Published online 10 June 2011 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* ***W499–W504***
doi:10.1093/nar/gkr413

Detecting selection in immunoglobulin sequences

Mohamed Uduman[1], Gur Yaari[2], Uri Hershberg[3], Jacob A. Stern[4], Mark J. Shlomchik[5,6] and Steven H. Kleinstein[1,2,*]

[1]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, [2]Department of Pathology, Yale University School of Medicine, New Haven, CT 06520, [3]School of Biomedical Engineering, Science and Health Systems, Drexel University, Philadelphia, PA 19104, [4]Brown University, Providence, RI 02912, [5]Department of Laboratory Medicine and [6]Department of Immunobiology, Yale University School of Medicine, PO Box 208035, New Haven, CT 06520, USA

Received March 4, 2011; Revised April 18, 2011; Accepted May 7, 2011

ABSTRACT

The ability to detect selection by analyzing mutation patterns in experimentally derived immunoglobulin (Ig) sequences is a critical part of many studies. Such techniques are useful not only for understanding the response to pathogens, but also to determine the role of antigen-driven selection in autoimmunity, B cell cancers and the diversification of pre-immune repertoires in certain species. Despite its importance, quantifying selection in experimentally derived sequences is fraught with difficulties. The necessary parameters for statistical tests (such as the expected frequency of replacement mutations in the absence of selection) are non-trivial to calculate, and results are not easily interpretable when analyzing more than a handful of sequences. We have developed a web server that implements our previously proposed Focused binomial test for detecting selection. Several features are integrated into the web site in order to facilitate analysis, including V(D)J germline segment identification with IMGT alignment, batch submission of sequences and integration of additional test statistics proposed by other groups. We also implement a *Z*-score-based statistic that increases the power of detecting selection while maintaining specificity, and further allows for the combined analysis of sequences from different germlines. The tool is freely available at http://clip.med.yale.edu/selection.

INTRODUCTION

During the course of an immune response, B cells that initially bind antigen with low affinity through their immunoglobulin (Ig) receptor are modified through cycles of somatic hypermutation and affinity-dependent selection to produce high-affinity memory and plasma cells. This affinity maturation is a critical component of T cell-dependent adaptive immune responses. It helps guard against rapidly mutating pathogens and underlies the basis for many vaccines. Somatic hypermutation is a process unique to B cells responding to antigen that results in a mutation rate that is 7–8 orders of magnitude above normal background, thus introducing about one point mutation per cell per division in the Ig receptor (1,2). Understanding the somatic hypermutation process also has applications far beyond pathogen responses. It has been found to occur in autoimmune responses, and in several proto-oncogenes (3). We have also demonstrated that somatic hypermutation can act genome-wide and thus represents a risk for genomic instability (4).

The ability to detect selection from mutated Ig sequences is a critical part of many studies. Current methods are based on comparing the observed frequency of replacement (i.e. non-synonymous) mutations to their expected frequency under the null hypothesis of no selection (5–8). Elevated levels indicate positive selection, while decreased levels indicate negative selection with significance determined by a binomial test. It is common to look for negative selection in the framework regions (FWRs), which provide the structural backbone of the receptor, and positive selection in the complementary determining regions (CDRs), where most contact residues for antigen binding are found. As the intrinsic biases of somatic hypermutation can give the appearance of selection (9), a significant challenge for these methods is calculating the expected frequency of replacements under the null hypothesis of no selection. We previously developed the Focused binomial test for detecting selection that improved upon existing methods by fully accounting for microsequence specificity and base substitution bias in somatic hypermutation (10). The Focused test also corrects for the decrease in specificity due to cross-talk

*To whom correspondence should be addressed. Tel: +1 203 785 6685; Fax: +1 203 785 6486; Email: steven.kleinstein@yale.edu

The authors with it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

in other methods by using a carefully derived null model of mutation.

Our web site implementing the Focused test has been increasingly used by a large number of groups since the initial publication of the method in 2008 (10). Here, we present an improved web server that includes many of the most requested features from users, such as V(D)J germline segment identification with IMGT alignment and batch submission of sequences. For comparison, we also integrate results from the previously proposed Local binomial (5) and multinomial (7) tests. Note that we have shown that the multinomial test is mathematically equivalent to a much simpler binomial test, referred to in this article as the Global binomial test. Along with these features, we have implemented an improved *Z*-score-based statistic that increases sensitivity and allows for the combined analysis of multiple-independent Ig sequences or clones.

METHODS

Preparing and submitting sequences

The input consists of the mutated sequences to be analyzed along with their associated germlines. In many cases, the experimental results consist only of the mutated sequences so the first step in preparing the data is to define the germline sequence so that individual mutations can be identified. There are numerous databases, such as IMGT/GENE-DB (11) and VBASE2 (12), that provide curated lists of Ig V(D)J germline gene segments. There are also several online tools that can infer the most likely germline rearrangements, including: SoDA (13), iHMMune-align (14) and IMGT/V-QUEST (15). Separate tests are carried out for detecting selection in the CDR and FWR. By default, the web server assumes that the sequences are aligned to conform with the IMGT unique numbering system. This allows standard definitions of CDR and FWR to be used. However, this is not required and users can choose a custom numbering to define these regions. All the sequences in a single input use the same CDR and FWR boundaries. The user may enter or upload their sequences in FASTA format. Multiple sequences sharing the same germline can be grouped together by placing the germline sequence first followed by the test sequence(s). Multiple groups of sequences sharing different germline sequences can also be placed in a single input file by placing an additional '>' at the start of the header line in all the germline sequences. Users also have the option of having the germline sequence(s) appear after each set of mutated sequences. To expedite the steps of germline determination and IMGT alignment, we provide an option for users to input a set of mutated Ig sequences without associated germlines. These sequences are processed using SoDA (13) to identify the V(D)J germline gene segments and align the input and germline using IMGT numbering. Users are given the option to download the resulting FASTA-formatted alignment as a text file or directly proceed to the selection analysis.

Analyzing sequences for detecting selection

Calculating the observed number of mutations. By comparing the input sequence to its associated germline, our program identifies the mutations and determines the number of replacement (R) and silent (S) mutations. Each mutation is considered independently in its germline context when determining whether it is an R or S. R and S mutations that fall into CDR and FWR are tabulated individually using the boundaries indicated by the user. A checkbox is provided to indicate that sequences are clonally related. In this case, each group of sequences associated with a germline is analyzed as a single unit and only unique mutations are used in the analysis. That is, the same base substitution occurring at the same position in multiple sequences is counted only once.

Estimating the expected frequencies of mutations. Having computed the observed number of mutations, the next step is to compute the expected number of mutations under the null hypothesis of no selection. Expectations are computed independently for each germline sequence in the input as previously described (10). A significant advantage of the Focused test over previous methods is that our null model fully accounts for the effects of microsequence specificity (16) and also introduces the well-characterized transition bias of somatic hypermutation (17,18). Briefly, the expected number of R mutations in the CDR ($\bar{R}_{\mathrm{CDR}}$) is the sum of the product of an individual point mutation falling in the CDR and the probability the base substitution results in an R mutation [Equation (1)]:

$$\bar{R}_{\mathrm{region}} = \sum_i \sum_b f_{\vec{GL}}(i) \cdot M_{GL[i]\to b} I_{\vec{GL}}(i,b) \qquad (1)$$

where i is summed over all positions in the region (i.e. CDR or FWR) and b over all possible nucleotides ($\{A,C,T,G\}$). In this equation, $\vec{GL}$ is a vector containing the nucleic content of each position in the germline sequence, $f_{\vec{GL}}(i)$ is the mutability index for position i in germline $\vec{GL}$ [as explained in (10)], $M_{a\to b}$ is the relative rate in which nucleotide a mutates to b (while $M_{a\to a} = 0$) and $I_{\vec{GL}}(i,b)$ is an indicator function that is 1 in cases where a mutation in position i from a to b results in a replacement mutation and 0 otherwise.

The binomial framework for detecting selection. The Local binomial, multinomial and Focused binomial tests for selection all determine whether the observed number of R mutations (x), in either the CDR or FWR, is significantly different than the expected number ($\bar{R}_{\mathrm{CDR}}\bar{R}_{\mathrm{FWR}}$) out of n observed mutations. The expected frequency (x/n) under the null hypothesis of no selection is denoted by p. For $x/n \le p$ (an indication of negative selection), the significance of the test is calculated as the probability of observing x or fewer R mutations by adding half the probability density function (P) at x to the cumulative distribution function (F) at $(x-1)$. The P-value of

Table 1. Definition of x, n and p for each of the implemented tests to detect selection in the CDR

Test	x	n	p
Focused	R_{CDR}	$R_{CDR}+S_{CDR}+S_{FWR}$	$\frac{\tilde{R}_{CDR}}{\tilde{R}_{CDR}+\tilde{S}_{CDR}+\tilde{S}_{FWR}}$
Local	R_{CDR}	$R_{CDR}+S_{CDR}$	$\frac{\tilde{R}_{CDR}}{\tilde{R}_{CDR}+\tilde{S}_{CDR}}$
Global	R_{CDR}	$R_{CDR}+S_{CDR}+R_{FWR}+S_{FWR}$	$\frac{\tilde{R}_{CDR}}{\tilde{R}_{CDR}+\tilde{S}_{CDR}+\tilde{R}_{FWR}+\tilde{S}_{FWR}}$

Equivalent tests for detecting selection in the FWR are obtained by swapping CDR and FWR in the definitions.

$x/n > p$ (an indication of positive selection) is one minus that of negative selection.

$$P_{\text{Binom}} = \Theta(x - pn) - \{F(x-1|n,p) + 0.5 \cdot P(x|n,p)\} \quad (2)$$

where $P(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x}$, $F(x|n,p) = \sum_{i=0}^{x} P(x|n,p)$ and $\Theta(x)$ is 1 for $x > 0$ and 0 elsewhere. Table 1 defines these parameters for each implemented test.

Using Z-scores to compute P-values. The standard way of calculating P-value (P_{Binom}) described in the previous section is conservative in the sense that the resulting specificity is often greater than the cutoff used to obtain it. This limits sensitivity and is particularly a problem when the total number of mutations per sequence is relatively low (~10), as is common for many Ig data sets, but is still significant for dozens of mutations. To correct this problem, we have implemented a Z-score-based method for computing the P-value. The Z score is defined as follows: for a random variable x_i corresponding to the x for sequence or clone i as defined in Table 1, the associated z_i score (which is itself a random variable) is defined as:

$$z_i \equiv (x_i - \mu_i)/\sigma_i \quad (3)$$

where $\mu_i = p_i n_i$ and $\sigma_i = \sqrt{p_i(1-p_i)n_i}$ are the mean and standard deviation of x_i, respectively. In order to obtain a P-value from the z_i, we use the normal approximation and get:

$$P_z = \Theta(z_i) - \text{erf}(z_i) \quad (4)$$

where erf is the Gauss error function; the reader is referred to Figure 1 to see graphically the different P-values definitions [normal approximation and the traditional way of defining it for a binomial distribution, Equation (2)]. We call this the Focused Z test (P_z) to distinguish it from the Focused binomial test (P_{Binom}).

Detecting selection in groups of sequences. Our web server also implements a test for detecting selection in a group of independent sequences. This can be helpful to improve sensitivity if the sequences do not exhibit statistically significant selection when analyzed individually. Previous tools did not allow for such analysis since grouping P-values computed by Equation (2) from sequences with different p's and n's cannot be done for a fixed P-value cutoff in a simple way. However, this can be done using the Z-score approach, through the application of

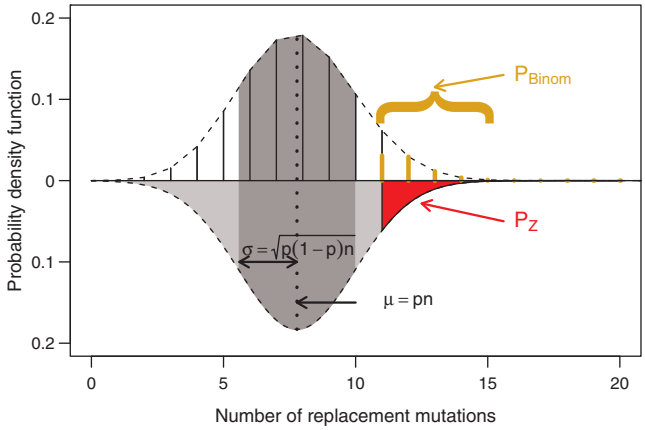

Figure 1. Graphic representation of the calculation for P_{Binom}, the binomial-based test (projected upwards) and, P_Z, the Z-score based test (projected downwards). The binomial distribution depicted by the bars corresponds to $x = 11$, $n = 20$ and $p = 0.389$ (typical parameters from the simulations presented in Figures 2 and 3). P_{Binom} [Equation (2)] is the sum of the yellow bars, while P_Z is calculated by approximating the binomial distribution with a normal distribution with the same mean (dotted line) and variance (1 SD interval around the mean is shaded in darker gray) and taking the integral (area under the curve) from $x = 11$ to ∞ (shaded red).

Stouffer's Z-score method (19), which has proven to be superior to alternatives (20). The resulting P value tests whether the collection of observed mutations came from binomial distributions with associated probabilities p_i. To analyze G-independent sequences, Z is defined as the mean of the previously computed Z-scores z_i, $i \in 1 \ldots G$ from Equation (3). Since the x_i's in this equation are independent, the expected mean and variance of Z are 0 and $\frac{1}{G}$, respectively. Using the normal approximation once again, the resulting P value for detecting selection is:

$$P_{Z,G} = \Theta(Z) - \text{erf}\left(\sqrt{G}Z\right) \quad (5)$$

By default, this test is applied to compute P values for each group of sequences associated with a single germline, as well as for the set of all sequences provided in the input. In order to compare with the original Focused binomial test, a checkbox allows the use of P_{Binom} for individual sequences. If the user indicates that sequences are clonally related then the set of clones is analyzed as a group (in which case, G will be the number of germlines provided in the input).

PERFORMANCE RESULTS

We have previously shown, using a simulation-based validation strategy, that the Focused binomial test provides the best trade-off between sensitivity and specificity compared with other available methods (10). Furthermore, we found that the Focused binomial test is able to detect selection in experimentally derived Ig sequences that have undergone affinity maturation, while maintaining good specificity on non-functional Ig sequences where selection is not a force (10). An independent study has corroborated these findings (21). As

described below, we use a simulation-based validation strategy to evaluate the performance characteristics of our updated *Z*-score-based methods.

Simulation of somatic hypermutation and selection

Simulation enables us to produce synthetic sequence data with a prescribed number of somatic mutations subject to varying amounts of positive and/or negative selection pressures. The simulation is initiated with a single IMGT formatted Ig V germline sequence. Mutations are introduced one-by-one along the entire length of the sequence (excluding gaps) in two steps. First, the position is chosen stochastically based on the micro-sequence specificity of each nucleotide. Second, the particular substitution is determined accounting for transition bias. Selection (ξ) is implemented separately in the CDR and FWR as uniformly increasing or decreasing by a log factor the expected probability of the expected frequency of R mutations in each region:

$$\xi_{\text{region}} \equiv \log\left[\left(\frac{\pi_{\text{region}}}{1-\pi_{\text{region}}}\right)\Big/\left(\frac{p_{\text{region}}}{1-p_{\text{region}}}\right)\right] \quad (6)$$

where p_{region} is the expected frequency of R mutations defined by the Local test (Table 1) and π_{region} is the actual probability applied to the simulated sequences in the region of interest (i.e. CDR or FWR). For example, ξ_{CDR} values of -1, 0 and 1 yield synthetic data with negative, neutral and positive selection in the CDR, respectively.

Using *Z*-scores improves sensitivity

To evaluate performance, a synthetic data set of mutated sequences was generated. This data set included 10 000 sequences with strong positive selection in the CDR ($\xi_{\text{CDR}} = 1$), and 10 000 sequences with neutral selection in the CDR ($\xi_{\text{CDR}} = 0$), allowing us to evaluate sensitivity and specificity, respectively. All sequences included strong negative selection in the FWR ($\xi_{\text{FWR}} = -1$). Each simulation starts with the *IgHV*7–3 germline sequence and introduces between 5 and 35 mutations per sequence, stochastically determined to reflect the numbers seen in experimental data. Results starting with other germline segments are similar. Plotting the fraction of sequences predicted as being positively selected for varying alpha cut-off values in the first data set against the second produces a receiver operating characteristic (ROC) curve. The ROC curves (Figure 2) for both the Focused binomial test (yellow) and the Focused *Z* test (red) are comparable, confirming the validity of the *Z*-score-based method. However, the position on the ROC curve where the α cutoff is 0.05 for the Focused binomial test (yellow *X*) falls short of the expected 1 − specificity of 0.05, indicated by the dashed line. Applying the *Z*-score-based method corrects for this discrepancy leading to an improvement in sensitivity of ~10%, as shown by the red X marking the same position on the ROC curve for the Focused *Z* test. The accuracy of the Focused Z test at $\alpha = 0.05$ is 0.72 while it is 0.68 for the Focused binomial test. Along with the Focused test, a checkbox on our web site also allows users to include results from the Local and Global (also known as the multinomial) tests (Table 1). However, it is important to note that these tests, as originally proposed, did not fully include for microsequence specificity or substitution biases when computing the expected frequencies of mutations, whereas our implementation accounts for these intrinsic biases. In addition, *P*-values for all the approaches are calculated using the *Z*-score-based method. Only results from the Focused test are output by default since this method exhibits better sensitivity than the Local test. Furthermore, we strongly caution against using the Global test because it fails to maintain the defined

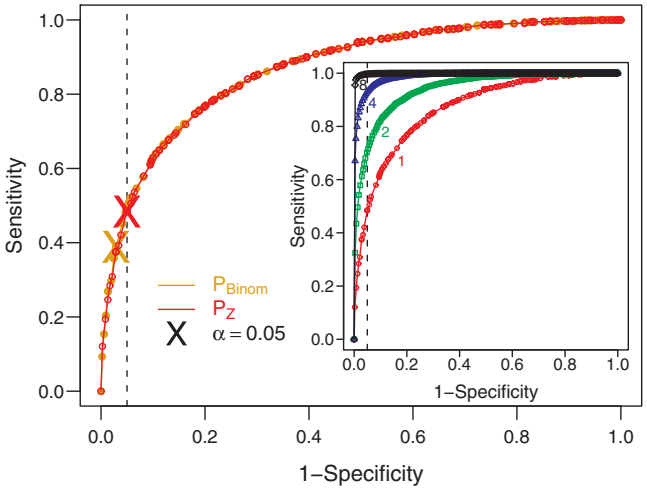

Figure 2. ROC curves comparing P_{Binom} and P_Z on simulated data. Sensitivity is based on detecting selection in simulated sequences with $\xi_{\text{CDR}} = 1$ and $\xi_{\text{FWR}} = -1$. Specificity is based on detecting selection in simulated sequences with $\xi_{\text{CDR}} = 0$ and $\xi_{\text{FWR}} = -1$. The dotted line indicates the expected specificity at $\alpha = 0.05$. The position of this cutoff on the P_{Binom} and P_Z curves is indicated by yellow and red Xs, respectively. The inset shows the ROC curves from $P_{Z,G}$ calculated for independent sequences grouped into sizes (G) of: 1,2,4 and 8.

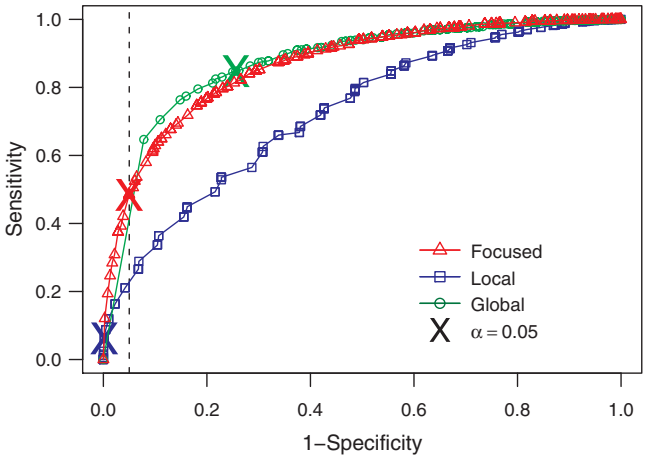

Figure 3. P_Z ROC curves comparing Focused (red triangle), Local (blue square) and Global (green circle) tests for detecting selection. The respective colored *X*'s indicates the positions on the ROC curves that corresponds to the $\alpha = 0.05$ cutoff. The sensitivity and specificity are computed on the same simulated data set from Figure 2.

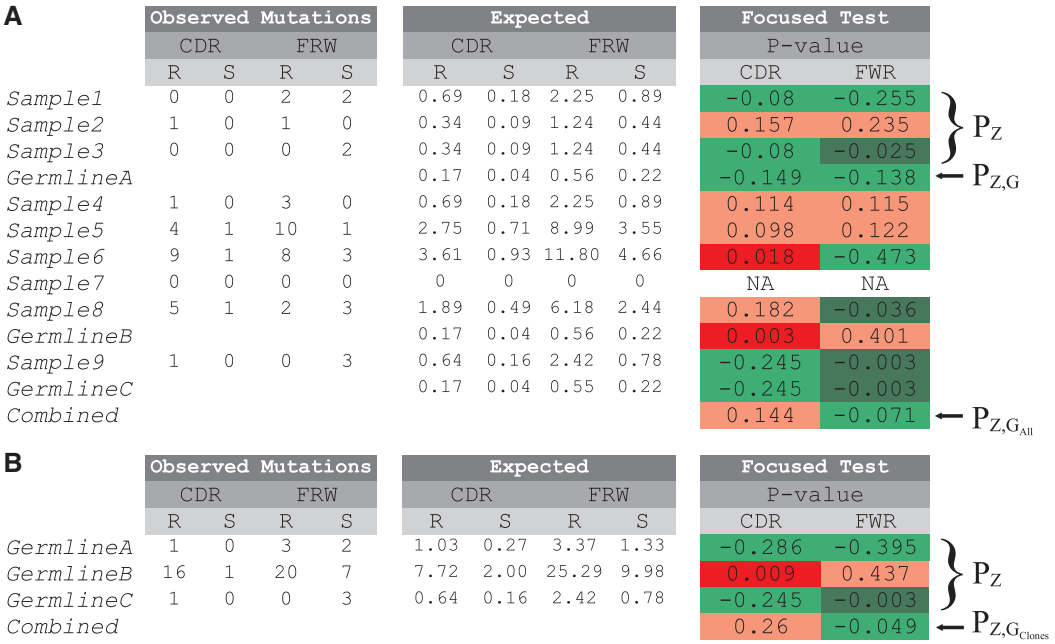

Figure 4. Example output from the website. (**A**) When sequences are independent, P_Z is output for each individual sequence and $P_{Z,G}$ is output for the set of sequences associated with each germline. (**B**) If the sequences are clonally related, then P_Z is output for each clone. Results from each germline (A) or clone (B) are further combined to produce a final $P_{Z,G}$ for the entire data set (labeled as $P_{Z,G_{All}}$ and $P_{Z,G_{Clones}}$, respectively). The resulting P values for the different tests in each region (CDR and FWR) are color coded: red for positive selection and green for negative selection, with light and dark colors indicating non-significant ($|P| > 0.05$) and significant ($|P| \leq 0.05$) selection, respectively. Note that P values less than zero are indicative of negative selection. P_Z and $P_{Z,G}$ are indicated on the side of the plot for selected cases.

specificity [Figure 3 and (10)], but nevertheless make its output available for comparative studies.

Grouping independent sequences improves senstivity

To see how analyzing sequences as a group would effect performance, we randomly sample 10 000 groups of G sequences and compute the $P_{Z,G}$ for each group. This is representative of testing for selection acting collectively on groups of independent Ig sequences. Using these sets of grouped sequences, we produced the ROC curves shown in Figure 2 (inset). It is evident that combining even small numbers of independent sequences can dramatically increase sensitivity without compromising specificity.

IMPLEMENTATION

The web interface makes use of PHP, JavaScript, CSS and AJAX technologies. All the statistics are computed in the back-end using R version 2.9.0 (22). The web site may be accessed using any modern web browser, such as Mozilla Firefox, Google Chrome, Safari and Internet Explorer.

DISCUSSION

The ability to analyze mutation patterns in Ig sequences and detect antigen-driven selection is critical to understanding adaptive immunity. We have presented a web site that makes available our previously published Focused test, along with the Local and Global tests based on statistics proposed by other groups (5–7). Consistent with previous results from our own group and others (10,21), a simulation-based validation found that the Focused test exhibited the best performance. The web site offers an integrated pipeline where users can carry out V(D)J identification with IMGT alignment using SoDA (13), quantify the mutational load in their sequences and analyze the mutations for evidence of positive and/or negative selection (Figure 4). The ability to carry out batch processing and analyze related sequences as a single group will be critical to gain insights from large-scale data sets generated by emerging techniques such as expression cloning (23) and deep sequencing of B cell repertoires (24).

ACKNOWLEDGEMENTS

We would like to thank Thomas Kepler and Supriya Munshaw for providing us with SoDA.

FUNDING

National Institutes of Health (R03AI092379-01) (in part). Funding for open access charge: National Institutes of Health (R03AI092379-01).

Conflict of interest statement. None declared.

REFERENCES

1. McKean,D., Huppi,K., Bell,M., Staudt,L., Gerhard,W. and Weigert,M. (1984) Generation of antibody diversity in the

immune response of BALB/c mice to influenza virus hemagglutinin. *Proc. Natl Acad. Sci. USA*, **81**, 3180–3184.
2. Kleinstein,S.H., Louzoun,Y. and Shlomchik,M.J. (2003) Estimating hypermutation rates from clonal tree data. *J. Immunol.*, **171**, 4639–4639.
3. Odegard,V.H. and Schatz,D.G. (2006) Targeting of somatic hypermutation. *Nat. Rev. Immunol.*, **6**, 573–583.
4. Liu,M., Duke,J.L., Richter,D.J., Vinuesa,C.G., Goodnow,C.C., Kleinstein,S.H. and Schatz,D.G. (2008) Two levels of protection for the b cell genome during somatic hypermutation. *Nature*, **451**, 841–845.
5. Shlomchik,M.J., Aucoin,A.H., Pisetsky,D.S. and Weigert,M.G. (1987) Structure and function of anti-DNA autoantibodies derived from a single autoimmune mouse. *Proc. Natl Acad. Sci. USA*, **84**, 9150–9154.
6. Chang,B. and Casali,P. (1994) The CDR1 sequences of a major proportion of human germline ig VH genes are inherently susceptible to amino acid replacement. *Immunol Today*, **15**, 367–373.
7. Lossos,I.S., Okada,C.Y., Tibshirani,R., Warnke,R., Vose,J.M., Greiner,T.C. and Levy,R. (2000) Molecular analysis of immunoglobulin genes in diffuse large b-cell lymphomas. *Blood*, **95**, 1797–1803.
8. Bose,B. and Sinha,S. (2005) Problems in using statistical analysis of replacement and silent mutations in antibody genes for determining antigen-driven affinity selection. *Immunology*, **116**, 172–183.
9. Dunn-Walters,D.K. and Spencer,J. (1998) Strong intrinsic biases towards mutation and conservation of bases in human IgVH genes during somatic hypermutation prevent statistical analysis of antigen selection. *Immunology*, **95**, 339–345.
10. Hershberg,U., Uduman,M., Shlomchik,M.J. and Kleinstein,S.H. (2008) Improved methods for detecting selection by mutation analysis of ig v region sequences. *Int. Immunol*, **20**, 683–694.
11. Giudicelli,V., Chaume,D. and Lefranc,M.-P. (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and t cell receptor genes. *Nucleic Acids Res.*, **33**, D256–D261.
12. Retter,I., Althaus,H.H., Mnch,R. and Mller,W. (2005) VBASE2, an integrative v gene database. *Nucleic Acids Res*, **33**, D671–D674.
13. Volpe,J.M., Cowell,L.G. and Kepler,T.B. (2006) SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. *Bioinformatics*, **22**, 438–444.
14. Gata,B.A., Malming,H.R., Jackson,K.J.L., Bain,M.E., Wilson,P. and Collins,A.M. (2007) iHMMune-align: hidden markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics*, **23**, 1580–1587.
15. Brochet,X., Lefranc,M.-P. and Giudicelli,V. (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.*, **36**, W503–W508.
16. Shapiro,G.S., Ellison,M.C. and Wysocki,L.J. (2003) Sequence-specific targeting of two bases on both DNA strands by the somatic hypermutation mechanism. *Mol. Immunol.*, **40**, 287–295.
17. Smith,D.S., Creadon,G., Jena,P.K., Portanova,J.P., Kotzin,B.L. and Wysocki,L.J. (1996) Di- and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive b cells. *J. Immunol.*, **156**, 2642–2652.
18. 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. *J. Immunol.*, **164**, 1971–1976.
19. Hedges,L.V. and Olkin,I. (1985) *Statistical Methods for Meta-analysis/Larry V. Hedges, Ingram Olkin*. Academic Press, Orlando, pp. 347–359, Bibliography: Includes index.
20. Whitlock,M.C. (2005) Combining probability from independent tests: the weighted z-method is superior to fisher's approach. *J. Evol. Biol.*, **18**, 1368–1373.
21. MacDonald,C.M., Boursier,L., D'Cruz,D.P., Dunn-Walters,D.K. and Spencer,J. (2010) Mathematical analysis of antigen selection in somatically mutated immunoglobulin genes associated with autoimmunity. *Lupus*, **19**, 1161–1170.
22. R Development Core Team. (2010) *R: A Language and Environment for Statistical Computing*, Vienna, Austria.
23. Wrammert,J., Smith,K., Miller,J., Langley,W.A., Kokko,K., Larsen,C., Zheng,N.-Y., Mays,I., Garman,L., Helms,C. *et al.* (2008) Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature*, **453**, 667–671.
24. Weinstein,J.A., Jiang,N., White,R.A., Fisher,D.S. and Quake,S.R. (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science*, **324**, 807–810.