Immunoglobulin heavy chains are sufficient to determine most B cell clonal relationships[1]

Running title: Heavy chains can determine clones with confidence

Julian Q. Zhou*, Steven H. Kleinstein*†§

* Interdepartmental Program in Computational Biology and Bioinformatics, Yale University

† Department of Pathology and Department of Immunobiology, Yale School of Medicine

§ Corresponding author:

Tel: +1 (203) 785-6685

Fax: +1 (203) 785-6486

Email: steven.kleinstein@yale.edu

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

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 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. The V regions are formed via V(D)J recombination. 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 (2).

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 (3). An ensuing challenge is to computationally infer B cell clonal relationships (4). This step is of great importance as the assessment of repertoire properties such as diversity (5) depends on proper identification of clones, as does the reconstruction of B cell clonal lineage (6) for tracing isotype switching (7) and antigen-specific (8) antibodies. To infer clones, differences at the sequence

nucleotide level, especially the high diversity in the CDR3 region, can serve as “fingerprints” (9). Likelihood-based (10) and distance-based (11–14) approaches exist. For instance, cells sharing the same IGHV and IGHJ genes, and whose heavy chain junctional sequences are sufficiently similar based on a fixed (11–13) or adaptive (14) distance threshold, may be clustered as clones. For validation, existing methods used simulated and experimental heavy chain sequences (10, 13, 14), measuring the fractions of sequences inferred to be clonally unrelated and related of being, respectively, truly unrelated and related (specificity and sensitivity). Recently, Nouri & Kleinstein reported both metrics at over 96% based on simulated data (14).

The majority of current BCR repertoire studies utilizes bulk sequencing (15), during which VH:VL pairing is lost (16). In the absence of VH:VL pairing, computational methods for identifying clones have focused on heavy chain BCR data. This is justified under the assumption that heavy chain junctional diversity alone should be sufficiently high such that, even without light chains, the likelihood of clonally unrelated cells being clustered together will be negligibly small (13). This reasoning has yet to be rigorously tested with experimental data. Recent breakthroughs in single-cell BCR sequencing technology have enabled the recovery of native VH:VL pairing (17, 18). We now have the opportunity to investigate the extent to which inclusion of light chains impacts the ability to accurately detect B cell clonal relationships.

Using single-cell VH:VL paired BCR data, we assessed the performance of heavy chain-based computational methods for identifying clones by measuring the extent to which the inferred 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.

Materials and Methods

Single cell immune profiling datasets

Four human datasets (**Supplemental Fig. 1A**) published by 10x Genomics for public use on 8/1/2018 were accessed (https://support.10xgenomics.com/single-cell-vdj/datasets) on 11/3/2018. Two datasets were sorted and produced by direct Ig enrichment of, respectively, CD19+ B cells isolated from PBMCs from a healthy donor, and GM12878 B-lymphoblastoid cell line. They contain VH:VL paired reads for individual cells. The other two datasets were unsorted and produced by V(D)J+5' gene expression profiling of, respectively, PBMCs from a healthy donor, and squamous non-small cell lung carcinoma (NSCLC) cells from a fresh surgical resection. These contain gene expression measurements and Ig enrichment with VH:VL pairing. All datasets were outputted by 10x Genomics via Cell Ranger (v2.2.0). We used "filtered contigs" and "filtered gene-barcode matrices" for Ig and gene expression respectively.

A fifth dataset (**Supplemental Fig. 1B**) from Croote et al. contains BCR contigs covering full-length V(D)J segments reconstructed from single-cell RNA-seq of FACS-sorted CD19+ B cells from six food-allergic individuals (19).

Germline V(D)J gene annotation of BCR contigs

Germline V(D)J gene annotation was performed using IMGT/HighV-QUEST and IgBLAST (v1.10.0). The germline reference used was IMGT release 201839-3. The 10x Genomics datasets also contained annotations by Cell Ranger. IMGT/HighV-QUEST annotations were used as final annotations post-filtering.

102

103 *Filtering of BCR contigs and cells*

104 For all datasets, only productively-rearranged BCR contigs with valid V and J gene annotations,
105 consistent chain annotation (excluding such contigs with IGHV and IGK/LJ), and junctions with
106 nucleotide lengths being a multiple of 3 were used. A contig must meet all abovementioned
107 criteria based on annotations from all programs used. Furthermore, only cells with exactly one
108 heavy chain contig paired with at least one light chain contig were examined. From the two
109 unsorted 10x Genomics datasets with gene expression, we considered only cells displaying a
110 transcriptomic profile consistent with being a B cell. Taking into account of the high dropout rate
111 of single cell RNA-seq, B cells were defined as any cell with non-zero log-normalized
112 expression for any one of these genes: pan-B cell markers CD19, CD24, and CD72 (20),
113 plasmablast markers CD38 and MKI67 (21), and the isotype-encoding genes.

114

115 *Heavy chain-based B cell clonal clustering*

116 For each dataset, on a per-subject basis, we identified clones using distance-based methods. We
117 used, separately, spectral clustering (SCOPer v0.1.999) (14) and hierarchical clustering (13)
118 (Change-O v0.4.3) (22). Both methods first partitioned cells into groups sharing the same
119 combination of IGHV gene, IGHJ gene, and heavy chain junction length (heavy chain VJL
120 combination), where junction is defined as the IMGT-numbered codon 104 (conserved Cys) to
121 codon 118 (conserved Phe/Trp) (23). 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**) (13).
126
127 *Calculation of the number and frequency of nucleotide mutations*
128 The number and frequency of nucleotide mutations were calculated based on IGH/K/LV
129 positions leading up to the junction region using the "calcObservedMutations" function from
130 SHazaM (v0.1.10) (22). To calculate the number of IGHV mutations shared pairwise between
131 cells from the same clone, we counted the number of positions at which mutations involving the
132 same nucleotide change were observed in both cells.

133 **Results**

134

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. (19) (**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 (14) 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 (13) 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) between the
174 levels of correlation for misclustered clones (0.761) and for accurate clones (0.769), suggesting
175 that erroneous barcoding was unlikely a concern. In addition, we considered the possibility of
176 poor germline V/J gene annotation for the light chains leading to a false appearance of light
177 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 (13). 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 (12, 14). Hershberg & Prak suggested a minimum threshold
213 of four shared mutations for inferring clones (12). 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 (12), 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.

Discussion

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

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

282 paired with an identical heavy chain rearrangement. In these scenarios, the heavy chain-based
283 clonal clustering accurately reflects the underlying biology. However, we believe that these
284 scenarios are unlikely as the chance that cells with identical heavy chains but different light
285 chains would developmentally share the same temporal and/or spatial trajectories, in addition to
286 being sampled together for sequencing, is expected to be very small.
287
288 When examining light chain consistency of cells from an inferred clone, if more than one light
289 chain sequence was associated with a cell, we checked every sequence present for match against
290 the clonal majority VJL combination. We did not, however, consider the possible complication
291 where there was partial sampling in such a cell. Hypothetically, should two clonally related B
292 cells with dual light chains each have a different light chain sequenced, our analysis would have
293 considered a heavy chain-based clone containing these cells misclustered. However, such B cells
294 have been reported to be rare, especially outside autoimmunity, with dual-κ and dual-λ cells
295 estimated to occupy about 2-10% (24, 25) and 0.2% (26) of the normal murine repertoire,
296 consistent with the percentages of cells with multiple light chains observed in the single-cell
297 datasets here (1.1%-7.4%; all cells from the Croote et al. dataset had exactly one light chain).
298
299 Our results suggest several ways that heavy chain-based clonal clustering could be improved. For
300 example, heavy chain junctions were more similar to each other in accurate clones than in
301 misclustered clones, suggesting that the clustering threshold was perhaps too lenient for the
302 latter. We also observed that accurate clones shared more mutations in their V segment. Some
303 likelihood-based methods, such as *partis*, implicitly take into consideration shared mutations via
304 the use of a multi-Hidden Markov Model that simultaneously emits multiple sequences (10).

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

Acknowledgments

The authors thank Dr. Nima Nouri for critical reading of this manuscript.

325 **References**

326

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

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

331 3. Robins, H. 2013. Immunosequencing: applications of immune repertoire deep sequencing.
332 *Curr. Opin. Immunol.* 25: 646–652.

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

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

338 6. Hoehn, K. B., G. Lunter, and O. G. Pybus. 2017. A Phylogenetic Codon Substitution Model
339 for Antibody Lineages. *Genetics* 206: 417–427.

340 7. Horns, F., C. Vollmers, D. Croote, S. F. Mackey, G. E. Swan, C. L. Dekker, M. M. Davis, and
341 S. R. Quake. 2016. Lineage tracing of human B cells reveals the in vivo landscape of human
342 antibody class switching. *eLife* 5: e16578.

343 8. Trück, J., M. N. Ramasamy, J. D. Galson, R. Rance, J. Parkhill, G. Lunter, A. J. Pollard, and
344 D. F. Kelly. 2015. Identification of Antigen-Specific B Cell Receptor Sequences Using Public
345 Repertoire Analysis. *J. Immunol.* 194: 252–261.

346 9. Dunn-Walters, D., C. Townsend, E. Sinclair, and A. Stewart. Immunoglobulin gene analysis
347 as a tool for investigating human immune responses. *Immunol. Rev.* 284: 132–147.

348 10. Ralph, D. K., and F. A. Matsen Iv. 2016. Likelihood-Based Inference of B Cell Clonal
349 Families. *PLOS Comput. Biol.* 12: e1005086.
350 11. Kepler, T. B., S. Munshaw, K. Wiehe, R. Zhang, J.-S. Yu, C. W. Woods, T. N. Denny, G. D.
351 Tomaras, S. M. Alam, M. A. Moody, G. Kelsoe, H.-X. Liao, and B. F. Haynes. 2014.
352 Reconstructing a B-Cell Clonal Lineage. II. Mutation, Selection, and Affinity Maturation. *Front.*
353 *Immunol.* 5.
354 12. Hershberg, U., and E. T. L. Prak. 2015. The analysis of clonal expansions in normal and
355 autoimmune B cell repertoires. *Phil Trans R Soc B* 370: 20140239.
356 13. Gupta, N. T., K. D. Adams, A. W. Briggs, S. C. Timberlake, F. Vigneault, and S. H.
357 Kleinstein. 2017. Hierarchical Clustering Can Identify B Cell Clones with High Confidence in Ig
358 Repertoire Sequencing Data. *J. Immunol.* 198: 2489–2499.
359 14. Nouri, N., and S. H. Kleinstein. 2018. A spectral clustering-based method for identifying
360 clones from high-throughput B cell repertoire sequencing data. *Bioinforma. Oxf. Engl.* 34: i341–
361 i349.
362 15. Nielsen, S. C. A., and S. D. Boyd. Human adaptive immune receptor repertoire analysis—
363 Past, present, and future. *Immunol. Rev.* 284: 9–23.
364 16. Georgiou, G., G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, and S. R. Quake.
365 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat.*
366 *Biotechnol.* 32: 158–168.
367 17. DeKosky, B. J., G. C. Ippolito, R. P. Deschner, J. J. Lavinder, Y. Wine, B. M. Rawlings, N.
368 Varadarajan, C. Giesecke, T. Dörner, S. F. Andrews, P. C. Wilson, S. P. Hunicke-Smith, C. G.
369 Willson, A. D. Ellington, and G. Georgiou. 2013. High-throughput sequencing of the paired
370 human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* 31: 166–169.

371 18. Busse, C. E., I. Czogiel, P. Braun, P. F. Arndt, and H. Wardemann. 2014. Single-cell based
372 high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *Eur. J.*
373 *Immunol.* 44: 597–603.
374 19. Croote, D., S. Darmanis, K. C. Nadeau, and S. R. Quake. 2018. High-affinity allergen-
375 specific human antibodies cloned from single IgE B cell transcriptomes. *Science* 362: 1306–
376 1309.
377 20. LeBien, T. W., and T. F. Tedder. 2008. B lymphocytes: how they develop and function.
378 *Blood* 112: 1570–1580.
379 21. Fink, K. 2012. Origin and Function of Circulating Plasmablasts during Acute Viral
380 Infections. *Front. Immunol.* 3.
381 22. Gupta, N. T., V. Vander Heiden, M. Uduman, D. Gadala-Maria, G. Yaari, and S. H.
382 Kleinstein. 2015. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire
383 sequencing data. *Bioinformatics* 31: 3356–3358.
384 23. Lefranc, M.-P., C. Pommié, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-
385 Contet, and G. Lefranc. 2003. IMGT unique numbering for immunoglobulin and T cell receptor
386 variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.* 27: 55–77.
387 24. Casellas, R., Q. Zhang, N.-Y. Zheng, M. D. Mathias, K. Smith, and P. C. Wilson. 2007. Igκ
388 allelic inclusion is a consequence of receptor editing. *J. Exp. Med.* 204: 153–160.
389 25. Velez, M.-G., M. Kane, S. Liu, S. B. Gauld, J. C. Cambier, R. M. Torres, and R. Pelanda.
390 2007. Ig Allotypic Inclusion Does Not Prevent B Cell Development or Response. *J. Immunol.*
391 179: 1049–1057.
392 26. Pelanda, R. 2014. Dual immunoglobulin light chain B cells: Trojan horses of autoimmunity?
393 *Curr. Opin. Immunol.* 27: 53–59.

394

395 **Figure Legends**

396

397 **Figure 1**. Performance of heavy chain-based **(A)** spectral clustering with adaptive threshold and
398 **(B)** hierarchical clustering with fixed threshold. Solid and shaded bars show percentages of non-
399 singleton inferred clones in which clonal members all carried light chains with the same V-J-
400 junction-length (VJL) combination. Numbers at the top indicate the denominators. A background
401 distribution was generated by permuting VH:VL pairings 100 times while maintaining inferred
402 clonal lineage structures. Hollow bars show average percentages of accurate clones across
403 permutations with standard errors. * denotes empirical one-sided, Bonferroni-corrected p-value <
404 0.05.

405

406 **Figure 2**. Characteristics of misclustered clones. **(A)** Heavy chain junction lengths of accurate
407 clones versus misclustered clones. **(B)** The maximum pairwise distance between heavy chain
408 junction sequences in accurate clones, versus the minimum pairwise distance between cells
409 expressing inconsistent light chains in misclustered clones. **(C)** The minimum pairwise shared
410 IGHV mutations in accurate clones, versus the maximum pairwise shared IGHV mutations
411 between cells expressing inconsistent light chains in misclustered clones.

1 **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)

2

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

2 **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

3

Figure 1

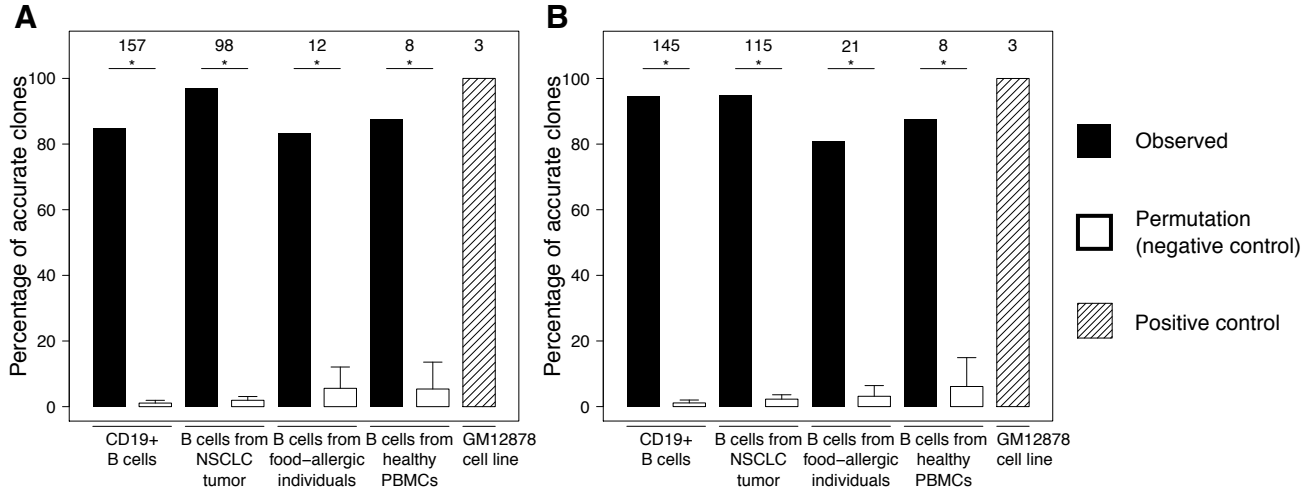
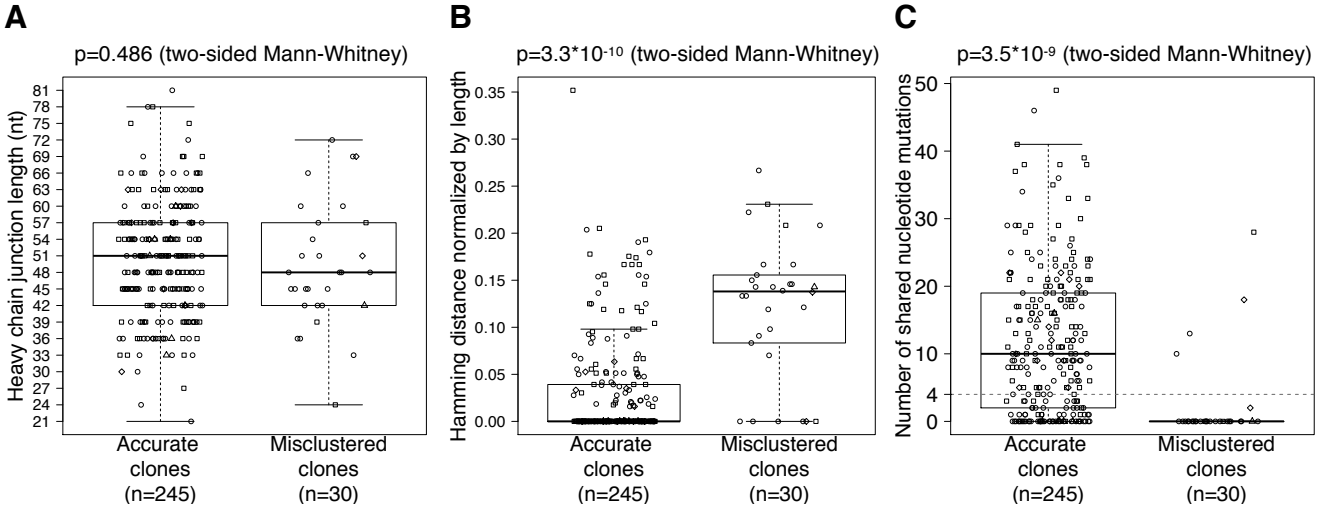

A
B
157
98
12
8
3
145
115
21
8
3
*
Percentage of accurate clones
0
20
40
60
80
100
CD19+ B cells
B cells from NSCLC tumor
B cells from food−allergic individuals
B cells from healthy PBMCs
GM12878 cell line
Observed
Permutation (negative control)
Positive control

Figure 2


A
p=0.486 (two-sided Mann-Whitney)
Heavy chain junction length (nt)
81
78
75
72
69
66
63
60
57
54
51
48
45
42
39
36
33
30
27
24
21
Accurate clones (n=245)
Misclustered clones (n=30)
B
p=3.3*10-10 (two-sided Mann-Whitney)
Hamming distance normalized by length
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Accurate clones (n=245)
Misclustered clones (n=30)
C
p=3.5*10-9 (two-sided Mann-Whitney)
Number of shared nucleotide mutations
50
40
30
20
10
4
0
Accurate clones (n=245)
Misclustered clones (n=30)
○ CD19+ B cells
□ B cells from NSCLC tumor
△ B cells from healthy PBMCs
◇ B cells from food−allergic individuals