. Janeway CA, Murphy K. Janeway's Immunobiology. 8th Revised edition. Taylor & Francis; 2011.

40. Mroczek ES, Ippolito GC, Rogosch T, Hoi KH, Hwangpo TA, Brand MG, et al. Differences in the composition of the human antibody repertoire by B cell subsets in the blood. B Cell Biol. 2014;5: 96. doi:10.3389/fimmu.2014.00096

41. Safonova Y, Lapidus A, Lill J. IgSimulator: a versatile immunosequencing simulator. Bioinformatics. 2015; btv326.

42. Khan TA, Friedensohn S, Vries ARG de, Straszewski J, Ruscheweyh H-J, Reddy ST. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. Sci Adv. 2016;2: e1501371. doi:10.1126/sciadv.1501371

43. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. Proc Natl Acad Sci. 2013;110: 13463–13468. doi:10.1073/pnas.1312146110

44. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-free profiling of immune repertoires. Nat Methods. 2014;11: 653–655. doi:10.1038/nmeth.2960

45. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9: 1735–1780.

46. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol. 2016;12: 878. doi:10.15252/msb.20156651

47. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33: 831–838. doi:10.1038/nbt.3300

48. Miles JJ, Silins SL, Burrows SR. Engineered T cell receptors and their potential in molecular medicine. Curr Med Chem. 2006;13: 2725–2736.

49. Jardine JG, Kulp DW, Havenar-Daughton C, Sarkar A, Briney B, Sok D, et al. HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen. Science. 2016;351: 1458–1463. doi:10.1126/science.aad9195

50. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods. 2015;12: 380–381. doi:10.1038/nmeth.3364

51. Greiff V, Menzel U, Haessler U, Cook SC, Friedensohn S, Khan TA, et al. Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. BMC Immunol. 2014;15: 40. doi:10.1186/s12865-014-0040-5

579 52. Menzel U, Greiff V, Khan TA, Haessler U, Hellmann I, Friedensohn S, et al. Comprehensive Evaluation and
580   Optimization of Amplicon Library Preparation Methods for High-Throughput Antibody Sequencing. PLoS
581   ONE. 2014;9: e96727. doi:10.1371/journal.pone.0096727

582 53. Team RDC. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2009.
583   Available: http://www.R-project.org

584 54. Rossum GV, Drake FLJ. The Python Language Reference Manual. Network Theory Ltd; 2011.

585 55. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2009.
586   Available: http://ggplot2.org

587 56. Neuwirth E. RColorBrewer: ColorBrewer Palettes [Internet]. 2014. Available: https://CRAN.R-
588   project.org/package=RColorBrewer

589 57. Gu Z. ComplexHeatmap: Making Complex Heatmaps [Internet]. 2016. Available:
590   https://github.com/jokergoo/ComplexHeatmap

591 58. Bischl B, Lang M, Mersmann O, Rahnenführer J, Weihs C. BatchJobs and BatchExperiments: Abstraction
592   Mechanisms for Using R in Batch Environments. J Stat Softw. 2015;64: 1–25.

593 59. Analytics R, Weston S. doParallel: Foreach Parallel Adaptor for the "parallel" Package [Internet]. 2015.
594   Available: https://CRAN.R-project.org/package=doParallel

595 60. Lefranc M-P, Giudicelli V, Ginestoux C, Bodmer J, Müller W, Bontrop R, et al. IMGT, the international
596   ImMunoGeneTics database. Nucleic Acids Res. 1999;27: 209–212. doi:10.1093/nar/27.1.209

597 61. Li S, Lefranc M-P, Miles JJ, Alamyar E, Giudicelli V, Duroux P, et al. IMGT/HighV QUEST paradigm for T
598   cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nat Commun.
599   2013;4. doi:10.1038/ncomms3333

600 62. Soberón J, Llorente J. The Use of Species Accumulation Functions for the Prediction of Species Richness.
601   Conserv Biol. 1993;7: 480–488. doi:10.1046/j.1523-1739.1993.07030480.x

602 63. Loo MPJ van der. The stringdist package for approximate string matching. R J. 2014;6: 111–122.

603 64. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006;Complex
604   Systems: 1695.

605 65. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning
606   in Python. J Mach Learn Res. 2011;12: 2825–2830.

Acknowledgments

Figure 1

Figure 1 Immunogenomic analysis of public and private antibody repertoires.

(A) We asked whether there are immunogenomic differences that predetermine a clonal sequence's (CDR3) public or private status within a 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, Figure 6F). Private clones are those distinct to each individual. We defined antibody and T cell clones based on 100% CDR3 (complementarity determining region 3) identity. For statistical power, we used six large-scale datasets (Supplementary Table 1) comprising different B-cell populations, species (humans, mice) and immune antigen receptors (B/T cell receptor).

(B) To answer our question, we decomposed public and private immune repertoires in conventional low-dimensional features (e.g., CDR3 amino acid usage, Figures 2 and 3) or novel high-dimensional features (CDR3 sequence decomposition into subsequences of length k (k-mers) separated by a gap of length m, Figures 4 and 5). Leveraging supervised machine learning (support vector machines), we tested whether low and high-dimensional features can detect immunogenomic differences between public and private repertoires (see Methods) 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 as well as 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 found to be generalizable across datasets produced in different laboratories with different library preparation and high-throughput sequencing (HTS) protocols.

Figure 2
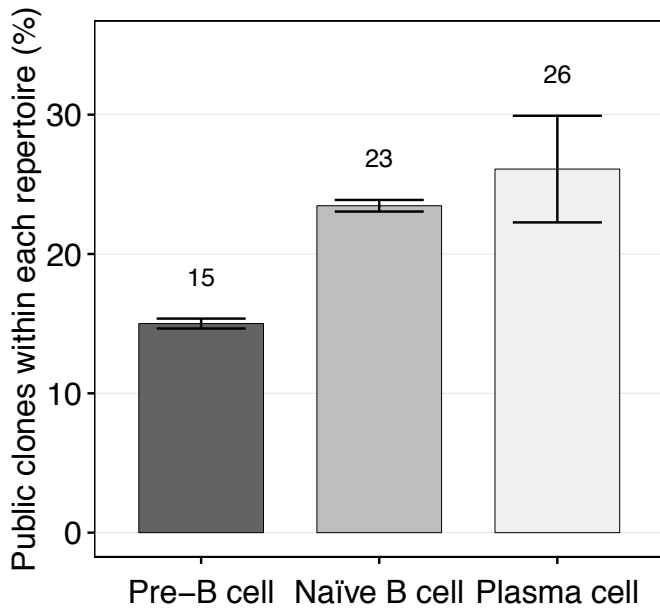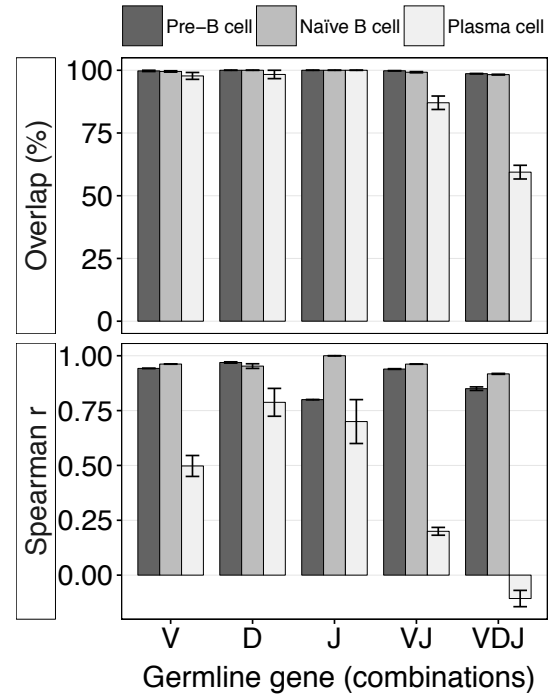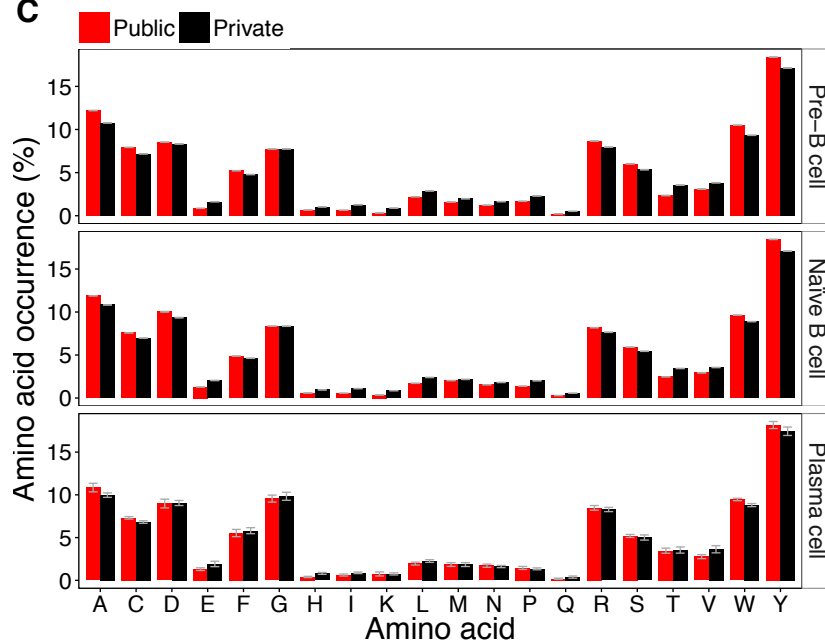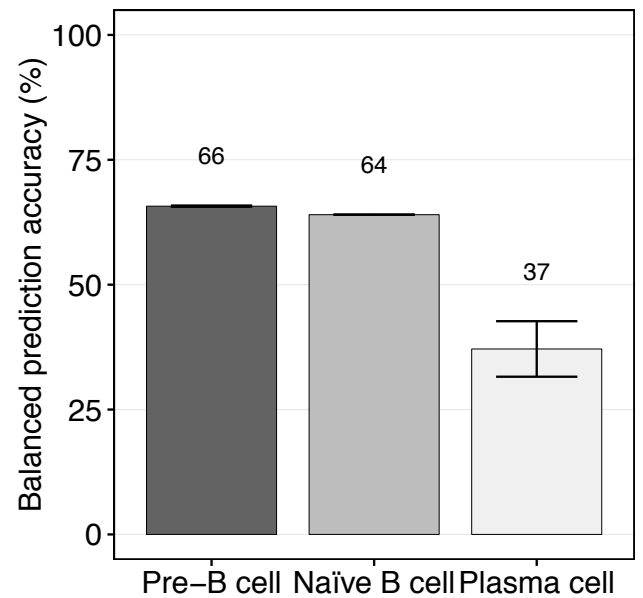
Figure 2 Public and private clone repertoires do neither differ predictively in germline gene usage nor amino acid composition

(A) Public clones represent 15–26% of murine antibody repertoires throughout B-cell ontogeny. Public clones were defined as being shared in at least two mice (see *Methods*).

(B) Overlap and Spearman correlation of V, D, J germline genes and their respective combinations (V-J, and V-D-J) between private and public clones by B-cell population.

(C) Relative amino acid composition of public (red) and private clones (black). Differences between public and private clones were not significant (Kolmorogov-Smirnov test, p>0.05).

(D) SVM-based discrimination of public and private clones based on CDR3 amino acid composition (see Methods). Balanced prediction accuracy was defined as the mean of specificity (detection rate of public clones) and sensitivity (detection rate of private clones). Barplots show mean±s.e.m.

Figure 3
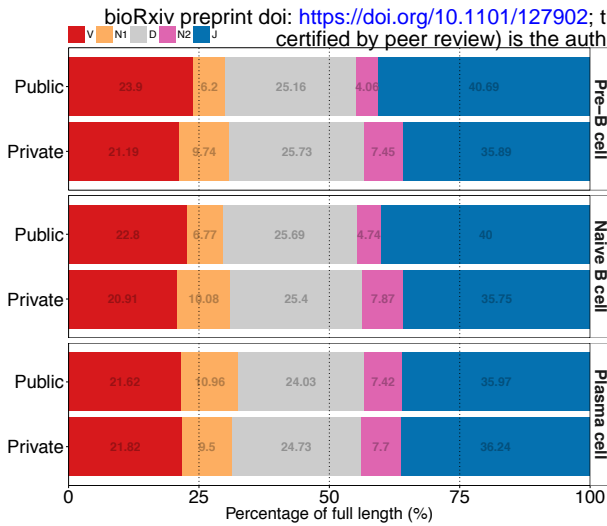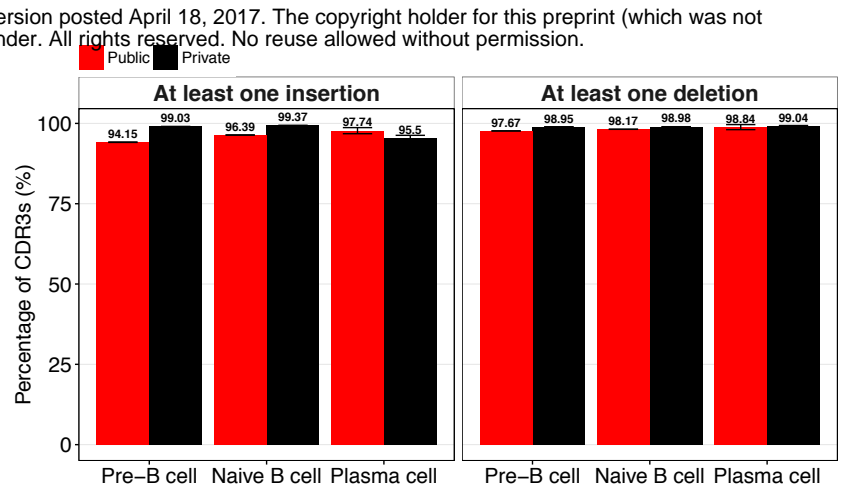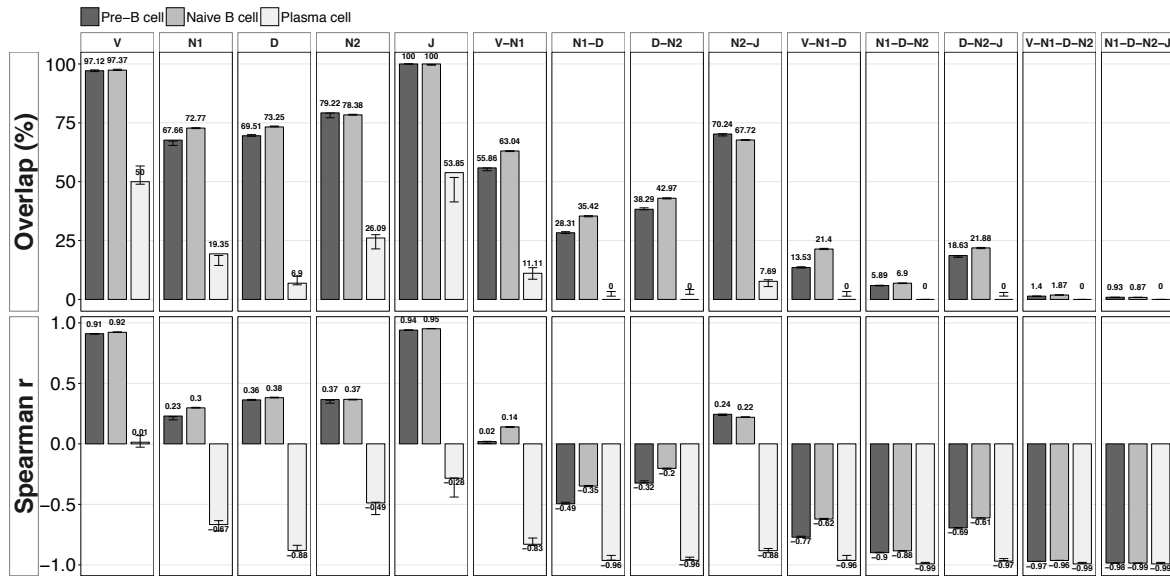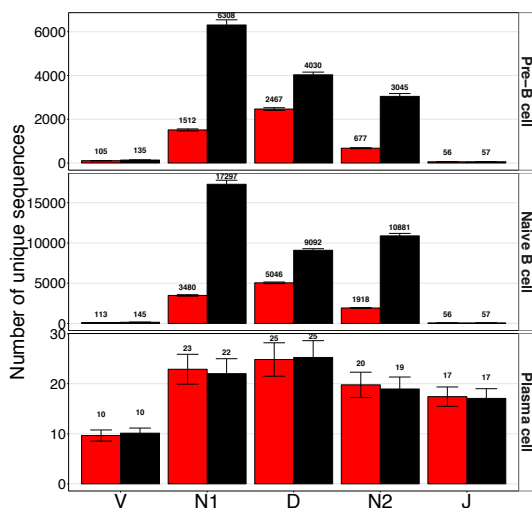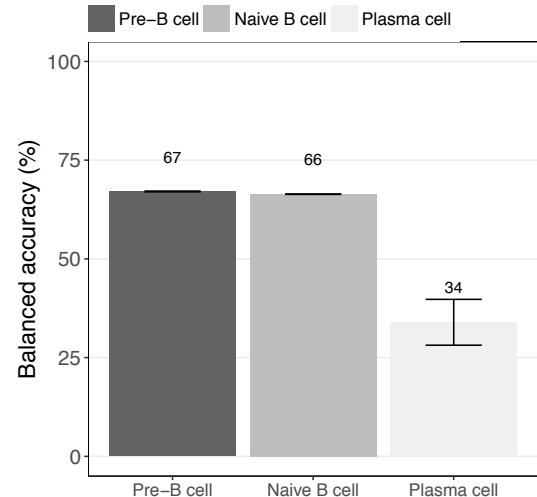
Figure 3 CDR3 subregion length does not predict a clone's public/private status.

(A) Normalized CDR3 subregion (V, N1, D, N2, J) lengths (median) of public and private clones by B-cell population.

(B) Frequency 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, J subregions (species richness) of public and private clones by B-cell population. Species richness of private clones CDR3 subregions was obtained by accounting for private and public clones size differences (bootstrapping, see *Methods*).

(E) SVM-based prediction of public and private clones based on V, N1, D, N2, J subregion composition (Figure 3A, see *Methods*). Balanced (prediction) accuracy was defined as the mean of specificity (detection rate of public clones) and sensitivity (detection rate of private clones). Barplots show mean±s.e.m.

Figure 4
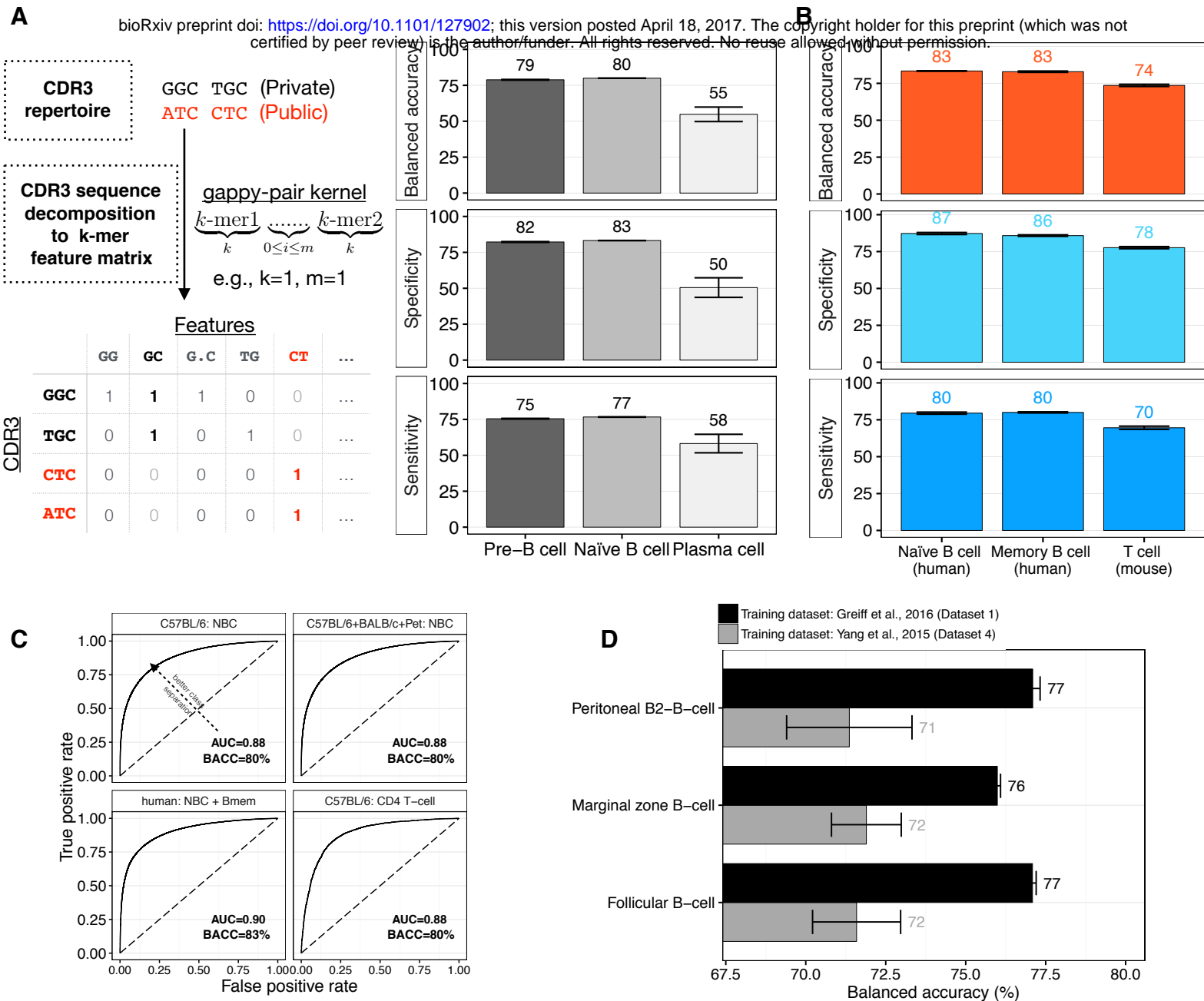
Figure 4 Public and private clones can be predicted with 80% accuracy using high-dimensional CDR3 sequence decomposition.

(A) Specificity (detection rate of public clones), sensitivity (detection rate of private clones) and balanced accuracy (mean of specificity and sensitivity) for public vs. private clones SVM discrimination by B-cell population. For each repertoire, a dataset composed of equal numbers of public and private clones (nucleotide sequences, length equilibrated) was assembled (*Methods*, Supplementary Table 1). Subsequently, as displayed in the insert, the gappy pair kernel function decomposes each CDR3 sequence into features made of two *k*-mers separated by a gap of maximal length *m*. The maximal number of features is $4^{(2 \times k)} \times (m+1) = 8192$ for nucleotide sequences ($k=3$, $m=1$) and $20^{(2 \times k)} \times (m+1) = 800$ for amino acid sequences ($k=1$, $m=1$). Based on this decomposition, a feature matrix of dimension #CDR3s times #Features is constructed. Each row of the feature matrix thus 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. The optimal parameter combinations ($k=3/m=1$ for nucleotide, $k=1/m=1$ for amino acid sequences) was determined by cross-validation on the training dataset (*Methods*).

(B) Prediction accuracy of public vs. private clones of human naïve and memory B-cells, and murine T cells. SVM parameters were identical to those used in (A).

(C) Public clones were accumulated across mice by B/T-cell populations (nBC, CD4), strain (nBC: C57BL/6, BALB/c, pet) or across B-cell populations (human naïve and memory B cells) in order to subsequently perform SVM-based classification as described in (A). Sizes of aggregated SVM-datasets ranged between ≈5x10⁴ (CD4 T cell) and 3x10⁶ (nBC: C57BL/6, BALB/c, pet) clones. ROC curves show excellent classification results (AUC [area under the ROC curve] ≈ 0.90).

(D) SVM-based prediction of public vs. private clones across experimental studies. 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: ≈2'400 clones, Supplementary Table 1). Barplots show mean±s.e.m.

Figure 5
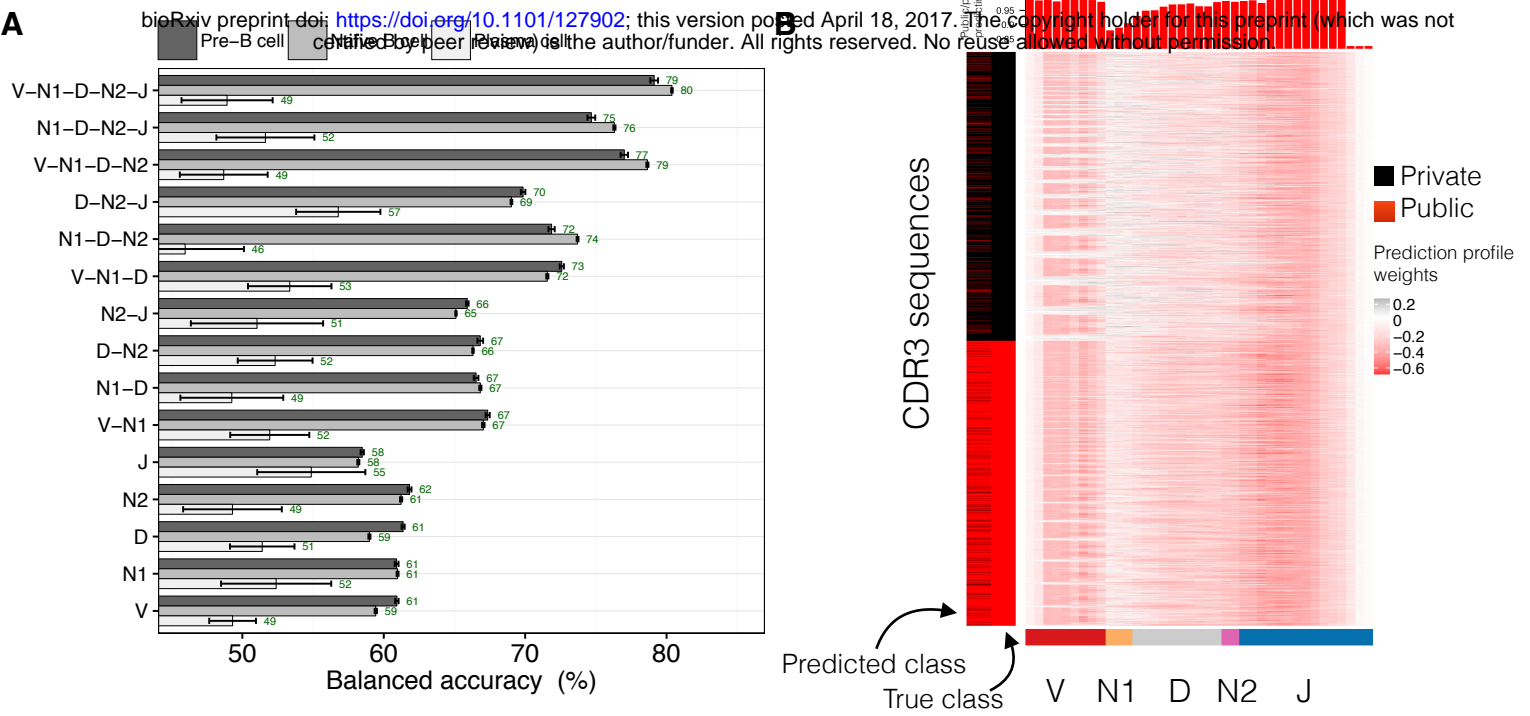
Figure 5 The N1-D-N2 subregions dominate the classification accuracy of public clones.

(A) Balanced accuracy of public and private clone discrimination using sequence-kernel-basel SVM analysis. For each combination of CDR3 subregion, gappy pair kernel parameters (k, m, cost) were determined by cross-validation. Barplots show mean±s.e.m.

(B) Exemplary visualization of prediction profiles of one test dataset (nBC) of CDR3s (rows) of length 39 (nt). Prediction profiles were computed as means of feature weights at each CDR3 position (1–39, see Methods). Positions colored red (<0) count towards "public" prediction of the respective CDR3s, whereas black-colored ones (>0) bias prediction towards the "private" clone status. Barplots indicate the percentage of private (black) or public predicting weights at each of the 39 positions. Color bars indicate median length of V (red), N1 (orange), D (grey), N2 (purple), J subregions (blue, Figure 3A). Prediction profiles across all CDR3 lengths as well as quantitative prediction profile analysis are given in Supplementary Figures 8 and 9, respectively.

Supplementary Figure 1
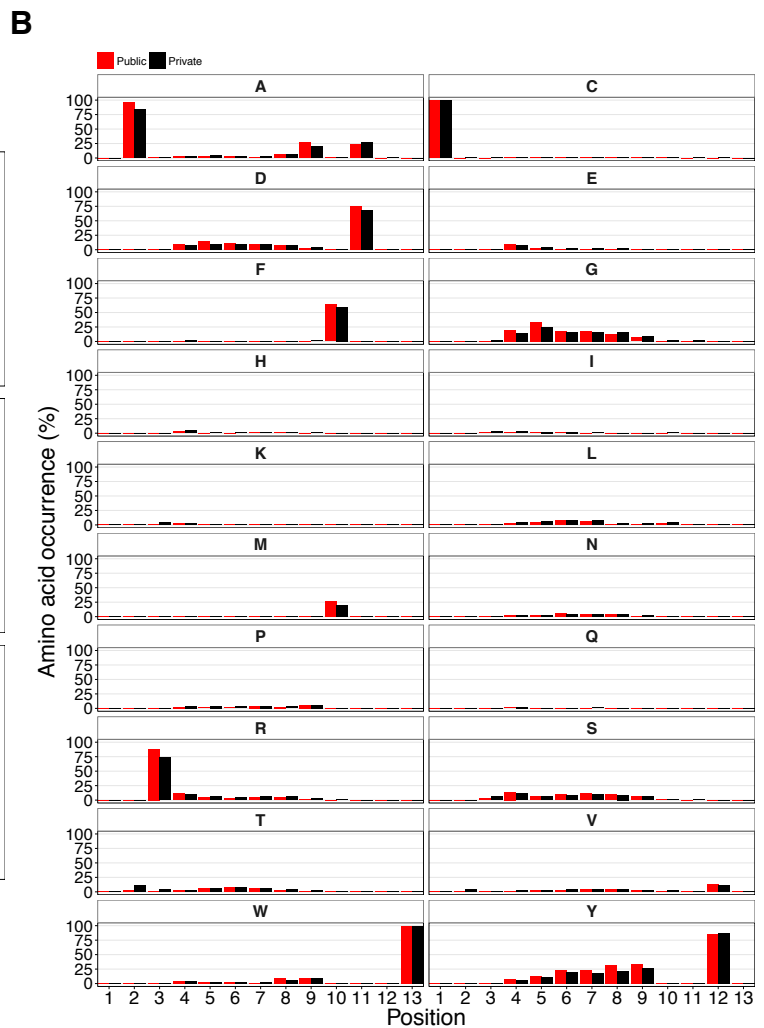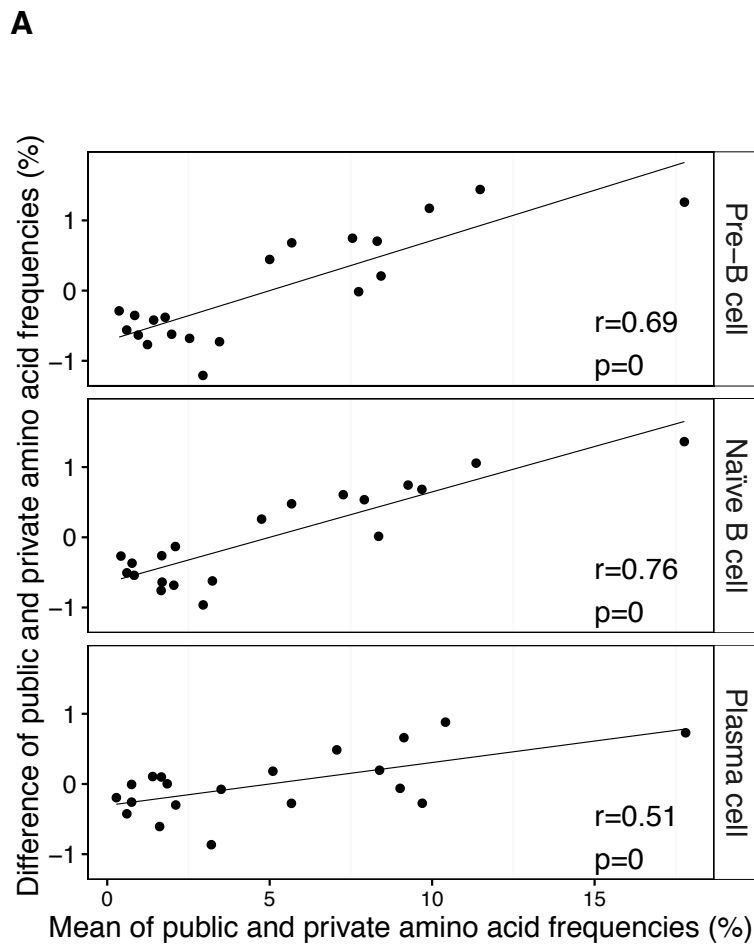
Supplementary Figure 1 The CDR3 amino acid composition at each CDR3 sequence position differs slightly between public and private clones (Dataset 1).

(A) Related to Figure 2D, the *difference* in public and private amino acid CDR3 frequencies is shown as a function of the *mean* of public and private CDR3 amino acid frequencies (by B-cell stage). The Pearson correlation coefficient and its p-value is shown for each B-cell stage .

(B) The amino acid composition of public and private clones is exemplarily shown at each position for CDR3s of amino acid length 13 (preBC repertoire). Barplots show mean±s.e.m.

Supplementary Figure 2
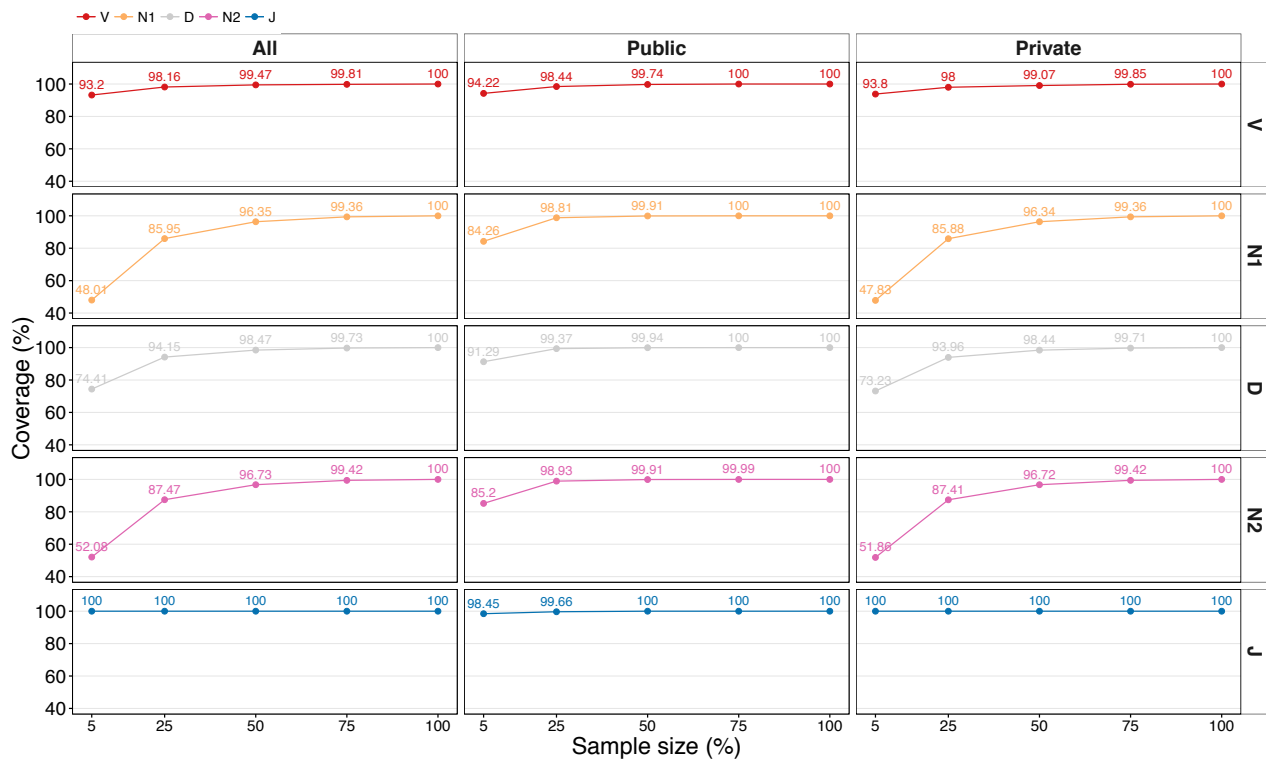
Supplementary Figure 2 Technological coverage of each subregion is achieved at 25–50% of sequencing reads (Dataset 1). To estimate the technological coverage of each CDR3 subregion (V, N1, D, N2, J), bootstrapping was performed. 5, 25, 50, 75 and 100% of the full diversity of each subregion was sampled to subsequently compare the bootstrapped diversity to the total number of unique sequences (coverage).
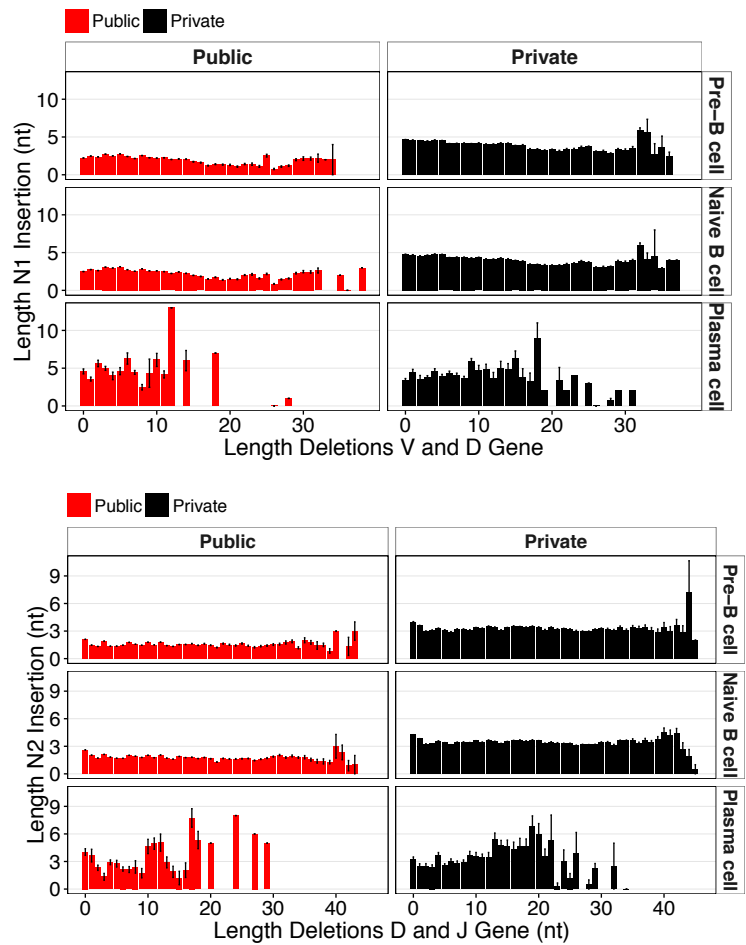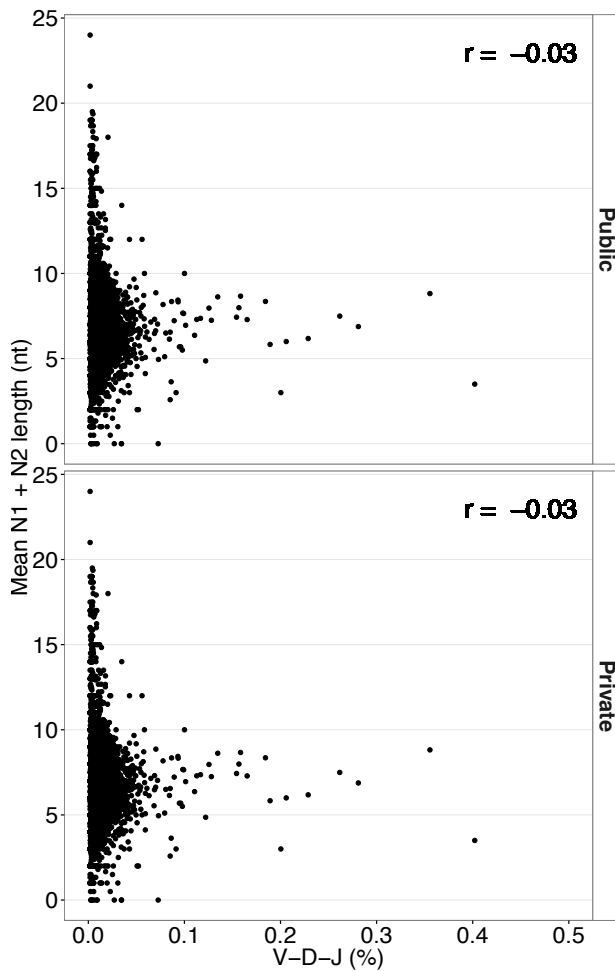
Supplementary Figure 3
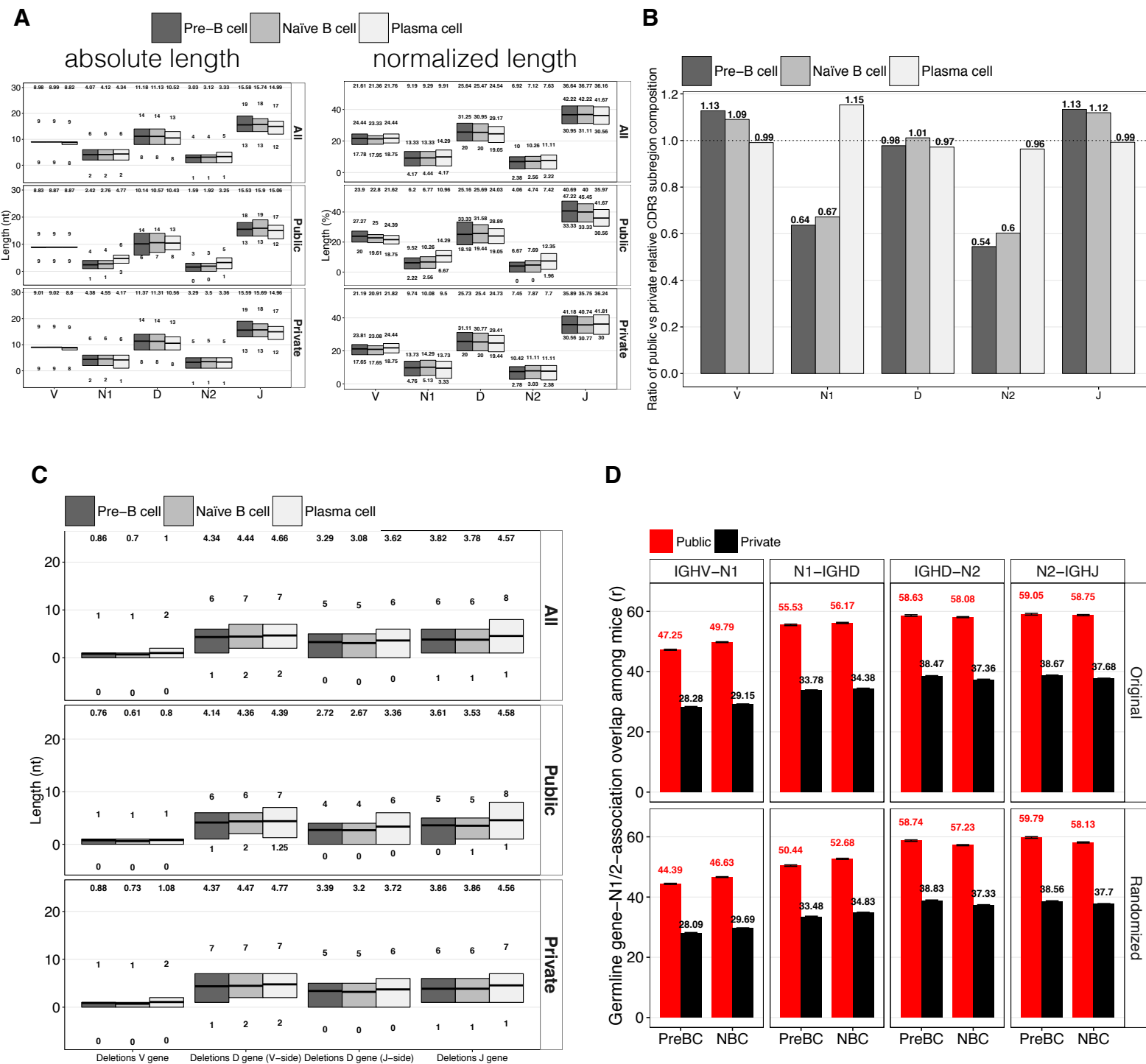
Supplementary Figure 3 Length comparison of CDR3 subregions and deletions between public and private clones (Dataset 1).

(A) Absolute and normalized CDR3 subregion lengths by B-cell population for public, private and all clones (irrespective of public/private status). Differences between public and private clones of preBC and nBC are significant (p<0.05).

(B) Ratios of normalized CDR3 subregions lengths (Figure 3B). Largest deviations from 1:1 ratio between public and private clones observed in preBC for the N1 and N2 subregions (36 and 46 percentage points, respectively).

(C) Absolute length of V, D, J deletions by B-cell population. Differences between public and private clones of preBC and nBC are significant (p<0.05).

(D) Nucleotide insertions (N1/N2) were aggregated by germline (IGHV/D/J, panels). Subsequently, the overlap of N1/N2 insertions was compared across mice, averaged and displayed by B-cell population (original). "Randomized" means the randomization of the association of germline genes and N1/N2 insertions. Barplots show mean±s.e.m.

Supplementary Figure 4
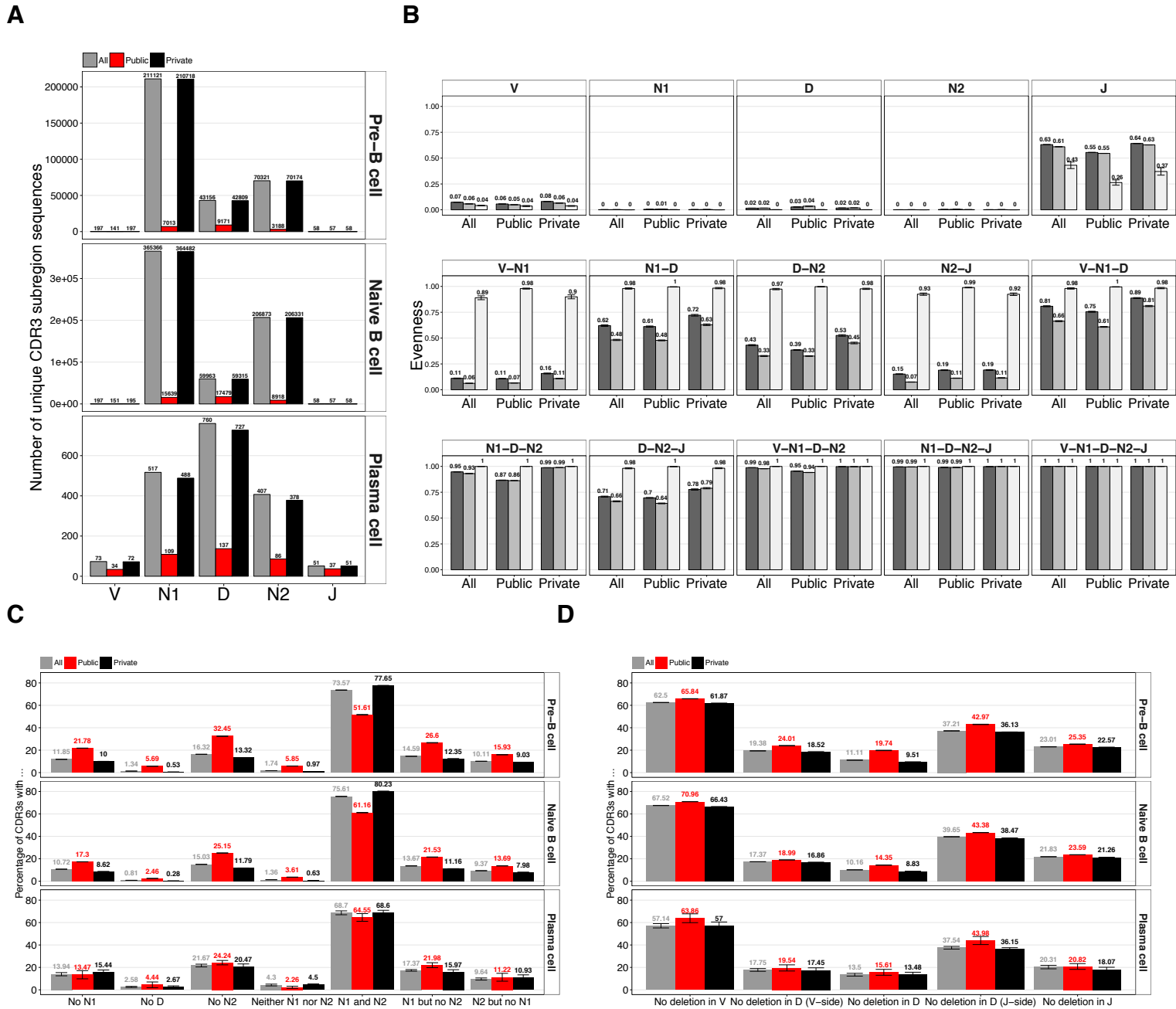
Supplementary Figure 4 Comparison of CDR3 subregion (and deletions) diversity and occurrence between public and private clones (Dataset 1).

(A) Total number of unique CDR3 subregion (V, N1, D, N2, J) sequences by B-cell population (non-bootstrapped, see Figure 3D for bootstrapped version).

(B) The Shannon evenness (see Methods) of CDR3 subregions and their combinations by B-cell population.

(C) Frequency of N1, N2 insertions across cases by B-cell population.

(D) Frequency of deletions across cases by B-cell population. Notably, deletions consistently occurred more often in private clones with the strongest contrast found in pre-B cells (private: 90.48% (100%-9.52%) vs. public: 80.38% (100%-19.62%)). Barplots show mean±s.e.m.

Supplementary Figure 5

A

B

Supplementary Figure 5 N1 and N2 subregion length is neither associated with germline gene usage nor with deletion length (Dataset 1).

(A) The frequency of each V-D-J combination was plotted against its corresponding mean N1+N2 length. Pearson correlation is shown.

(B) Association of N1/N2 insertions and deletions (V/D, and D/J) by B-cell population and clonal status (public/private). Barplots show mean±s.e.m.
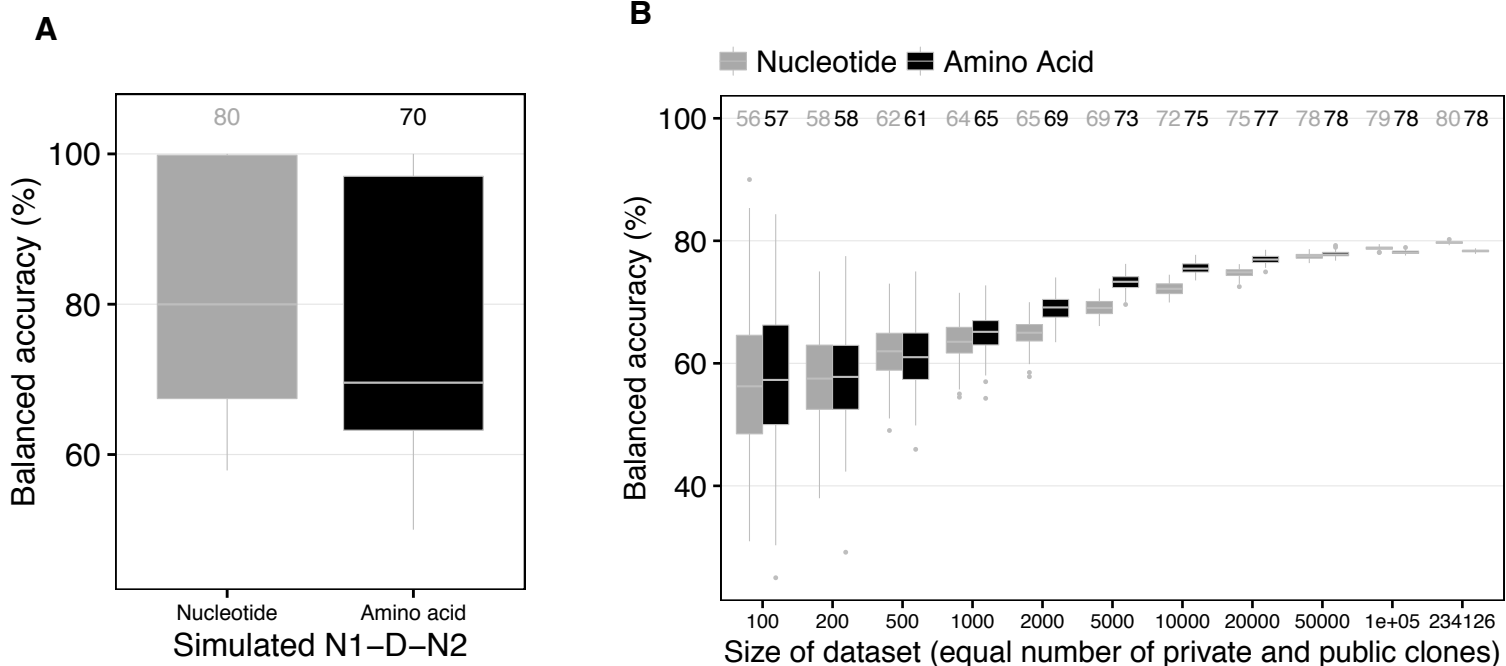
Supplementary Figure 6

A

Alternative mouse strains

B **Sequence randomization**

C **Label shuffling**

Discrimination of public/public and private/private clones across datasets

D

E **CDR3 subregion shuffling**

F **Alternative public clone definition**
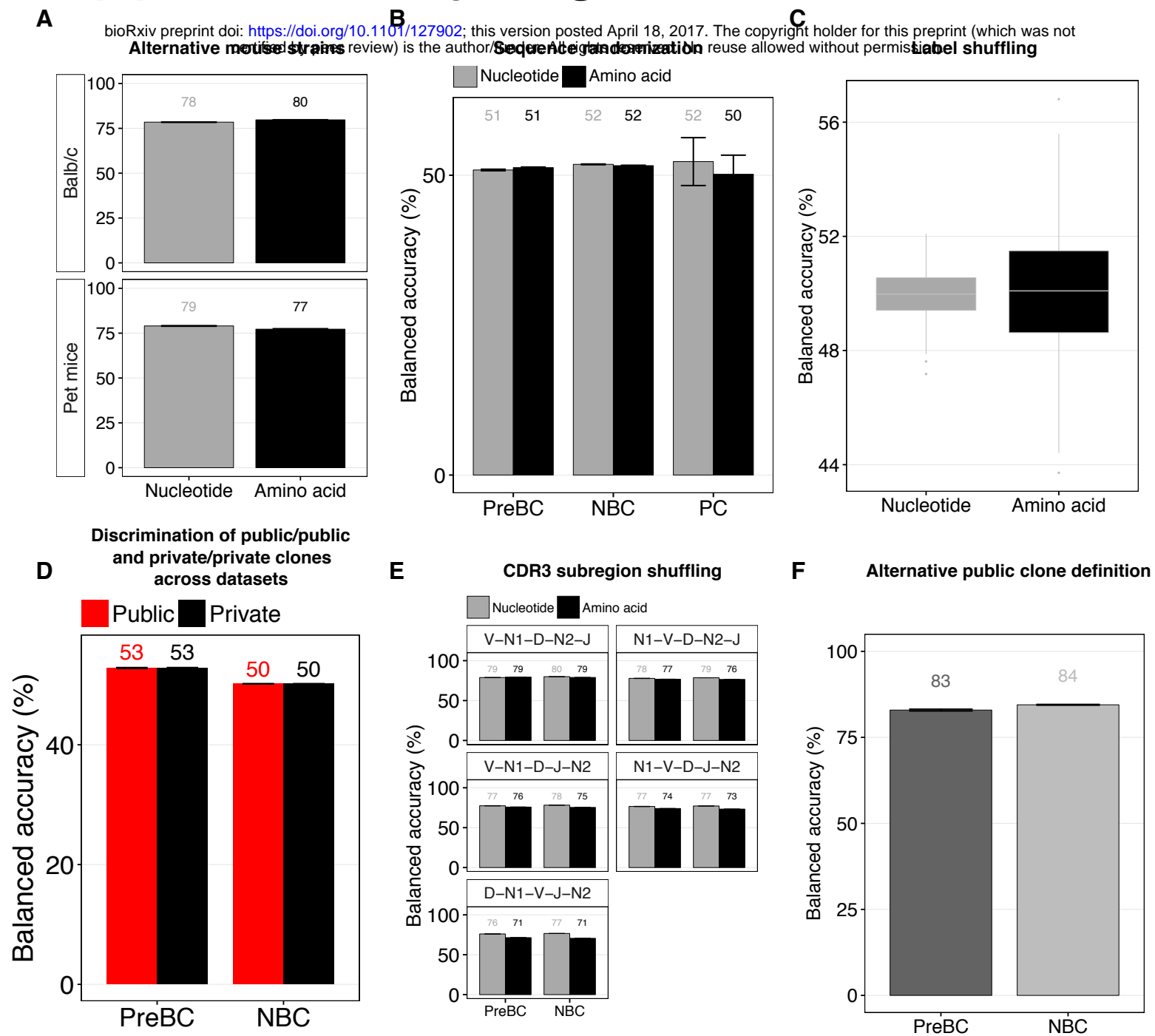
Supplementary Figure 6 SVM-based discrimination of amino acid CDR3s and negative SVM controls.

(A) Analogously to Figure 4A, SVM-based discrimination was performed on public and private amino acid CDR3 sequences of nBC from Balb/c (Dataset 2) and pet mice (Dataset 3).

(B) Balanced accuracy of SVM-based discrimination of randomized public and private CDR3 sequences (nucleotide, amino acid) by B-cell population (Dataset 1). CDR3 sequences were randomized by nucleotide/amino acid shuffling. SVM was performed as described for Figure 4A.

(C) Balanced accuracy of SVM-based discrimination performed on an nBC sample of Dataset 1 of which the labels (public, private) were shuffled in a random order. SVM was performed as described for Figure 4A.

(D) Balanced accuracy of SVM-based discrimination of public vs. public and private vs. private clones of repertoires of identical B-cell population. SVM was performed as described for Figure 4A. This was to confirm that public and private clones across mice do not differ from one another.

(E) Balanced accuracy of SVM-based discrimination of preBC and nBC repertoires (Dataset 1) of which the CDR3 subregions were shuffled as indicated.

(F) Validation that public/private clone balanced prediction accuracy is independent of public clone definition. In contrast to the public clone definition adopted for Dataset 1, 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: 4'682 ± 657 (preBC, mean ± sd), 28'249 ± 6736 (nBC)). Subsequently, SVM-based discrimination of public and private amino acid was carried out analogously to that described for Figure 4A. Barplots show mean±s.e.m.

Supplementary Figure 7

Supplementary Figure 7 Removal of differences in N1, D, N2 subregion length between public and private clones does not influence prediction accuracy and prediction accuracy increases as a function of dataset size.

(A) Prediction accuracy of sequence-kernel-based SVM-classification of simulated public and private N1-D-N2 sequences. In order to test whether the length differences of public and private N1, D, N2 subregions (Figure 3A) impact prediction accuracy, N1, D, N2 subregion sequences of public and private clones from one nBC repertoire (Dataset 1) were randomly assembled to N1-D-N2 simulated sequences. These simulations were performed 1'500 times: for each simulation run, the individual N1, D, N2 subregion lengths were held constant for both public and private clones and SVM parameters were chosen as for Figure 4A.

(B) Public and private clone balanced accuracy as a function of dataset size. From the largest 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 Figure 4A). For each simulated dataset, the number of public and private clones drawn was kept identical.
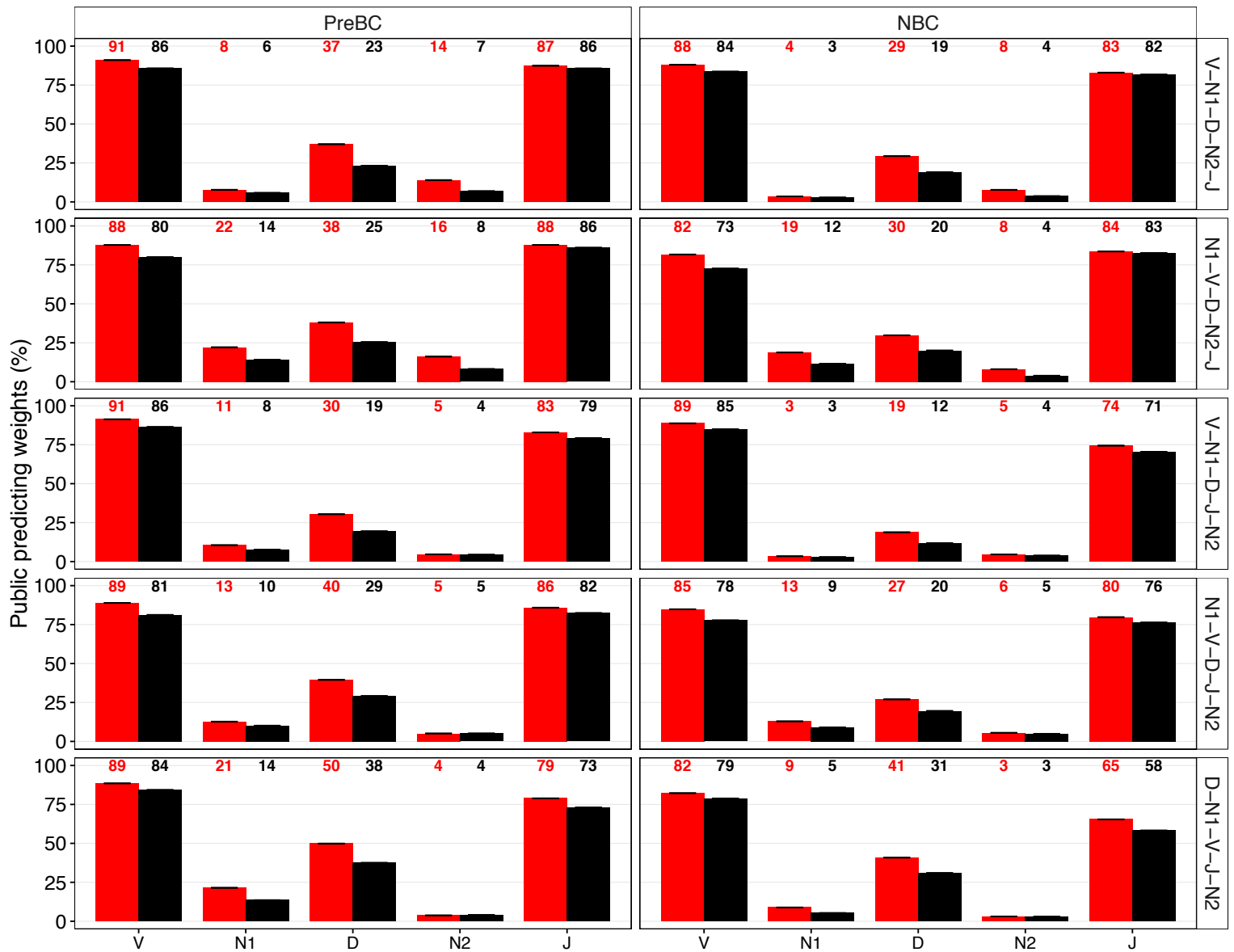
Supplementary Figure 8
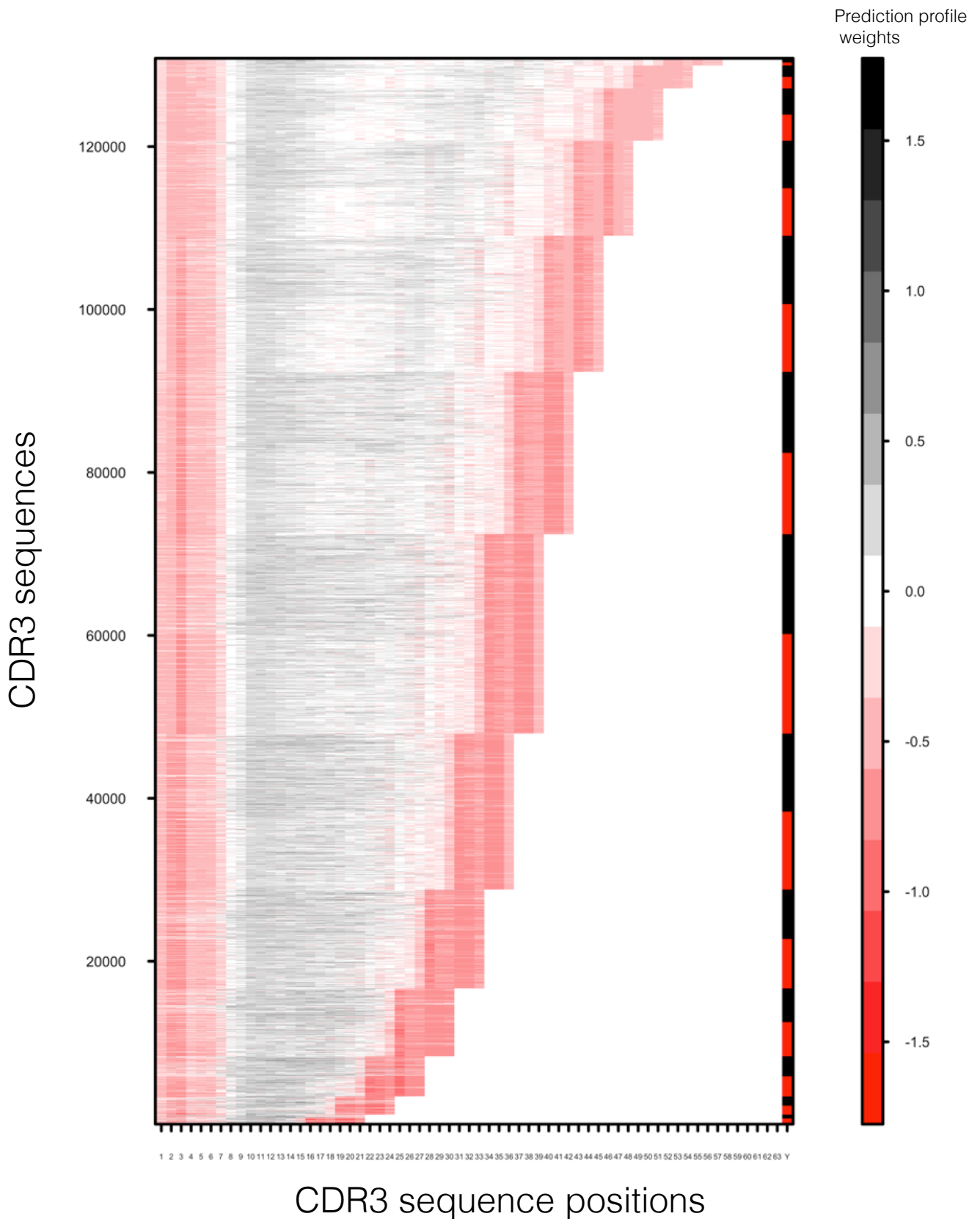
Supplementary Figure 8 Prediction profiles are shown for all public and private CDR3 sequences of an exemplary naïve B-cell repertoire (Dataset 1). See Figure 5 and Methods for explanation of prediction profiles. Legend: red: public clone, black: private clone.

Supplementary Figure 9

Supplementary Figure 9 Quantitative prediction profile analysis by V, N1, D, N2, J subregion. For each of the five SVM cases shown in Supplementary Figure 6E, the percentage of public-predicting SVM weights (<0) was determined by subregion (V, N1, D, N2, J), public/private status and B-cell population across all CDR3 lengths (not solely CDR3 length 39 as shown in Figure 5B). Barplots show mean±s.e.m.

Supplementary Figure 10
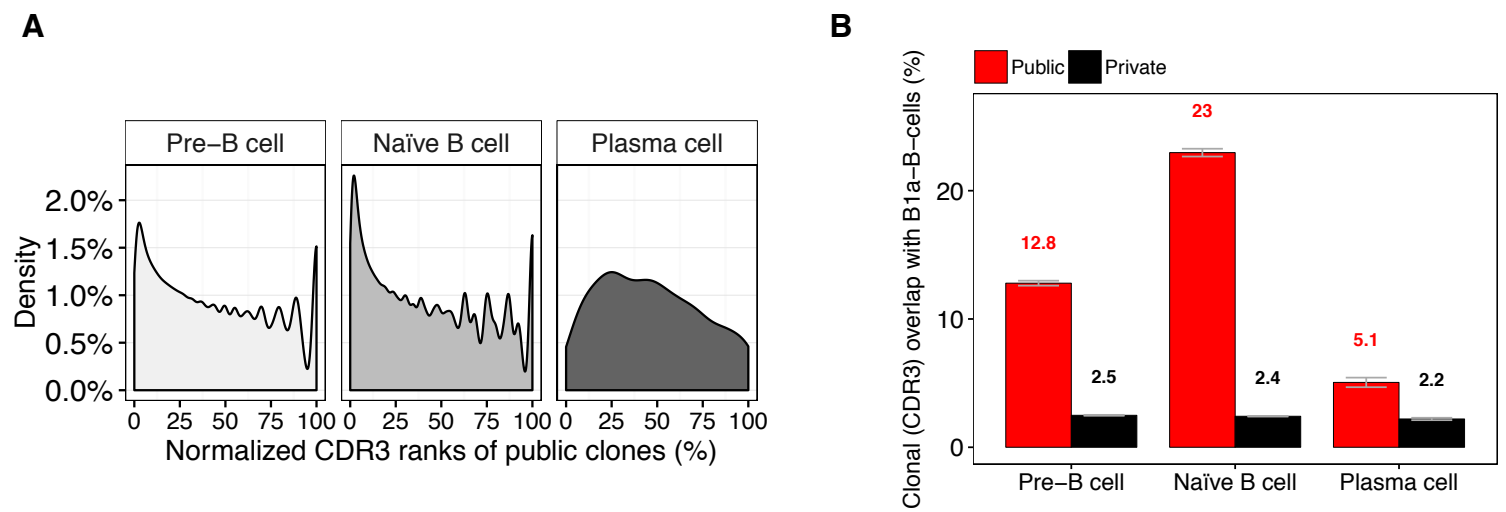
Supplementary Figure 10 Murine B2-B-cell public clones are biased towards higher frequencies and are enriched in natural antibody producing B1a-B-cells.

(A) Mean normalized rank of public CDR3s among all clones within a repertoire. PreBC and nBC public clones are more likely to occur at higher frequency then expected at random (uniform distribution).

(B) PreBC, nBC and PC (Dataset 1) public and private clone overlap with B1a-B-cells (Dataset 1). Differences between public and private clones were significant (p<0.05). Barplots show mean±s.e.m.

Supplementary Table 1

Data origin	Cell type	#CDR3
Dataset 1 **mouse (C57BL/6J), B cell** **(Greiff et al., 2016)**	pre-B cell, (IgM) (19) naïve B cell, (IgM) (19) plasma cell, (IgG) (19)	48'682 ± 10'493 177'197 ± 28'393 51 ± 30
Dataset 2 **mouse (Balb/c), B cell** **(Greiff et al., 2016)**	naïve B cell, (IgM) (4)	244'067 ± 11'293
Dataset 3 **mouse (pet shop), B cell** **(Greiff et al., 2016)**	naïve B cell (IgM) (3)	34'218 ± 1'207
Dataset 4 **mouse (C56BL/6J), B cell** **(Yang et al., 2015)**	B-1a B cells (43) marginal zone B cell (7) follicular B cell (5) pB2 B cell (5)	2'867 ± 1'742 2'987 ± 1'151 2'295 ± 406 1'519 ± 429
Dataset 5 **human (healthy), B cell** **(DeWitt et al., 2016)**	naïve B cell, (IgM) (3) memory B cell, (IgM, IgG) (3)	289'598 ± 36'627 35'221 ± 5'163
Dataset 6 **mouse (C57BL/6J), T cell** **(Madi et al., 2015)**	CD4 T cell (28)	2'621 ± 1'777

Supplementary Table 1 Size of CDR3-length equilibrated datasets used for SVM classification. For each of the six datasets used in this study (see Methods), a dataset of CDR3-length equilibrated sequences was constructed consisting of 50% public and 50% private CDR3 sequences. Mean and standard deviation across all samples of a given dataset and B/T cell population are displayed. Numbers in brackets indicate sample size.

Supplementary Table 2

	Total number (38 preBC/ nBC samples)	Extrapolated theoretical diversity	Coverage of extrapolated theoretical diversity (%)
V	198	198	100
N1	$10^{5.4}$	$10^{7.4}$	71
D	$10^{4.7}$	$10^{4.7}$	100
N2	$10^{5.2}$	$10^{6.9}$	74
J	58	58	100

Supplementary Table 2 CDR3 subregion diversity was sampled with >70% coverage. 1st column: cumulative (across preBC and nBC, Dataset 1) species richness of each CDR3 subregion, 2nd column: extrapolation of theoretical diversity using nonlinear regression (see Methods), 3rd column: percentage ratio of 1st and 2nd column. The approximate naïve murine nucleotide CDR3 diversity can be determined by calculating the product of the entries in the 2nd column $\approx 198 \times 10^{7.4} \times 10^{4.7} \times 10^{6.9} \times 58 \approx 10^{23}$.