Bioinformatics, 34, 2018, i341–i349
doi: 10.1093/bioinformatics/bty235
ISMB 2018

A spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data

Nima Nouri[1] and Steven H. Kleinstein[1,2,*]

[1]Department of Pathology, Yale School of Medicine, New Haven, CT 06520, USA and [2]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511, USA

*To whom correspondence should be addressed.

Abstract

Motivation: B cells derive their antigen-specificity through the expression of Immunoglobulin (Ig) receptors on their surface. These receptors are initially generated stochastically by somatic rearrangement of the DNA and further diversified following antigen-activation by a process of somatic hypermutation, which introduces mainly point substitutions into the receptor DNA at a high rate. Recent advances in next-generation sequencing have enabled large-scale profiling of the B cell Ig repertoire from blood and tissue samples. A key computational challenge in the analysis of these data is partitioning the sequences to identify descendants of a common B cell (i.e. a clone). Current methods group sequences using a fixed distance threshold, or a likelihood calculation that is computationally-intensive. Here, we propose a new method based on spectral clustering with an adaptive threshold to determine the local sequence neighborhood. Validation using simulated and experimental datasets demonstrates that this method has high sensitivity and specificity compared to a fixed threshold that is optimized for these measures. In addition, this method works on datasets where choosing an optimal fixed threshold is difficult and is more computationally efficient in all cases. The ability to quickly and accurately identify members of a clone from repertoire sequencing data will greatly improve downstream analyses. Clonally-related sequences cannot be treated independently in statistical models, and clonal partitions are used as the basis for the calculation of diversity metrics, lineage reconstruction and selection analysis. Thus, the spectral clustering-based method here represents an important contribution to repertoire analysis.

Availability and implementation: Source code for this method is freely available in the **SCOPe** (Spectral Clustering for clOne Partitioning) R package in the Immcantation framework: www.immcantation.org under the CC BY-SA 4.0 license.

Contact: steven.kleinstein@yale.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

B cell receptors (BCRs, also referred to as Immunoglobulins, (Igs)) are expressed by B cells and serve as the primary means for specific detection of foreign antigens. BCRs are comprised of two identical heavy and light chain proteins. BCRs exhibit extensive naive sequence diversity, which is generated through a somatic gene rearrangement process termed V(D)J recombination (Tonegawa, 1983). For heavy chain rearrangements, V(D)J recombination brings together one Variable (V) region gene with one Diversity (D) gene and one Joining (J) gene. For light chain rearrangements V genes are rearranged directly to J genes. Further diversity is generated at the junctions between these joining gene segments by N-addition, P-addition and exonucleolytic nibbling (Murphy, 2011). The large number of possible V(D)J gene segments, combined with junctional diversity, result in a theoretical diversity of $> 10^{14}$. During T-dependent responses, antigen-activated B cells undergo rapid proliferation and further diversification of their BCR by somatic hypermutation (SHM), an enzymatically-driven process introducing point substitutions into the Ig locus at a rate of ~1/1000 bp/cell division (Kleinstein *et al.*, 2003; McKean *et al.*, 1984). This clonal expansion and diversification is coupled to selection for binding to specific antigen, resulting in affinity maturation of the immune response (Bannard and Cyster, 2017; Victora and Nussenzweig, 2012). In healthy human adults, the average mutation frequency in the memory B cell population can reach close to 10% (Vander Heiden *et al.*, 2017). The total collection of BCRs in an individual, tissue, or cell subset is referred to as the 'repertoire' of the given cell population.

Rapid improvements in next-generation sequencing (NGS) technologies have revolutionized our ability to carry out large-scale adaptive immune receptor repertoire sequencing (AIRR-Seq) experiments (Boyd and Joshi, 2015). AIRR-Seq (Rubelt *et al.*, 2017) is increasingly being applied to profile the BCR repertoire to gain insights into the adaptive immune response in healthy individuals and in those with a wide range of diseases, including auto-immunity, infection, allergy, cancer and aging (Boyd *et al.*, 2009; Hershberg *et al.*, 2014; Logan *et al.*, 2011; Wang *et al.*, 2014). As NGS technologies continue to improve, these repertoire experiments are producing ever larger datasets, with tens- to hundreds-of-millions of sequences. A key computational challenge in the analysis of these data is partitioning the sequences to identify members of a clone, which arise from a shared V(D)J recombination event (naive cell) but can express different BCRs due to the accumulation of mutations introducing by SHM. Accurate identification of clonal relationships is important, as these clonal groups form the basis for a wide range of repertoire analysis, including diversity analysis, lineage reconstruction and detection of antigen-specific sequences (Yaari and Kleinstein, 2015).

Several methods have been proposed to partition a set of BCR sequences into clonally-related groups. These methods can broadly be divided into probabilistic (Kepler, 2013; Ralph and Matsen, 2016) and distance-based models (Glanville *et al.*, 2011; Gupta *et al.*, 2017; Nouri and Kleinstein, 2017). Probabilistic models work by calculating the likelihood of sharing a B cell ancestor and subsequently inferring the clones. The distance-based methods work by leveraging the high diversity of the junction region (i.e. complementarity determining region 3, plus the conserved flanking amino acid residues) as a 'fingerprint' to infer the B cell clones. In general, the likelihood-based approaches are computationally demanding, and distance-based approaches are more widely used for identifying clonally-related sequences. Many studies simply choose a threshold and define any sequences with junction region sequence similarity below this fixed threshold to be clonally-related; e.g. 10% and 15% are common thresholds (Hershberg and Prak, 2015; Jiang *et al.*, 2013; Meng *et al.*, 2017; Vander Heiden *et al.*, 2017). Glanville *et al.* (Glanville *et al.*, 2011) offered a subject-specific method of choosing this threshold using the distribution of distances of each sequence to its nearest (non-identical) neighbor. This distance distribution tends to be bi-modal, and it is assumed that the lower mode represents sequences with clonal relatives and the higher mode represents those without clonal relatives (i.e. singletons) in the dataset. In this case, a reasonable choice of distance threshold is one that separates these two modes (Jiang *et al.*, 2013). Building on this insight, Gupta *et al.* (Gupta *et al.*, 2017) developed an automated method to analyze this distance-to-nearest distribution and choose a threshold. They also analyzed different linkage approaches for clustering the sequences into groups (complete, average and single) as well as different measures of distance between junction regions. Gupta *et al.* found that applying a fixed distance threshold with single-linkage hierarchical clustering using Hamming distance normalized by junction length detected clones with high confidence on several simulated and experimental datasets. Recently, we further extended this hierarchical clustering-based technique by developing an approach to estimate the study-specific sensitivity and specificity for any choice of distance threshold, thus providing a quantitative basis for choosing a fixed threshold value for partitioning (Nouri and Kleinstein, 2017). Our method works by modeling the distance-to-nearest distribution as a mixture of two univariate curves, and then fitting the parameters of those curves [Fig. 1, panels A: (Stern *et al.*, 2014), B: (Parameswaran *et al.*, 2013), C: (Vander Heiden *et al.*, 2017) and D: (de Bourcy *et al.*, 2017)].

Despite their wide use, existing distance-based methods have several shortcomings. First, not all datasets exhibit bi-modality in their distance-to-nearest distribution making it difficult to justify the choice of a fixed threshold. This occurs, for example, in datasets composed of many expanded BCR clones which results in the absence of a second large distance peak [Fig. 1, panels E: (Stern *et al.*, 2014), F: (Meng *et al.*, 2017), G and H: (Ralph and Matsen, 2016)]. The single-linkage hierarchical clustering-based method described above will not work in these cases. Second, the use of a fixed threshold for partitioning sequences into distinct clones does not allow for widely varying levels of clonal diversification. In the case where this threshold is applied to the dendrogram produced by single-linkage clustering, the assessment of cluster quality is reduced to a single similarity between a pair of sequences irrespective of all others. This can result in long chains of sequences where every sequence is close to one other, but the entire set is widely dissimilar (the so-called 'chaining phenomenon' of single-linkage). For such a method to work well, the distribution of maximum-distance-within clusters (compactness) and minimum-distance-between clusters (isolation) should be distinguishable. In datasets where these two distributions overlap, the inferred clones are highly sensitive to the fixed threshold placed on the similarity measure, thus leading to heterogeneity in the inferred clones.

In this paper, we present a new technique for inferring BCR clones based on spectral clustering with an adaptive threshold to determine the local sequence neighborhood. This method does not require bi-modality in the distance-to-nearest distribution. Furthermore, by allowing the required level of junction sequence similarity to vary in different local neighborhoods, the inferred clones exhibit more homogeneity. We show that this approach performs better than a fixed threshold method on simulated and experimental datasets, and is also a reliable partitioning method when existing approaches fail to work. Finally, we show that the spectral clustering-based technique offers a significant improvement in computational speed.

2 Materials and methods

In following sections, we briefly discuss the general framework of clonal inference methods used in this study. As with many previous approaches, we first partition sequences into groups that share the same V gene, J gene and junction length (Boyd *et al.*, 2009; Glanville *et al.*, 2011; Jiang *et al.*, 2013; Stern *et al.*, 2014; Tsioris *et al.*, 2015). Next, each of these groups is subject to the spectral and hierarchical clustering-based methods.

2.1 Spectral clustering-based method

The spectral clustering-based method proceeds in five steps, as follows (an overview of the approach is shown in Supplementary Fig. S1A–D):

i. **Compute the similarity matrix**: Given a set of BCR sequences $\{x_1, x_2, \ldots, x_n\}$ we generate a symmetric matrix with entries s_{ij} defined by the Hamming distance between the junction regions of sequences x_i and x_j.

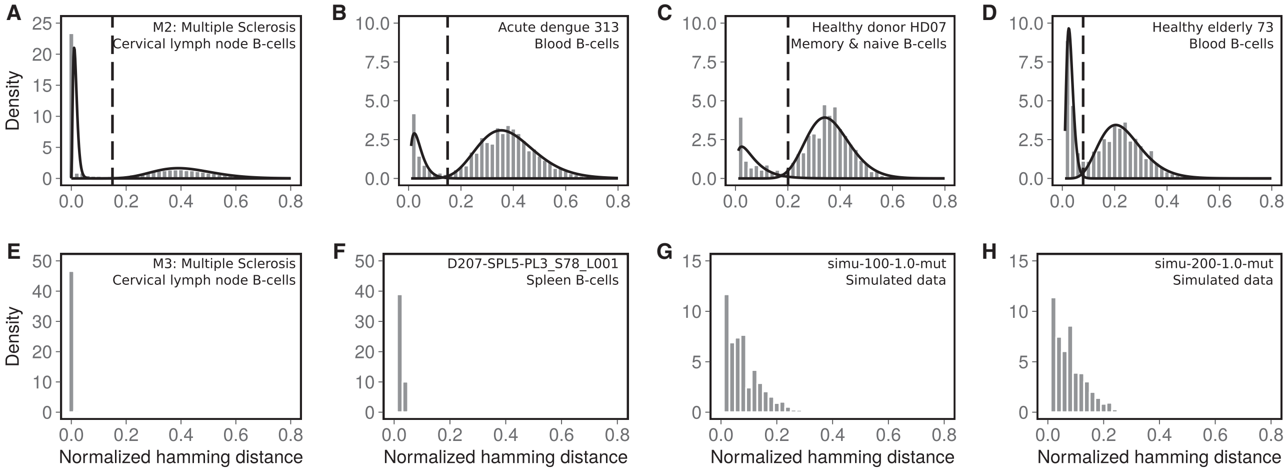

Fig. 1. The distance-to-nearest distribution can be bimodal or unimodal. BCR sequencing data was obtained for (**A**, **E**) cervical lymph node B cells from multiple sclerosis patients (Stern *et al.*, 2014), (**B**) peripheral blood B cells from a patient with an acute dengue virus infection (Parameswaran *et al.*, 2013), (**C**) sorted memory and naive B cells from a healthy donor blood sample (Vander Heiden *et al.*, 2017), (**D**) peripheral blood B cells from a healthy, older adult donor (de Bourcy *et al.*, 2017), (**F**) splenic B cells from an organ donor (Meng *et al.*, 2017) and (**G**, **H**) simulated data from Ralph and Matsen (Ralph and Matsen, 2016). Within each dataset, the nucleotide Hamming distance (normalized by junction length) from each sequence to every other sequence with the same V gene annotation, J gene annotation and junction length was calculated and the nearest (non-zero) neighbor was identified. Bi-modal 'distance-to-nearest' distributions (A, B, C, D) were fit to a mixture model of Gamma distributions (solid lines) in order to determine the optimal threshold (vertical dashed lines; Nouri and Kleinstein, 2017)

ii. **Compute the kernel matrix**: Given the (n, n) similarity matrix, we generate a fully connected graph such that its elements represent the local neighborhood relationship of each sequence to all other sequences (i.e. edges between sequences in local neighborhoods are connected with relatively high positive weights, while edges between far away sequences have smaller positive weights). This is implemented using a Gaussian kernel matrix with elements $k_{ij} = \exp(-s_{ij}^2/2\sigma_i\sigma_j)$, where the parameters σ_i and σ_j (SD) control the width of the neighborhoods corresponding to the sequences x_i and x_j, respectively. The SD is computed such that the width of neighborhood varies in different parts of the graph capturing a dynamic threshold among only those junction segments which have shown higher similarity than the other sequences. To calculate the scale parameter σ_i, the rank-ordered set of distances corresponding to the ith row of the similarity matrix s is examined to find the first largest gap in distance values. This gap is flagged as the neighborhood width. Finally, we compute the scale parameter σ_i associated with ith sequence as the SD of distances within this neighborhood (Fig. 2).

iii. **Compute the Laplacian matrix**: Given the (n, n) kernel matrix we generate the graph Laplacian defined as $L = D - K$, where D is a diagonal matrix defined as $D_{ii} = \sum_j A_{ij}$ (Mohar *et al.*, 1991; Mohar, 1997). Subsequently, we calculate the eigenvectors and eigenvalues of this matrix using the `eigen` function from **base** R package (version 3.4.3).

iv. **Determine the number of clusters**: Given the set of eigenvalues $\{0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n\}$ we infer the number of clusters, k, such that all eigenvalues $\lambda_1, \ldots, \lambda_k$ are very small ($\simeq 0$), but λ_{k+1} is relatively large. Therefore, the rank-ordered eigenvalues are examined to find the first largest gap where $\lambda_{k+1} > 0$, while the eigenvalue $\simeq 0$ has multiplicity up to kth eigenvalue. Then, the value k is used as the number of clusters (Von Luxburg, 2007).

v. **Clonal inference**: Given the number of clusters k, we perform k-means Euclidean distance-based clustering, using the `kmeans` function from **stats** R package (version 3.4.3), over the k eigenvectors associated with the smallest k eigenvalues to find the appropriate clones.

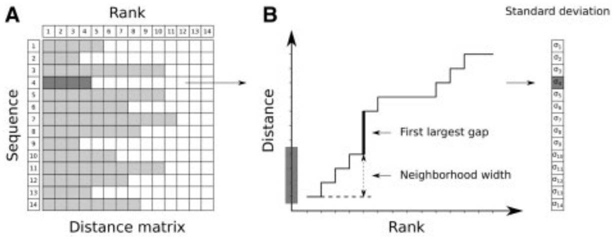

Fig. 2. Schematic overview for determination of scale parameter used in the spectral clustering-based method. (**A**) The Hamming distance of each unique sequence (rows) to every other sequence with the same V gene, J gene and junction length is determined and rank-ordered (columns). (**B**) For each row *i*, consecutive elements are examined to find the first largest gap in distance values, which is used to define the neighborhood width. The scale parameter σ_i associated to the *i*th sequence is determined as the SD of distances within this neighborhood (shaded areas in A)

2.2 Hierarchical clustering-based method

The hierarchical clustering-based method applied herein is described in Gupta *et al.* (Gupta *et al.*, 2017) and Nouri *et al.* (Nouri and Kleinstein, 2017); an overview of the approach is shown in Supplementary Figure S1E–H. Specifically, we use the bygroup subcommand of `DefineClones.py` in the **Change-O** package (version 0.3.9; Gupta *et al.*, 2015) and the `findThreshold` function from the **SHazaM** R package (version 0.1.9) with the default parameters.

3 Results

3.1 The spectral clustering-based method has high sensitivity and specificity

We first characterized the performance of the spectral clustering-based method on simulated data, where clonal relationships are known with certainty. Specifically, we used the simulated datasets

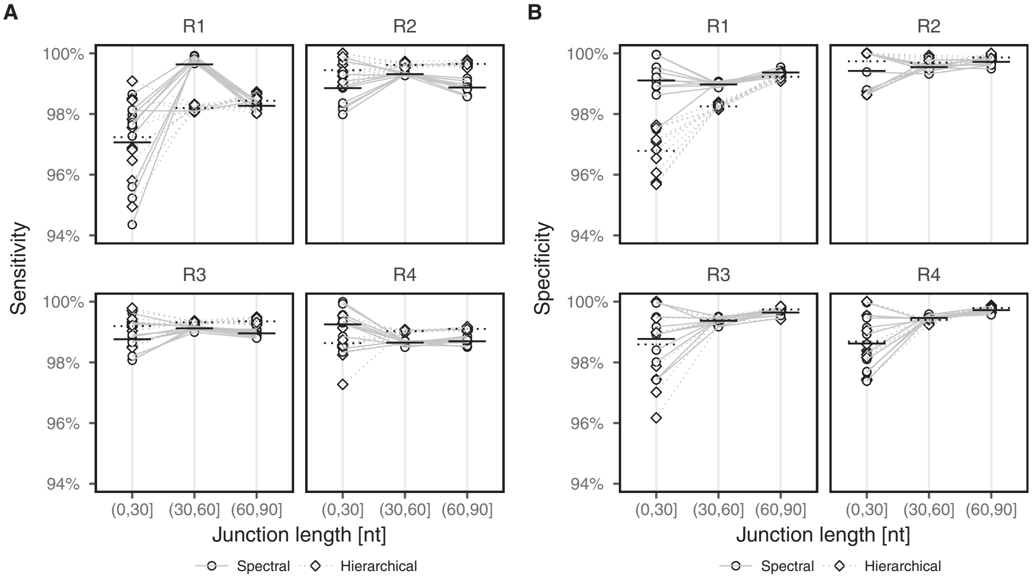

Fig. 3. The clustering-based method identifies clones with high confidence on simulated data. The spectral (circles) and hierarchical (diamonds) clustering-based methods were applied to identify clonally related sequences in 40 simulated datasets (10 datasets from each of 4 simulated individuals R1–R4) generated by Gupta *et al.* (Gupta *et al.*, 2017). Performance was assessed by calculating (**A**) sensitivity and (**B**) specificity on three junction-length domains. Mean performance is indicated by the solid bars (spectral) and dashed bars (hierarchical)

from Gupta *et al.* (Gupta *et al.*, 2017). These simulations start with a set of observed lineage tree topologies from lymph node samples from each of four individuals [M2, M3, M4 and M5 from Stern *et al.* (Stern *et al.*, 2014)], and generate a simulated dataset for each individual (R1, R2, R3 and R4, respectively) by randomly selecting a new germline sequence for every lineage and then stochastically re-introducing mutations along the lineage branches. This process was repeated 10 times for each individual to create a collection of 40 simulated datasets; R1.1–R1.10 ~ 86k, R2.1–R2.10 ~ 55k, R3.1–R3.10 ~ 85k and R4.1–R4.10 ~ 130k total reads. Using these data, the sensitivity of the method is defined as the fraction of all sequence pairs from the same clone that were correctly inferred by the method, while specificity is defined as the fraction of pairs of unrelated sequences that were successfully inferred by the method to be in different clones. As we previously found that specificity decreased as a function of junction length (Gupta *et al.*, 2017), we characterized performance separately over three different junction length domains: (i) $(0, 30]$ ($< 5\%$ of the sequences), (ii) $(30, 60]$ (~65% of the sequences) and (iii) $(60, 90]$ (~30% of the sequences) nucleotides.

The spectral clustering-based method achieved high sensitivity and specificity across all simulated datasets and junction lengths (Fig. 3). For junction lengths > 30, the sensitivity and specificity were $> 98\%$. Performance was significantly more variable for junction lengths in the $(0, 30]$ range. The decreased number of sequences in this range causes a small number of false-positives or false-negatives to have an outsized impact on the performance measure. Nevertheless, only moderately decreased performance was seen for junction lengths < 30, particularly sensitivity in R1, with sensitivity and specificity remaining $> 98\%$ for most simulations.

We next compared the performance of our new method with a current state-of-the-art method that uses single-linkage hierarchical clustering with a fixed distance threshold to partition sequences into clones (Nouri and Kleinstein, 2017; Methods Section). In this case, the threshold was chosen to maximize the average predicted sensitivity and specificity by fitting a mixture model to the distribution of distances of every sequence to its nearest, non-identical neighbor. Like the spectral clustering-based method, we found that the hierarchical clustering-based method achieved sensitivity and specificity $> 98\%$ for junction lengths > 30 and that this performance was moderately worse for junction lengths < 30. However, while performance was similar overall, the spectral clustering-based method had noticeably better sensitivity for junction lengths $(30, 60]$(nt) in R1, and better specificity for junction lengths $(0, 30]$ in R1 (Fig. 3). Overall, these results show that the spectral clustering-based method can identify clones with high confidence, on par or better than a state-of-the-art distance-based method on simulated data.

3.2 More shared mutations and higher homogeneity compared with hierarchical clustering

We next sought to compare the performance of the spectral and hierarchical clustering-based methods on experimental data. As a first comparison, we used the lymph node samples from the four individuals that the simulation was based on (Stern *et al.*, 2014). These are M2 (~100k total reads), M3 (~150k total reads), M4 (~200k total reads) and M5 (~400k total reads). Notably, while the spectral clustering-based method was able to partition the sequences in all four samples, the hierarchical clustering-based method failed to converge on an optimized threshold at which to cut the hierarchy for three of the samples (M3, M4 and M5). This is because only a single peak (at close distances) is apparent in the distance-to-nearest distributions for these samples, suggesting they are composed of highly expanded clones with few singleton sequences (Fig. 1E). In contrast, there is a clear bi-modal distribution in the sample from M2, which allows the mixture modeling optimization to work (Fig. 1A). Since we could not optimize the distance threshold for M3–M5, we chose to use a threshold of 0.15 normalized Hamming distance for these samples, which is consistent with many previous human studies (Hershberg and Prak, 2015; Meng *et al.*, 2017) and the same as the optimized threshold found for M2.

Most of the clones identified by the spectral and hierarchical clustering-based methods were identical, or highly overlapping, across all four individuals (Fig. 4A). The degree of overlap between

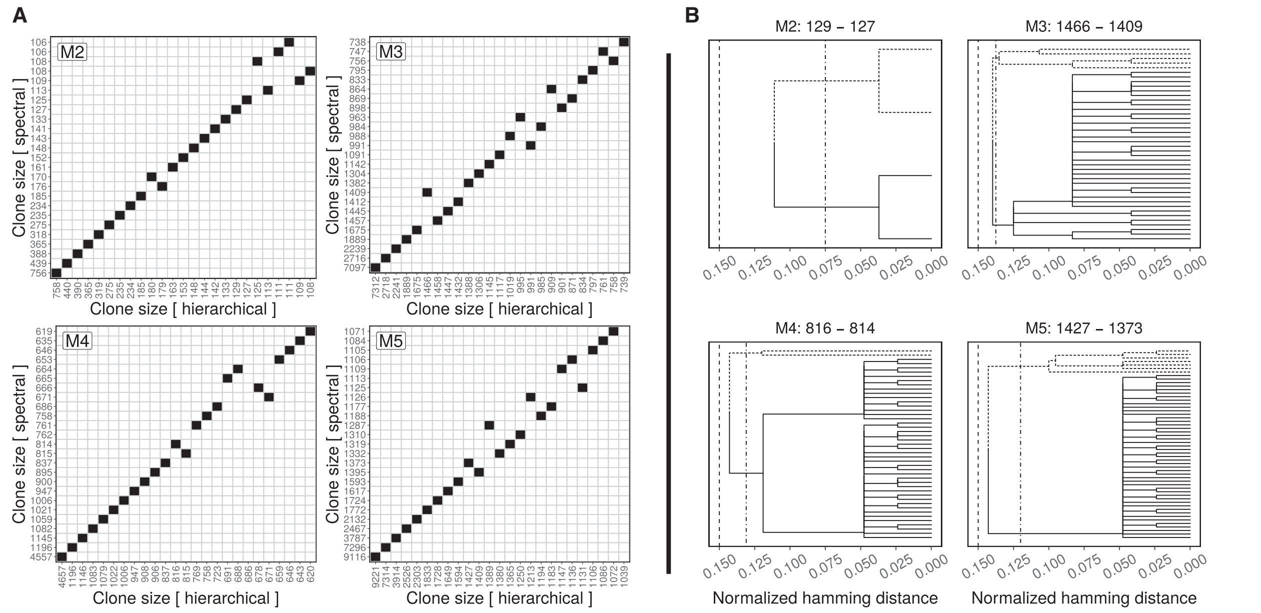

Fig. 4. Clones identified by the spectral clustering-based method are more homogeneous. The spectral and hierarchical clustering-based methods were applied independently to identify clones in cervical lymph node samples from four multiple sclerosis patients (M2, M3, M4 and M5) obtained from (Stern *et al.*, 2014). (**A**) Comparison of clone sizes between pairs of 25 largest inferred clones via hierarchical (*x*-axis) and spectral (*y*-axis) clustering-based methods. Clones with any overlapping membership are indicated in black. (**B**) Dendrogram trees representative of cases where the two methods differed in each of the four individuals. The spectral clustering-based method implied a smaller threshold (vertical dot-dash line) for these clones that removed outlying branches (dashed branches), thus creating a more homogeneous clone compared to the fixed threshold at 0.15 (vertical dashed line) used by the hierarchical clustering-based method. Panel titles indicate total sequences, while leaves represent unique sequences in each clone

the top 25 clones inferred by the two methods, quantified as the size of their intersection divided by the average of their sizes (Ralph and Matsen, 2016), was greater than 0.9 for all clones. We can also measure the consistency between the spectral and hierarchical clustering-based methods by the correlation coefficient between the rows and columns of the similarity matrix. If we consider each row and column as samples of random variables generated by hierarchical and spectral clustering techniques, respectively, then the sample correlation coefficient lies between values –1 (i.e. perfectly anticorrelated) and 1 (i.e. perfectly correlated). Measures of correlation ~0.98 were obtained in all individuals. In cases where the two methods differed, the spectral clustering-based method tended to identify more homogeneous clones. In contrast, the hierarchical clustering-based method shows evidence of the well-known chaining phenomenon and can include sequences which are less similar to others within a clone, thus resulting in more heterogeneity within the clones (Fig. 4B).

The underlying clonal relationships among the sequences in the experimental datasets are not known with certainty. However, we reasoned that true clones would share more SHMs (either because they occurred early in the clonal expansion, or they were positively selected during affinity maturation), and that a higher frequency of shared mutations should indicate better performance. To carry out this analysis, we defined a shared mutation as a single nucleotide mutation that is precisely replicated across all sequences in the clone. We first compared the total number of shared mutations in the 50 largest clones produced by the each of the clustering methods with a set of negative controls (randomly sampled sequences) that share the same V gene annotation and have the same size as the given clone (we generated 100 negative controls for each of the 50 largest clones inferred by each clustering technique from all four samples). Since the precise V gene sequence (up to the start of the junction) is not used to partition sequences into clones, we counted shared mutations in this region. In M2–M5, no shared mutations were found in the negative controls, while 26–155 and 20–125 shared mutations were found using the spectral and hierarchical clustering-based methods across the four datasets, respectively (Fig. 5E–H). These results indicate that both clustering-based methods identify clones that are highly unlikely to be random.

When comparing the spectral and hierarchical clustering-based methods, we found that most of the clones were highly overlapping and contained equivalent numbers of shared mutations (Fig. 5A–D). However, a small number of clones differed and, overall, the spectral clustering-based method produced a larger number of shared mutations regardless of whether sharing was measured over the entire clone (Fig. 5E–H) or pair-wise between each sequence in the clone (Supplementary Fig. S2C). Inspection of the dendrogram from these cases showed that spectral clustering implied a smaller distance threshold for these clones that resulted in the removal of outlying branches, thus creating a more homogeneous lineage (Fig. 6). In general, the scale parameters σ used in the spectral clustering-based method were systematically lower than fixed threshold used in the hierarchical clustering-based method (Supplementary Fig. S2A), thus spectral clustering generated more smaller clones (Supplementary Fig. S2B).

We further tested the methods using experimental data from BCR sequencing of PBMCs from 58 individuals with acute dengue virus infection (Parameswaran *et al.*, 2013). These samples contained ~1–10k total reads (we excluded two datasets with total reads < 1k sequences). A bimodal distribution in the distance-to-nearest distribution was evident in all samples (Fig. 1B), and the hierarchical clustering-based method successfully converged on an optimized threshold to cut the hierarchy in every sample. Once again, both clustering-based methods identified more shared mutations than a negative control, which produced no shared mutations in ~99% of cases (we generated 100 negative controls for each of

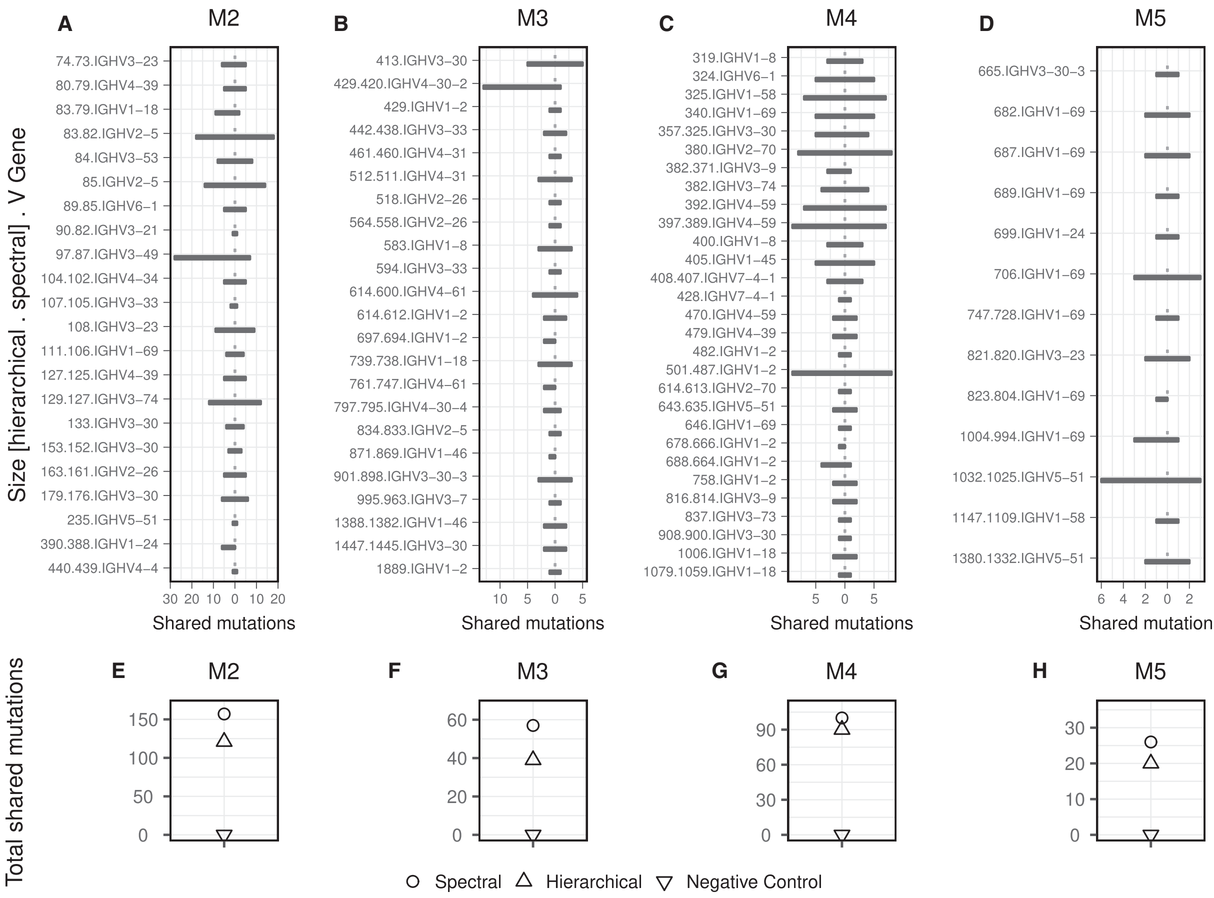

Fig. 5. The spectral clustering-based method identifies clones with more shared mutations. (**A**–**D**) The number of shared mutations in the V segment (up to the start of the junction) was determined for the 50 largest clones (covering ~30% of the total reads) inferred by spectral clustering (black bars towards left), hierarchical clustering (black bars towards right) and negative controls (grey bars). Results are shown for the same four experimental datasets shown in Figure 4. Note that fewer than 50 clones are shown because some pairs of largest inferred clones did not overlap or no shared mutation was observed for either method. (**E**–**H**) The total number of shared mutations across all clones identified by the spectral (circles) and hierarchical (triangles pointing up) clustering-based methods, as well as a negative control (triangles pointing down) was determined for each subject M2–M5

the 50 largest clones inferred by each clustering technique from all 58 individuals). In addition, in every case where the spectral and hierarchical clustering-based method differed, the clone identified by the spectral method had a larger number of shared mutations (Fig. 7 and Supplementary Fig. S3C). The scale parameters σ was also systematically lower than the fixed thresholds (Supplementary Fig. S3A) and spectral clustering generated more smaller clones (Supplementary Fig. S3B). Overall, these results suggest that the spectral clustering-based method identifies true clonal relationships and is likely to be more specific compared to the hierarchical clustering-based method on real experimental data.

3.3 Spectral clustering performs with high confidence on repertoires with few singletons

Current automated methods for choosing a fixed threshold to identify clones in the hierarchal clustering-based method depend on the bi-modality of the distance-to-nearest distribution. However, this distribution can sometimes be unimodal, as seen earlier for the lymph node repertoires from M3–M5 (Fig. 1E). In these cases, the hierarchical clustering-based method fails to produce a partition. In contrast, the spectral clustering-based method returns a partition with clones that included many shared mutations in the V gene (Fig. 5). Since the true clonal relationships are not known for these experimental data, we next sought to characterize the performance of the spectral clustering-based method on simulated data displaying similar unimodal properties.

For the analysis here, we used three simulated datasets (sim-50-1.0-mut, sim-100-1.0-mut and sim-200-1.0-mut) by Ralph and Matsen (Ralph and Matsen, 2016) each containing ~10k total reads with ~10% mean mutation frequency. Inspection of the distance-to-nearest distribution of these datasets shows that they are unimodal with a large peak at small distances and no apparent second peak (Fig. 1G–H). As expected from the lack of bi-modality, the optimized threshold cannot be computed which results in the failure of the hierarchical clustering-based method. To confirm that a fixed threshold was not appropriate for these data, we compared the maximum-distance-within clones (compactness) and minimum-distance-between clones (isolation) distributions (Fig. 8A–C). These two distributions are widely overlapping, meaning no fixed threshold will adequately separate sequences into clones. This contrasts with the simulated datasets from Gupta *et al.* (Gupta *et al.*, 2017; e.g. R1.1 in Fig. 8D), where the compactness and isolation conditions are clearly fulfilled. Despite low compactness and isolation in the Ralph and Matsen simulations, the spectral clustering-based method has high sensitivity (~97% sim-50-1.0-mut, ~98% sim-100-1.0-mut, ~96% sim-200-1.0-mut) and specificity (~97% sim-50-1.0-mut, ~98% sim-100-1.0-mut, ~100% sim-200-1.0-mut). The true and inferred clone sizes are highly correlated (Fig. 9A–C). The degree of overlap between the top 25 true and inferred clones,

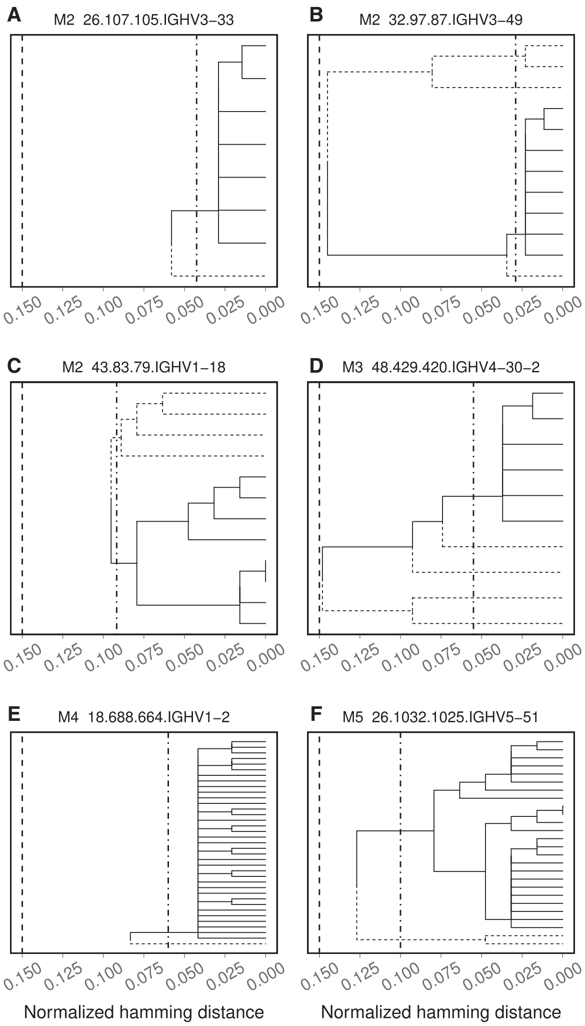

Fig. 6. Clones identified by the spectral clustering-based method are more homogeneous. Representative examples of dendrogram trees from clones where the spectral and hierarchical clustering-based methods found differing numbers of shared mutations in (**A**, **B**, **C**) M2, (**D**) M3, (**E**) M4 and (**F**) M5 (see details in Fig. 5). Dendrogram leaves are unique sequences in the clone found by both clustering-based methods (connected by solid lines) or only by the hierarchical clustering-based method (connected by dashed lines). Each panel also shows the fixed threshold of 0.15 normalized Hamming distance used by the hierarchical clustering-based method (vertical dashed lines) and the threshold necessary to reproduce the clone identified by the spectral clustering-based method (vertical dot-dash lines)

quantified as the size of their intersection divided by the average of their sizes (Ralph and Matsen, 2016), was greater than 0.9. Furthermore, we found an overall correlation of ∼0.98 for all three datasets. Thus, the spectral clustering-based method is able to identify clones with high confidence in repertoires that display a unimodal distance-to-nearest distribution, and do not meet the compactness and isolation criteria necessary for a fixed threshold to work well.

3.4 The spectral clustering-based method is computationally efficient

The hierarchical clustering-based method can be computationally expensive, particularly the optimization of the distance threshold. On a Linux computer with a 2.20 GHz Intel processor and 32 GB RAM, we found that partitioning a dataset of ∼130k sequences took > 90 min (Fig. 10). Even a smaller dataset of ∼54k sequences took over 25 min. In contrast, the spectral clustering-based technique completed in less than 20 min in both cases. Even in cases where the hierarchical clustering-based method failed to return a partition, the spectral clustering-based method efficiently identify clones (taking < 5 min for ∼10k sequences). Overall, on multiple simulated and experimental datasets spanning a wide range of sizes, from less than 10k to more than 100k sequences, the spectral clustering-based method was associated with significant time savings compared to the hierarchical clustering-based method (Fig. 10).

4 Conclusion

B cell clones [the descendants of a common ancestor sharing the same V(D)J recombination event] are a fundamental unit of analysis for adaptive immune responses. Here, we present a computationally efficient method to identify B cell clones from AIRR-Seq datasets. This method is based on spectral clustering of the junction sequences of BCRs that share the same V gene, J gene and junction length. It uses an adaptive threshold that analyzes sequences in a local neighborhood. While previous methods that use a fixed threshold offer the advantage of being easy to explain (e.g. members of a clone have junctions with at least 90% sequence similarity), a single, fixed threshold is not appropriate for repertoires where the distribution of inter-clonal distances overlaps the distribution of intra-clonal distances. The spectral clustering-based method presented herein can identify clones even in samples that have a unimodal distance-to-nearest distribution.

Validation of methods for partitioning sequences into clones presents several challenges. Gold standard experimental data, where clonal relationship between different sequences is known with

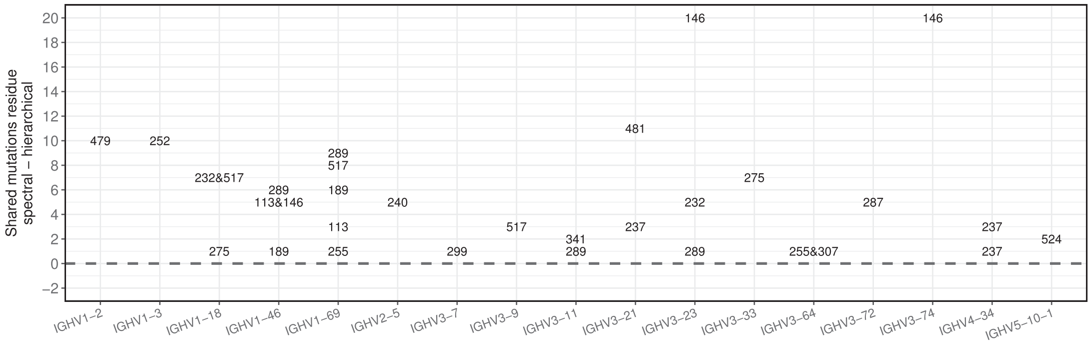

Fig. 7. The spectral clustering-based method identifies clones with more shared mutations in subjects with acute dengue virus infections. The spectral and hierarchical clustering-based methods were applied to peripheral blood B cell repertoires from 58 subjects with acute dengue virus infections (Parameswaran *et al.*, 2013). The total number of shared mutations in the V segment (up to the start of the junction) was determined for clones that were among the 50 largest inferred by both the spectral clustering and hierarchical clustering methods (covering ∼25% of the total reads), and the difference between the methods was calculated. Each number represents a single clone, with the number specifying the individual where the clone was observed and the *x*-axis label indicating the V gene used by the clone

certainty, do not exist for human immune responses. Simulations offer a mechanism to generate data where the underlying clonal groups are known, but the critical parameters to drive such these models are unknown. Indeed, different simulations can produce widely different distance distributions (for example, see Fig. 8). To some extent, this mirrors the diversity seen in different types of immune responses, variability in the tissues profiled and whether bulk or sorted B cells were sequenced. Naive B cells from peripheral blood of healthy young adults are expected to look very different from germinal center B cells from a lymph node. Here, we used both simulated and human experimental datasets to validate our methods. By using simulations from two different groups with different properties (Gupta *et al.*, 2017; Ralph and Matsen, 2016), we could show that the spectral clustering-based method can identify clones with high confidence either with or without bi-modality in the distance-to-nearest distribution. To evaluate performance on experimental data, we measured the number of shared mutations in each clone, reasoning that real clones should share more mutations. The spectral clustering method explicitly models the local relationships among the most similar sequences by defining a local neighborhood, thus leading to increase homogeneity within the inferred clones. This leads to more shared mutations in the structure of BCR clone lineage trees.

Computational efficiency is an important consideration. Some steps in the fully automated hierarchical clustering-based method used here, in particular choosing the distance threshold for cutting the dendrogram, can be slow (Nouri and Kleinstein, 2017). Runtime can be improved by randomly sub-sampling sequences from the repertoire, but it would be at the potential expense of excluding some genuine information. The spectral clustering-based method exhibits faster performance by removing this computationally expensive step, so that in repertoires containing ∼150k sequences the clones can be inferred in ∼20 mins. Run times can be further improved by distributing the computation across many processing cores. In our current implementation, the parallelization is achieved by distributing the clonal inference process from each group of sequences (with same V gene, J gene and junction length) across cores dynamically.

SHM mainly introduces point substitutions into the BCR sequence. However, insertions and deletions (indels) can also be introduced at a low frequency. Smith *et al.* (Smith *et al.*, 1996) reported $< 2\%$ of all somatic mutations are single-base insertions or

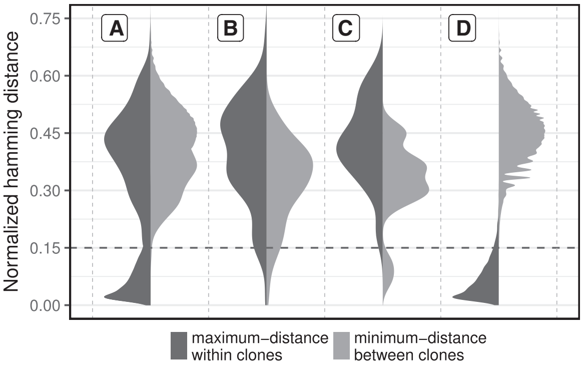

Fig. 8. Compactness and isolation properties vary across repertoire datasets. For each dataset, the maximum-distance-within clones (compactness, black) and minimum-distance-between clones (isolation, grey) were calculated across all clones. Results are shown for simulated datasets from Ralph and Matsen (Ralph and Matsen, 2016) (**A**) sim-50-1.0-mut, (**B**) sim-100-1.0-mut, (**C**) sim-200-1.0-mut, and Gupta *et al.* (Gupta *et al.*, 2017) (**D**) R1.1. The horizontal dashed line indicates the threshold of 0.15 normalized Hamming distance used by the hierarchical clustering-based method

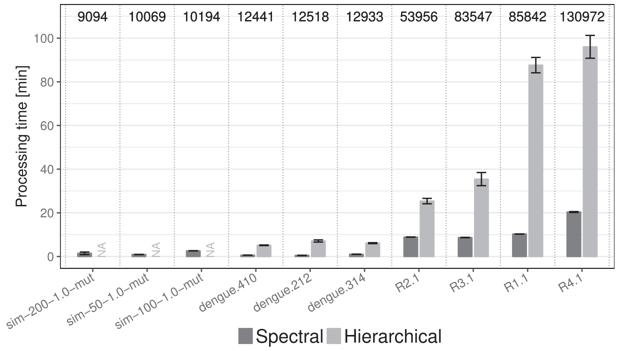

Fig. 10. The spectral clustering-based method is computationally efficient. The running times for spectral (black bars) and hierarchical (grey bars) clustering-based methods were measured for several experimental (Parameswaran *et al.*, 2013) and simulated (Gupta *et al.*, 2017; Ralph and Matsen, 2016) datasets spanning a wide range of sizes (total number of sequences indicated above each column). NA's indicate datasets in which the hierarchical clustering-based method failed to converge on a threshold. Error bars indicate the SEM calculated from 20 bootstrap replicates (with replacement) from the original dataset. Evaluation was carried out on a Linux computer with a 2.20 GHz Intel processor with 32 GB RAM

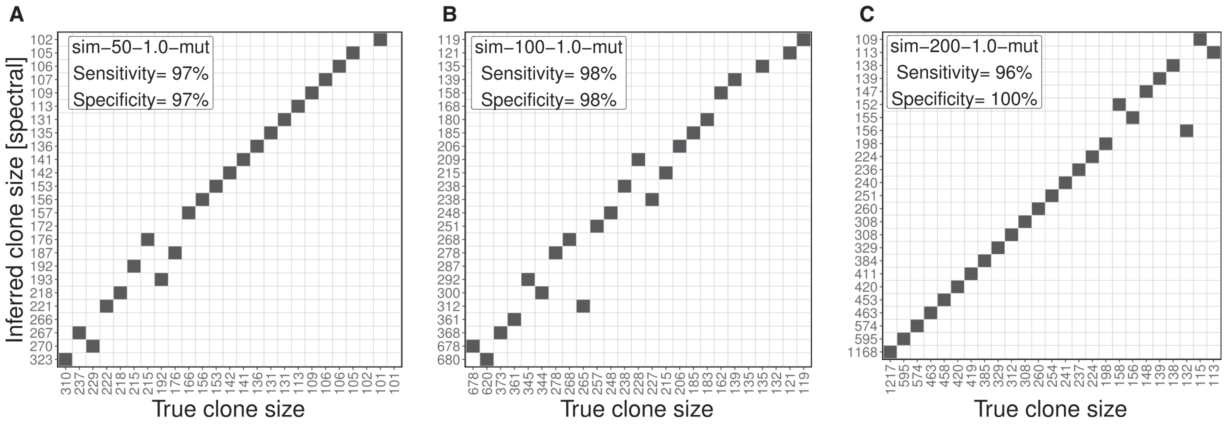

Fig. 9. The spectral clustering-based method identifies clones with high sensitivity and specificity in repertoires with unimodal distance-to-nearest distributions. The spectral clustering-based method was applied to identify clones in simulated data from Ralph and Matsen (Ralph and Matsen, 2016). The clone sizes of the 25 largest inferred clones (*y*-axis) were compared with their true sizes (*x*-axis) in (**A**) sim-50-1.0-mut, (**B**) sim-100-1.0-mut and (**C**) sim-200-1.0-mut. Clones with any overlapping membership are indicated in black

deletions in the introns between the joining and constant region genes. Briney *et al.* (Briney *et al.*, 2012) reported that in-frame insertions and deletions were happened in ~2.0% and ~2.0–3.0%, respectively, of all sequences in memory B cells. Hwang *et al.* (Hwang *et al.*, 2017) reported an estimated frequency of < 1% deletion mutations in VRC01-class broadly neutralized HIV antibodies. The method developed here initially groups sequences that share a V gene, J gene and junction length. Thus, if cells have accumulated an indel in the junction region during affinity maturation, these will not be grouped together properly. In principle, it is possible to remove the restriction of a common junction length, and cluster all sequences sharing the same V gene and J gene. In this case, the distance used in this study (i.e. Hamming distance) would need to be replaced with another string metric that accounts for differing sequence lengths. One possibility is the Levenshtein distance, which measures the edit distance between two sequences with different lengths. The Levenshtein distance is the minimum number of single character edits (insertions, deletions or substitutions) required to change one sequence into the other. Proper tuning the costs associated with insertion, deletion and substitution will an important area of further study here.

Partitioning BCR sequences into clonal groups is a key step in AIRR-Seq analysis. In this study, we have developed a spectral clustering-based method that uses an adaptive threshold to tune the required level of similarity among sequences in different local neighborhoods. This method improves current distance based methods for inferring clonal relationships on both simulated and human experimental data. An implementation of this methodology is freely available as part of the **SCOPe** (Spectral Clustering for clOne Partitioning) R package in the Immcantation framework (www.immcantation.org).

Acknowledgements

The authors thank Dr. Yuval Kluger, Department of Pathology at Yale School of Medicine, for useful comments on this work. The authors are also grateful to Dr. Jason Vander Heiden, Department of Neurology at Yale School of Medicine, for a careful reading of this manuscript.

Funding

This work was supported by the HPC facilities operated by, and the staff of, the Yale Center for Research Computing. This work was supported in part by the National Institutes of Health (NIH) under award number R01AI104739.

Conflict of Interest: none declared.

References

Bannard,O. and Cyster,J.G. (2017) Germinal centers: programmed for affinity maturation and antibody diversification. *Curr. Opin. Immunol.*, **45**, 21–30.

Boyd,S.D. and Joshi,S.A. (2015) High-throughput DNA sequencing analysis of antibody repertoires. In: Crowe,J. (eds) *Antibodies for Infectious Diseases*. American Society of Microbiology, Washington, DC, pp. 345–362.

Boyd,S.D. *et al.* (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci. Trans. Med.*, **1**, 12ra23–12ra23.

Briney,B.S. *et al.* (2012) Location and length distribution of somatic hypermutation-associated dna insertions and deletions reveals regions of antibody structural plasticity. *Genes Immun.*, **13**, 523–529.

de Bourcy,C.F. *et al.* (2017) Phylogenetic analysis of the human antibody repertoire reveals quantitative signatures of immune senescence and aging. *Proc. Natl. Acad. Sci.*, **114**, 1105–1110.

Glanville,J. *et al.* (2011) Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl. Acad. Sci.*, **108**, 20066–20071.

Gupta,N.T. *et al.* (2015) Change-o: a toolkit for analyzing large-scale b cell immunoglobulin repertoire sequencing data. *Bioinformatics*, **31**, 3356–3358.

Gupta,N.T. *et al.* (2017) Hierarchical clustering can identify B cell clones with high confidence in ig repertoire sequencing data. *J. Immunol.*, **198**, 2489–2499.

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

Hershberg,U. *et al.* (2014) Persistence and selection of an expanded b-cell clone in the setting of rituximab therapy for Sjögren's syndrome. *Arthritis Res. Therapy*, **16**, R51.

Hwang,J.K. *et al.* (2017) Sequence intrinsic somatic mutation mechanisms contribute to affinity maturation of VRC01-class HIV-1 broadly neutralizing antibodies. *Proc. Natl. Acad. Sci.*, **114**, 8614–8619.

Jiang,N. *et al.* (2013) Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Trans. Med.*, **5**, 171ra19–171ra19.

Kepler,T.B. (2013) Reconstructing a B-cell clonal lineage. I. statistical inference of unobserved ancestors. *F1000Research*, **2**, 103.

Kleinstein,S.H. *et al.* (2003) Estimating hypermutation rates from clonal tree data. *J. Immunol.*, **171**, 4639–4649.

Logan,A.C. *et al.* (2011) High-throughput vdj sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proc. Natl. Acad. Sci.*, **108**, 21194–21199.

McKean,D. *et al.* (1984) Generation of antibody diversity in the immune response of balb/c mice to influenza virus hemagglutinin. *Proc. Natl. Acad. Sci.*, **81**, 3180–3184.

Meng,W. *et al.* (2017) An atlas of B-cell clonal distribution in the human body. *Nat. Biotechnol.*, **35**, 879.

Mohar,B. (1997) Some applications of laplace eigenvalues of graphs. In: Hahn G., Sabidussi G. (eds) *Graph Symmetry. NATO ASI Series (Series C: Mathematical and Physical Sciences)*, Vol. 497. Springer, Dordrecht.

Mohar,B. *et al.* (1991) The laplacian spectrum of graphs. *Graph Theor. Combinatorics Appl.*, **2**, 871–898.

Murphy,K. *et al.* (2011) An introduction to immunobiology and innate immunity. In: Murphy,K. (ed.) *Janeway's Immunobiology*. Garland Science, New York, NY, pp. 1–98.

Nouri,N. and Kleinstein,S.H. (2017) Performance-optimized partitioning of clonotypes from high-throughput immunoglobulin repertoire sequencing data. *bioRxiv*, page 175315.

Parameswaran,P. *et al.* (2013) Convergent antibody signatures in human dengue. *Cell Host Microbe*, **13**, 691–700.

Ralph,D.K. and Matsen,F.A., IV. (2016) Likelihood-based inference of B cell clonal families. *PLoS Comput. Biol.*, **12**, e1005086.

Rubelt,F. *et al.* (2017) Adaptive immune receptor repertoire community recommendations for sharing immune-repertoire sequencing data. *Nat. Immunol.*, **18**, 1274.

Smith,D.S. *et al.* (1996) Di-and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive B cells. *J. Immunol.*, **156**, 2642–2652.

Stern,J.N. *et al.* (2014) B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci. Trans. Med.*, **6**, 248ra107–248ra107.

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

Tsioris,K. *et al.* (2015) Neutralizing antibodies against west nile virus identified directly from human B cells by single-cell analysis and next generation sequencing. *Integrative Biol.*, **7**, 1587–1597.

Vander Heiden,J.A. *et al.* (2017) Dysregulation of B cell repertoire formation in myasthenia gravis patients revealed through deep sequencing. *J. Immunol.*, **198**, 1460–1473.

Victora,G.D. and Nussenzweig,M.C. (2012) Germinal centers. *Annu. Rev. Immunol.*, **30**, 429–457.

Von Luxburg,U. (2007) A tutorial on spectral clustering. *Stat. Comput.*, **17**, 395–416.

Wang,C. *et al.* (2014) Effects of aging, cytomegalovirus infection, and ebv infection on human b cell repertoires. *J. Immunol.*, **192**, 603–611.

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