Sequence analysis

Somatic hypermutation analysis for improved identification of B cell clonal families from next-generation sequencing data

Nima Nouri $^{1,2}$ and Steven H. Kleinstein $^{1,2,3,*}$

$^1$Department of Pathology, Yale School of Medicine, New Haven, CT 06511, USA.
$^2$Center for Medical Informatics, Yale School of Medicine, New Haven, CT 06511, USA.
$^3$Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511, USA.

*To whom correspondence should be addressed.

Abstract

**Motivation:** Adaptive immune receptor repertoire sequencing (AIRR-Seq) offers the possibility of identifying and tracking B cell clonal expansions during adaptive immune responses. Members of a B cell clone are descended from a common ancestor and share the same initial V(D)J rearrangement, but their BCR sequence may differ due to the accumulation of somatic hypermutations (SHMs). Clonal relationships are learned from AIRR-seq data by analyzing the BCR sequence, with the most common methods focused on the highly diverse CDR3 region. However, clonally related cells often share SHMs which have been accumulated during affinity maturation. Here, we investigate whether shared SHMs in the V and J segments of the BCR can be leveraged along with the CDR3 sequence to improve the ability to identify clonally related sequences. We develop independent distance functions that capture shared mutations and CDR3 similarity, and combine these in a spectral clustering framework. Using simulated data, we show that this model improves both the sensitivity and specificity for identifying clonal relationships.

**Availability:** Source code for this method is freely available in the SCOPer (Spectral Clustering for clOne Partitioning) R package (version 0.2 or newer) in the Immcantation framework: [www.immcantation.org](http://www.immcantation.org) under the CC BY-SA 4.0 license.

**Contact:** steven.kleinstein@yale.edu

1 Introduction

B cells recognize pathogens through their B cell receptor (BCR). The ability to recognize and initiate a response to a wide variety of pathogens depends upon a large population of B cell lymphocytes each of which expresses a particular receptor for antigen. The diversity of the BCRs (also referred to as Immunoglobulins, (Igs)) is a result of genetic recombination and diversification mechanisms. BCRs are comprised of two identical heavy (IGH) and light (IGL) chain proteins. For IGH-chains diversity is initially created in the germline via recombination of variable IGHV, diversity IGHD, and joining IGHJ genes (termed the V(D)J recombination process (Tonegawa, 1983)). Diversity in IGH is further increased by addition of P- and N-nucleotides at the IGHV/IGHD and IGHD/IGHJ boundaries (Alt and Baltimore, 1982; Lafaille et al., 1989; Murphy, 2011). For IGL, the IGLV gene is rearranged directly to IGLJ gene. The region where IGHV, IGHD and IGHJ come together in IGH (or IGLV and IGLJ for IGL) is termed the CDR3, and this high diversity region is often involved in antigen-binding (Xu and Davis, 2000).

During T-dependent responses, antigen-activated B cells undergo clonal expansion and acquire additional diversity through somatic hypermutation (SHM), an enzymatically-driven process introducing point substitutions into the BCR locus at a rate of $\sim 1/1000$ bp/cell division (McKean et al., 1984; Wood et al., 2001). B cells that acquire mutations that improve their ability to bind the pathogen are preferentially expanded leading to affinity maturation of the B cell population over time. Therefore, SHMs have important consequences for the kinetics (antibody-response), quality (antigen-specificity), and size (response-signature) of the B cell clones as the fundamental building blocks of immune repertoires (Kepler and Perelson, 1993).

Accurate identification of clonal relationships is important, as these clonal groups form the basis for a wide range of repertoire analysis, including diversity analysis (Robins et al., 2013; Meng et al., 2017; Rosenfeld et al., 2018), lineage reconstruction and detection of antigen-specific sequences (Yaari and Kleinstein, 2015; Tsioris et al., 2015) and effector functionality (McKean et al., 1984; Sablitzky et al., 1985). One way to monitor and track the B cell clonal lineages is to perform large-scale sampling of bulk B cell populations, amplifying, and sequencing the expressed antibody gene rearrangements by next-generation sequencing (NGS) (Metzker, 2010). Recent studies by NGS have greatly expanded our understanding of B cell clonal lineage development in high-throughput Adaptive Immune Receptor Repertoire sequencing (AIRR-seq) data (Boyd...
Fig. 1. A B cell lineage tree showing the relationships between clonally-related cells. The germline sequence (diamond) is shown at the root of lineage, and is connected by a single branch to the most recent common ancestor (MRCA) (square). This branch consists of mutations that are shared across all members of a clone. Several sub-branches descend from the MRCA to inferred sequences (triangles) carrying mutations that are shared by a subset of clone members. Finally, the inferred sequences are connected to observed sequences (circles) through mutations that are unique to each given observed sequence. Shared and unique mutations are marked at each branch by horizontal lines and arrowheads, respectively.

Antibody diversity is largely dominated by the IGH-chain (Xu and Davis, 2000). The IGH-chain owes this diversity to the: (1) use of an IGHD gene, which IGL-chains lack, (2) addition of short palindromic (P) nucleotides at the IGHV-IGHD and IGHD-IGHJ joints (Lafaille et al., 1989), (3) insertion of non-templated (N) nucleotides at the IGHV-IGHD and IGHD-IGHJ joints by terminal deoxynucleotidyl transferase (TdT) (Alt and Baltimore, 1982), and (4) higher rates of SHM than IGL-chains (Wood et al., 2001). The IGH-chain junction region (i.e. complementarity determining region 3, CDR3, plus the conserved flanking amino acid residues) commonly serves as an identifier for clonality inference methodologies. For instance, sequences whose junctions are identical or have a high degree of homology (measured by string distance at the nucleotide level) are often classified as belonging to the same clone (Hershberg and Prak, 2015). However, to avoid grouping together highly homologous yet distinct sequences, some studies also separate these groups by their constituent IGHV- and IGHJ-gene annotations (Zhang et al., 2015). Many methods also assume that members of a clone share the same junction length, because SHMs introduced into the BCR sequence are predominantly point substitutions (McKean et al., 1984; Kleinstein et al., 2003). In a different approach, probabilistic models have also been developed to calculate the likelihood of sharing a common B cell ancestor and subsequently infer clonal grouping (Kepler, 2013; Ralph and Matsen IV, 2016). Kepler (2013) reconstructs a B-cell clonal lineage using the posterior distribution over clone members possible ancestors, and Ralph and Matsen IV (2016) infers the clonally related sequences using a multi-hidden Markov Model (multi-HMM). However, both methodologies have complexities that become substantially expensive for large sequencing datasets. Overall, in practice, a common approach is to infer clones among sequences with high junction region similarity, as well as identical junction length and IGHV- and IGHJ-gene usage (referred to here as a distance-based model) (Hershberg and Prak, 2015).

While distance-based strategies are common among current studies, clonal relationship inference solely based on the similarity of the junction region does not leverage the potential information in the V and J segments. It has been suggested that incorporating shared SHMs in these regions could improve distance-based clonal inference (Zhou and Kleinstein, 2019). Members of an expanded B cell clone often share specific somatic mutations and, sometimes, combinations of mutations across BCR. Mutations may be shared among two or more members of a clone as a simple result of being passed down during cell division, or may be positively selected as part of the affinity maturation process (Clarke et al., 1985; Blier and Bothwell, 1987; Diamond et al., 1992; Coker et al., 2003; Furuta et al., 2017). This hierarchy of shared mutations can be considered as the “glue” binding all the members of a B cell clone together and shaping its lineage tree (Figure 1). This additional IGH-chain information could be leveraged to refine clonal relationships.

In this study, we investigated whether shared SHM patterns in the V and J segments of the BCR can be leveraged along with the CDR3 sequence to improve the ability to identify clonally related sequences. This model is implemented in the new version of SCOPer. The first version of SCOPer, a spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data, was presented in Nouri and Kleinstein (2018b). In the following sections, we discuss the main steps of the methodology and explain our implementation of the recent improvements upon the original framework. We further examine the performance of SCOPer using simulated and experimental datasets.

2 Method

The clonal inference procedure by SCOPer is performed as follows (Figure 2). First, BCR sequences IGHV and IGHJ genes are identified. This can be done using various publicly available tools such as IMGT/HighV-QUEST (Alamyat et al., 2012) or IgBLAST (Ye et al., 2013). Then, sequences are partitioned into groups that share the same IGHV-gene and IGHJ-gene (gene-level grouping). The gene-level grouping is based on the assumption that the identity of germline gene (the clone members unmutated common ancestor) cannot change through affinity maturation. Sequences are further assumed to evolve only through point mutation (no indels), so a sub grouping level is also applied in order to force sequences in the same clonal family to have identical junction region length (length-level grouping). Henceforth, we refer to such a group as “VJ(ℓ)-group”. Next,
each VJ(\ell)-group is retrieved for inferring the BCR clonal relationships using spectral clustering-based approach.

2.1 Distance matrix calculation

The distance-based step of SCOPer is focused on the sequencing reads’ junction region. At this step, we generate a symmetric and positive pair-wise similarity matrix \( X_{ij} \) defined by the Hamming distance between the junction regions corresponding to the \( i^{th} \) and \( j^{th} \) sequences from a given VJ(\ell)-group. This is called the “junction-targeted” distance matrix. The Hamming distance is defined as the number of positions at which the corresponding nucleotides are different. The distance matrix can also be generated from CDR3 region by excluding the three-nucleotide prefix and suffix from both ends of the junction (i.e. converting junction segment to CDR3 region). Henceforth, this is called a “CDR3-targeted” distance matrix.

2.2 Stress matrix calculation

The stress-based step of SCOPer is focused on the IGHV and IGJH regions. We develop a model in which the occurrence of a mutation at the same nucleotide position of a pair of sequences (the so-called “pair-wise shared mutation”) will be used, accompanying with distance-based step, to infer their clonal relationship. These shared mutations can take place early in the clonal expansion, or be positively selected during affinity maturation, and can be inherited by all, a subset, or at least a pair of the descendants of a common ancestor (Yaari et al., 2015). The model is implemented so that a pair of sequences with a higher shared mutation rate is more likely to belong to the same clone, whereas a pair of sequences with a lower shared mutation rate are considered more independent from each other. Recall from continuum mechanics, the pair-wise shared mutations can be loosely referred to as “stress” which expresses the internal forces that clonally related sequences exert on each other.

We begin with identification of the pair-wise mutations. First, depending on the type of distance matrix calculated in the previous step (i.e., junction- or CDR3-targeted) the junction or CDR3 region of the sequences and germlines are masked. Then, for each VJ(\ell)-group a single effective germline is generated by building the effective sequence of all germlines (allele-grouping). This effective germline is deterministic such that if a position contains different nucleotides, the effective will be an IUPAC (International Union of Pure and Applied Chemistry) character representing all of the nucleotides present. Therefore, the effective germline captures all of the information contained in its constituents. Finally, in each VJ(\ell)-group, pairs of sequences are compared with the group effective germline to identify mutations. Using the effective germline ensures a fair comparison among all pairs of sequences in a given group.

We continue with a categorical approach to classify the identified pair-wise mutations (Figure 3). For each pair of \( i^{th} \) and \( j^{th} \) sequences the mutations at each position are flagged with a binary variable and categorized in three classes: (1) a single mutation which occurs only in one of the sequences, \( \alpha_{ij}^{(n)} \), (2) two unique mutations which occur in both sequences, \( \beta_{ij}^{(n)} \), and (3) a shared mutation which occurs in both sequences, \( \gamma_{ij}^{(n)} \). Here, the parameter \( n \) indicates the position of each nucleotide along the sequence string. The binary variables are retrieved to create two matrices. One of the matrices accumulates the total number of mutations:

\[
T_{ij} = \frac{1}{\nu_{ij}} \sum_{n} \left( \alpha_{ij}^{(n)} + 2\beta_{ij}^{(n)} + \gamma_{ij}^{(n)} \right),
\]  

(1)

A second matrix accumulates the shared mutations:

\[
H_{ij} = \frac{2}{\nu_{ij}} \sum_{n} \gamma_{ij}^{(n)}. 
\]  

(2)

Here, \( T_{ij} \) is a positive value and always larger than or equal to positive value \( H_{ij} \). The term \( \nu_{ij} \), average number of informative positions (\( \subset \{A,C,G,T\} \}) in \( i^{th} \) and \( j^{th} \) sequences, is a normalizing factor used to prevent the bias toward pairs of sequences with less non-ACGT positions.

We note that mutational biases have been reported (Elharnati et al., 2015; Yeap et al., 2015) both in the bases that are targeted (Bet et al., 1993; Shapiro et al., 2003) as well as the substitutions that are introduced (Smith et al., 1996; Cowell and Kepler, 2000). These intrinsic biases, combined with the particular codon usage and base composition in BCR sequences, have critical influence on B cell clonal expansion. The SHM biases have been summarized by hot- and cold-spot preferential targeting models (e.g., the WRC/GYW and WA/TW hot-spots, and SYC/GRS cold-spots, where mutation position is underlined). Hence, the influence of a pair-wise shared mutation in the identification of clonal relationships should be constrained based on the micro-sequence context (e.g., a 5-mer motif in which a mutation occurs at the central position). This is because the high likelihood of capturing a shared mutation in hot-spots may bias the clonal inference process. This concern can be addressed by taking advantage of the “5SF” targeting model (a SHM targeting model that produces background likelihood of a particular mutation, based on the surrounding sequence context as well as the mutation itself) for each of the 1024 possible 5-mer motifs (Yaari et al., 2013). Using the 5SF model, the occurrence of a shared mutation at a SHM hot-spot position is considered to be less influential than a shared mutation at a cold- or neutral-spot position, in clonal relationship inference process. Thus, a matrix is generated whose elements are calculated by averaging the mutabilities (an effective mutability) of the germlines 5-mer motifs in which a shared mutation occurred at the central position (Figure 3):

\[
M_{ij} = \prod_{n} \left( 1 - \mu_{ij}^{(n)} \right).
\]  

(3)

Each mutability (\( \mu \)) is subtracted from one to reverse the scaling direction, so that the SHMs at hot-spots become less influential.

We finalize the stress-based step by calling equations 1, 2, and 3 to calculate the stress between \( i^{th} \) and \( j^{th} \) sequences:

\[
S_{ij} = M_{ij} \times N(T_{ij} - H_{ij}|\sigma_T) \times (1 - N(H_{ij}|\sigma_H)).
\]  

(4)
Fig. 4. The stress-based model pulls together clonally-related sequences to improve the B cell clonal inference process. (A) V(D)J recombination generates a set of highly diverse (unmutated) sequences with large distances between independent clones (inter-clonal diversity). Some sequences from independent clones could end up with CDR junction distances and shared SHMs profile similar (dashed-lines), and may lead to false positives in the clonal relationship inference process. (B) The stress between pairs of sequences, expressed via shared mutations, acts as a spring that pulls clonally-related sequences toward each other resulting in a more accurate distinction of local neighborhoods. Black circles indicate observed sequences, while white circles indicate germlines (GL1 and GL2).

Here, \( N(x|\sigma) = \exp(-x^2/2\sigma^2) \) is a continuous Gaussian probability distribution, where parameter \( \sigma_T \) and \( \sigma_H \) are the standard deviations of the \( T \) and \( H \) matrices capturing the variability of total and shared SHM events in each VHV(\( \ell \))-group, respectively. It is important to note that for different VHV(\( \ell \))-groups the level of similarity that indicates common clonality may be different. Therefore, using the Gaussian probability distribution, built upon the given VHV(\( \ell \))-group, will make the model capable of adapting itself to the local level of mutation frequency. We further note that, the stress becomes non-zero only if the number of pair-wise shared mutations is non-zero (\( H_{ij} \neq 0 \)). Conversely, the stress is forced to zero by the third term of Eq. 4, which accumulates more shared mutations will have higher stress, and if the two pairs have the same number of shared mutations, then the pair which accumulates less non-shared mutations, will have higher stress. (Note that \( T_{ij} \) is always larger than or equal to \( H_{ij} \)).

2.3 Graph composition and local scaling

The graph construction at the core of SCOPer relies on a quantitative notion of adaptive local neighborhoods in the dataset, which are encoded by a symmetric Kernel function. The Kernel function is used to capture intrinsic data geometries that approximate underlying manifold models from the data. To construct the kernel graph, first, we generate a weighted-distance matrix in the form of,

\[
W_{ij} = \begin{cases} 
X_{ij} & \text{distance-based model} \\
(1 - S_{ij}) X_{ij} & \text{stress-based model} 
\end{cases}
\]

The model is named “stress-based” when shared mutations are involved in partitioning clones, otherwise it is named “distance-based”. In stress-based model, each stress value \( S_{ij} \) is subtracted from one to reverse the scaling direction, so that the pair of sequences with higher stress become closer to each other, thereby more likely to belong to the same clone. The stress-based model from Equation 5 can be loosely thought of as Hooke’s Law (\( W = \kappa X \), where \( \kappa = 1 - S \)), which rules the attraction force between a pair of sequences using a “spring” with proportionality factor \( \kappa \) (see Figure 4). In the subsequent step, we generate a fully connected graph Kernel using a Gaussian similarity function in the form of,

\[
K_{ij} = \exp(-W_{ij}^2/w_i w_j).
\]

Here, parameters \( w_i \) and \( w_j \) are the scaling distances corresponding to the \( i^{th} \) and \( j^{th} \) sequences, respectively, which control the width of local neighborhoods allowing the level of similarity to vary in different parts of the graph. In this way, the local neighborhoods are determined for each sequence, instead of selecting an universal scaling parameter for all. The width of each local neighborhood is identified by a single weighted-distance value such that sequences inside the neighborhood are more similar to each other than the outsider sequences. In order to determine the sequence-to-sequence scaling parameters a self-tuning framework (Zelnik-Manor and Perona, 2005) (the so-called distance-gap procedure) is incorporated into SCOPer. The distance-gap procedure determines the scale parameter \( w_i \) corresponding to the \( i^{th} \) sequence by seeking a relatively large gap in the set of weighted-distances from \( i^{th} \) sequence to the rest of the sequences. The distance-gap pipeline performs as follows. First, the set of weighted-distances corresponding to the \( i^{th} \) row of the matrix \( W \) is retrieved. Then, a binned Gaussian kernel density estimate of the weighted-distances is generated using the \texttt{density} function from the \texttt{stats} R package. Next, the set of extrema of the continuous density distribution is flagged by finding the weighted-distances at which the first derivative of the distribution is zero while the second derivative is positive, indicating a local minimum following a local maximum. Recall from univariate Calculus that the first and second derivative for some function \( f(x) \) corresponds to the slope of the tangent line and curvature of \( f \) at point \( x \), respectively. Finally, the scale parameter \( w_i \) associated with \( i^{th} \) sequence is determined as the closest smaller weighted-distance to the extremum with the lowest density value. If such an extremum were not found, the scale parameter \( w_i \) is simply determined as the first largest gap of the rank-ordered set of entries corresponding to the \( i^{th} \) row of the matrix \( W \).

Local scaling is especially useful when the classification of the B cell repertoire contains multiple scales (e.g., if one clone is tight, while another
one is sparse). By means of local scaling, the junction sequence similarities between different clones are lower than the similarities within any single clone. Therefore, edges between sequences in local neighborhoods are connected with relatively high kernels (i.e., $K_{ij} \to 1$), while edges between far away sequences have smaller kernels (i.e., $K_{ij} \to 0$). This is an important advantage of this methodology, by allowing the level of sequence similarity to vary in different local neighborhoods (a biologically plausible assumption), over the methodologies, e.g., hierarchical clustering-based, that partition sequences using an universal (fixed) level of similarity over all the sequences.

### 2.4 Spectral decomposition and clustering

Having defined a scheme to set the graph scale parameters automatically, following with the calculation of the graph Kernel matrix $\mathbf{K}$, the last unknown free parameter in the model is the number of clones $k$, which is determined by the eigen-decomposition of the Laplacian matrix. First, the Laplacian matrix $L = D - K$ is calculated, where $D$ is the diagonal matrix with its $i$th diagonal element being the sum of $i$th row of $K$. Then, the Laplacian matrix is eigen-decomposed with eigenvalues $\{0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m\}$ and corresponding eigenvectors $\{\psi_i\}_{i=1}^m$, where $m$ indicates the number of sequences. Then, the number of clones $k$ is determined by finding the largest gap within the eigenvalue spectrum (the so-called “eigen-gap” procedure) at which adding another clone does not give much better modeling of the data. Finally, we perform $k$-means Euclidean distance-based clustering over the $k$ eigenvectors $\{\psi_i\}_{i=1}^k$ associated with the smallest $k$ eigenvalues to find the members of each clone.

## 3 Bulk B cell simulation and library preparation

Each simulated dataset was generated using the AbSim R package (version 0.2.6) in a B cell single-lineage fashion (Yermanos et al., 2017). Each B cell clone simulation begins with a random selection from sets of IGHV, IGHD, and IGHJ germline sequences (Giudicelli et al., 2004) to produce a unique V(D)J recombination event. Then, clones are made by introducing mutations using a local nucleotide context-dependent model (i.e., SSF model from Yaari et al. (2013)), along a phylogenetic tree in which branching events occur stochastically. This process was repeated to create a collection of 25 simulated datasets. The size of each repertoire was sampled from a normal distribution (mean equal to 500k and standard deviation equal to 50k) and the clone sizes were sampled from a gamma distribution (shape equal to 0.75, scale equal to 0.75, and amplitude sampled from a normal distribution with mean equal to 1k and standard deviation equal to 0.1k). The remaining parameters were set as default. After simulation was done, the gene segments, including junction segment, of each simulated sequence were identified using IgBLAST version 1.13.0 (Ye et al., 2013). Then, the outputs were retrieved and tab-delimited database files were generated using the command line tool MakeDb from Change-O (version 0.4.5) (Gupta et al., 2015). Quality checks were also undertaken to remove non-productive sequences. Specifically, each sequence was checked to satisfy a set of constraints that the: (1) whole sequence be annotated as functional, (2) whole sequence contains no stop codons, and (3) junction is in-frame (i.e. the length is modulo 3). Sequences which did not meet these criteria were excluded. At this point, sequences that are identical (i.e. copies that were generated coincidentally) are grouped together into “unique sequences”. The simulated datasets were further processed using the SHazaM framework (www.immcantation.org) resulting in new columns containing VJ(ℓ)-group identifiers, mutation frequencies, and distance-to-nearest values (i.e., distribution of normalized Hamming distances from each junction sequence to its nearest non-identical neighbor in a given VJ(ℓ)-group). Finally, the outcome was a single tab-delimited file per each simulated dataset containing the metadata information associated with each sequence to be used as input to the clonal inference pipeline.

Table 1 presents an overview of 25 simulated datasets used in this study. Furthermore, the global metrics of the BCR simulated repertoires, including: (1) junction length distribution, (2) distance-to-nearest distribution, (3) clonal relative abundance distribution, (4) clone size distribution, (5) mutation frequency distribution, (6) number of clones per VJ(ℓ)-group, (7) average pair-wise SHM for clone, and (8) negative-control test, are presented in Supplementary Figures 1-25A-H, respectively.

### 4 Results

#### 4.1 Pair-wise shared SHM are enriched in B cell clones

Clonally related cells will share SHMs that were accumulated by common ancestors over the course of clonal expansion. However, cells from distinct clones are also expected to share mutations at some positions, such as SHM hot-spots. Therefore, we sought to evaluate the degree to which pair-wise shared mutations were enriched in B cell clones. For each simulated dataset, the pair-wise shared SHM matrix $H$ was generated for each B cell clone by comparing the IGHV and IGHD regions of each pair of sequences with the relevant germline sequence. Then, the average of the upper
triangular elements was calculated (note that $H$ is a symmetric matrix). We found that pair-wise shared SHMs could be identified in $\sim 95\%$ of non-singleton B cell clones (i.e., clones with more than one member) across all simulated datasets. The non-singleton clones without shared mutations tended to be small (with $< 5$ members), so the chance of observing pair-wise shared mutations is lower (Supplementary Figures 1-25C).

We next sought to test whether this high rate of pair-wise SHM sharing was specific to clonally-related sequences. We generated a set of artificial clones (negative controls) by randomly sampling sequences across known clones. Specifically, for each clone from the 100 of artificial clones (negative controls) by randomly sampling sequences pair-wise shared mutations is lower (Supplementary Figures 1-25C).

To test this hypothesis, we compare the performance of distance-based model (equation 5) using either the junction-targeted (termed as ham–junc) or CDR3-targeted (termed as ham–cdr3) methodologies. Using simulated data, performance was quantified using the measures of sensitivity, specificity, and precision. The sensitivity (true positive rate) of each method is defined as the fraction of all sequence pairs from the same clone that were correctly inferred by the method, while specificity (true negative rate) is defined as the fraction of pairs of unrelated sequences that were successfully inferred by the method to be in different clones. Finally, the precision (positive predictive value) of each method is defined by measuring how often inferred clonal relative sequence pairs are truly clonally related. We found that both models inferred the clonal relationships with high sensitivity, specificity, and precision with values of $\geq 94.0\%$ on average across all simulated datasets. However, each of the measures of accuracy were significantly ($p < 0.001$) improved when distance was based on the CDR3 region, rather than the junction region (Figure 5). Thus, the conserved positions flanking the CDR3 should not be used to define the distance between sequences.

We next asked whether incorporating shared SHMs into the procedure lead to even better performance. We thus characterized the performance of stress-based model (equation 5) using CDR3-targeted (termed as ham–shm–cdr3) methodology. Including shared SHM with the stress-based model improved measures of sensitivity, specificity, and precision to $\geq 97\%$ on average across all simulated datasets. For the sake of completeness, we also characterized the performance of stress-based model using junction-targeted (termed as ham–shm–junc) methodology. Consistent with our analysis of the distance-based method, we found that using the junction rather than the CDR3 region led to a significant ($p < 0.001$) decrease in performance (Figure 5).

These results indicate that the best performance within the spectral clustering-based framework is achieved when the stress-based model was accompanied with a CDR3-targeted distance method. Overall, when the original SCOPer methodology (ham–junc) is compared to the new stress-based model (ham–shm–cdr3), a $\sim 3\%$ improvement in the sensitivity, $\sim 2.5\%$ improvement in the specificity, and $\sim 1\%$ improvement in the precision was achieved on average across all simulated datasets (Figure 5).

To better understand how the stress-based method improves the performance of the clonal relationship inference, we examined its operation in detail using one of the identified VJ(3)-groups with 42 unique sequences. As these are simulated data, we know that these sequences

Fig. 5. Integrating information from CDR3 similarity (distance-based) and shared mutations in the V and J segments (stress-based) improves clonal relationship inference. The spectral clustering-based framework was applied to identify clonally-related sequences in 25 simulated datasets (diamonds) generated via AbSim R package (Yermans et al., 2017) (Table 1). Performance was assessed by calculating (A) sensitivity, (B) specificity, and (C) precision via applying the distance-based approach on the junction (ham–junc) and CDR3 (ham–cdr3) regions, as well as the integrated distance- and stress-based approaches on the junction (ham–shm–junc) and CDR3 (ham–shm–cdr3) regions. Mean performance is indicated by the solid bars, while the error bars define one standard deviation. The asterisks (+) indicate $p < 0.001$ by paired t-test (note: the t-test has been performed only for the cases of interest.).

4.2 The stress-based method improves the sensitivity, specificity, and precision of clonal relationship inference

The original distance-based model for identification of B cell clones used by SCOPer measures distance using the junction region of the BCR (Nouri and Kleinstein, 2018b). The junction includes the CDR3 along with the two flanking amino acids (one 5' that is encoded by IGHV, and one 3' that is encoded by IGKJ) (Lefranc, 2014). As the two flanking positions are highly conserved, we sought to determine whether they were necessary to include in the distance measure. Indeed, we hypothesized that including these positions could even lead to decreased performance, as they are likely to be identical across independent clones and will have increasing influence on the distance for clones with shorter junction lengths.
are comprised of two clones, one consisting of 41 sequences, and the other of only one sequence. Comparing the clonal relationships using the CDR3-targeted distance-based method alone (ham-cdr3) and the CDR3-targeted distance-based method along with stress-based method (ham-shm-cdr3), we find that both methodologies inferred two clones. However, CDR3-targeted distance-based method failed to accurately infer the clonal relationships of one of the sequences, which resulted in one false positive and multiple false negatives (Figure 6A). On the other hand, when the stress among sequences was expressed using the pair-wise SHMs (on average \(23 \pm 8\) mutations were counted per pair, from which \(7 \pm 5\) mutations were shared), the clonally-related sequences were pulled toward each other whereas the singleton remained separated, thereby the performance of the local scaling procedure was improved (Figure 6B). Hence, the ham-shm-cdr3 method resulted in no false relationships in this particular case (Figure 6C).

Along with simulated data, we also evaluated the performance of the ham-shm-cdr3 method by estimating specificity using experimental BCR sequencing data from 58 individuals with acute dengue infection (Parameswaran et al., 2013). By definition, clones cannot span different individuals. To estimate specificity, we combine data from multiple individuals, use our proposed method to identify clonal relationships, and then count the frequency of clones that are (incorrectly) inferred to be shared across individuals (Gupta et al., 2017). We use the procedure proposed in (Nouri and Kleinstein, 2018a). First, one of the individuals (the dataset with largest number of unique sequences) was chosen as the “base”. Next, a single sequence was chosen randomly from each of the remaining individuals and added to the sequencing data from the base individual. Specificity was then defined by how often the sequences from non-base individuals were correctly determined to be singletons. Any grouping of these sequences into larger clones must be a false positive. This procedure was then repeated for 100 cycles. The results indicated that the ham-shm-cdr3 method has a high specificity with a value of \(\sim 96.0\%\) on average across all cycles. Thus, combining shared SHMs in the V and J segments of the BCR can be leveraged along with the CDR3 sequence to identify clonally related sequences with high specificity in experimental data.

4.3 The SCOPer algorithm is efficiently parallelized

Computational efficiency is an important property, considering the recent growth in the size of typical BCR repertoires (Soto et al., 2019; Briney et al., 2019). Using the distance-based method we found that clonal partitioning \(\sim 480 \pm 85\) k simulated sequences (the average repertoire size used in this study) took \(\sim 30 \pm 5\) min, but when the stress-based model was involved the partitioning took \(\sim 160 \pm 15\) min. This assessment was performed on a Linux computer with a 2.20 GHz Intel processor and 32 GB RAM. There are two main factors that drive this increased computational cost. In our current implementation, clonal inference is performed on the set of unique sequences (i.e., sequences with distinct nucleotide sequences). When using a distance-based model that considers only the junction or CDR3, the chance of having identical sequences in each VH(\(\ell\)) group is high (on average across all simulated datasets \(\sim 60\%\) of CDR3s are unique per each VH(\(\ell\)) group). This decreases the computational cost of the algorithm. In contrast, when using the stress-based model, the V and J segments are also relevant, allowing fewer sequences to be combined into identical groups (i.e., leading to more unique sequences). The computational cost increases with this increasing number of sequences \(n\). Specifically the eigen-decomposition algorithm, which is scaled by \(O(3n^2)\) (we note that the targeted matrix, to be spectrally decomposed, is symmetric which improves the computational cost significantly). Furthermore, the pair-wise SHM analysis brings additional computational complexity. For instance, the computational complexity of generating the pair-wise shared SHM matrix \(H\) algorithm is \(O(n^2)\). This run time will be summed up by the pair-wise distance matrix \(X\) with the same computational complexity. However, the SCOPer distributed implementation facilitates the clonal inference process by parallelizing the computation and greatly reducing the running time. In our current implementation, the parallelization is achieved by distributing the clonal inference process from each VH(\(\ell\)) group of sequences across processing cores dynamically. The parallelization is possible on cores from a single workstation or on high-performance computing (HPC) cluster facilities. For instance, using only five cores in parallel decreased the running time to \(\sim 44 \pm 11\) min, a \(\sim 4\)-fold improvement, for partitioning \(\sim 480 \pm 85\) k sequences. Our benchmarks across all simulated data sets demonstrate good scalability resulting in a speedup, defined as the time it takes the stress-based algorithm to execute with one processor divided by...
the time it takes to execute in parallel, that is linear to the number of cores (<10) utilized (Figure 7).

5 Conclusion

B cell clonal diversity is introduced through two main mechanisms. The first occurs during maturation in the bone marrow by random joining of germline-encoded V, D, and J heavy chain genes (or V and J light chain genes) combined with the action of exonucleases and terminal deoxynucleotidyl transferase, which add diversity at the recombination boundaries. This diversity acts as a fingerprint that can be used to separate distinct clones based on the distance between their junction (or CDR3) nucleotide sequence (inter-clonal diversity). Subsequently, upon encountering cognate antigen, B cells can enter a germinal center and undergo further diversification through SHM and affinity maturation. The accumulation of SHMs has the effect of splitting out the sequences of B cell clonal variants around their initial points of creation (intra-clonal diversity). A significant challenge in the clonal relationships inference problem is to define meaningful metrics which can leverage inter-clonal diversity to recognize sequences that are part of independent clones (specificity), while also modeling intra-clonal diversity to recognize the variants that are clonally-related (sensitivity).

We developed an unsupervised learning algorithm based on spectral clustering that provides a framework for the inference of B cell clonal relationships. This method combines CDR3 similarity with shared SHM profiles in the V and J segments to capture both inter- and intra-clonal diversification. We showed that the inclusion of pair-wise shared SHM patterns improves the methods ability to identify clonally related sequences. Overall, the method determines B cell clones by: (1) common IGHV- and IGHJ-gene calls, (2) identical or similar CDR3 nucleotide sequences, and (3) shared somatic hypermutation patterns in the V and J regions.

In the absence of gold standard experimental data with known clonal relationship between sequences, the validation was performed using B cell simulations which offer a mechanism to generate data where the underlying clonal groups are known. However, using experimental data we also reported a measure of specificity based on the frequency of clones that are predicted to be shared across individuals.

The influence of SHM hot- and cold-spot biases in the clonal inference process have been incorporated using an SHM targeting model (first term in equation 4). The analysis described here uses the SSF targeting model for SHM that was previously constructed (Yaari et al., 2013). However, while hot- and cold-spot biases are generally conserved across individuals, these intrinsic biases can be altered by age (Hoehn et al., 2019), and may also differ across species (Cui et al., 2016). Clonal identification could be improved by using a data-specific targeting model that can be built using toolkits available in the Immcantation framework (www.immcantation.org). The SSF model seeks to avoid the biases introduced by selection, and rather seeks to capture only the intrinsic biases introduced by the activation-induced cytidine deaminase (AID) binding preferences and error-prone DNA repair in a 5-mer micro-sequence context (Yaari et al., 2013). Future improvements to the SHM targeting model, such as including effects beyond motif-specificity (MacCarthy et al., 2009), may also improve clonal relationship inference.

While the methodology presented here was developed and tested for sequencing data from the H chain only, cutting-edge technologies, including single-cell sequencing, provide paired IGH- and IGL-chain data (DeKosky et al., 2015; Macosko et al., 2015; Briggs et al., 2017). These paired data can be incorporated into the proposed method by extending the criteria for the initial grouping of sequences (i.e., VJ(ℓ)-groups) to include the same IGHV-gene, IGHJ-gene, IGH-CDR3 length, IGLV-gene, IGLJ-gene, and IGL-CDR3 length. BCR clonal inference can then be carried out as before on these more refined groups. The low diversity of the IGL-chain junction region makes it unlikely that including this region in the clustering will provide a significant performance improvement (Zhou and Kleinstein, 2019).

The definition of clone used in this work is based on the assumption that SHM introduces only point substitutions into the BCR sequence. However, it has been shown that insertions and deletions (indels) can also be introduced at a low frequency (<2-3%) (Smith et al., 1996; Ohlin and Borrebaek, 1998; Wilson et al., 1998; de Wildt et al., 1999; Briney et al., 2012; Hwang et al., 2017). Distance functions that allow for sequences of different lengths could be used to identify clonally related sequences that differ by indels (leading, for example, to sequences with different CDR3 lengths). However, these must be rigorously tested.

The methods described in this study have been implemented in the SCOPer (Spectral Clustering for eOne Partitioning) R package, which provides a computational framework to explore multiple approaches to infer clonal relationships in AIRR-seq data. This implementation of SCOPer is freely available as part of the Immcantation framework (www.immcantation.org) under the CC BY-SA 4.0 license. The input and output formats of SCOPer conform to the Change-O (Gupta et al., 2015) and AIRR (Vander Heiden et al., 2018) file standard, and thus the method can be used seamlessly as part of the Immcantation tool suite, including methods for B cell clonal lineage reconstruction, lineage topology analysis, clonal diversity analysis, and other advanced repertoire analyses linked to the clonal landscape.

6 Data availability

The simulated data are accessible at http://clip.med.yale.edu/papers/Nouri2019A.
Acknowledgements

This work was supported in part by the National Institutes of Health (NIH) under award number R01AI104739 and a National Library of Medicine (NLM) Fellowship. We thank the Yale Center for Research Computing for guidance and use of the research computing infrastructure. We wish to acknowledge Alina Aleksandrova for a careful reading of the manuscript. The authors also thank Susanna Marquez for useful comments related to the development of the code.

Author contributions

N.N. and S.H.K. conceived and designed the project. N.N. implemented the method. Both authors wrote and edited the manuscript.

Competing interests

S.H.K. receives consulting fees from Northrop Grumman.

References

Alamyar, E., Duroux, P., Lefranc, M.-P., and Giudicelli, V. (2012). 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. In Immunogenetics, pages 569–604. Springer.

Alt, F. W. and Baltimore, D. (1982). Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-JH fusions. Proceedings of the National Academy of Sciences, 79(13), 4118–4122.

Betz, A. G., Rada, C., Pannell, R., Milstein, C., and Neuberger, M. S. (1993). Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechanism: clustering, polarity, and specific hot spots. Proceedings of the National Academy of Sciences, 90(6), 2385–2388.

Blier, P. and Bothwell, A. (1987). A limited number of B cell lineages generates the heterogeneity of a secondary immune response. The Journal of Immunology, 139(12), 3996–4006.

Boyd, S. D. and Joshi, S. A. (2015). High-throughput dna sequencing analysis of antibody repertoires. In Antibodies for Infectious Diseases, pages 345–362. American Society of Microbiology.

Briggs, A. W., Goldfless, S. J., Timberlake, S., Belmont, B. J., Clouser, C. R., Koppstein, D., Sok, D., Heiden, J. V. A., Tamminen, M. V., Kleinstein, S. H., et al. (2017). Tumor-infiltrating immune repertoires captured by single-cell barcoding in emulsion. bioRxiv, page 134841.

Briney, B., Inderbitzin, A., Joyce, C., and Burton, D. R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature, 566(7744), 393.

Briney, B. S., Willis, J. R., and Crowe, J. (2012). Location and length distribution of somatic hypermutation-associated dna insertions and deletions reveals regions of antibody structural plasticity. Genes and immunity, 13(7), 523–529.

Clarke, S. H., Huppi, K., Ruezinsky, D., Staudt, L., Gerhard, W., and Weigert, M. (1985). Inter-and intrachromosomal diversity in the antibody response to influenza hemagglutinin. Journal of Experimental Medicine, 161(4), 687–704.

Coker, H. A., Durham, S. R., and Gould, H. J. (2003). Local somatic hypermutation and class switch recombination in the nasal mucosa of allergic rhinitis patients. The Journal of Immunology, 171(10), 5602–5610.

Cowell, L. G. and Kepler, T. B. (2000). The nucleotide-replacement spectrum under somatic hypermutation exhibits microsequence dependence that is strand-symmetric and distinct from that under germline mutation. The Journal of Immunology, 164(4), 1971–1976.

Cui, A., Di Niro, R., Vander Heiden, J. A., Briggs, A. W., Adams, K., Gilbert, T., O’Malley Connor, K. C., Vignaletu, F., Shlomchik, M. J., and Kleinstein, S. H. (2016). A model of somatic hypermutation targeting in mice based on high-throughput Ig sequencing data. The Journal of Immunology, 197(9), 3566–3574.

de Wildt, R. M., van Venrooij, W. I., Winter, G., Hoet, R. M., and Tomlinson, I. M. (1999). Somatic insertions and deletions shape the human antibody repertoire. Journal of molecular biology, 294(3), 701–710.

DeKosky, B. J., Kojima, T., Rodin, A., Charab, W., Ippolito, G. C., Ellington, A. D., and Georgiou, G. (2015). In-depth determination and analysis of the human paired heavy-and light-chain antibody repertoire. Nature medicine, 21(1), 86.

Diamond, B., Katz, J. B., Paul, E., Aranow, C., Lustgarten, D., and Scharff, M. D. (1992). The role of somatic mutation in the pathogenic anti-DNA response. Annual review of immunology, 10(1), 731–757.

Elhamati, Y., Sethna, Z., Marcou, Q., Callian Jr, C. G., Mora, T., and Walczak, A. M. (2015). Inferring processes underlying B-cell repertoire diversity. Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1676), 20140243.

Furuta, M., Ueno, M., Fujimoto, A., Hayami, S., Yasukawa, S., Kojima, F., Arihiro, K., Kawakami, Y., Wardell, C. P., Shiraiishi, Y., et al. (2017). Whole genome sequencing discriminates hepatocellular carcinoma with intrahepatic metastasis from multi-centric tumors. Journal of hepatology, 66(2), 363–373.

Giudicelli, V., Chaume, D., and Lefranc, M.-P. (2004). IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. Nucleic acids research, 32(suppl._2), W435–W440.

Glanville, J., Kuo, T. C., von Büdingen, H.-C., Guey, L., Berka, J., Sundar, P. D., Huerta, G., Mehta, G. R., Oksenberg, J. R., Hauser, S. L., et al. (2011). Naïve antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proceedings of the National Academy of Sciences, 108(50), 20066–20071.

Gupta, N. T., Vander Heiden, J. A., Udmum, M., Gadala-Maria, D., Yaari, G., and Kleinstein, S. H. (2015). Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. Bioinformatics, 31(20), 3356–3358.

Gupta, N. T., Adams, K. D., Briggs, A. W., Timberlake, S. C., Vignaletu, F., and Kleinstein, S. H. (2017). Hierarchical clustering can identify B cell clones with high confidence in ig repertoire sequencing data. The Journal of Immunology, 198(6), 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(1676), 20140239.

Hoehn, K., B., Vander Heiden, J. A., Zhou, J., Q., Lunter, G., Pybus, O. G., and Kleinstein, S. (2019). Repertoire-wide phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging and vaccination. BioRxiv, page 558825.

Hwang, J. K., Wang, C., Du, Z., Meyers, R. M., Kepler, T. B., Neuberg, D., Kwong, P. D., Mascola, J. R., Joyce, M. G., Bonsignori, M., et al. (2017). Sequence intrinsic somatic mutation mechanisms contribute to affinity maturation of VRC01-class HIV-1 broadly neutralizing antibodies. Proceedings of the National Academy of Sciences, 114(32), 8614–8619.

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

Kepler, T. B. and Perelson, A. S. (1993). Somatic hypermutation in B cells: an optimal control treatment. Journal of theoretical biology, 164(1), 37–64.
Zhou, J. Q. and Kleinstein, S. H. (2019). Immunoglobulin heavy chains are sufficient to determine most B cell clonal relationships. bioRxiv, page 665760.