Immunoglobulin heavy chains are sufficient to determine most B cell clonal relationships

Running title: Heavy chains can determine clones with confidence

Julian Q. Zhou[1], Steven H. Kleinstein[1,2 *]

[1] Interdepartmental Program in Computational Biology and Bioinformatics, Yale University

[2] Department of Pathology and Department of Immunobiology, Yale School of Medicine

[*] To whom correspondence should be addressed (steven.kleinstein@yale.edu)

Abstract

B cell clonal expansion is vital for adaptive immunity. High-throughput B cell receptor (BCR) sequencing enables investigating this process, but requires computational inference to identify clonal relationships. This inference usually relies on only the BCR heavy chain, as most current protocols do not preserve heavy:light chain pairing. The extent to which paired light chains aids inference is unknown. Using human single-cell paired BCR datasets, we assessed the ability of heavy chain-based clonal clustering to identify clones. Of the expanded clones identified, <20% grouped cells expressing inconsistent light chains. Heavy chains from these misclustered clones contained more distant junction sequences and shared fewer V segment mutations than the accurate clones. This suggests that additional heavy chain information could be leveraged to refine clonal relationships. Conversely, light chains were insufficient to refine heavy chain-based clonal clusters. Overall, the BCR heavy chain alone is sufficient to identify clonal relationships with confidence.

Introduction

B cell-mediated immunity relies on immunoglobulin (Ig) antibodies produced as a result of B cell clonal expansion. A B cell receptor (BCR) is the membrane-bound form of an antibody, and is made up of heavy and light polypeptide chains paired in a heterodimeric fashion. Each chain contains a variable (V) region, and together the V regions from the heavy and light chains form the antigen-binding sites (1). The V regions are formed via V(D)J recombination (2). In human, this shuffling process brings together one gene each from numerous IGHV, IGHD, and IGHJ genes for the heavy chain V (VH) region; and one gene each from either IGKV and IGKJ genes, or IGLV and IGLJ genes for, respectively, the κ or the λ light chain V (VL) region. Enzyme-mediated editing of the V(D)J junctions and the pairing of heavy and light chains inject additional diversity (1). During adaptive immune responses, B cells proliferate and further diversify via somatic hypermutation (SHM), forming clones consisting of cells which originated from the same V(D)J recombinant events, yet whose BCRs differ at the nucleotide level. As a result, each BCR is largely unique, with recent estimate suggesting 10^{16}-10^{18} unique paired antibodies in the circulating repertoire (3).

Adaptive Immune Repertoire Receptor sequencing (AIRR-seq) allows for high-throughput profiling of the diverse BCR repertoire via full-length V(D)J sequencing in bulk (4). An ensuing challenge is to computationally infer B cell clonal relationships (5–7). This step is of great importance as the assessment of repertoire properties such as diversity (8) and mutational load (9) depends on proper identification of clones, as does the reconstruction of B cell clonal lineage (10, 11) for tracing isotype switching (12, 13) and the development of antigen-specific (14) and/or broadly neutralizing (15) antibodies. To infer clones, differences at the sequence

45 nucleotide level, especially the high diversity in the CDR3 region, can serve as "fingerprints"
46 (16). Likelihood-based (17) and distance-based (18–21) approaches exist. For instance, cells
47 sharing the same IGHV and IGHJ genes, and whose heavy chain junctional sequences are
48 sufficiently similar based on a fixed (18–20) or adaptive (21) distance threshold, may be
49 clustered as clones. For validation, existing methods used simulated and experimental heavy
50 chain sequences (17, 20, 21), measuring the fractions of sequences inferred to be clonally
51 unrelated and related of being, respectively, truly unrelated and related (specificity and
52 sensitivity). Recently, Nouri & Kleinstein reported both metrics at over 98% (21).
53
54 The majority of current BCR repertoire studies utilizes bulk sequencing (22, 23), during which
55 VH:VL pairing is lost (24, 25). In the absence of VH:VL pairing, computational methods for
56 identifying clones have focused on heavy chain BCR data. This is justified under the assumption
57 that heavy chain junctional diversity alone should be sufficiently high such that, even without
58 light chains, the likelihood of clonally unrelated cells being clustered together will be negligibly
59 small (20). This reasoning has yet to be rigorously tested with experimental data. Recent
60 breakthroughs in single-cell BCR sequencing technology have enabled the recovery of native
61 VH:VL pairing (26–28) and have been applied to study vaccine responses (26, 29, 30), antibody
62 structure (31), and autoimmunity (32). We now have the opportunity to investigate the extent to
63 which inclusion of light chains impacts the ability to accurately detect B cell clonal relationships.
64
65 Using single-cell VH:VL paired BCR data, we assessed the performance of heavy chain-based
66 computational methods for identifying clones by measuring the extent to which the inferred
67 clonal members expressed consistent light chains sharing the same V and J genes and junction

length. We conclude that clonal members of the majority of the inferred clones exhibited light chain consistency. For the majority of the accurately inferred heavy chain-based clones, light chain information did not lead to further clonal clustering with greater granularity. At least some of the information gained from paired light chain data was apparent when considering the pattern of shared mutations in the heavy chain V segment, which is not considered in current distance-based clonal clustering methods, thus offering the potential for further improvements in heavy chain-based clonal inference.

75 **Materials and Methods**

76 *Single cell immune profiling datasets*

77 Four human datasets (**Supplemental Fig. 1A**) published by 10x Genomics on Aug 1, 2018 and
78 licensed under Creative Commons Attribution were downloaded
79 (https://support.10xgenomics.com/single-cell-vdj/datasets) on Nov 3, 2018. Two datasets were
80 sorted and produced by 10x Chromium direct Ig enrichment of, respectively, CD19+ B cells
81 isolated from PBMCs from a healthy donor, and GM12878 B-lymphoblastoid cell line. They
82 contain VH:VL paired reads for individual Ig-expressing cells. The other two datasets were
83 unsorted and produced by 10x Chromium V(D)J+5' gene expression profiling of, respectively,
84 PBMCs from a healthy donor, and squamous non-small cell lung carcinoma (NSCLC) tumor
85 cells from a fresh surgical resection. These contain single-cell 5' gene expression measurements,
86 and targeted enrichment with VH:VL pairing of Igs. All datasets were outputted by Cell Ranger
87 (v2.2.0), 10x Genomics' analysis pipeline. We used "filtered contigs" and "filtered gene-barcode
88 matrices" for Ig and gene expression respectively.

89

90 A fifth dataset (**Supplemental Fig. 1B**), as described by Croote et al., contains BCR contigs
91 covering full-length V(D)J segments reconstructed from single-cell RNA-seq of FACS-sorted
92 CD19+ lymphocytes from six food-allergic individuals (33).

93

94 *Germline V(D)J gene annotation of BCR contigs*

95 Germline V(D)J gene annotation was performed using IMGT/HighV-QUEST (34) and IgBLAST
96 (v1.10.0) (35). The germline reference used was IMGT/V-QUEST reference directory sets
97 release 201839-3 (Sept 26, 2018). The 10x Genomics datasets also contained annotations by Cell

98 Ranger's native algorithm. IMGT/HighV-QUEST annotations were used as final annotations
99 post-filtering.
100
101 *Filtering of BCR contigs and cells*
102 For all datasets, only productively-rearranged BCR contigs with valid V and J gene annotations,
103 consistent chain annotation (as opposed to, for example, IGHV1 and IGKJ2 for the same contig),
104 and junctions with nucleotide lengths being a multiple of 3 were used. A contig must meet all
105 abovementioned criteria based on annotations from all programs used (IMGT/HighV-QUEST,
106 IgBLAST, and, where applicable, Cell Ranger). Furthermore, only cells with exactly one heavy
107 chain contig paired with at least one light chain contig were examined. From the two unsorted
108 10x Genomics datasets with gene expression, we considered only cells displaying a
109 transcriptomic profile consistent with being a B cell. Taking into account of the high dropout rate
110 of single cell RNA-seq (36), B cells were defined as any cell with non-zero log-normalized
111 expression for any one of these genes: pan-B cell markers CD19, CD24, and CD72 (37),
112 plasmablast markers CD38 and MKI67 (38), and the isotype-encoding genes.
113
114 *Heavy chain-based B cell clonal clustering*
115 For each dataset, on a per-subject basis, we identified clones among the analyzed cells using
116 distance-based methods. We used, separately, spectral clustering, implemented via SCOPer
117 (v0.1.999 as of Nov 5, 2018) (21); and hierarchical clustering (20), implemented via SHazaM
118 (v0.1.10) and Change-O (v0.4.3) (39). Both methods first partitioned cells into groups sharing
119 the same combination of IGHV gene, IGHJ gene, and heavy chain junction length (heavy chain
120 VJL combination), where junction is defined as the IMGT-numbered codon 104 (conserved Cys)

121 to codon 118 (conserved Phe/Trp) (40). Within each group, based on distances among the heavy
122 chain junction sequences, a threshold was used to cluster cells within that group into clones. For
123 spectral clustering, adaptive thresholds were chosen by an unsupervised machine learning
124 algorithm. For hierarchically clustering, a subject-specific, fixed threshold was chosen upon
125 inspection of distance-to-nearest-neighbor plots (**Supplemental Fig. 1C**) (20, 41).
126
127 *Calculation of the number and frequency of nucleotide mutations*
128 The number and frequency of nucleotide mutations were calculated using the
129 "calcObservedMutations" function from SHazaM (v0.1.10) (39). For heavy chains, IMGT-
130 numbered IGHV positions 1 through 312 (40) were used. For light chains, IGK/LV positions
131 leading up to the junction region were used. To calculate the number of IGHV mutations, if any,
132 shared pairwise between cells from the same clone, we counted the number of positions at which
133 mutations involving the same nucleotide change were observed in both cells.

134 **Results**

135 We performed clonal relationship inference for five single-cell, VH:VL paired, human BCR
136 datasets, using only the heavy chain sequence from each cell. The datasets included four publicly
137 available ones from 10x Genomics and one described by Croote et al. (33) (**Materials &**
138 **Methods**). Of these, the B-lymphoblastoid GM12878 cell line dataset served as positive control,
139 as any clone present in a cell line culture can be expected to comprise genetically identical clonal
140 members. A distance-based spectral clustering method (21) was applied to identify clones for
141 each dataset. The datasets contained between 3 and 157 non-singleton clones (i.e., clones
142 containing at least two cells) (**Table 1**). Due to the small number of clones in each of the six
143 food-allergic individuals in the Croote et al. dataset, we aggregated those results for display after
144 performing analysis on a per-subject basis.

145

146 *Heavy chain-based clonal clustering is accurate for over 80% of clones*

147 To assess the extent to which heavy chain-based clonal clustering captures the underlying
148 biological truth of B cell clonal relationships, we examined whether cells clustered into the same
149 clone based on their heavy chains alone expressed consistent light chains. Specifically, within
150 the same clone, cells that are truly clonally related should carry light chains comprised of the
151 same combination of IGK/LV gene, IGK/LJ gene, and junction sequences of identical lengths
152 (hereafter referred to as the V-J-junction-length, or VJL, combination). An inferred clone was
153 considered accurate if all of its clonal members carried light chains with the same VJL
154 combination, and "misclustered" otherwise. Using spectral clustering with adaptive thresholds,
155 83% to 97% of the inferred clones were accurate (**Figure 1A**). Another distance-based
156 hierarchical clustering method using a fixed distance threshold (20) yielded similar results

157 (**Figure 1B**), and therefore we focus on presenting results from spectral clustering hereafter. To
158 test the possibility that the observed accuracy arose by chance, we randomly permuted the
159 VH:VL pairings of the cells, while maintaining their heavy chain-based clustering structures.
160 Across 100 permutations, only 1% to 6% (SD 1% to 8%) of the inferred clones were accurate by
161 chance (**Figure 1A**). Overall, these results show that heavy chain-based clonal clustering can
162 determine clonal relationships with reasonable confidence (>80%) in terms of light chain
163 consistency.
164
165 We next investigated the possibility that the observed level of confidence was deflated by factors
166 unrelated to the clonal clustering method itself. We considered the possibility that the
167 misclustered clones (**Supplemental Fig. 2A**) arose due to erroneous barcoding during
168 sequencing preparation, which has the potential to link together heavy and light chains from
169 unrelated cells. We reasoned that incorrectly paired heavy and light chains would show a
170 decreased correlation in their SHM frequencies relative to correctly paired chains. Thus, we
171 computed the Pearson correlation coefficient between IGHV and IGK/LV mutation frequencies
172 for cells expressing non-majority light chain VJL combinations from misclustered clones. We
173 found no significant difference (p=0.926, Fisher's r-to-z transformation and z-test) (42) between
174 the levels of correlation for misclustered clones (0.761) and for accurate clones (0.769),
175 suggesting that erroneous barcoding was unlikely a concern. In addition, we considered the
176 possibility of poor germline V/J gene annotation for the light chains leading to a false appearance
177 of light chain inconsistency. Since higher SHM frequencies are associated with increasing V(D)J
178 annotation errors, we compared the light chain mutation frequency in misclustered and accurate
179 clones. We found that the average IGK/LV mutation frequency across cells expressing non-

180 majority light chain VJL combinations in misclustered clones was not significantly higher than
181 that across cells in accurate clones (p=0.999; **Supplemental Fig. 2B**). Overall, these results
182 suggest that the observed confidence for accurately identifying clones using heavy chain-based
183 clonal clustering was not deflated by a false appearance of light chain inconsistency created by
184 erroneous barcoding or incorrect light chain germline gene annotation.
185
186 *Characteristics of misclustered clones suggest room for improvement in heavy chain-based*
187 *clustering*
188 Current distance-based, heavy chain-based clonal clustering methods utilize information
189 confined to VJL combination and distances between junction sequences to identify clonally-
190 related sequences. We investigated the characteristics of the heavy chains of cells from
191 misclustered clones to determine whether there was additional information in the heavy chains
192 that could improve the clustering. Gupta et al. noted that shorter heavy chain junction lengths
193 rendered lower and possibly insufficient diversity for effectively distinguishing clonal members
194 from non-clonal ones (20). However, we found no significant difference between the heavy
195 chain junction lengths of accurate clones and those of misclustered clones (p=0.486; **Figure 2A**).
196 Thus, it is unlikely that heavy chain junction length could serve as an effective indicator for
197 misclustered clones.
198
199 A central component of distance-based clonal clustering is the choice of a distance threshold that
200 determines how dissimilar the junction sequence can be before it is unlikely for cells to be
201 clonally related. To determine whether a better choice of distance threshold had the potential to
202 correct misclustered clones, we compared the maximum pairwise distance between heavy chain

203 junction sequences of cells in accurate clones, with the minimum pairwise distance between cells
204 carrying light chains with different VJL combinations in misclustered clones. Cells with
205 inconsistent light chains in misclustered clones had significantly more dissimilar heavy chain
206 junction sequences compared with cells in accurate clones ($p=3.3*10^{-10}$; **Figure 2B**). This
207 implies that some of the misclustered clones could have been corrected by using a numerically
208 lower (stricter) distance threshold, and that this lower threshold would not break apart the
209 accurate clones.
210
211 True B cell clones are expected to share mutations resulting from SHM followed by clonal
212 expansion and/or positive selection (19, 21). Hershberg & Prak suggested a minimum threshold
213 of four shared mutations for inferring clones (19). We investigated whether cells in misclustered
214 clones shared fewer mutations in their IGHVs compared to cells in accurate clones. To do so, we
215 compared the minimum number of shared mutations between cells in accurate clones, with the
216 maximum number of shared mutations between cells carrying light chains with different VJL
217 combinations in misclustered clones. We found that cells carrying inconsistent light chains in
218 misclustered clones shared significantly fewer mutations compared with cells in accurate clones
219 ($p=3.5*10^{-9}$; **Figure 2C**). Using Hershberg & Prak's threshold of four shared mutations (19), 26
220 of the 30 misclustered clones (87%) would not have been clustered together based on their heavy
221 chains (thus increasing specificity). On the other hand, within 4 of these misclustered clones, a
222 subset of cells with consistent light chains would also become separated (thus reducing
223 sensitivity). While the tradeoff between sensitivity and specificity needs further investigation,
224 these results suggest that the extent of shared mutations in the IGHV is a potentially useful
225 characteristic to consider in distance-based clonal clustering methods.

226

227 *Light chain information is insufficient for refining heavy chain-based clones*

228 Given the availability of paired light chains in the single-cell datasets that we analyzed, we
229 assessed the value added from that information. For misclustered clones (**Supplemental Fig.**
230 **2A**), this was trivial as any light chain inconsistency was immediately resolved by regrouping the
231 cells into smaller clusters based on their light chain VJL combinations. For accurate clones,
232 while the cells express consistent light chains, it is possible that these clusters may still contain
233 multiple true clones grouped together. We thus investigated the extent to which further clustering
234 cells based on the similarity of their light chain junctions (analogous to heavy chain-based
235 clustering) would further split the accurate clones. When clustering using the heavy chains, a
236 threshold around 0.2 normalized Hamming distance tended to separate clonally related cells from
237 unrelated ones (**Supplemental Fig. 1C**). However, applying this same threshold to cluster light
238 chains within accurate clones added virtually no information. The light chain junction regions of
239 cells in accurate clones were highly similar and significantly more so compared with their heavy
240 chain counterparts (Bonferroni-corrected p's<0.001) (**Supplemental Fig. 2C**). In all but four of
241 the accurate clones, the light chain junction sequence of a clonal member was at most 0.2
242 normalized Hamming distance away from the junction sequence of another clonal member most
243 similar to itself (**Supplemental Fig. 2C**). In other words, clonal members tended to carry light
244 chains with junction sequences that were at least 80% similar to each other.

245

246 We next investigated whether there would be further clustering based on light chain junctions at
247 lower distance thresholds ranging, in increasing order of stringency, from 0.15 to 0.05
248 normalized Hamming distance, while bearing in mind that one could always artificially yield

249 further clustering by imposing an increasingly stricter clustering threshold. At each clustering
250 threshold, we determined the percentage of heavy chain-based clones that were further clustered
251 on the basis of distances between their light chain junction sequences (**Table 2**). On average, 5%
252 of the heavy chain-based accurate clones inferred via spectral clustering were further clustered at
253 0.15, the most lenient threshold explored. Even at 0.05, the strictest threshold explored, only
254 23.2% of the accurate clones were further clustered. This threshold is approaching the mean light
255 chain SHM frequency, which ranges from 0.01 to 0.05 across the datasets, raising questions as to
256 whether such further clustering is artificial rather than biological. Overall, light chain
257 information does not support clonal clustering with greater granularity for the majority of heavy
258 chain-based accurate clones.

259 **Discussion**

260 In this study, we investigated the accuracy of heavy chain-based clonal inference. With single-
261 cell VH:VL paired BCR datasets, we performed B cell clonal inference using only the heavy
262 chains, effectively treating the datasets as if bulk-sequenced and unpaired. Over 80% of the
263 inferred clones were accurate as defined by light chain consistency. Within the majority of these
264 accurate clones, an additional round of clustering using the light chain sequence failed to yield
265 finer resolution (<10% at a threshold of 0.1 normalized Hamming distance). Including a
266 requirement that members of a clone share mutations between heavy chain sequences would
267 have corrected 87% of the misclustered clones, though at the expense of also breaking up the
268 accurately clustered part in 13% of these misclustered clones. Overall, while there remains
269 additional information from the heavy chain that could be leveraged for improvement, we found
270 heavy-chain based clustering alone capable of identifying clonal relationships with reasonable
271 confidence.

272

273 Cells from four of the misclustered clones (13%) carried closely related heavy chains sharing a
274 reasonable number of mutations in the V segment, yet they expressed light chains with different
275 VJL combinations. There are several possibilities for how such clones could arise. During B cell
276 maturation, the heavy chain rearranges first and the cell proliferates before the light chain
277 rearranges (1). Thus, it is possible that the misclustered clonal relationships we detect represent
278 daughter cells of the same heavy chain VDJ-rearranged B cell that developed different light
279 chains. Alternatively, a group of identical, autoreactive, immature B cells could have undergone
280 receptor editing, during which independent rearrangements gave rise to different light chains
281 paired with an identical heavy chain rearrangement. In these scenarios, the heavy chain-based

282 clonal clustering accurately reflects the underlying biology. However, we believe that these
283 scenarios are unlikely as the chance that cells with identical heavy chains but different light
284 chains would developmentally share the same temporal and/or spatial trajectories, in addition to
285 being sampled together for sequencing, is expected to be very small.
286
287 When examining light chain consistency of cells from an inferred clone, if more than one light
288 chain sequence was associated with a cell, we checked every sequence present for match against
289 the clonal majority VJL combination. We did not, however, consider the possible complication
290 where there was partial sampling in such a cell. Hypothetically, should two clonally related B
291 cells with dual light chains each have a different light chain sequenced, our analysis would have
292 considered a heavy chain-based clone containing these cells misclustered. However, such B cells
293 have been reported to be rare, especially outside autoimmunity, with dual-κ and dual-λ cells
294 estimated to occupy about 2-10% (43, 44) and 0.2% (45) of the normal murine repertoire,
295 consistent with the percentages of cells with multiple light chains observed in the single-cell
296 datasets here (1.1%-7.4%; all cells from the Croote et al. dataset had exactly one light chain).
297
298 Our results suggest several ways that heavy chain-based clonal clustering could be improved. For
299 example, heavy chain junctions were more similar to each other in accurate clones than in
300 misclustered clones, suggesting that the clustering threshold was perhaps too lenient for the
301 latter. We also observed that accurate clones shared more mutations in their V segment. Some
302 likelihood-based methods, such as *partis*, implicitly take into consideration shared mutations via
303 the use of a multi-Hidden Markov Model that simultaneously emits multiple sequences (17).
304 While computationally slower compared to the distance-based methods explored here, future

305 studies should explore the potential benefit of such methods that use both junction distance and
306 shared mutation patterns.
307
308 A limitation of this study is that all but one of the datasets contained relatively small numbers of
309 B cells, and none was sorted for specific subsets, such as memory B cells, that would enrich for
310 expanded clones. As a result, only a small number of the inferred clones contained multiple cells,
311 and were thus suitable for analysis. As high-quality, single-cell B cell datasets of higher
312 throughput with sorting for relevant B cell subsets emerge, similar analyses could be performed,
313 leading to better estimates of performance.
314
315 Clonal relationship inference is a crucial early step fundamental to computational BCR repertoire
316 analyses. As the volume of unpaired BCR data continues to grow as studies take advantage of
317 the relatively low cost of bulk BCR sequencing, current clonal clustering methods can determine
318 most clonal relationships with reasonable confidence using heavy chains only, and their
319 performance may continue to improve by taking into account of additional characteristics such as
320 the number of shared mutations in the heavy chains.

Acknowledgments

The authors thank Dr. Nima Nouri for critical reading of this manuscript. This work was supported by the High Performance Computing facilities operated by, and the staff of, the Yale Center for Research Computing.

Funding

This work was supported in part by the National Institutes of Health (NIH) under award number R01AI104739.

329 **References**

330 1. Murphy, K., and C. Weaver. 2017. *Janeway's Immunobiology*, 9th Edition. Garland Science,
331 New York, NY.

332 2. Tonegawa, S., C. Steinberg, S. Dube, and A. Bernardini. 1974. Evidence for somatic
333 generation of antibody diversity. *Proc. Natl. Acad. Sci. U. S. A.* 71: 4027–4031.

334 3. Briney, B., A. Inderbitzin, C. Joyce, and D. R. Burton. 2019. Commonality despite exceptional
335 diversity in the baseline human antibody repertoire. *Nature* 566: 393–397.

336 4. Robins, H. 2013. Immunosequencing: applications of immune repertoire deep sequencing.
337 *Curr. Opin. Immunol.* 25: 646–652.

338 5. Greiff, V., E. Miho, U. Menzel, and S. T. Reddy. 2015. Bioinformatic and Statistical Analysis
339 of Adaptive Immune Repertoires. *Trends Immunol.* 36: 738–749.

340 6. Yaari, G., and S. H. Kleinstein. 2015. Practical guidelines for B-cell receptor repertoire
341 sequencing analysis. *Genome Med.* 7: 121.

342 7. Chaudhary, N., and D. R. Wesemann. 2018. Analyzing Immunoglobulin Repertoires. *Front.*
343 *Immunol.* 9.

344 8. Gibson, K. L., Y.-C. Wu, Y. Barnett, O. Duggan, R. Vaughan, E. Kondeatis, B.-O. Nilsson, A.
345 Wikby, D. Kipling, and D. K. Dunn-Walters. 2009. B-cell diversity decreases in old age and is
346 correlated with poor health status. *Aging Cell* 8: 18–25.

347 9. Wang, C., Y. Liu, L. T. Xu, K. J. L. Jackson, K. M. Roskin, T. D. Pham, J. Laserson, E. L.
348 Marshall, K. Seo, J.-Y. Lee, D. Furman, D. Koller, C. L. Dekker, M. M. Davis, A. Z. Fire, and S.
349 D. Boyd. 2014. Effects of Aging, Cytomegalovirus Infection, and EBV Infection on Human B
350 Cell Repertoires. *J. Immunol.* 192: 603–611.

351 10. Hoehn, K. B., G. Lunter, and O. G. Pybus. 2017. A Phylogenetic Codon Substitution Model
352 for Antibody Lineages. *Genetics* 206: 417–427.
353 11. Davidsen, K., and F. A. I. Matsen. 2018. Benchmarking Tree and Ancestral Sequence
354 Inference for B Cell Receptor Sequences. *Front. Immunol.* 9.
355 12. Looney, T. J., J.-Y. Lee, K. M. Roskin, R. A. Hoh, J. King, J. Glanville, Y. Liu, T. D. Pham,
356 C. L. Dekker, M. M. Davis, and S. D. Boyd. 2016. Human B-cell isotype switching origins of
357 IgE. *J. Allergy Clin. Immunol.* 137: 579-586.e7.
358 13. Horns, F., C. Vollmers, D. Croote, S. F. Mackey, G. E. Swan, C. L. Dekker, M. M. Davis,
359 and S. R. Quake. 2016. Lineage tracing of human B cells reveals the in vivo landscape of human
360 antibody class switching. *eLife* 5: e16578.
361 14. Trück, J., M. N. Ramasamy, J. D. Galson, R. Rance, J. Parkhill, G. Lunter, A. J. Pollard, and
362 D. F. Kelly. 2015. Identification of Antigen-Specific B Cell Receptor Sequences Using Public
363 Repertoire Analysis. *J. Immunol.* 194: 252–261.
364 15. Zhu, J., X. Wu, B. Zhang, K. McKee, S. O'Dell, C. Soto, T. Zhou, J. P. Casazza, N. C. S.
365 Program, J. C. Mullikin, P. D. Kwong, J. R. Mascola, L. Shapiro, J. Becker, B. Benjamin, R.
366 Blakesley, G. Bouffard, S. Brooks, H. Coleman, M. Dekhtyar, M. Gregory, X. Guan, J. Gupta, J.
367 Han, A. Hargrove, S. Ho, T. Johnson, R. Legaspi, S. Lovett, Q. Maduro, C. Masiello, B.
368 Maskeri, J. McDowell, C. Montemayor, J. Mullikin, M. Park, N. Riebow, K. Schandler, B.
369 Schmidt, C. Sison, M. Stantripop, J. Thomas, P. Thomas, M. Vemulapalli, and A. Young. 2013.
370 De novo identification of VRC01 class HIV-1–neutralizing antibodies by next-generation
371 sequencing of B-cell transcripts. *Proc. Natl. Acad. Sci.* 110: E4088–E4097.
372 16. Dunn-Walters, D., C. Townsend, E. Sinclair, and A. Stewart. Immunoglobulin gene analysis
373 as a tool for investigating human immune responses. *Immunol. Rev.* 284: 132–147.

374 17. Ralph, D. K., and F. A. Matsen Iv. 2016. Likelihood-Based Inference of B Cell Clonal
375 Families. *PLOS Comput. Biol.* 12: e1005086.
376 18. Kepler, T. B., S. Munshaw, K. Wiehe, R. Zhang, J.-S. Yu, C. W. Woods, T. N. Denny, G. D.
377 Tomaras, S. M. Alam, M. A. Moody, G. Kelsoe, H.-X. Liao, and B. F. Haynes. 2014.
378 Reconstructing a B-Cell Clonal Lineage. II. Mutation, Selection, and Affinity Maturation. *Front.*
379 *Immunol.* 5.
380 19. Hershberg, U., and E. T. L. Prak. 2015. The analysis of clonal expansions in normal and
381 autoimmune B cell repertoires. *Phil Trans R Soc B* 370: 20140239.
382 20. Gupta, N. T., K. D. Adams, A. W. Briggs, S. C. Timberlake, F. Vigneault, and S. H.
383 Kleinstein. 2017. Hierarchical Clustering Can Identify B Cell Clones with High Confidence in Ig
384 Repertoire Sequencing Data. *J. Immunol.* 198: 2489–2499.
385 21. Nouri, N., and S. H. Kleinstein. 2018. A spectral clustering-based method for identifying
386 clones from high-throughput B cell repertoire sequencing data. *Bioinforma. Oxf. Engl.* 34: i341–
387 i349.
388 22. Hou, X.-L., L. Wang, Y.-L. Ding, Q. Xie, and H.-Y. Diao. 2016. Current status and recent
389 advances of next generation sequencing techniques in immunological repertoire. *Genes Immun.*
390 17: 153–164.
391 23. Nielsen, S. C. A., and S. D. Boyd. Human adaptive immune receptor repertoire analysis—
392 Past, present, and future. *Immunol. Rev.* 284: 9–23.
393 24. Georgiou, G., G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, and S. R. Quake.
394 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat.*
395 *Biotechnol.* 32: 158–168.

396 25. Friedensohn, S., T. A. Khan, and S. T. Reddy. 2017. Advanced Methodologies in High-
397 Throughput Sequencing of Immune Repertoires. *Trends Biotechnol.* 35: 203–214.
398 26. DeKosky, B. J., G. C. Ippolito, R. P. Deschner, J. J. Lavinder, Y. Wine, B. M. Rawlings, N.
399 Varadarajan, C. Giesecke, T. Dörner, S. F. Andrews, P. C. Wilson, S. P. Hunicke-Smith, C. G.
400 Willson, A. D. Ellington, and G. Georgiou. 2013. High-throughput sequencing of the paired
401 human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* 31: 166–169.
402 27. Busse, C. E., I. Czogiel, P. Braun, P. F. Arndt, and H. Wardemann. 2014. Single-cell based
403 high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *Eur. J.*
404 *Immunol.* 44: 597–603.
405 28. Zheng, G. X. Y., J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo,
406 T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G.
407 Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson,
408 R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J.
409 Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H.
410 Bielas. 2017. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8:
411 14049.
412 29. Lavinder, J. J., Y. Wine, C. Giesecke, G. C. Ippolito, A. P. Horton, O. I. Lungu, K. H. Hoi,
413 B. J. DeKosky, E. M. Murrin, M. M. Wirth, A. D. Ellington, T. Dörner, E. M. Marcotte, D. R.
414 Boutz, and G. Georgiou. 2014. Identification and characterization of the constituent human
415 serum antibodies elicited by vaccination. *Proc. Natl. Acad. Sci.* 111: 2259–2264.
416 30. Tan, Y.-C., L. K. Blum, S. Kongpachith, C.-H. Ju, X. Cai, T. M. Lindstrom, J. Sokolove, and
417 W. H. Robinson. 2014. High-throughput sequencing of natively paired antibody chains provides

418 evidence for original antigenic sin shaping the antibody response to influenza vaccination. *Clin.*
419 *Immunol.* 151: 55–65.
420 31. DeKosky, B. J., O. I. Lungu, D. Park, E. L. Johnson, W. Charab, C. Chrysostomou, D.
421 Kuroda, A. D. Ellington, G. C. Ippolito, J. J. Gray, and G. Georgiou. 2016. Large-scale sequence
422 and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc.*
423 *Natl. Acad. Sci.* 113: E2636–E2645.
424 32. Roy, B., R. S. Neumann, O. Snir, R. Iversen, G. K. Sandve, K. E. A. Lundin, and L. M.
425 Sollid. 2017. High-Throughput Single-Cell Analysis of B Cell Receptor Usage among
426 Autoantigen-Specific Plasma Cells in Celiac Disease. *J. Immunol.* 199: 782–791.
427 33. Croote, D., S. Darmanis, K. C. Nadeau, and S. R. Quake. 2018. High-affinity allergen-
428 specific human antibodies cloned from single IgE B cell transcriptomes. *Science* 362: 1306–
429 1309.
430 34. Li, S., M.-P. Lefranc, J. J. Miles, E. Alamyar, V. Giudicelli, P. Duroux, J. D. Freeman, V. D.
431 A. Corbin, J.-P. Scheerlinck, M. A. Frohman, P. U. Cameron, M. Plebanski, B. Loveland, S. R.
432 Burrows, A. T. Papenfuss, and E. J. Gowans. 2013. IMGT/HighV QUEST paradigm for T cell
433 receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat.*
434 *Commun.* 4: 2333.
435 35. Ye, J., N. Ma, T. L. Madden, and J. M. Ostell. 2013. IgBLAST: an immunoglobulin variable
436 domain sequence analysis tool. *Nucleic Acids Res.* 41: W34–W40.
437 36. Neu, K. E., Q. Tang, P. C. Wilson, and A. A. Khan. 2017. Single-Cell Genomics:
438 Approaches and Utility in Immunology. *Trends Immunol.* 38: 140–149.
439 37. LeBien, T. W., and T. F. Tedder. 2008. B lymphocytes: how they develop and function.
440 *Blood* 112: 1570–1580.

441 38. Fink, K. 2012. Origin and Function of Circulating Plasmablasts during Acute Viral
442 Infections. *Front. Immunol.* 3.
443 39. Gupta, N. T., V. Vander Heiden, M. Uduman, D. Gadala-Maria, G. Yaari, and S. H.
444 Kleinstein. 2015. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire
445 sequencing data. *Bioinformatics* 31: 3356–3358.
446 40. Lefranc, M.-P., C. Pommié, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-
447 Contet, and G. Lefranc. 2003. IMGT unique numbering for immunoglobulin and T cell receptor
448 variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.* 27: 55–77.
449 41. Glanville, J., T. C. Kuo, H.-C. von Büdingen, L. Guey, J. Berka, P. D. Sundar, G. Huerta, G.
450 R. Mehta, J. R. Oksenberg, S. L. Hauser, D. R. Cox, A. Rajpal, and J. Pons. 2011. Naive
451 antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation.
452 *Proc. Natl. Acad. Sci.* 108: 20066–20071.
453 42. Cohen, J., P. Cohen, S. G. West, and L. S. Aiken. 2003. *Applied Multiple*
454 *Regression/Correlation Analysis for the Behavioral Sciences*, Third Edition. Lawrence Erlbaum
455 Associates, Mahwah, New Jersey.
456 43. Casellas, R., Q. Zhang, N.-Y. Zheng, M. D. Mathias, K. Smith, and P. C. Wilson. 2007. Igκ
457 allelic inclusion is a consequence of receptor editing. *J. Exp. Med.* 204: 153–160.
458 44. Velez, M.-G., M. Kane, S. Liu, S. B. Gauld, J. C. Cambier, R. M. Torres, and R. Pelanda.
459 2007. Ig Allotypic Inclusion Does Not Prevent B Cell Development or Response. *J. Immunol.*
460 179: 1049–1057.
461 45. Pelanda, R. 2014. Dual immunoglobulin light chain B cells: Trojan horses of autoimmunity?
462 *Curr. Opin. Immunol.* 27: 53–59.
463

464 **Table 1: Numbers of heavy chain-based clones and clone sizes**

Dataset	Total number of clones	Number of non-singleton clones	Clone size of non-singleton clones (number of clones)
CD19+ B cells	8268	157	2 (136); 3(15); 4 (4); 5 (2)
B cells from NSCLC tumor	1247	98	2 (81); 3 (9); 4 (4); 6 (2); 8 (1); 14 (1)
B cells from food-allergic individuals	952	12	2 (10); 3 (1); 5 (1)
B cells from healthy PBMCs	1105	8	2 (6); 3 (2)
GM12878 cell line	5	3	2 (1); 5 (1); 790 (1)

465

466 **Table 2: Percentage of accurate heavy chain-based clones**

467 **further clustered based on light chains**

Clustering threshold for light chain junctions	*0.05*	*0.10*	*0.15*
Dataset	Percentage of heavy chain-based clones undergoing further clustering		
CD19+ B cells	20.3	6.8	3.0
B cells from NSCLC tumor	32.6	18.9	16.8
B cells from food-allergic individuals	40.0	10.0	0.0
B cells from healthy PBMCs	0.0	0.0	0.0
Average	23.2	8.9	5.0

468

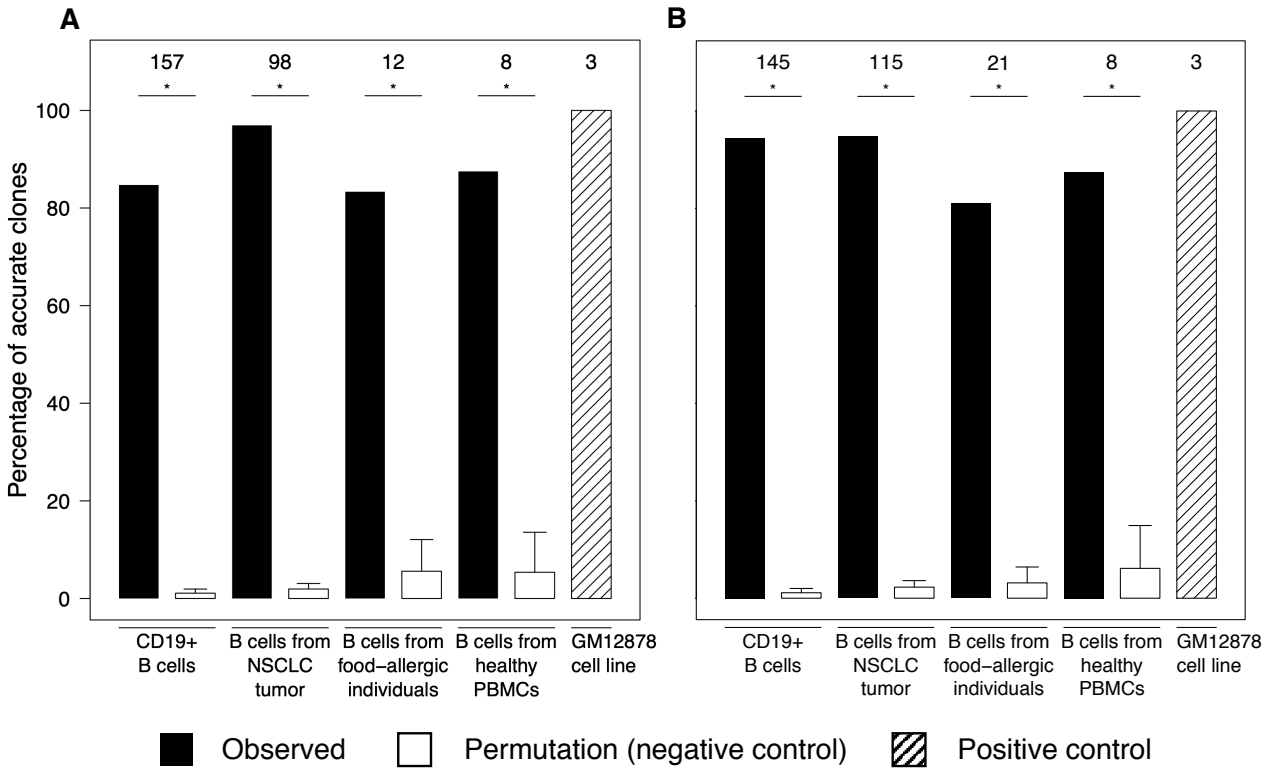
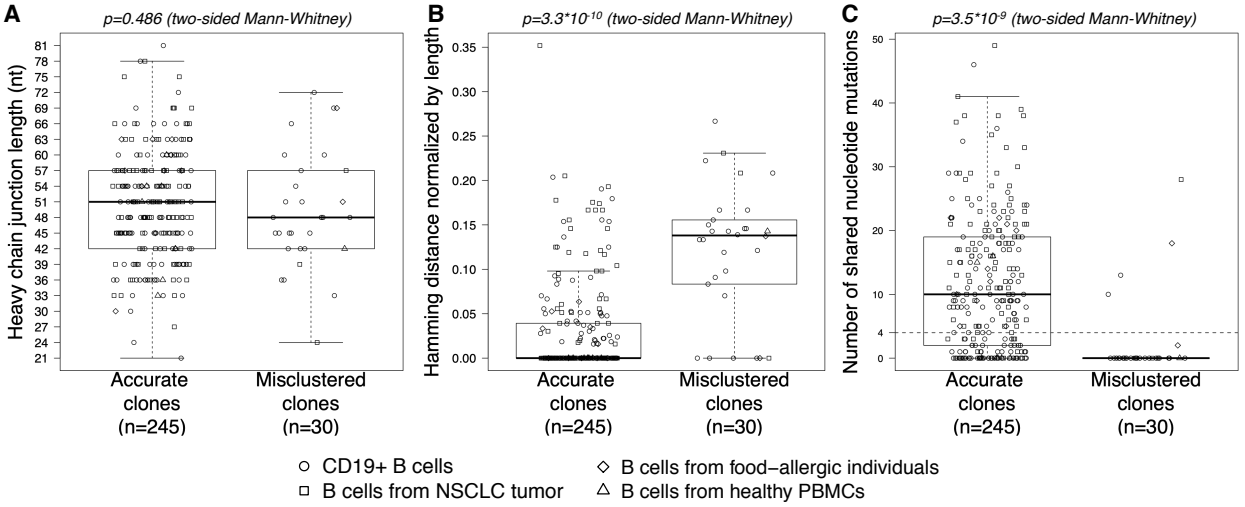


Figure 1. Performance of heavy chain-based **(A)** spectral clustering with adaptive threshold and **(B)** hierarchical clustering with fixed threshold. Solid and shaded bars show percentages of non-singleton inferred clones in which clonal members all carried light chains with the same V-J-junction-length (VJL) combination. Numbers at the top indicate the denominators. A background distribution was generated by permuting VH:VL pairings 100 times while maintaining inferred clonal lineage structures. Hollow bars show average percentages of accurate clones across permutations with standard errors. * denotes empirical one-sided, Bonferroni-corrected p-value < 0.05.



478

479 **Figure 2**. Characteristics of misclustered clones. **(A)** Heavy chain junction lengths of accurate

480 clones compared with misclustered clones. **(B)** The maximum pairwise distance between heavy

481 chain junction sequences in accurate clones, compared with the minimum pairwise distance

482 between cells expressing inconsistent light chains in misclustered clones. **(C)** The minimum

483 pairwise shared heavy chain V segment mutations in accurate clones, compared with the

484 maximum pairwise shared heavy chain V segment mutations between cells expressing

485 inconsistent light chains in misclustered clones.