Published in final edited form as:

J Immunol. 2017 March 15; 198(6): 2489–2499. doi:10.4049/jimmunol.1601850.

Hierarchical clustering can identify B cell clones with high confidence in Ig repertoire sequencing data[1]

Namita T. Gupta[*], **Kristofor D. Adams**[†], **Adrian W. Briggs**[†], **Sonia C. Timberlake**[†], **Francois Vigneault**[†], and **Steven H. Kleinstein**[*,‡,§]

[*]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520

[†]AbVitro, Boston, MA 02210

[‡]Departments of Immunobiology and Pathology, Yale School of Medicine, New Haven, CT 06520

Abstract

Adaptive immunity is driven by the expansion, somatic hypermutation, and selection of B cell clones. Each clone is the progeny of a single B cell responding to antigen, with diversified Ig receptors. These receptors can now be profiled at large-scale by next-generation sequencing. Such data provide a window into the micro-evolutionary dynamics that drive successful immune responses and the dysregulation that occurs with aging or disease. Clonal relationships are not directly measured, but must be computationally inferred from these sequencing data. While several hierarchical clustering-based methods have been proposed, they vary in distance and linkage methods and have not yet been rigorously compared. Here we use a combination of human experimental and simulated data to characterize the performance of hierarchical clustering-based methods for partitioning sequences into clones. We find that single linkage clustering has high performance, with specificity, sensitivity, and positive predictive value (PPV) all over 99%, whereas other linkages result in a significant loss of sensitivity. Surprisingly, distance metrics that incorporate the biases of somatic hypermutation do not outperform simple Hamming distance. Although errors were more likely in sequences with short junctions, using the entire dataset to choose a single distance threshold for clustering is near optimal. Our results suggest that hierarchical clustering using single linkage with Hamming distance identifies clones with high confidence and provides a fully automated method for clonal grouping. The performance estimates we develop provide important context to interpret clonal analysis of repertoire sequencing data and allow for rigorous testing of other clonal grouping algorithms.

Introduction

The capacity of B cells to modify their antibodies or immunoglobulin (Ig) receptors to adapt in response to pathogenic challenges is a key mechanism that protects us from infection. An initial diversity of ~10^7 unique Ig molecules (1) stems from somatic recombination of gene

[1]S.H.K. was supported by NIH R01AI104739 and N.T.G. was supported by NIH R01AI104739, United States-Israel BSF 2013395, and PhRMA R12886. The Yale HPC facilities are funded by NIH grants RR19895 and RR029676-01.

[§]Corresponding author: Tel: +1 (203) 785-6685, Fax: +1 (203) 785-6486, steven.kleinstein@yale.edu.

segments in the B cell Ig gene locus compounded by stochastic nucleotide insertions and deletions at the junctions of these segments. Upon activation, these naïve B cells diversify further by undergoing clonal expansion with somatic hypermutation (SHM) in the Ig gene (approximately one point mutation per 1000 bp per cell division (2, 3)) followed by selection for higher affinity B cells. This micro-evolutionary process known as affinity maturation results in B cells with diversified Ig receptors that are clonal relatives of the original activated B cell. In healthy human adults, class-switched memory B cells express Ig receptors that are ~7% mutated (4). Our ability to profile this adaptive immune response has dramatically improved through the application of next-generation sequencing, which allows for measurement of tens to hundreds of millions of B cell receptors (5). However, the identification of sequences that belong to the same B cell clone in these data remains a significant challenge (6).

Adaptive immune receptor repertoire sequencing (AIRR-Seq) is being widely used for both basic science and clinical studies (7–9). Statistical properties of the repertoire, such as diversity or mutational load, are being used to gain insights into the dysregulation that occurs with aging or disease (4, 8, 10–21). Properly identifying clones is central to the calculation of many of these properties. For example, clone size distributions are the basis for several diversity measures, such as species richness, Shannon entropy, and the Gini-Simpson index (6) that parallel diversity measures in ecology (22). Diseases such as chronic lymphocytic leukemia are characterized by low diversity that is driven by the dominance of a small number of clones (23), and repertoire sequencing has been used to improve minimal residual disease detection for lymphoid cancers (8, 10). Responses to drugs such as rituximab have also been measured by changes in repertoire diversity in autoimmune disease (11, 12), characterizing treatment regimens that lead to successful remission or result in persistent clonal expansions. Decreases in repertoire diversity have been associated with aging (4, 13, 14). In subjects with seasonal allergies, the IgE repertoire is the least diverse compared to other isotypes in blood and nasal biopsies, indicating a focused immune response (16).

Analysis of diversity within a clonal lineage also has several applications. Reconstruction of the B cell clonal lineages using methods such as maximum parsimony or likelihood (24) allows tracing somatic mutations through the Ig sequences and helps in understanding the evolution of neutralizing antibodies (17, 18). Lineage relationships have also been used to gain insight into the mechanisms underlying isotype switching (25, 26) and to show that B cell clones in the central nervous system in subjects with multiple sclerosis are first activated in the periphery (15). Identifying clones that include sequences with known antigen specificities has also been used to reveal novel antigen-specific sequences (19, 20). Thus, clonal partitioning of AIRR-Seq data is central to a wide range of applications.

Despite its importance, there is no consensus on the best method for grouping Ig sequences into B cell clones. Most current approaches leverage the high diversity of the junction region (*i.e.,* where the V, D, and J gene segments join) as a “fingerprint” to identify each B cell clone (27). Since it is unlikely that two separate recombination events would lead to identical junctions, sequences with junction regions that are “similar enough” are determined to share a common B cell ancestor (*i.e.*, be clonally related) rather than to have

arisen independently. Probabilistic models have been developed to calculate likelihood of sharing a B cell ancestor and subsequently infer clonal grouping (28, 29). However, these algorithms have run times that scale exponentially, which is computationally intractable for large sequencing datasets (29). In practice, most studies cluster sequences based on junction sequence similarity (13, 15, 18, 21, 30–32).

While many clustering approaches exist, hierarchical clustering is the most widely used framework for grouping clonally-related sequences. Hierarchical clustering requires a measure of distance between pairs of sequences, and a choice of linkage to define the distance between groups of sequences. Since hierarchical clustering produces a tree defining the relationships between all sequences, it is also necessary to specify a method to cut the hierarchy in order to identify discrete clonal groups. In practice, most studies first split the sequences using some similarity requirement on the germline gene segments (*e.g.*, identical V and J gene segments, and junction length), and then apply hierarchical clustering on the junction sequence of these smaller groups (8, 15, 18, 21, 30–34). Several distance metrics have been proposed, including Hamming distance, which is simply the absolute count of differences between two amino acid (31, 33) or nucleotide (21, 32) sequences, normalized edit distance (30), and a metric that incorporates hot/cold-spot biases in SHM targeting (15, 18). In addition to metrics defining distance between two sequences, linkage methods define how distance is calculated between groups of sequences. Different clonal grouping algorithms use single (15, 18, 21, 32), average (30), or complete (13) linkage. The threshold at which the hierarchy is cut to define clusters of clonally related sequences has also been determined in several ways. Chen et al. (30) propose a fixed threshold that is manually identified based on when the rate of cluster merging events changes for a gold standard dataset (30). Glanville, et al. (31) introduced a method based on the observed bimodal distribution of distances from each sequence to its nearest neighbor. In this case, the first mode is assumed to represent sequences with clonal relatives in the data (near neighbors), while the second mode is taken to represent sequences without clonal relatives in the data (distant neighbors). The threshold is then selected to be the value that separates the two modes of this distribution (21, 31). As of yet, there has not been an in-depth evaluation of performance of hierarchical clustering-based clonal grouping algorithms including a comparison of the different distance and linkage methods on AIRR-Seq data.

In this paper, we carry out a comparative analysis of distance metrics and linkage methods for hierarchical clustering-based clonal grouping of Ig heavy chain sequences. A combination of experimental and simulation-based criteria is used to evaluate the performance of these algorithms, including estimates of specificity, sensitivity, and positive predictive value (PPV). Overall, we find that single-linkage hierarchical clustering with nucleotide Hamming distance has excellent performance, with specificity, sensitivity, and PPV all over 99%. Implementations of all clonal grouping methods, along with extensive documentation, are available through the Change-O and SHazaM packages (35) as part of the Immcantation tool suite (http://immcantation.readthedocs.io) for AIRR-Seq analysis.

Materials & Methods

Human B cell receptor repertoire sequencing data

Three B cell receptor repertoire sequencing datasets (Healthy, Dengue, and West Nile virus (WNV)) were used to measure the performance of clonal grouping methods. The 'Healthy' dataset was composed of sequences from peripheral blood mononuclear cells (PBMCs) isolated from healthy adult subjects (n = 27) as previously described (14). Sequences were filtered as described in (14) including removal of non-IGH artifacts, sequences with V-gene insertion or deletions, and chimeric sequences. The 'Dengue' dataset was composed of sequences from PBMCs isolated from subjects with acute Dengue infection (n = 42) as described previously (36). Sequences were filtered as described in (36) including removal of reads containing insertions and deletions as determined by alignment against the V and J germline repertoires. The 'WNV' dataset was composed of sequences from PBMCs and sorted plasma, memory, and naïve B cells isolated from subjects recently infected with WNV (n = 7) as previously described (18). Sequences were filtered as described in (18) including filtering based on sequence quality using pRESTO (37) version 0.4. In each case, processed sequencing data was obtained from the authors. Germline gene segments were inferred for each sequence by using IMGT/HighV-QUEST (38). The 'Healthy' dataset was run through IMGT/HighV-QUEST on December 21, 2014, 'Dengue' was run through IMGT/HighV-QUEST March 12, 2015, and the 'WNV' dataset was run through IMGT/HighV-QUEST on March 21, 2014. Sequences identified as non-functional by IMGT/HighV-QUEST were removed using the changeo-clt toolkit version 0.2.0 (35).

Two additional B cell receptor repertoire sequencing datasets from healthy adult subjects were used as a source of naïve B cell receptor sequences for the lineage simulations. The first was composed of sequences from PBMCs and sorted naïve B cells isolated from healthy control subjects (n = 4) as part of a study of Myasthenia Gravis described in Vander Heiden, et al. (Submitted). The second was composed of sequences from total RNA isolated from blood samples of healthy adult subjects (n=3) as part of an influenza vaccination study described in Laserson et al. (39). In this case the samples, which were originally sequenced using Roche 454, were re-sequenced using Illumina MiSeq and published for the first time here (see details below). Identical sequences from the same sample were counted once, but identical sequences from different samples were counted independently. Both datasets are available on the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) under BioProject accession numbers PRJNA349143 (influenza vaccination study) and PRJNA338795 (myasthenia gravis study).

Library preparation and BCR sequencing of healthy subject sequences from Laserson, et al

The blood samples collected in the influenza vaccination study by Laserson et al (39) were re-sequenced using the Illumina MiSeq platform as previously described (18, 40). Briefly, RNA was reverse-transcribed into cDNA using a biotinylated oligo dT primer. An adaptor sequence was added to the 3′ end of all cDNA, which contains the Illumina P7 universal priming site and a 17-nucleotide unique molecular identifier (UMI). Products were purified using streptavidin-coated magnetic beads followed by a primary PCR reaction using a pool

of primers targeting the IGHA, IGHD, IGHE, IGHG, IGHM, IGKC and IGLC regions, as well as a sample-indexed Illumina P7C7 primer. The immunoglobulin-specific primers contained tails corresponding to the Illumina P5 sequence. PCR products were then purified using AMPure XP beads. A secondary PCR was then performed to add the Illumina C5 clustering sequence to the end of the molecule containing the constant region. The number of secondary PCR cycles was tailored to each sample to avoid entering plateau phase, as judged by a prior quantitative PCR analysis. Final products were purified, quantified with Agilent Tapestation and pooled in equimolar proportions, followed by high-throughput paired-end sequencing on the Illumina MiSeq platform. For sequencing, the Illumina 600 cycle kit was used with the modifications that 325 cycles was used for read 1, 6 cycles for the index reads, 300 cycles for read 2 and a 10% PhiX spike-in to increase sequence diversity.

Read processing of healthy subject sequences from Laserson, et al

MiSeq reads were demultiplexed using Illumina software. Positions with less than Phred quality 5 were masked with Ns. Isotype-specific primers and unique molecular barcodes (UMI) were identified in the amplicon and trimmed using pRESTO (37) MaskPrimers-cut. Read 1 and read 2 consensus sequences were generated separately for each mRNA from reads grouped by UMI, which represent PCR replicates arising from a single initiating mRNA molecule. UMI read groups were aligned with MUSCLE (41), and pRESTO was used to construct a consensus sequence using BuildConsensus, requiring >=60% of called PCR primer sequences agree for the read group, maximum nucleotide diversity of 0.1, using majority rule on indel positions, and masking alignment columns with low posterior (consensus) quality. Paired end consensus sequences were then stitched in two rounds. First, ungapped alignment of each read pair's consensus sequence termini was optimized using a Z-score approximation and scored with a binomial p-value as implemented in pRESTO AssemblePairs-align. For read pairs failing to stitch this way, stitching was attempted using the human BCR germline V exons to scaffold each read prior to stitching or gapped read-joining, using pRESTO AssemblePairs-reference. Positions with posterior consensus quality less than Phred 5 were masked again with Ns. All pRESTO tools used were version 0.5.1 in conjunction with Python 3.4. Germline gene segments were inferred using IgBLAST version 1.4.0 (42) with the IMGT/GENE-DB (43) reference sequences from June 7, 2014 and output was parsed with changeo-clt (35) MakeDb version 0.3.0. Duplicate sequences were collapsed and only those heavy chain sequences with at least two reads supporting the sequence were retained for further analysis.

Simulation of B cell clonal lineages

Each simulated clone was generated by introducing mutations into an experimentally observed naïve B cell receptor sequence according to an observed lineage tree topology (ie, branching pattern). Lineage tree topologies were previously derived based on sequencing data from lymph node samples collected as part of a published study of Multiple Sclerosis (15). The set of 7103, 4066, 8244, and 14782 lineage topologies from four subjects (referred to here as R1, R2, R3 and R4, respectively) were each used as the basis for 10 simulations resulting in 40 total simulated datasets. To generate a simulated dataset, the root of each lineage was randomly chosen (without replacement) from a large pool of un-mutated

sequences from healthy subjects obtained from Vander Heiden, et al. (Submitted) and Laserson, et al. (39) (described above). Mutations were then added to the sequence in order to match the experimentally-observed mutation counts of each branch in the lineage tree according to the human S5F (hS5F) targeting model (44). The simulated sequences then had germline gene segments inferred using IgBLAST version 1.4.0 (42) with the IMGT/GENE-DB (43) reference sequences from May 2, 2016 and output was parsed with changeo-clt (35) MakeDb version 0.3.3.

Distance metrics

Hamming distance was defined as the absolute count of letter changes between nucleotide junction sequences (ham) or amino acid junction sequences (aa). The 5-mer distance metrics were all based on the hS5F targeting and substitution models described in (44), which estimates: (1) the relative probability of a nucleotide position being targeted for somatic mutation, and (2) the probability of mutating to each of the three other possible nucleotides, based on the two nucleotides up- and downstream. This probability (p) was transformed into a distance (d) using the formula: $d = -\log_{10}(p)$. The distance between two junction sequences was defined to be the sum of distances between each nucleotide position. For a given mutation between two junction sequences, the hS5F-min model took its distance to be the minimum of mutating from nucleotide n_1 to n_2 and from n_2 to n_1 at that position. The hS5F-avg model took the distance of the mutated position to be the average of mutating from n_1 to n_2 and from n_2 to n_1.

The human S1F (hS1F) model is equivalent to hS5F-min, but used the human symmetric substitution matrix based on the single mutated nucleotide described in (44). The m1n model is equivalent to hS5F-min, but used the mouse symmetric substitution matrix based on the single mutated nucleotide described in (45). Both the human and mouse substitution matrices are 4x4 matrices whose rows correspond to the conditional probabilities of each of the four nucleotides to mutate to any of the other nucleotides. Hence, the diagonal of each matrix is 0, and each row sums to 1. To estimate distances between sequences based on each single nucleotide substitution matrix, we first approximated each matrix by a symmetric matrix both with respect to time (*i.e.*, equal likelihood of mutating from nucleotide n_1 to n_2 and from n_2 to n_1), and with respect to strand (*e.g.*, equal likelihood of mutating from C to T and from G to A). The symmetric matrix has three free parameters that were estimated by minimizing the sum of squares between this matrix and the original substitution matrix. The fitted matrix was normalized to ensure that each row sums to 1. Using the symmetric substitution matrix, a distance between two sequences was computed by going over all substitution events and summing over the logarithms of the corresponding conditional probabilities from the symmetric substitution matrix.

When normalizing by length for any distance metric, these distances were divided by the length of the junction region.

Implementation of clonal grouping algorithms

Clonal grouping algorithms were implemented and are made available in the change-o-clt toolkit (35) (version 0.3.1 or newer). Sequences were first grouped by shared V gene, J gene

and junction length. Within these groups of sequences, hierarchical clustering was performed using the bygroup subcommand of DefineClones.py with the specified distance metric and linkage type. The resulting hierarchy was then trimmed into flat clusters at a fixed threshold determined using an automated method based on analyzing the “distance-to-nearest” profile. For each sequence, the distance to its nearest un-identical junction was calculated using the SHazaM R package (35) (version 0.1.3 or newer). The ideal bandwidth for the fourth derivative kernel density estimate of these distances was then estimated using the unbiased cross-validation method (46) of the fourth derivative of the kernel density estimate (47, 48) from the kedd R package (49) (version 1.0.3). This bandwidth was used to calculate a binned kernel density estimate of the distances with a Gaussian kernel using the KernSmooth R package (version 2.23–15). The minimum between the two modes of the resulting bimodal distribution of distances was then calculated by finding the first value at which the first derivative was zero while the second derivative was positive, indicating a local minimum following a local maximum. If such a minimum were not found an error would have been returned, but this was not the case for any of the analyses herein.

Specificity, positive predictive value, and sensitivity

Performance was characterized by considering the binary classification task of defining the relationship between all pairs of sequences (s_1 and s_2) with the same junction length. These classifications were then pooled together for the entire dataset. If s_1 and s_2 were known to be unrelated (termed condition negative) but were grouped into the same cluster (termed test positive), this was counted as a false positive. If they were grouped into different clusters (termed test negative), this was a true negative. If s_1 and s_2 were known to be related (termed condition positive) but were grouped into different clusters, this was counted as a false negative. If they were grouped into the same cluster, this was counted as a true positive. These relationships are outlined in Figure 1.

In the case of experimental data, two sequences were known to be unrelated if they were derived from two separate individuals. Therefore, false positives were defined as sequences from different individuals being grouped together in a clone, while true negatives were defined as sequences from different individuals that were grouped into separate clones. Specificity was then calculated by dividing the number of true negative classifications by the number of condition negative classifications. In other words, specificity was defined as the fraction of pairs of unrelated sequences that were successfully inferred by the algorithm to be unrelated.

For the simulated datasets, the precise clonal membership of each sequence was known, yielding the intuitive definition of false positive and false negative classification. Positive predictive value (PPV) was calculated by dividing true positive classifications by test positive classifications. In other words, PPV was the fraction of predicted clonal relationships that were actually true. Sensitivity was calculated by dividing true positive classifications by condition positive classifications. In other words, sensitivity was defined to be the fraction of actual clonal relationships that were successfully inferred by the algorithm.

Shannon entropy calculation

The Shannon entropy of clonally related sequences (within clones) was calculated for true clones having at least two members. Entropy was calculated for each of the first 24 nucleotide positions of the junction within each clone and averaged across clones having junction length <30 nt and 51 nt. To calculate Shannon entropy of clonally unrelated sequences (between clones), the most mutated sequence was selected from each true clone. These mutated sequences were then placed into groups sharing the same V gene, J gene, and junction length. Groups with only one sequence were discarded. Entropy was calculated for each of the first 24 nucleotide positions of the junction within each group and averaged across groups having junction length <30 nt and 51 nt. Error bars represent standard error of the mean. The calculations were made on all of the simulated datasets pooled together.

Results

The problem of clonal grouping takes a set of B cell receptor sequences as input and returns a partition of that set into subsets (clonal groups) that each represent an independent clonal lineage. Here we investigate hierarchical clustering-based algorithms which infer a dendrogram based on pairwise sequence distances and then cut the dendrogram at a fixed distance (or “threshold”) to predict groups of clonally-related sequences. To evaluate the performance of these clonal grouping algorithms, we consider three metrics: specificity, sensitivity, and positive predictive value (PPV) (Figure 1).

Specificity quantifies how frequently unrelated sequences are correctly separated into different clonal groups. In most experimental datasets, the exact clonal relationships between sequences are unknown. However, to estimate specificity we can take advantage of the fact that B cell clones cannot span multiple individuals (by definition). Using this knowledge, specificity is defined based on how frequently sequences from separate individuals are incorrectly inferred to be clonal relatives. This measure is used to quantify performance on three human Ig AIRR-Seq datasets (referred to as Healthy, Dengue, and WNV as detailed in Methods), each of which contains samples from multiple individuals.

Sensitivity represents inclusivity of an algorithm by measuring how often clonally related sequences are grouped together. PPV is a complementary metric that quantifies the precision of an algorithm by measuring how often inferred clonal relatives are truly clonally related. The calculations for sensitivity and PPV require knowledge of true clonal relationships and thus cannot be estimated from current human experimental datasets. For these measures, performance was evaluated using 40 simulated datasets based on four experimentally observed sets of clonal lineage structures (referred to as R1-R4 as detailed in Methods). In the following sections, we use these performance metrics on the human experimental and simulation data to evaluate the choice of distance metrics, linkage methods and threshold parameters in clustering-based clonal grouping algorithms.

Automated determination of clonal distance thresholds

A key step in hierarchical clustering-based clonal grouping involved choosing a threshold at which to cut the dendrogram, thus forming discrete groups of clonally-related sequences. In

previous work (21, 31), this threshold has often been fixed at a single value determined by manual inspection of a histogram of nearest-neighbor distances (the so-called “distance-to-nearest” plot (6)). These histograms are typically bimodal and the threshold is selected to separate these two modes (Figure 2A). This choice is motivated by the intuition that the smaller peak represents the distance between sequences within a clone (intra-clonal distance), while the larger peak represents the distance between sequences in different clones (inter-clonal distance). Inspection of the nearest-neighbor distance distributions for the Healthy, Dengue, and WNV experimental datasets used in this study showed that they are each clearly bimodal. However, they differed in the values that best separated the two modes (Figure 2B). This result indicates that the distance threshold for clonal grouping is dataset-specific and must be re-computed for each study.

Manual determination of the clustering threshold is problematic because inspecting a distribution by eye is time-consuming and imprecise. We therefore sought to develop an automated analytic procedure for inferring the clustering threshold that mimics the widely used manual approach. Since the histograms generated from real data are rarely smooth, we first smooth the empirical distributions using a binned Gaussian kernel density estimator using a procedure that is well-suited for bimodal distributions (48) (see Methods for details). Next, we computationally determine the minimum between the two peaks of the smoothened distribution and define this value to be the clustering threshold (Figure 2A). This method placed the threshold at intuitive locations in the Healthy, Dengue and WNV experimental datasets (Figure 2B). This method for automated determination of the clustering threshold enables efficient application of clonal grouping algorithms under many parameter settings and on many different datasets.

We next applied the automated threshold to assess the performance of clonal grouping methods on experimental and simulated datasets. Hierarchical clustering using the nucleotide-based Hamming distance metric with single linkage was an effective approach. The mean specificity of the algorithm was over 99% on experimental data (Figure 3A), the mean sensitivity was ~99% on simulated datasets (Figure 3B), and PPV was over 99% on simulated datasets (Figure 3C). In contrast, using amino acid-based Hamming distance – which has been used in some previous studies (31, 33) – had significantly worse sensitivity (Supplementary Figure 1). One potential shortcoming of using the Hamming distance metric is that mutations in short junctions are penalized more heavily than mutations in longer junctions. Since junction regions vary widely in length (33–81 nucleotides, 95% range from experimental datasets) and the clustering algorithm uses a fixed threshold, this bias could lead to suboptimal performance. In an attempt to address this issue, we and others have used a length-normalized Hamming distance metric, in which Hamming distance is divided by the length of the junction. This length-normalization had minimal effect on specificity in the experimental data (Figure 3A), but significantly improved sensitivity (Figure 3B) and PPV (Figure 3C) in the simulated data ($p<10^{-4}$, paired t-test). Thus, length-normalization of the distance metric is an important step in clonal grouping algorithms.

Single linkage has highest sensitivity with minimal compromise of PPV

Hierarchical and other agglomerative clustering algorithms require a method for determining the distance between two sets of points (in this case, sequences). The most common linkage methods include single, average, and complete linkage (50). Single linkage defines the inter-set distance as the minimum distance between all pairs of points from the given sets. This generally results in larger and more heterogeneous clusters (50). Complete linkage defines the inter-set distance as the maximum distance between all pairs of points from the given sets, and generally results in smaller and more homogeneous clusters (50). Average linkage defines the inter-set distance as the average distance between all pairs of points from the given sets, thus providing a compromise between single and complete linkage.

As expected, single linkage had the lowest specificity followed by average and then complete linkage (Figure 4A). However, these differences were small, and specificity was over 99% in all cases. A similar ranking was found for PPV, with complete and average linkage significantly improving performance relative to single linkage ($p < 10^{-4}$, paired t-test; Figure 4C). Once again, however, the absolute performance differences were small, with all three approaches exhibiting a mean PPV of over 99%. As specificity and PPV both reflect the accuracy of clonal grouping, we conclude that all of the linkage methods are accurate. In contrast, single linkage exhibited significantly higher sensitivity for clonal grouping relative to both average and complete linkage ($p < 10^{-4}$, paired t-test; Figure 4B). In this case, the sensitivity differences were large, with single linkage having a mean sensitivity of 99% compared to 88% for average linkage and 60% for complete linkage. Overall, these results show that single linkage is significantly better at capturing the breadth of true clonal relationships, with only a modest reduction in accuracy.

Incorporating SHM biases does not significantly improve clonal grouping

While the Hamming distance between two sequences is quick and easy to compute, it does not account for the intrinsic targeting and substitution biases in SHM (44). It is well established that Activation-induced cytidine deaminase (AID) and the error prone DNA repair pathways that drive B cell diversification frequently target specific DNA motifs (termed hot-spots), while others are rarely mutated (termed cold-spots). There is also a substitution bias such that transition mutations are significantly more frequent than transversions. Weighing all mutations equally undervalues the less probable mutations because two sequences are less likely to be part of a clone if they differ by mutations that occur less frequently (i.e., transversion mutations at cold-spot positions).

To test whether accounting for the intrinsic biases of SHM could improve the performance of clonal grouping algorithms, we implemented four previously proposed SHM models that account for these biases in different ways (see Methods for details). The first two models (hS5F-min and hS5F-avg) incorporate the human S5F targeting (mutability and substitution) models that incorporate the effects of the two nucleotides up- and downstream of a mutation (44). For each pair of nucleotides (n_1 and n_2) that differ between two junctions being compared, the hS5F-avg metric assumes that each one has an equal probability of having been present in the most recent common ancestor. Thus, the distance is taken as the average of mutating from n_1 to n_2 and from n_2 to n_1. The hS5F-min metric assumes that the ancestral

base is the one that leads to the most likely mutation, and thus uses the minimum distance at each nucleotide position. The second two models (hS1F and m1n) ignore mutability, but account for substitution bias using a model that depends only on the targeted base (*i.e.*, ignoring surrounding nucleotides). As these models are symmetric, there is no assumption of which nucleotide was ancestral. Surprisingly, we found no significant performance differences for any of the distance metrics in experimental (Figure 5A) or simulated datasets (Figure 5B,C). These results support the use of the more efficient nucleotide Hamming distance metric. Overall, we find that hierarchical clustering using length-normalized nucleotide Hamming distance with single linkage performs well with mean sensitivity, specificity, and PPV all over 99%.

Sequences with short junctions have high false positive rate

We next investigated the dependence of performance on junction length to better understand the source of errors in clonal grouping. Junction length was minimally correlated with specificity in the experimental datasets (r = 0.1, Pearson's correlation; Figure 6A). Similarly, there was no correlation of junction length with sensitivity in the simulated datasets (r = 0.02, Pearson's correlation; Figure 6B). In contrast, there was a strong positive correlation of junction length with PPV in the simulated datasets (r = 0.4, Pearson's correlation), with a mean PPV of 99.1% for sequences with shorter junctions (junctions shorter than 30 nt represented by at least 0.001% of the sequences in the repertoire) compared to a mean PPV of 99.8% for sequences with longer junctions (Figure 6C). For sequences with shorter junction lengths, the right peak in the nearest-neighbor distance distributions (interpreted as distances between unrelated sequences) begins to overlap the left peak (interpreted as distances between clonally-related sequences). This pattern of decreasing inter-clonal distances as junction lengths decrease was also apparent considering nearest-neighbors across individuals (Figure 7). Thus, it appears that the distance threshold that effectively separates clonal members with longer junctions begins to group together unrelated sequences with shorter junctions. These results raise the possibility that using a single distance threshold to separate clonal groups across the entire data set may not be optimal for sequences with shorter junctions.

To determine if using multiple distance thresholds could improve performance, we assessed precision (PPV) and recall (sensitivity) across a range of distance thresholds using sequences of varying junction lengths. We selected the shortest junction length represented by at least 0.001% of total sequences (24 nt), the overall average junction length (51 nt), and the longest junction length with a distinguishable spread in precision across distance thresholds (81 nt) as example junction lengths with which to assess performance. When considering all junction lengths as one group, the automated threshold appears close to optimal in trading off between PPV and sensitivity, with both over 99% (Figure 8A). The same holds true when considering the average (51 nt; Figure 8C) and longer junction lengths (81 nt; Figure 8D). Interestingly, the single threshold chosen on the entire data set still provided a near optimal trade-off in performance for sequences with shorter (24 nt) junctions, although peak sensitivity was lower for some of the simulated repertoires (Figure 8B). Thus, using a junction length-specific threshold is unlikely to improve performance.

The inability to separate unrelated sequences with shorter junction lengths implies a lack of diversity between clones. Indeed, sequences with short junctions had lower nucleotide diversity than sequence with longer junctions (Figure 9). In other words, unrelated sequences with short junctions were more similar on a per nucleotide basis than unrelated sequences with longer junctions. This difference in diversity was not spread evenly across the junction region, but only became apparent after the first ~7 nt of the junction region, which are generally derived directly from the V gene segment (43). As expected, this was in contrast to nucleotide diversity within clones, which was low across all junction lengths (Figure 9). Overall, these results show that sequences with shorter junctions have a lower diversity than expected (given their length), making it difficult to separate clonally related and unrelated sequences.

Although sequences with longer junctions can be grouped into clones with high sensitivity and PPV, false positive assignments are still present. One reason underlying these errors is the use of the IGHJ6 gene, which is over-represented in false positives with junctions at least 30 nt in length ($p<10^{-3}$, Chi-squared test). The IGHJ6 gene extends an extra ten nucleotides into the junction region relative to all other IGHJ genes (43) and clones that use this J gene would thus be more similar to each other than clones using other J genes.

Discussion

AIRR-Seq enables large-scale characterization of the Ig repertoire with the potential for significant basic science and clinical insights. Effective population-level analysis of these data often relies on first identifying groups of clonally related sequences. While hierarchal clustering-based approaches are widely applied in current studies, estimates of their performance and the tradeoffs inherent in the choice of distance or linkage method are lacking. In this study, we carry out an in-depth comparison of hierarchical clustering-based clonal grouping algorithms using an automated analysis pipeline, along with experimental and simulated validation datasets. The analysis pipeline has three stages. First, sequences are separated by V gene, J gene, and junction length. Second, sequences in these groups are assembled into a hierarchy as defined by the distance metric and linkage method. Finally, the hierarchy is partitioned into discrete clones at a fixed distance threshold. While previous applications of this framework relied on a manual process to choose the distance threshold, we minimized human imprecision by developing an automated method to select a customized threshold for any dataset based on analysis of the “distance-to-nearest” distribution.

Quantitative evaluation of the clonal grouping methods was based on a combination of human experimental and simulated Ig heavy chain data using the common performance measures of specificity, sensitivity and positive predictive value (PPV) (29, 30, 34). Experimental data was used to estimate specificity based on the fact that, by definition, B cell clones cannot span different individuals. Sensitivity and related measures like PPV cannot be estimated from human experimental data, since current approaches do not allow unequivocal identification of members of a clone. However, some murine model systems now allow identification of individual clones through Brainbow color labelling of individual B cells prior to affinity maturation (51). In this study, we used simulated data – where all

clonal relationships are known explicitly – to estimate sensitivity and PPV along with specificity. Simulations have previously been used to validate performance of a probabilistic clonal grouping algorithm (29) and to benchmark other repertoire analysis tools (52, 53). Unlike previous approaches, our validation framework does not rely on an underlying model of clonal expansion and affinity maturation. Rather, the lineage tree topologies are taken from experimental datasets and are overlaid with new root sequences and somatic mutation patterns to more closely mimic observed repertoire structures. Furthermore, sensitivity and PPV were calculated on the dataset as a whole, which is less biased by clone size than the per-read averages calculated in previous studies (29).

Hierarchical clustering methods relate sequences to each other, but do not split sequences into discrete clusters. Thus, a critical step in clonal grouping based on hierarchical clustering is determining the threshold used to partition the sequences into clusters (each representing a single clone). Previous approaches used a fixed threshold such as 80% nucleotide similarity (32), or relied on manual inspection of the “distance-to-nearest” plot to generate a study-specific threshold (31). “Distance-to-nearest” plots are generally bimodal, with the two peaks interpreted as clonally related (small distance peak) and unrelated (larger distance peak) sequences. Previously the clustering threshold has been determined manually by looking for the distance that best separates these two peaks. However, this process is time-consuming, can be subjective, and there are often multiple possible thresholds that provide equivalent separation between the peaks. To minimize human bias and to enable rapid evaluation of a range of parameter choices for this study, we developed an automated method to mimic the manual approach. Other general methods that have been used for determining the number of clusters in other types of data include the silhouette (54) or v-fold cross validation (55), but these require many rounds of clustering for optimization and are computationally intractable for the large size of AIRR-Seq datasets. The gap statistic (56) is also not applicable since it requires a null distribution of expected within-cluster dispersion, which is unknown and would require several assumptions to simulate for Ig sequences. Thus, the automated threshold inference based on the “distance-to-nearest” plot proved most feasible for the data type and is supported by biological intuition.

Hierarchical clustering is an agglomerative (or “bottom up”) method. Each sequence starts as its own cluster, and the closest pair of clusters is merged together until all sequences are connected. Closeness is defined by a distance metric. Many previous studies used Hamming distance (21, 32), which simply counts the number of differences between two junction sequences. Others attempted to incorporate the intrinsic biases of somatic hypermutation to account for the presence of hot- and cold-spots (15, 18). Here we found that incorporating the targeting and substitution biases of SHM into the distance metric did not significantly improve performance compared to nucleotide Hamming distance. It is possible that more sophisticated distance measures could play a more important role under conditions different from those investigated here. For example, when the mutation frequency is low, a different metric may better capture the importance of each individual mutation in determining the separation between clones. However, the current results suggest that the additional assumptions and computational cost of more complex distance metrics are unlikely to provide substantial performance improvements.

While distance is measured between pairs of individual sequences, the linkage method defines how to calculate the closeness between clusters that contain multiple sequences. We evaluated the tradeoffs in the most common linkage methods: single, average and complete. Single linkage is generally considered to be the most inclusive and we found that it provides the best overall performance with specificity, sensitivity, and PPV all over 99%. However, the appropriate choice of linkage may depend on the biological question being addressed. Complete linkage offers a higher PPV, but at the cost of a significant loss of sensitivity. This may be appropriate for research questions that are highly dependent on the accuracy of calling sequences as part of the same clone. For example, studies that attempt to link small numbers of antigen-specific sequences with clonal relatives or establish migration patterns between compartments with infrequent overlaps may benefit from the high confidence in each clonal connection provided by complete linkage. Nevertheless, the high absolute performance of single linkage should be acceptable for most studies.

The specificity, sensitivity, and PPV of single linkage clustering with Hamming distance are all over 99%. However, the errors that are made by this algorithm are not random. We found that Ig sequences incorrectly grouped together as clonally-related had disproportionately short junction regions (here defined as less than 30 nt). Since the V gene extends into the junction region by approximately seven nucleotides (43), a higher fraction of the nucleotides in short junctions would be expected to have limited diversity compared with longer junctions, potentially limiting the ability to distinguish between clones. This could be particularly problematic when using a length normalized distance metric, which we showed was critical to achieve acceptable specificity. However, our analysis showed that the problem went beyond the V segment constituting a higher fraction of junction nucleotides. Clonally unrelated sequences with short junctions have less entropy on a per nucleotide basis compared to similar sequences with longer junctions. This lack of inter-clonal diversity could be due to a lower mutation frequency, the use of a restricted set of D genes, or fewer untemplated nucleotide additions between the germline gene segments. However, the entropy of clonally related sequences was comparable between short and longer junctions, suggesting that a uniformly lower mutation frequency is not responsible for the lower diversity in short junctions. Current algorithms for inferring germline gene segments still struggle with inference of the D gene (57), making it difficult to determine if the underlying cause of low diversity is due to D gene usage bias, fewer untemplated nucleotide additions, or another mechanism.

Despite the diversity differences between shorter and longer junctions, using a separate threshold to partition sequences with different junction lengths did not improve performance. As precision-recall curves showed, the single threshold selected by analyzing the entire dataset as a whole almost always optimized the trade-off between sensitivity and PPV for all junction lengths. While a few repertoires did have an alternate threshold with slightly improved performance, these thresholds were not evident from the “distance-to-nearest” distributions. It is possible a method other than hierarchical clustering could better separate clones with shorter junctions, but this would be a minor improvement as absolute performance of the single linkage hierarchical clustering with Hamming distance was high.

False positive clonal assignments still occur among sequences with longer junctions, but these appear to have a different underlying cause. In this case the lack of nucleotide diversity can be explained, at least in part, by an over-representation of the IGHJ6 gene. This gene extends an extra ten nucleotides into the junction region (43), causing sequences to appear more similar than others using different J genes. It is possible that a separate analysis of these sequences may improve performance. One possibility for better separating clones that do not have sufficient diversity in the junction region is to require shared mutations in the V or J region, although this would penalize clones that have few mutations overall. Likelihood-based approaches, such as Cloanalyst (28) or partis (29), may help to increase confidence in clones with short junctions or those using IGHJ6, although these approaches are too computationally intensive to use on full AIRR-Seq data sets. While it has been suggested that partis improves performance relative to hierarchal clustering (29), this study did not use dataset-specific distance thresholds and thus likely dramatically underestimated the performance of the clustering-based method.

The comparative analysis presented here suggests clear tradeoffs in the choice of distance and linkage methods. However, it is possible that different tradeoffs would become apparent in data with different clone size distributions, mutation frequencies, etc. The simulation data used to measure sensitivity and PPV were based on lineage tree topologies drawn from only four underlying repertoires. The simulations also assume that Ig sequences maintain the same junction length during clonal evolution, an assumption that was also made in the clustering algorithm. However, recent research indicates that a small percentage of SHM events may lead to changes in junction length within a clone (58). Insertions/deletions may be present in the junction due to sequencing errors, but the inclusion of UMIs followed by computational approaches for sequencing error-correction can reduce this impact (6, 59). Few clonal grouping methods deal with junction length differences, and while these effects are also not accounted for in the current study, their influence on performance is expected to be small. Another possible source of bias in the performance on experimental data is the potential presence of so-called “public clones,” or highly similar sequences across individuals. Such sequences may skew specificity estimates that were approximated on publicly available human experimental datasets based on the frequency of inferred groups that spanned individuals. Furthermore, this specificity measure depends on the frequency of highly similar sequences found across individuals, which may differ from the frequency of highly similar sequences found within an individual by chance. “Public clones” are especially prevalent in light chains due to the relative lack of diversity in the junction region, caused in part by the absence of the D gene segment. Clonal grouping using only light chain sequences is expected to have lower performance than the heavy chain results shown here. Future studies could benefit from using a larger number of AIRR-Seq datasets that span age, tissue, disease state, etc. in addition to simulations based on a larger number of underlying experimental Ig repertoires.

In summary, computational methods for grouping Ig sequences into B cell clones is a critical part of AIRR-seq studies, and allows for understanding the structure and affinity maturation of the Ig repertoire. Here we developed a framework for comparative analysis of clonal grouping approaches and determined that single linkage hierarchical clustering with length-normalized nucleotide Hamming distance performs well on both human experimental and

simulated datasets. This algorithm is available as part of the Change-O and SHazaM packages (35) in our Immcantation tool suite (http://immcantation.readthedocs.io).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Gur Yaari and Moriah Cohen for pre-processing of the influenza vaccination data used to generate the simulated data and for helpful comments on the manuscript. The authors also thank the HPC facilities operated by the Yale Center for Research Computing and Yale's W. M. Keck Biotechnology Laboratory.

References

1. Volpe JM, Kepler TB. Large-scale analysis of human heavy chain V(D)J recombination patterns. Immunome Res. 2008; 4:3. [PubMed: 18304322]
2. Kleinstein SH, Louzoun Y, Shlomchik MJ. Estimating hypermutation rates from clonal tree data. J Immunol. 2003; 171:4639–4649. [PubMed: 14568938]
3. McKean D, Huppi K, Bell M, Staudt L, Gerhard W, Weigert M. Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. Proc Natl Acad Sci. 1984; 81:3180–3184. [PubMed: 6203114]
4. Wu YCB, Kipling D, Dunn-Walters DK. Age-Related Changes in Human Peripheral Blood IGH Repertoire Following Vaccination. Front Immunol. 2012; 3:193. [PubMed: 22787463]
5. Boyd SD, Joshi SA. High-Throughput DNA Sequencing Analysis of Antibody Repertoires. Microbiol Spectr. 2014; 2 AID-0017-2014.
6. Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. Genome Med. 2015; 7:121. [PubMed: 26589402]
7. Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. Science. 2009; 324:807–10. [PubMed: 19423829]
8. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD, Nadeau KC, Egholm M, Miklos DB, Zehnder JL, Fire AZ. Measurement and Clinical Monitoring of Human Lymphocyte Clonality by Massively Parallel V-D-J Pyrosequencing. Sci Transl Med. 2009; 1:12ra23–12ra23.
9. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology. 2012; 135:183–91. [PubMed: 22043864]
10. Logan AC, Gao H, Wang C, Sahaf B, Jones CD, Marshall EL, Buno I, Armstrong R, Fire AZ, Weinberg KI, Mindrinos M, Zehnder JL, Boyd SD, Xiao W, Davis RW, Miklos DB. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. Proc Natl Acad Sci. 2011; 108:21194–21199. [PubMed: 22160699]
11. Hershberg U, Meng W, Zhang B, Haff N, St Clair EW, Cohen PL, McNair PD, Li L, Levesque MC, Luning Prak ET. Persistence and selection of an expanded B-cell clone in the setting of rituximab therapy for Sjogren's syndrome. Arthritis Res Ther. 2014; 16:R51. [PubMed: 24517398]
12. Boletis JN, Marinaki S, Skalioti C, Lionaki SS, Iniotaki A, Sfikakis PP. Rituximab and mycophenolate mofetil for relapsing proliferative lupus nephritis: A long-term prospective study. Nephrol Dial Transplant. 2009; 24:2157–2160. [PubMed: 19179411]
13. Ademokun A, Wu YCC, Martin V, Mitra R, Sack U, Baxendale H, Kipling D, Dunn-Walters DK. Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. Aging Cell. 2011; 10:922–30. [PubMed: 21726404]
14. Wang C, Liu Y, Xu LT, Jackson KJ, Roskin KM, Pham TD, Laserson J, Marshall EL, Seo K, Lee JY, Furman D, Koller D, Dekker CL, Davis MM, Fire AZ, Boyd SD. Effects of aging, cytomegalovirus infection, and EBV infection on human B cell repertoires. J Immunol. 2014; 192:603–11. [PubMed: 24337376]

15. Stern JNH, Yaari G, Vander Heiden JA, Church GM, Donahue WF, Hintzen RQ, Huttner AJ, Laman JD, Nagra RM, Nylander A, Pitt D, Ramanan S, Siddiqui BA, Vigneault F, Kleinstein SH, Hafler DA, O'Connor KC. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. Sci Transl Med. 2014; 6:248ra107–248ra107.
16. Wu YCB, James LK, Vander Heiden JA, Uduman M, Durham SR, Kleinstein SH, Kipling D, Gould HJ. Influence of seasonal exposure to grass pollen on local and peripheral blood IgE repertoires in patients with allergic rhinitis. J Allergy Clin Immunol. 2014; 134:604–612. [PubMed: 25171866]
17. Zhu J, Ofek G, Yang Y, Zhang B, Louder MK, Lu G, McKee K, Pancera M, Skinner J, Zhang Z, Parks R, Eudailey J, Lloyd KE, Blinn J, Alam SM, Haynes BF, Simek M, Burton DR, Koff WC, Mullikin JC, Mascola JR, Shapiro L, Kwong PD. Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. Proc Natl Acad Sci. 2013; 110:6470–5. [PubMed: 23536288]
18. Tsioris K, Gupta NT, Ogunniyi AO, Zimnisky RM, Qian F, Yao Y, Wang X, Stern JNH, Chari R, Briggs AW, Clouser CR, Vigneault F, Church GM, Garcia MN, Murray KO, Montgomery RR, Kleinstein SH, Love JC. Neutralizing antibodies against West Nile virus identified directly from human B cells by single-cell analysis and next generation sequencing. Integr Biol. 2015; 7:1587–1597.
19. Zhu J, Wu X, Zhang B, McKee K, O'Dell S, Soto C, Zhou T, Casazza JP, Mullikin JC, Kwong PD, Mascola JR, Shapiro L. De novo identification of VRC01 class HIV-1-neutralizing antibodies by next-generation sequencing of B-cell transcripts. Proc Natl Acad Sci. 2013; 110:E4088–97. [PubMed: 24106303]
20. Truck J, Ramasamy MN, Galson JD, Rance R, Parkhill J, Lunter G, Pollard AJ, Kelly DF. Identification of Antigen-Specific B Cell Receptor Sequences Using Public Repertoire Analysis. J Immunol. 2015; 194:252–261. [PubMed: 25392534]
21. Jiang N, He J, Weinstein JA, Penland L, Sasaki S, He XS, Dekker CL, Zheng NY, Huang M, Sullivan M, Wilson PC, Greenberg HB, Davis MM, Fisher DS, Quake SR. Lineage structure of the human antibody repertoire in response to influenza vaccination. Sci Transl Med. 2013; 5:171ra19.
22. Hill M. Diversity and evenness: a unifying notation and its consequences. Ecology. 1973; 54:427–432.
23. van Dongen JJM, Langerak AW, Bruggemann M, Evans PAS, Hummel M, Lavender FL, Delabesse E, Davi F, Schuuring E, Garcia-Sanz R, van Krieken JHJM, Droese J, Gonzalez D, Bastard C, White HE, Spaargaren M, Gonzalez M, Parreira A, Smith JL, Morgan GJ, Kneba M, Macintyre EA. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. Leukemia. 2003; 17:2257–2317. [PubMed: 14671650]
24. Nei, M., Kumar, S. Molecular evolution and phylogenetics. Oxford University Press; 2000.
25. Horns F, Vollmers C, Croote D, Mackey SF, Swan GE, Dekker CL, Davis MM, Quake SR. Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching. Elife. 2016:5.
26. Looney TJ, Lee J-Y, Roskin KM, Hoh RA, King J, Glanville J, Liu Y, Pham TD, Dekker CL, Davis MM, Boyd SD. Human B-cell isotype switching origins of IgE. J Allergy Clin Immunol. 2016; 137:579–586.e7. [PubMed: 26309181]
27. Hershberg U, Luning Prak ET. The analysis of clonal expansions in normal and autoimmune B cell repertoires. Philos Trans R Soc Lond B Biol Sci. 2015; 370:20140239. [PubMed: 26194753]
28. Kepler TB. Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors. F1000Research. 2013; 103:1–15.
29. Ralph D, Matsen FA. Likelihood-based inference of B-cell clonal families. 2016:1–46. arXiv.
30. Chen Z, Collins AM, Wang Y, Gaeta BA. Clustering-based identification of clonally-related immunoglobulin gene sequence sets. Immunome Res. 2010; 6:S4.
31. Glanville J, Kuo TC, von Büdingen HC, Guey L, Berka J, Sundar PD, Huerta G, Mehta GR, Oksenberg JR, Hauser SL, Cox DR, Rajpal A, Pons J. Naive antibody gene-segment frequencies

are heritable and unaltered by chronic lymphocyte ablation. Proc Natl Acad Sci. 2011; 108:20066–71. [PubMed: 22123975]

32. Jiang N, Weinstein JA, Penland L, White RA, Fisher DS, Quake SR. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. Proc Natl Acad Sci. 2011; 108:5348–5353. [PubMed: 21393572]
33. Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. Blood. 2010; 116:1070–8. [PubMed: 20457872]
34. Briney B, Le K, Zhu J, Burton DR. Clonify: unseeded antibody lineage assignment from next-generation sequencing data. Sci Rep. 2016; 6:23901. [PubMed: 27102563]
35. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. Bioinformatics. 2015; 31:3356–3358. [PubMed: 26069265]
36. Parameswaran P, Liu Y, Roskin KM, Jackson KKL, Dixit VP, Lee JY, Artiles KL, Zompi S, Vargas MJ, Simen BB, Hanczaruk B, McGowan KR, Tariq MA, Pourmand N, Koller D, Balmaseda A, Boyd SD, Harris E, Fire AZ. Convergent Antibody Signatures in Human Dengue. Cell Host Microbe. 2013; 13:691–700. [PubMed: 23768493]
37. Vander Heiden JA, Yaari G, Uduman M, Stern JNH, O'Connor KC, Hafler DA, Vigneault F, Kleinstein SH. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics. 2014; 30:1930–1932. [PubMed: 24618469]
38. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT® tools for the nucleotide analysis of immunoglobulin (IG) and t cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. Methods Mol Biol. 2012; 882:569–604. [PubMed: 22665256]
39. Laserson U, Vigneault F, Gadala-Maria D, Yaari G, Uduman M, Vander Heiden JA, Kelton W, Taek Jung S, Liu Y, Laserson J, Chari R, Lee J-H, Bachelet I, Hickey B, Lieberman-Aiden E, Hanczaruk B, Simen BB, Egholm M, Koller D, Georgiou G, Kleinstein SH, Church GM. High-resolution antibody dynamics of vaccine-induced immune responses. Proc Natl Acad Sci. 2014; 111:4928–4933. [PubMed: 24639495]
40. Di Niro R, Lee SJ, Vander Heiden JA, Elsner RA, Trivedi N, Bannock JM, Gupta NT, Kleinstein SH, Vigneault F, Gilbert TJ, Meffre E, McSorley SJ, Shlomchik MJ. Salmonella Infection Drives Promiscuous B Cell Activation Followed by Extrafollicular Affinity Maturation. Immunity. 2015; 43:120–131. [PubMed: 26187411]
41. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. [PubMed: 15034147]
42. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 2013; 41:1–7. [PubMed: 23143271]
43. Giudicelli V, Chaume D, Lefranc MP. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res. 2005; 33:D256–61. [PubMed: 15608191]
44. Yaari G, Vander Heiden JA, Uduman M, Gadala-Maria D, Gupta N, Stern JNH, O'Connor KC, Hafler DA, Laserson U, Vigneault F, Kleinstein SH. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. Front Immunol. 2013; 4:358. [PubMed: 24298272]
45. Shapiro GS, Ellison MC, Wysocki LJ. Sequence-specific targeting of two bases on both DNA strands by the somatic hypermutation mechanism. Mol Immunol. 2003; 40:287–95. [PubMed: 12943801]
46. Wand, MP., Jones, MC. Kernel Smoothing, Vol. 60 of Monographs on statistics and applied probability. Chapman and Hall; London: 1995.
47. Darlington R. Is kurtosis really "peakedness?". Am Stat. 1970
48. Hansen, B. Bandwidth selection for nonparametric distribution estimation. Univ. Wisconsin; 2004. manuscript
49. Guidoum, AC. R Packag. version 1.0.3. 2015. kedd: Kernel estimator and bandwidth selection for density and its derivatives.

50. Jain, AK., Dubes, RC. Algorithms for Clustering Data. Prentice-Hall, Inc; Upper Saddle River, NJ, USA: 1988.
51. Tas JMJ, Mesin L, Pasqual G, Targ S, Jacobsen JT, Mano YM, Chen CS, Weill JC, Reynaud CA, Browne EP, Meyer-Hermann M, Victora GD. Visualizing antibody affinity maturation in germinal centers. Science. 2016; 351:1048–54. [PubMed: 26912368]
52. Safonova Y, Lapidus A, Lill J. IgSimulator: a versatile immunosequencing simulator. Bioinformatics. 2015:1–2.
53. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. Nucleic Acids Res. 2012; 40:e134. [PubMed: 22641856]
54. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987; 20:53–65.
55. Statsoft, I. Electronic Statistics Textbook. 2013.
56. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B (Statistical Methodol. 2001; 63:411–423.
57. Munshaw S, Kepler TB. SoDA2: A Hidden Markov Model approach for identification of immunoglobulin rearrangements. Bioinformatics. 2010; 26:867–872. [PubMed: 20147303]
58. Yeap LS, Hwang JK, Du Z, Meyers RM, Meng FL, Jakubauskaitė A, Liu M, Mani V, Neuberg D, Kepler TB, Wang JH, Alt FW. Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. Cell. 2015; 163:1124–1137. [PubMed: 26582132]
59. Shugay M, Britanova OV, Merzlyak EM, Turchaninova Ma, Mamedov IZ, Tuganbaev TR, Bolotin Da, Staroverov DB, Putintseva EV, Plevova K, Linnemann C, Shagin D, Pospisilova S, Lukyanov S, Schumacher TN, Chudakov DM. Towards error-free profiling of immune repertoires. Nat Methods. 2014; 11:653–5. [PubMed: 24793455]

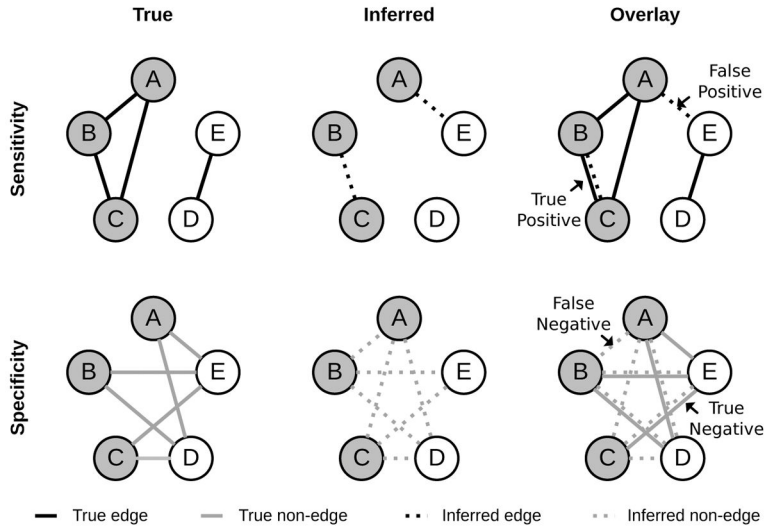

Figure 1. Overview of method for calculating the performance measures: sensitivity, PPV and specificity

Each node represents a sequence that belongs to either of two clones (grey or white clone fill color). In the top-left panel, sequences that are truly from the same clone are connected with solid lines. In the top-middle panel, sequences that are inferred to be from the same clone are connected with dashed lines. Both true and inferred relationships are shown in the top-right panel. In the bottom-left panel, sequences that are clonally unrelated are connected by solid lines. In the bottom-middle panel, sequences that are inferred to be clonally unrelated are connected by dashed lines. The overlap of true and inferred unrelated sequences is shown in the bottom-right panel. The true and inferred edges in the right panels representing clonally related (top) or unrelated (bottom) sequences are compared to assess performance. Sensitivity is defined by the number of true positive edges (dashed-solid double lines in the top-right panel) divided by the number of true edges (solid lines in the top-right panel). PPV is the number of true positive edges (dashed-solid double lines in the top-right panel) divided by the number of inferred edges (dashed lines in the top-right panel). Specificity is number of true negative edges (dashed-solid double lines in the bottom-right panel) divided by the number of true non-edges (solid lines in the bottom-right panel).

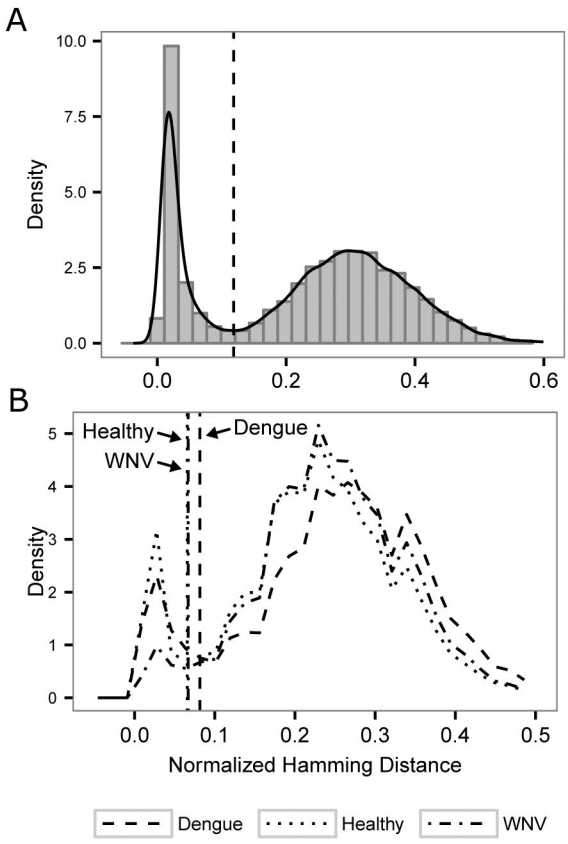

Figure 2. Analysis of the "distance-to-nearest" neighbor plot to define the distance threshold for partitioning clones

For each sequence, the length-normalized nucleotide Hamming distance to every other sequence was calculated, and the nearest (non-zero) neighbor was identified. (A) The histogram of nearest neighbor distances for a simulated dataset was fit using a density estimation of the distribution (solid line), and this fitting was then used to automatically infer a threshold that separated the two modes of the distribution (dotted vertical line). (B) Nearest neighbor distributions were calculated for the Dengue (solid line), Healthy (dashed line), and WNV (dotted line) experimental datasets. Inferred thresholds for each of these human data sets are indicated by the vertical lines.

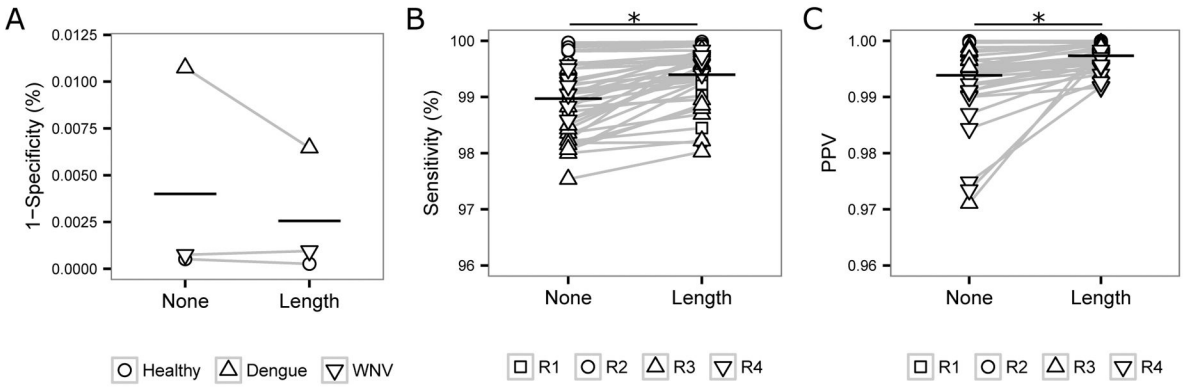

Figure 3. Length normalization of the distance measure increases performance

Single linkage hierarchical clustering was used to identify clonally-related sequences using a distance metric based on the absolute Hamming distance of the junction sequences (None), or the Hamming distance normalized by the length of the junction (Length). (A) Specificity was calculated using three human experimental data sets (Healthy, Dengue, and WNV). Sensitivity (B) and PPV (C) were calculated using 40 simulated datasets based on four experimentally observed sets of clonal lineage structures (R1–R4). Bars indicate mean performance. * $p < 0.0001$ by paired t-test.

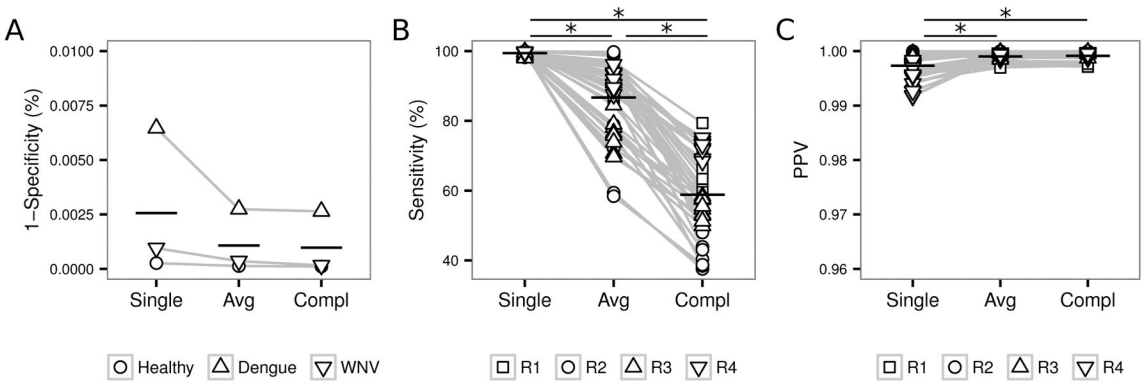

Figure 4. Single linkage clustering provides the highest sensitivity, with minimal loss of specificity or PPV

Hierarchical clustering was used to identify clonally-related sequences using length-normalized Hamming distance and Single (Single), Average (Avg) or Complete (Compl) linkage. (A) Specificity was calculated using three human experimental data sets (Healthy, Dengue and WNV). Sensitivity (B) and PPV (C) were calculated using 40 simulated datasets based on four experimentally observed sets of clonal lineage structures (R1–R4). Bars indicate mean performance. * $p < 0.0001$ by paired t-test.

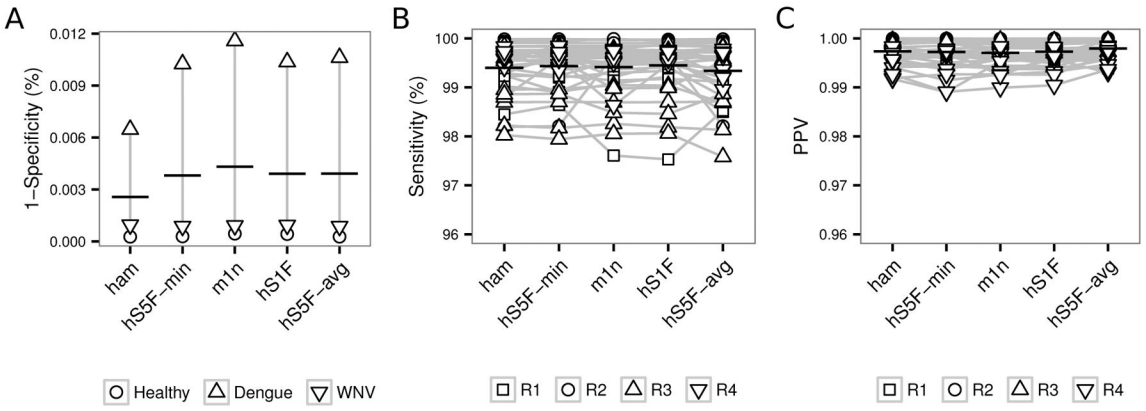

Figure 5. Including hot- and cold-spot biases in the distance measure does not significantly impact the performance of clonal grouping

Single linkage hierarchical clustering was used to identify clonally-related sequences using length-normalized nucleotide Hamming distance (ham), as well as other distance metrics that incorporated varying SHM biases as described in Materials and Methods: hS5F-min, m1n, hS1F, and hS5F-avg. (A) Specificity was calculated using three human experimental data sets (Healthy, Dengue and WNV). Sensitivity (B) and PPV (C) were calculated using 40 simulated datasets based on four experimentally observed sets of clonal lineage structures (R1–R4). Bars indicate mean performance.

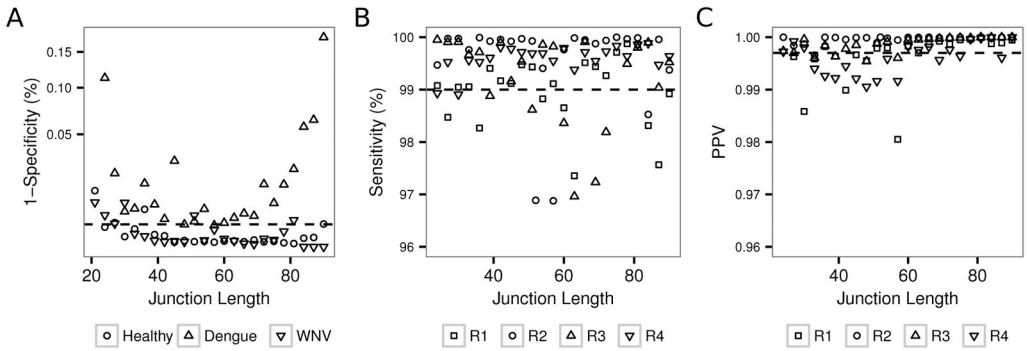

Figure 6. PPV is decreased among sequences with smaller junction lengths

Single linkage hierarchical clustering was used to identify clonally-related sequences using length-normalized nucleotide Hamming distance. (A) Specificity was calculated using three human experimental data sets (Healthy, Dengue and WNV). Sensitivity (B) and PPV (C) were calculated using 40 simulated datasets based on four experimentally observed sets of clonal lineage structures (R1–R4). Horizontal dashed lines are shown at an arbitrary value in each panel to highlight trends.

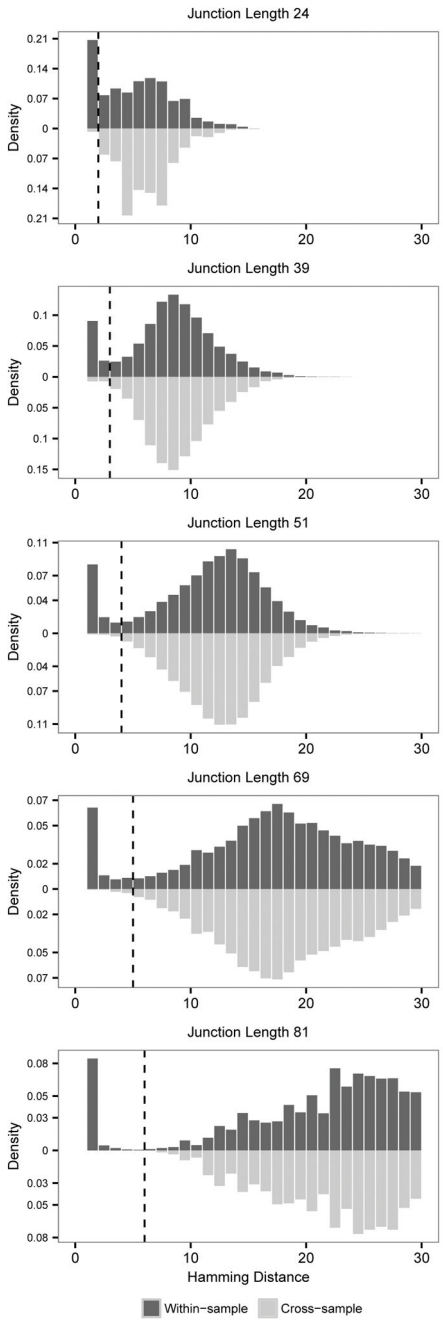

Figure 7. Peaks in the "distance-to-nearest" neighbor distribution begin to converge at small junction lengths

Nucleotide Hamming distance to nearest neighbor distributions were calculated for sequences with junctions of length 24, 39, 51, 69, and 81 nucleotides from the Healthy experimental dataset. Nearest neighbors were defined using sequences within the same subject (dark grey bars), or by using sequences from all other subjects (light grey bars). The single distance threshold inferred using normalized Hamming distance on all junction lengths in the Healthy dataset is shown by the dashed line in each distribution.

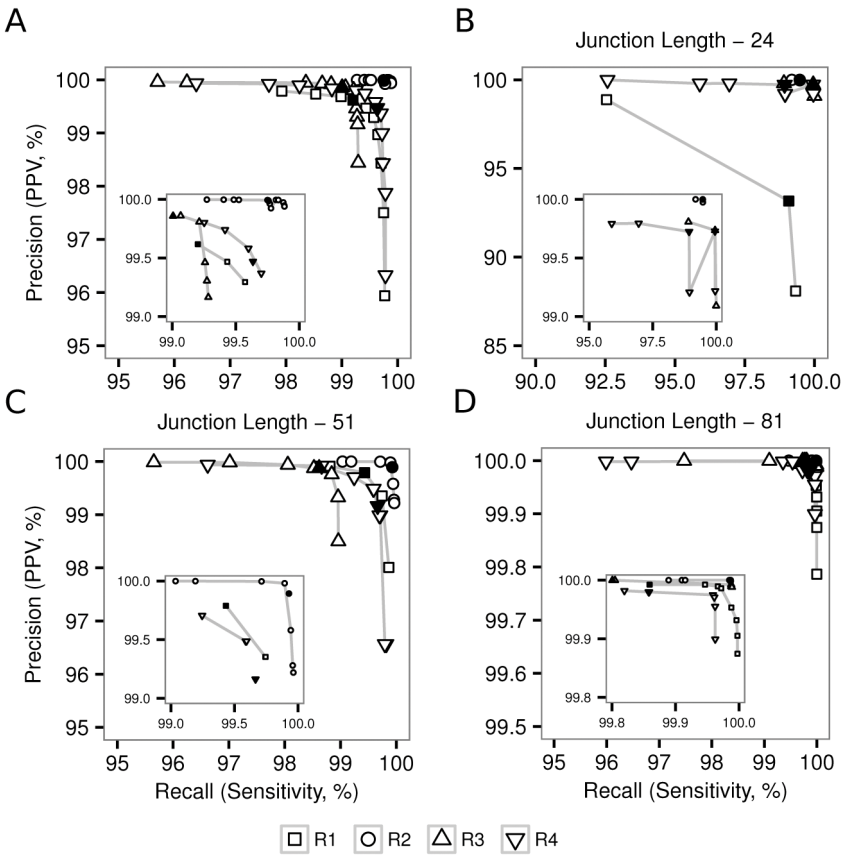

Figure 8. A single distance threshold is near-optimal for all junction lengths

Single linkage hierarchical clustering with length-normalized nucleotide Hamming distance was used to identify clonally-related sequences in 40 simulated datasets based on four experimentally observed sets of clonal lineage structures (R1–R4). Precision-recall curves were generated by varying the distance threshold from 0.10 to 0.20 at intervals of 0.01. The precision (PPV) and recall (sensitivity) of each run were averaged across the ten simulations of each repertoire in (A) sequences of all lengths, and sequences with junctions of length (B) 24, (C) 51 and (D) 81 nucleotides. The performance of the algorithm run with the inferred threshold is shown by filled points in all panels. Insets show the same data zoomed in on the upper right of the plot.

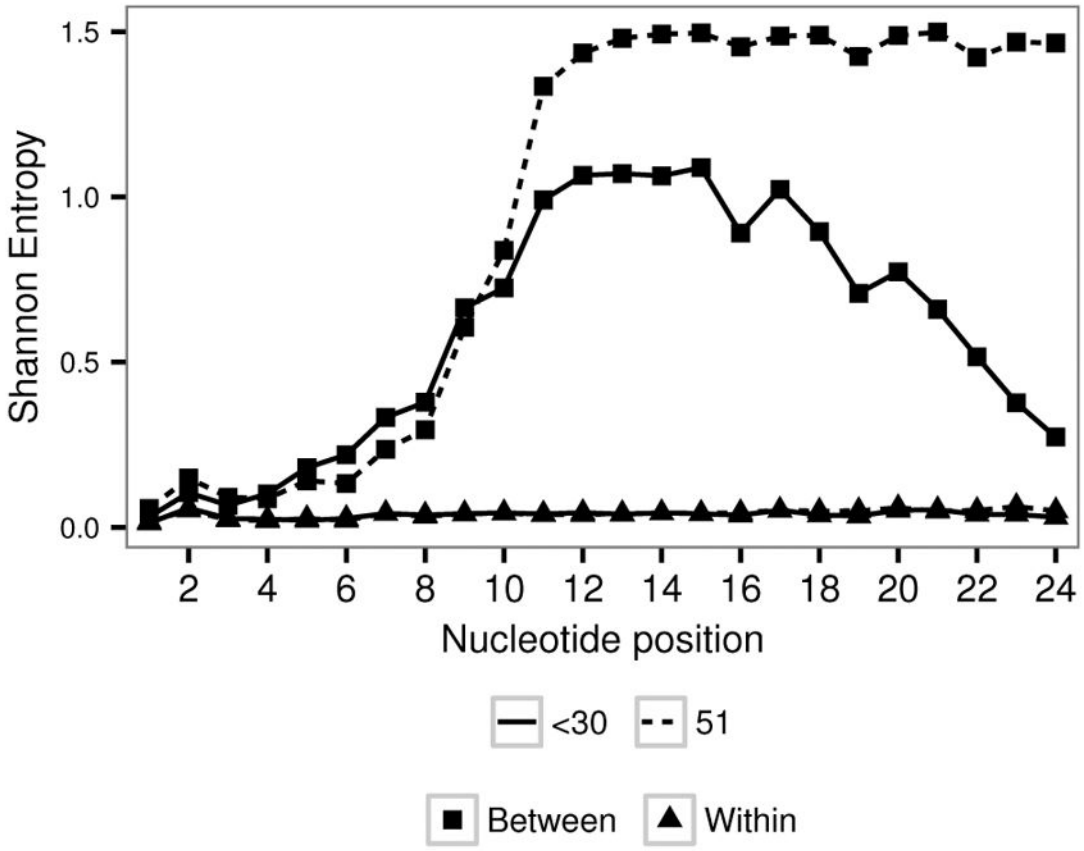

Figure 9. Unrelated sequences with shorter junctions have lower entropy per nucleotide

(A) Shannon entropy was calculated for each of the first 24 junction nucleotides of clonally related (triangle) and unrelated (square) sequences with junctions <30 nt in length (solid lines) and 51 nt in length (dashed lines). Error bars are standard error of the mean entropy across all sequences in the given group.