RESEARCH ARTICLE

Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors [v1; ref status: indexed, http://f1000r.es/z6]

Thomas B Kepler[1,2]

[1]Department of Microbiology, Boston University School of Medicine, Boston, MA, 02118, USA
[2]Department of Mathematics & Statistics, Boston University, Boston, MA, 02118, USA

v1 **First Published:** 03 Apr 2013, **2**:103 (doi: 10.12688/f1000research.2-103.v1)
Latest Published: 03 Apr 2013, **2**:103 (doi: 10.12688/f1000research.2-103.v1)

Abstract

One of the key phenomena in the adaptive immune response to infection and immunization is affinity maturation, during which antibody genes are mutated and selected, typically resulting in a substantial increase in binding affinity to the eliciting antigen. Advances in technology on several fronts have made it possible to clone large numbers of heavy-chain light-chain pairs from individual B cells and thereby identify whole sets of clonally related antibodies. These collections could provide the information necessary to reconstruct their own history - the sequence of changes introduced into the lineage during the development of the clone - and to study affinity maturation in detail. But the success of such a program depends entirely on accurately inferring the founding ancestor and the other unobserved intermediates. Given a set of clonally related immunoglobulin V-region genes, the method described here allows one to compute the posterior distribution over their possible ancestors, thereby giving a thorough accounting of the uncertainty inherent in the reconstruction.

I demonstrate the application of this method on heavy-chain and light-chain clones, assess the reliability of the inference, and discuss the sources of uncertainty.

Article Status Summary

Referee Responses

Referees	1	2	3
v1 published 03 Apr 2013	✓ report	? report	✓ report

1 **Deborah Dunn-Walters**, King's College London School of Medicine UK

2 **Austin Hughes**, University of South Carolina USA

3 **James Crowe Jr**, Vanderbilt University USA

Latest Comments

No Comments Yet

Corresponding author: Thomas B Kepler (tbkepler@bu.edu)

How to cite this article: Kepler TB (2013) Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors [v1; ref status: indexed, http://f1000r.es/z6] *F1000Research* 2013, **2**:103 (doi: 10.12688/f1000research.2-103.v1)

Grant information: This work was supported by NIH/NIAID research contract HHSN272201000053C to (TBK, PI) and a Vaccine Development Center grant in the Collaboration for AIDS Vaccine Discovery Program from the Bill and Melinda Gates Foundation (B. Haynes, PI).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: No competing interests were disclosed.

First Published: 03 Apr 2013, **2**:103 (doi: 10.12688/f1000research.2-103.v1)
First Indexed: 23 Apr 2013, **2**:103 (doi: 10.12688/f1000research.2-103.v1)

Background

During the course of an infection, the host's immune system produces antibody molecules that bind to molecular determinants (antigens) on the infectious agent, thereby neutralizing the agent and targeting it for removal by additional antimicrobial effectors. The heavy and light chain immunoglobulin (Ig) genes that encode the components of the antibody molecule result initially from the stochastic intrachromosomal rearrangement of gene segments arrayed in libraries of such gene segments[1]. These genes are further modified after the activation of the B cells that possess them through somatic hypermutation targeted to the rearranged Ig genes[2]. Those B cells whose Ig genes encode molecules with greater affinity for the eliciting antigen gain a proliferative and survival advantage. In this way, the overall affinity of the pool of serum antibodies increases, sometimes by two or more orders of magnitude. This *affinity maturation*[3] is an essential component of the establishment of humoral immunity, the basis for the large majority of successful vaccines[4].

A great deal has been learned about affinity maturation, particularly with regard to the mechanism of somatic hypermutation[5] and the dynamic organization of the cellular environment in which affinity maturation takes place[6,7] (see the recent review by Shlomchik and Weisel[8]), but the mechanism underlying the selective aspects of affinity maturation remains poorly understood. There is increasing interest in the manipulation of affinity maturation pathways in vaccinology[9] and thus in comparing the biophysical properties of mature antibodies to those of their inferred unmutated ancestors (UA)[10–18]. Little attention has been paid, however, to the uncertainties inherent in the inference of these UAs. Given the sensitive dependence of antibody-antigen interactions on single amino acid changes[19], estimating these uncertainties is essential. Under some circumstances, there may be more than one history consistent with prior knowledge that is supported by the data; having the means to determine these cases and provide a set of alternative UAs that as an ensemble cover a significant posterior probability could be valuable, as was shown by Alam *et al.* in a study of the affinity maturation of a broadly neutralizing anti-HIV-1 antibody[14].

The inference of ancestral rearrangements involves the alignment of two (light chain) or three (heavy chain) gene segments in tandem to the target mature Ig gene. The identities of the gene segments are not known in advance. Instead, there is a library of gene segments from which each segment is drawn stochastically; the identity of each segment is part of the inference. The problem is complicated by randomness in the location of the recombination points, where each gene segment begins or ends, because this condition implies that the alignments are not independent. Further challenges are encountered by the presence of nontemplated (N-) nucleotides added at random to the junctions between gene segments, and of course, by point mutations.

There is a well-developed literature on ancestor reconstruction in phylogenetics (see, for example, Pagel *et al.*, 2004[20]). This line of research has informed the development of my methods, but the problem at hand requires tools beyond those that have been developed by its practitioners. The difference between the previous phylogenetic methods and the method described here is that the former do not take into account the complex process through which the Ig ancestor is constructed. This process places a strong statistical constraint on what ancestral states are permissible. My method owes a great deal to this prior work but does not aim to improve upon it fundamentally. It simply extends a small part of its methods to a new domain of application.

Independent of this previous work from phylogenetics there are applied methods developed by computational immunologists. Indeed, computational methods developed to address the problem have been used for some time[21]. There are several different approaches and corresponding programs available online for carrying out these analyses, including iHMMune[22], V-Quest[23], Joinsolver[24], and SoDA and SoDA2[25,26]. None of these applications, however, provides either of two features essential for the systematic reconstruction of clonal histories. First, one must be able to use all of the information available in a set of clonally related Ig genes in a statistically principled manner. All currently available Ig alignment tools work with one sequence at a time. Second, one needs systematic uncertainty estimates on the UA. In order to say anything of interest about the UA and the clonal history, there must be some level of certainty that the inferred sequence really is the actual UA.

The method described here provides these features. It is based on a hierarchical model of Ig gene development that produces an analysis of the clonal history and posterior probabilities on the UA. The method uses the information available across all members of a clone in a consistent and powerful manner.

Methods

One starts with a query set Q of observed Ig variable-region gene sequences assumed to share descent from a common ancestor α. The task is to estimate the DNA sequence α or, more generally, a posterior probability on α. There are two distinct stochastic processes that together give rise to Q. The stochastic intrachromosomal rearrangement process transforms the germline configuration to the unmutated (naïve) ancestor. Somatic mutation transforms the naïve ancestor to the mature (mutated) antibodies that are observed. To each of these stochastic processes there corresponds a probability function, each of which, in turn, has a natural interpretation within the framework of Bayesian inference. The rearrangement process generates a distribution $P_0(\alpha)$ on unmutated ancestors. For each unmutated ancestor a, somatic mutation then generates the likelihood function $P(Q \mid \alpha)$ relating the ancestor to the observed query sequences. Once these functions are computed, Bayes' Theorem is used to compute the posterior probability on α given Q,

$$P(\alpha \mid Q) = \frac{P(Q \mid \alpha) P_0(\alpha)}{\sum_{\alpha'} P(Q \mid \alpha') P_0(\alpha')} \quad (1)$$

Parameterization of the recombination process

To avoid unnecessary complication, light chain sequences will be used for illustration. The extension to heavy chains is straightforward, but even for the simpler light chains the notation becomes clumsy and obscures the intuition behind the method. Heavy chain rearrangements involve an additional gene segment (DH) and two junctions rather than the one that light chain rearrangements have. Figure 1 illustrates the parameterization of a heavy-chain rearrangement

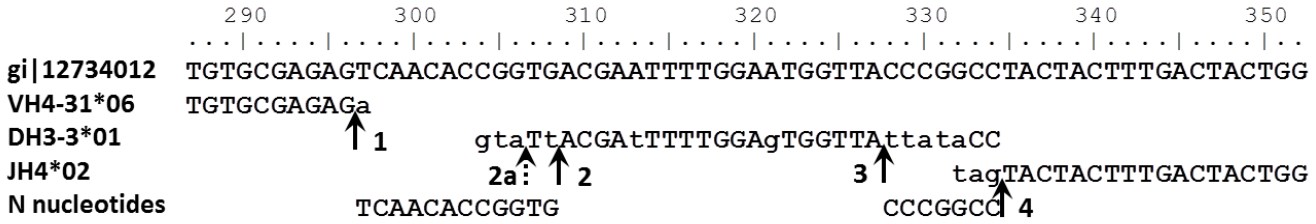

Figure 1. Illustration of parameters for the rearrangement model. Labelled vertical arrows indicate the positions of the recombination sites: 1) RV = 1; 2) RD1 = 5; 3) RD2 = 7; 4) RJ = 3. The dashed arrow 2a indicates a possible alternative recombination site: RD1 = 3. Lowercase letters in the gene-segment sequences indicate mismatches between the observed sequence and the gene segment. The last line shows N nucleotide sequences consistent with the observed sequence.

and provides a guide applicable to both heavy and light-chain rearrangements.

A light-chain rearrangement results from the selection of a V-gene segment V, the selection of a J-gene segment J, the specification of the recombination point in both of these segments R_V, R_J, and the sequence n of the N nucleotides randomly added to the junction between the gene segments. These elements are regarded as parameters in a statistical model: V and J are categorical parameters naming specific gene segments, R_V and R_J are integers, and n is a DNA sequence. R_V is defined as the position of the 3′- most V nucleotide included in the rearrangement; R_J is the position of the 5′- most J nucleotide included. The DNA sequence n may have length zero (meaning that the V and J segments are directly joined and no N nucleotides occur).

Each combination of parameter values generates a specific DNA sequence, although a given sequence may be generated by more than one set of parameter values. One computes the posterior distribution on these parameters, and uses it to generate posteriors probabilities on the quantities of interest, such as the nucleotides at each position of the founder gene.

Let $S(V, J, R_V, R_J, n)$ be the sequence generated by indicated arguments. Then the distribution on unmutated rearrangements is

$$P_0(\alpha) = \sum_{V,J,R_V,R_J,n} I[\alpha = S(V,J,R_V,R_J,n)]\,\pi(V,J,R_V,R_J,n) \qquad (2)$$

where I is the Boolean indicator: I [*true*] = 1, I [*false*] = 0 and π is the prior probability on rearrangement parameters.

V and J are taken to be independent and $\pi(V, J, R_V, R_J, n) = \pi(V, R_V)\ \pi(J, R_J)\ \pi(n)$. Although this assumption is not strictly true–there are small correlations among V, D, and J gene segments[27], the inclusion of these correlation would have very small effects on the resulting inference at the cost of substantial computational effort.

For the analyses in this paper I use gene-segment libraries derived from the IMGT reference libraries[28]. These libraries contain multiple alleles for each gene segment locus. Priors are assigned to the gene segments such that each gene segment locus has the same prior probability, regardless of the number of allelic variants present. Within a gene-segment locus, the distribution on alleles is uniform. When more prior information is available–for example, if one knows the allelic frequencies in the relevant population or knows precisely which alleles are carried by the subject–this information is easily accommodated in the prior probabilities.

The recombination sites are also assigned prior probabilities uniformly across their assumed range. The largest allowed value for R_V corresponds to the position just 3′ of the codon encoding the second invariant cysteine residue. The largest allowed value of R_J corresponds to the position just 5′ of the codon encoding the invariant tryptophan residue. For all gene segments, the smallest allowed value of the recombination points is -4, corresponding to four P nucleotides[29].

For N-nucleotide sequences, an improper prior is used, formally assigning a uniform distribution over all sequence lengths. The use of this uninformative prior is computationally convenient and has little consequence in practice, since ancestral sequences that have excessively long N regions will be judged very unlikely to give rise to the observed sequences and will not contribute substantially to inferences. The mechanics of this phenomenon will become clearer when I describe the computation of the likelihood and sequence alignment.

The likelihood function

The second probability function required is the likelihood, describing the probability that the query sequences Q arose from a given ancestor α by somatic mutation. The likelihood function depends implicitly on the multiple sequence alignment used as well as on the assumed phylogenetic tree. It is computationally infeasible to account completely for these additional sources of uncertainty. Indeed, it remains a significant challenge in the general case[30]. Fortunately, somatic hypermutation only infrequently creates insertions or deletions[31], which are the major cause of uncertainty in multiple sequence alignment. Rather than sum over many multiple alignments, for each gene segment I use the alignment with the maximum score as detailed below.

I assume that the complete multiple alignment A_C can be decomposed into a multiple sequence alignment A_Q among the query sequences in Q and the alignment A between A_Q and the UA. A_Q is estimated in advance and treated as given in subsequent computations. Then for each gene segment, the maximum likelihood alignment between it and A_Q is computed.

Every tree T can be represented by a tree T_1 with unit average branch length and a mutation rate μ taken to multiply each branch of T_1 to yield T. Although the estimated ancestor is insensitive to variation in the assumed tree[32], the estimate of uncertainty is clearly sensitive to the assumed overall mutation rate, i.e., to the overall scaling of the branch lengths.

The procedure is to iteratively estimate T_1 given the UA and the UA given T_1, integrating over μ at each stage. One starts with a simple tree T_1 invariant under permutations of the gene assignments to tips (a *palm tree*, with a branch from the root to the last common ancestor of Q and branches of equal length from the last common ancestor to each member of Q). Then, given T_1, estimate the posterior on the rearrangement parameters (integrating over μ), find the UA with maximum posterior likelihood, use this sequence at the root to re-estimate T_1, and continue iteratively until convergence is reached.

Although the pairwise alignments A_V, A_D, and A_J of the V, D and J gene segments to Q are not independent, they are conditionally independent given the recombination points. Therefore, the likelihood factorizes into components corresponding to gene segments as follows, using the light chain for the example,

$$\begin{aligned} &P(Q \mid V, J, R_V, R_J, n, A, T)\,\pi(V, J, R_V, R_J, n) = \\ &P(Q \mid V, R_V, A_V, T)\,\pi(V, R_V)\,P(Q \mid J, R_J, A_J, T)\,\pi(J, R_J) \\ &\times P(Q \mid n, T)\,\pi(n)\,f(R_V, R_J, A_V, A_J) \end{aligned} \tag{3}$$

The last function contains the dependence among the gene segment pairwise alignments. $f(R_V, R_J, A_V, A_J) = 1$ when the position of R_J in A_Q is 3' of the position of R_V in A_Q, that is, when the gene segments do not overlap. Otherwise, it is zero.

Sequence alignment and somatic mutation

The positions in the ancestor are assumed to evolve independently. For a single query sequence q, one has

$$\log P(q \mid \alpha, \lambda) = \sum_{i=1}^{L} \log M(q_i \mid \alpha_i, \lambda) \tag{4}$$

where q_i is the nucleotide at position i in the query, L is the length of q, α_i is the nucleotide at position i in the ancestor, and λ is the product of time and mutation rate, or branch length. The function M represents the substitution model. For this paper, I use the simple Jukes-Cantor form[33].

Within each component of the likelihood, the substitution model allows the computation of the likelihood for any sequence α placed at the root of T, conditional on T. Since the columns of the individual gene segment alignments are independent, the overall likelihood is the product of the likelihoods for each column in the alignment, each of which is given by taking the product of the likelihoods along each branch in T and summing over all combinations of nucleotides at the interior nodes[34].

Given a pair of nucleotide sequences with one taken to be derived from the other, an alignment between them is equivalent to an accounting of the mutations via which the derivation occurred. Given a substitution model, there is an alignment scoring scheme that corresponds to that substitution model, so that the score for any alignment is the log of the likelihood of the corresponding set of substitutions.

It is straightforward to generalize these observations to the alignment of a nucleotide sequence against a set of sequences, the multiple sequence alignment among which is presumed given. Let the set of nucleotides at position i in the alignment be denoted $\mathbf{q}_i$ and the nucleotide in the ancestor at position i be denoted α_i. The following pairwise alignment scoring scheme is obtained.

Match score–aligning the jth position in the ancestor against the ith position in the derived gene:

$$m(i, j) = \log M(\mathrm{q}_i \mid \alpha j, T) \tag{5}$$

Insertion score–aligning a gap in the ancestor against the ith position of the derived sequence:

$$I(i) = \log M(\mathrm{q}_i \mid -, T). \tag{6}$$

Deletion score–placing a gap at any position in the derived sequence:

$$d(x) = \log M(- \mid x, T), \tag{7}$$

where x is any nucleotide. To account for long deletions or insertions one could use an affine gap score, but in this paper just the simple gap penalties above are used.

Nontemplated nucleotides

In addition to the standard scoring elements for pairwise alignment, the alignment of rearranging antigen receptors requires an additional scoring element for the treatment of N nucleotides. The score for the assignment of a given nucleotide to a generic N nucleotide rather than to a specific N nucleotide state (A,G,T,C) is computed. Denoting by $\pi_N(x)$ the prior probability for a random N nucleotide to have state x, the score corresponding to the assertion that position i in the query sequence alignment is encoded by an N nucleotide is

$$N_i = \log \sum_{x \in \{A,G,T,C\}} M(\mathrm{q}_i \mid x, t)\pi_N(x) \tag{8}$$

For the analyses conducted in this paper, $\pi_N(x) = 1/4$ for all nucleotides x, though, again, the use of informative priors is straight forward.

With all the components of the scoring function in place, one is able to use dynamic programming to find the alignment that maximizes the alignment score.

Because of N nucleotides and increased uncertainty estimating DH gene segments, the complementarity determining region 3 (CDR3) is typically the region of lowest confidence. In addition, all three CDRs accumulate mutations more rapidly than the framework regions in both selected and unselected genes[27]. For these reasons,

Algorithm

The algorithm is schematized as follows.

(Preparation)
Align Q using multiple sequence alignment to give A_Q.
Assume an initial unit-length palm tree, T_1.
While not converged:

{
  Estimate rearrangement parameters given T_1.
  For each discretized value of μ
  {
    Compute the likelihood for each $\alpha_i \in \{A, C, G, T\}$ at each position i of A_Q.
    Align each gene segment V, (D), J in the gene segment library to A_Q, using Eqs. (5–8), computing the likelihood for the relevant parameters in each alignment.
    Compute the posterior on α conditional on μ using Eqs. (1, 2).
  }
  Compute the posterior on μ.
  Marginalize the posterior on α over μ.
  Add the modal (maximum posterior probability) UA α^* to Q.
  Estimate new tree T_1' with α^* at root.
  If $T_1' == T_1$ converged = true
  Else $T_1 = T_1'$
}

CDR3 is susceptible to having its true mutation rate underestimated. The heuristic employed here is to assume a mutation frequency two-fold higher in CDR3 than in the remainder of the gene. This value is consistent with the enhancement of mutation frequency measured in CDR1 and CDR2 where there is much greater confidence in the counting of mutations[35].

The foregoing method was implemented using CLUSTALW[36] within BioEdit to compute A_Q, PHYLIP's dnaml[37] for clonal tree estimation, and software I developed, ARPP UAI, for all other computations.

Reconstructing a B-cell clonal lineage: Antigen Receptor Probabilistic Parser (ARPP) software and Clone K/H FASTA files

4 Data Files

http://dx.doi.org/10.6084/m9.figshare.656793

Results

To examine the reliability of error estimation for the method, I identified two relatively large sets of clonally-related genes for testing. The first, Clone H, is a set of 84 heavy-chain genes[38] of common length 376 nucleotides (nt), with an average (± standard deviation) pairwise difference of 30.4 ± 9.4 nt and a maximum pairwise distance of 61 nt. Figure 2 shows the clonal phylogram for this

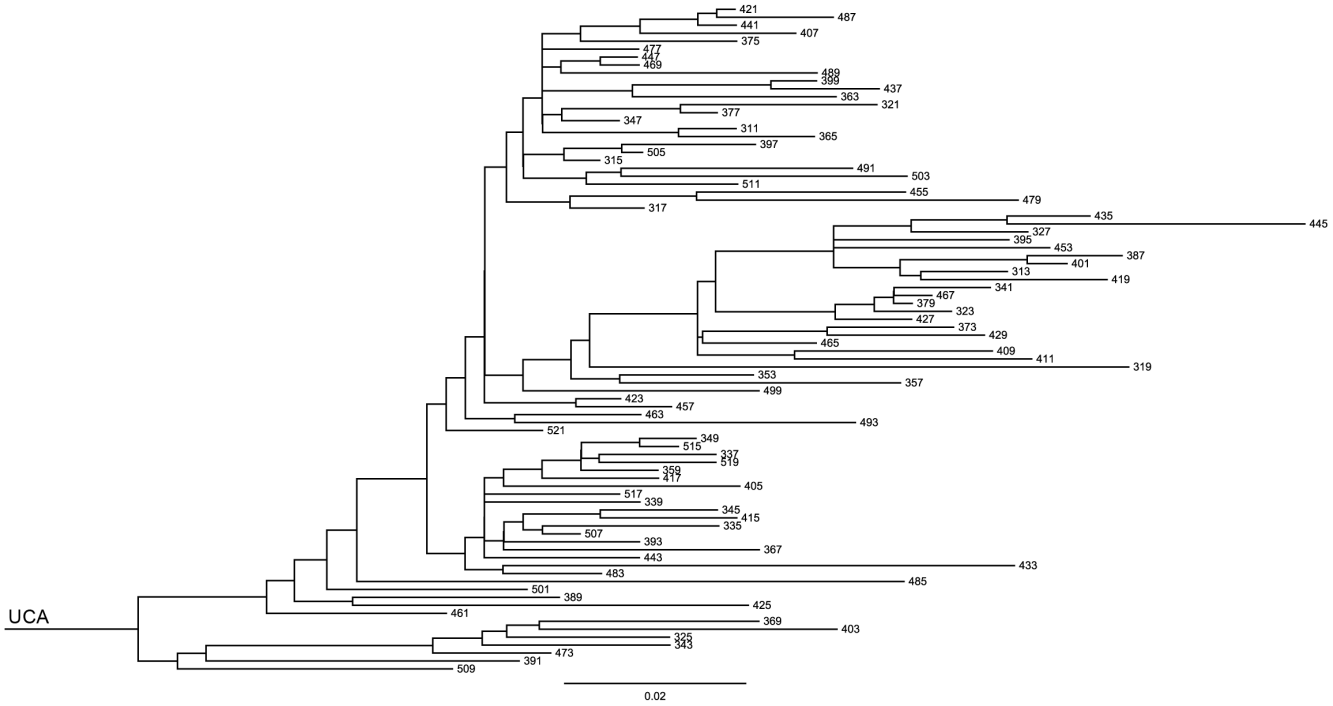

Figure 2. Phylogram of Clone H. The scale bar shows evolutionary distance, or expected number of mutations per position.

set of sequences. The second, Clone K, is a set of 12 kappa-chain sequences[16] of length 299, with an average of 12.2 ± 4.8 nt differences and a maximum pairwise distance of 21 nt.

I applied the inference procedure to Clone H and found that the VH gene segments with the greatest posterior probabilities are VH4-34*01 and VH4-34*03, with nearly identical posterior probabilities of 0.49 each. These two alleles differ from each other in two places. The majority of sequences in the alignment matches one of the alleles at one of these two informative sites but matches the other allele at the other informative site. The modal DH gene segment is DH6-6*01 with posterior probability 0.94. The modal JH gene segment is JH6-1*02 with posterior probability greater than 0.99. The most likely rearrangement has VH using as many as 7 p-nucleotides, no N nucleotides in the V-D junction, and 14 N nucleotides in the D-J junction (Figure 3). The observed sequences have an average mutation frequency of 8.0% compared to the UA.

The UA of Clone K is inferred to have been rearranged using VK1-39*1 with probability greater than 0.999 and to the JK1*1 with probability 0.98. No N nucleotides are required for the rearrangement. The observed sequences have an average mutation frequency of 5.6% compared to UA.

The inference procedure produces a posterior marginal probability mass function over nucleotides at each position of the UA. The probable error at each position is defined as one minus the maximum value of the posterior probability at that position. The total probable error is the sum of the probable errors over positions, and gives the expected number of mismatches between the inferred modal UA and the true UA.

To examine the reliability of the estimated probable error, I subsampled the sequence sets and performed the inference on each of the subsamples. For Clone H, ten pseudorandom samples for each size 1, 3, 9, and 27 were generated. For Clone K, UAs were estimated using each of the individual sequences alone. The resulting modal UAs for all sets were compared to the modal UAs inferred from the complete set.

For Clone H, the total probable error for the UA inferred from the complete set is 2.0. Figure 4 shows the results of these analyses for Clone H. The observed number of mismatches for each subsample is plotted against the total probable error for that subset. The distribution of probable error by nucleotide position shows that some uncertainty is attributable to uncertainty in the allele used in the ancestral rearrangement (Figure 3, position 273) and some is attributable to uncertainty in the N nucleotides and junctions (Figure 3, HCDR3).

For Clone K, the total probable error for UA inferred from the whole set is 0.07. For the 12 UAs obtained from individual sequences, the mean total probable error is 0.14 ± 0.05. There were no mismatches among the light-chain UAs.

Influence of prior distribution

To quantify the impact of the prior distributions on the inference, I performed the inference using the same sequence sets, but with

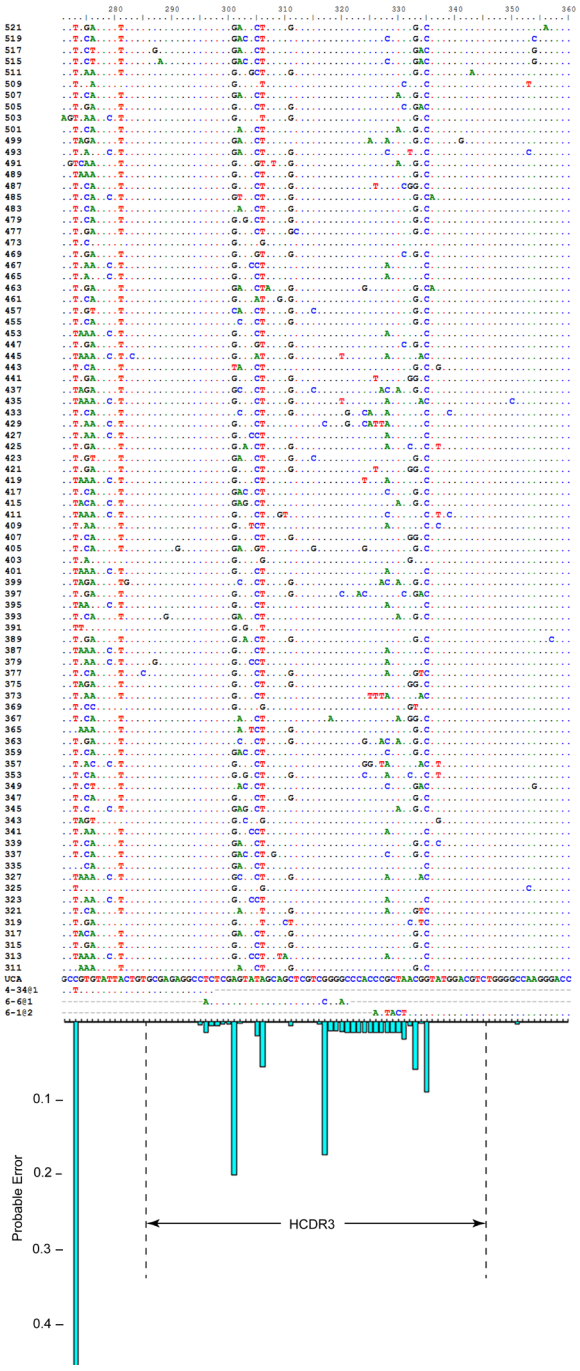

Figure 3. Nucleotide alignment and error profile. Nucleotide alignment of observed heavy-chain sequences, inferred unmutated ancestor, and modal gene segments, with the probable error (below), illustrating the influence of N nucleotides, junctions, and allelic ambiguity on uncertainty. The large probable error at position 273 is due to allelic ambiguity. A second position in FR1 has similar probable error due to allelic ambiguity (not shown). HCDR3 is indicated. The 84 sequences at the top of the alignment are fragments of the observed members of Clone H (naming is arbitrary). The 4 sequences at the bottom of the alignment are the modal unmutated ancestor (UA), and the modal gene segments. A dot in the sequence indicates a match to the UA at that position.

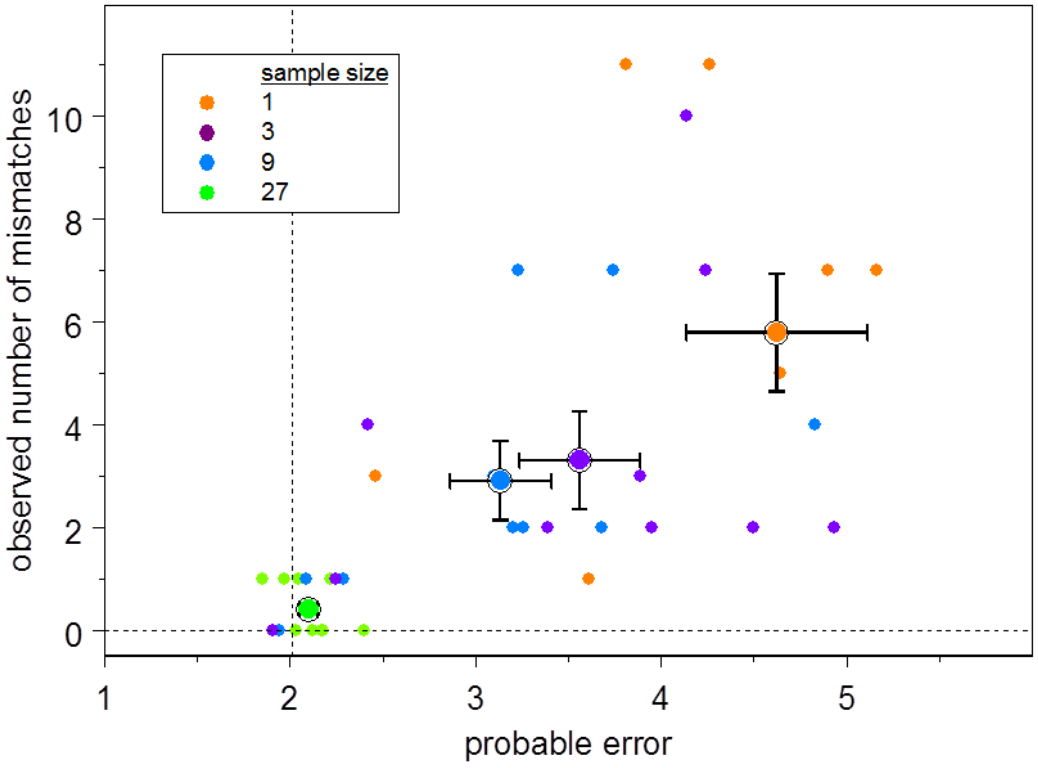

Figure 4. Observed number of mismatches vs. probable error. The number of mismatches between the modal unmutated ancestor (UA) for each subsample compared to the UA for the Clone H complete set vs. the estimated error summed over all positions for each Clone H subsample UA. Symbol color indicates subsample size as shown in the legend. The larger symbols indicate the means; the half-widths of the error bars are the standard errors of the means. The dashed vertical line indicates the total probable error using the complete 84-sequence set.

a simple uniform prior on nucleotides at each position rather than the prior based on knowledge of the rearrangement mechanism and gene segments. Under this model, the modal UA differs from that of the full rearrangement-based model in 11 positions for the heavy-chain clone, and in 10 positions for the light-chain clone. The total probable error for the heavy chains and light chains is 8.5 and 11.5, respectively for the model with uniform priors.

Discussion

I have developed a method for the inference of clonal history in sets of affinity-matured clonally-related immunoglobulin genes. The method allows one to compute posterior distributions on the rearrangement parameters, and hence marginal distributions on several elements, including the nucleotide sequence of the unmutated ancestor.

The probable error is strongly dependent on the interplay of N nucleotides and mutation frequency. This phenomenon occurs because nucleotides near the recombination junction are ambiguous with regard to their origin. A nucleotide that does not match the relevant gene segment at a position near the unknown recombination junction may have been encoded by the gene segment and mutated. Alternatively, it may have been encoded by an N nucleotide. The relative probabilities of these alternatives depend on the mutation frequency. If there are few mutations elsewhere in the gene (where they can be determined more reliably) the likelihood of a mismatch in the junction being due to a mutation is small.

The second major source of uncertainty is allelic diversity. It is often the case, as it is with Clone H, that mutation has destroyed the information required to distinguish which of two or more alleles was used. The greater part of the total uncertainty will be due to one of these two phenomena (Figure 3). This state of affairs also implies that the errors may be correlated, and the distribution of the total number of mismatches overdispersed, as is evident in Figure 4.

One expects the total uncertainty to be proportional to the distance from the root to the most recent common ancestor of the observed sequences (as long as that distance is not too large). Adding related sequences to a clonal set improves the inference to the extent that they push back the time of the most recent ancestor.

Where there are few N nucleotides and allelic polymorphism either not present or not obscured by mutations, the UA can be inferred

with great precision, even in the presence of significant levels of mutation, as is the case with Clone K.

Conclusions

Technology now provides immunologists with the means to reconstruct clonal histories, synthesize the unobserved ancestors, and retrace the steps of affinity maturation to provide deeper insight into the humoral immune response in general and into vaccine design in particular. But the value of the information obtained in this way is wholly dependent on the reliability of the inferential part of the reconstruction. If the ancestors and intermediates are misinferred, the reconstructed history will be potentially misleading.

The methods outlined here are intended to ensure reliable inference and to indicate when multiple histories must be considered.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by NIH/NIAID research contract HHSN272201000053C to (TBK, PI) and a Vaccine Development Center grant in the Collaboration for AIDS Vaccine Discovery Program from the Bill and Melinda Gates Foundation (B. Haynes, PI).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

I thank Grace Kepler, Barton Haynes, Larry Liao and the members of the Duke Human Vaccine Institute Antibodyome group for stimulating discussions.

References

1. Tonegawa S: **Somatic generation of antibody diversity.** *Nature.* 1983; **302**(5909): 575–581.
PubMed Abstract | Publisher Full Text
2. McKean D, Huppi K, Bell M, *et al.*: **Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin.** *Proc Natl Acad Sci U S A.* 1984; **81**(10): 3180–3184.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Eisen HN, Siskind GW: **Variations in affinities of antibodies during the immune response.** *Biochemistry.* 1964; **3**(7): 996–1008.
PubMed Abstract | Publisher Full Text
4. Amanna IJ, Slifka MK: **Contributions of humoral and cellular immunity to vaccine-induced protection in humans.** *Virology.* 2011; **411**(2): 206–215.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Muramatsu M, Nagaoka H, Shinkura R, *et al.*: **Discovery of activation-induced cytidine deaminase, the engraver of antibody memory**. In: Frederick WA, Tasuku H, editors. *Adv Immunol.* Academic press, 2007; **94**: 1–36.
PubMed Abstract | Publisher Full Text
6. Jacob J, Kassir R, Kelsoe G: ***In situ*** **studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. I. The architecture and dynamics of responding cell populations.** *J Exp Med.* 1991; **173**(5): 1165–1175.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Berek C, Berger A, Apel M: **Maturation of the immune response in germinal centers.** *Cell.* 1991; **67**(6): 1121–1129.
PubMed Abstract | Publisher Full Text
8. Shlomchik MJ, Weisel F: **Germinal centers.** *Immunol Rev.* 2012; **247**(1): 5–10.
PubMed Abstract | Publisher Full Text
9. Haynes BF, Kelsoe G, Harrison SC, *et al.*: **B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study.** *Nat Biotechnol.* 2012; **30**(5): 423–433.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Xiao X, Chen W, Feng Y, *et al.*: **Germline-like predecessors of broadly neutralizing antibodies lack measurable binding to HIV-1 envelope glycoproteins: implications for evasion of immune responses and design of vaccine immunogens.** *Biochem Biophys Res Commun.* 2009; **390**(3): 404–409.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Xiao X, Chen W, Yang F, *et al.*: **Maturation pathways of cross-reactive HIV-1 neutralizing antibodies.** *Viruses.* 2009; **1**(3): 802–817.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Dimitrov DS: **Therapeutic antibodies, vaccines and antibodyomes.** *MAbs.* 2010; **2**(3): 347–356.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Zhou T, Georgiev I, Wu X, *et al.*: **Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01.** *Science.* 2010; **329**(5993): 811–817.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Alam SM, Liao HX, Dennison SM, *et al.*: **Differential reactivity of germ line allelic variants of a broadly neutralizing HIV-1 antibody to a gp41 fusion intermediate conformation.** *J Virol.* 2011; **85**(22): 11725–11731.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Ma BJ, Alam SM, Go EP, *et al.*: **Envelope deglycosylation enhances antigenicity of HIV-1 gp41 epitopes for both broad neutralizing antibodies and their unmutated ancestor antibodies.** *PLoS Pathog.* 2011; **7**(9): e1002200.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Liao HX, Chen X, Munshaw S, *et al.*: **Initial antibodies binding to HIV-1 gp41 in acutely infected subjects are polyreactive and highly mutated.** *J Exp Med.* 2011; **208**(11): 2237–2249.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Bonsignori M, Hwang KK, Chen X, *et al.*: **Analysis of a clonal lineage of HIV-1 envelope V2/V3 conformational epitope-specific broadly neutralizing antibodies and their inferred unmutated common ancestors.** *J Virol.* 2011; **85**(19): 9998–10009.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Wu X, Zhou T, Zhu J, *et al.*: **Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing.** *Science.* 2011; **333**(6049): 1593–1602.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Braden BC, Goldman ER, Mariuzza RA, *et al.*: **Anatomy of an antibody molecule: structure, kinetics, thermodynamics and mutational studies of the antilysozyme antibody D1.3.** *Immunol Rev.* 1998; **163**(1): 45–57.
PubMed Abstract | Publisher Full Text
20. Pagel M, Meade A, Barker D: **Bayesian estimation of ancestral character states on phylogenies.** *Syst Biol.* 2004; **53**(5): 673–684.
PubMed Abstract | Publisher Full Text
21. Kepler TB, Borrero M, Rugerio B, *et al.*: **Interdependence of N nucleotide addition and recombination site choice in V(D)J rearrangement.** *J Immunol.* 1996; **157**(10): 4451–4457.
PubMed Abstract
22. Gaëta BA, Malming HR, Jackson KJ, *et al.*: **iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences.** *Bioinformatics.* 2007; **23**(13): 1580–1587.
PubMed Abstract | Publisher Full Text
23. Brochet X, Lefranc MP, Giudicelli V: **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.* 2008; **36**(Web Server issue): W503–W508.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Souto-Carneiro MM, Longo NS, Russ DE, *et al.*: **Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER.** *J Immunol.* 2004; **172**(11): 6790–6802.
PubMed Abstract
25. Volpe JM, Cowell LG, Kepler TB: **SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations.** *Bioinformatics.* 2006; **22**(4): 438–444.
PubMed Abstract | Publisher Full Text
26. Munshaw S, Kepler TB: **SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements.** *Bioinformatics.* 2010; **26**(7): 867–872.
PubMed Abstract | Publisher Full Text | Free Full Text
27. Volpe JM, Kepler TB: **Large-scale analysis of human heavy chain V(D)J recombination patterns.** *Immunome Res.* 2008; **4**: 3.
PubMed Abstract | Publisher Full Text | Free Full Text
28. Lefranc MP, Giudicelli V, Ginestoux C, *et al.*: **IMGT®, the international ImMunoGeneTics information system®.** *Nucleic Acids Res.* 2009; **37**(Database issue): D1006–D1012.
PubMed Abstract | Publisher Full Text | Free Full Text
29. Meier JT, Lewis SM: **P nucleotides in V(D)J recombination: a fine-structure**

analysis. *Mol Cell Biol.* 1993; **13**(2): 1078–1092.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Liu K, Raghavan S, Nelesen S, *et al.*: **Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees.** *Science.* 2009; **324**(5934): 1561–1564.
PubMed Abstract | Publisher Full Text

31. Wilson PC, de Bouteiller O, Liu YJ, *et al.*: **Somatic hypermutation introduces insertions and deletions into immunoglobulin V genes.** *J Exp Med.* 1998; **187**(1): 59–70.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Hanson-Smith V, Kolaczkowski B, Thornton JW: **Robustness of ancestral sequence reconstruction to phylogenetic uncertainty.** *Mol Biol Evol.* 2010; **27**(9): 1988–1999.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Jukes TH, Cantor CR: **Evolution of protein molecules.** *Mammalian Protein Metabolism.* Academic press. 1969; 21–132.
Reference Source

34. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol.* 1981; **17**(6): 368–376.
PubMed Abstract | Publisher Full Text

35. Cowell LG, Kim HJ, Humaljoki T, *et al.*: **Enhanced evolvability in immunoglobulin V genes under somatic hypermutation.** *J Mol Evol.* 1999; **49**(1): 23–26.
PubMed Abstract | Publisher Full Text

36. Larkin MA, Blackshields G, Brown NP, *et al.*: **Clustal W and Clustal X version 2.0.** *Bioinformatics.* 2007; **23**(21): 2947–2948.
PubMed Abstract | Publisher Full Text

37. Felsenstein J: **PHYLIP (Phylogeny Inference Package).** 3.6 ed. Seattle, WA: distributed by the author, Department of Genome Sciences, University of Washington, 2005.
Reference Source

38. Wilson PC, Wilson K, Liu YJ, *et al.*: **Receptor revision of immunoglobulin heavy chain variable region genes in normal human B lymphocytes.** *J Exp Med.* 2000; **191**(11): 1881–1894.
PubMed Abstract | Publisher Full Text | Free Full Text

Current Referee Status:

Referee Responses for Version 1

James Crowe Jr
Vanderbilt School of Medicine, Vanderbilt University, Nashville, TN, USA

Approved: 23 April 2013

Referee Report: 23 April 2013
Current high throughput DNA sequencing technologies, including those for amplicons such as PCR products of antibody gene transcripts, allow for the production of millions or billions of nucleotide sequence files. An intriguing finding that has emerged recently is that B cells of apparent clonal families encoding highly related antibody transcripts, representing what appear to be somatic variants, circulate in the peripheral blood and other tissues. After sequence alignment, it seems very intuitive to infer that highly related sequences actually derived in vivo from single B cell clones. However, as the author points out, there is uncertainty in the inferences made by alignment or conventional phylogenetic tools because of the nontemplated regions of recombined antibody genes, and the high frequency of somatically mutated residues in clones from memory B cells. Current computational tools for identification of likely germline gene segments used in the original recombination are reasonable, but currently there are not adequate tools to determine the likelihood as to whether particular recombined and somatically mutated sequences derived biologically from another less mutated sequence in the repertoire from a sample. This is the gap that the author attempts to fill with the tool described.

This tool likely has significant limitations, but it is important that such tools be developed and tested, with comparison to biological experiments. As sequencing technologies become ever more efficient, it is likely that increased sequencing depth will allow experimentalists to ‘fill in the gaps’ of these types of proposed phylogenies, offering some level of verification of the accuracy of the inferences. Expression and testing of binding of antibodies in intermediate nodes of these phylogenies could be used to experimentally validate the relevance of the inferences. This type of work is already ongoing in several laboratories aimed at rational vaccine and antibody design.

I am not a mathematician, so I cannot comment as to whether the statistical methods are really appropriate in this work. I can comment however that there are a number of limitations that arise from biological particulars of antibody gene repertoires that likely need to be accounted for in later iterations of this tool. The possible number and diversity of nontemplated junctional nucleotides is theoretically nearly infinite and position independent, but structural constraints limit the length and type of residues encoded in junctions. In fact, canonical structural configurations of the necks of the hypervariable loops (CDRs) likely limit the sequence diversity that can be observed in peripheral blood expressed antibodies after selection. I am not certain if these structural constraints can be used to constrain the inferred phylogenies generated, but that would be very helpful, since somatic variants are unlikely to violate the common structural determinants of the antibody paratopes in antigen-specific repertoires. Antibody genes contain more mutable codons than many other genes encoding proteins, so the likelihood of coding changes may need to be accommodated. Insertions and deletions occur with reasonable frequency in these genes during the process of somatic hypermutation. Some sequences that arise from somatic hypermutation stimulated by a foreign antigen may be eliminated due to autoreactivity or other selective pressures. I also

did not see any methods for dealing with sequencing errors, which are vexing in this context. All of these biologic phenomena affecting antibody repertoires make inference of antibody gene phylogenies especially challenging.

Nevertheless, I find it encouraging that new tools like this are being developed that can be tested, evolved and validated in this area. The sequencing technologies present the practical problem of inferring relationships between observed transcripts already, and laboratory experimentalists need practical tools like this for establishing limited sets of candidate genes to synthesize and study. As larger repertoires from more diverse sets of individuals are obtained, the relevance of these tools will become clear. It is especially intriguing to think that, with sufficient sequencing efforts, we may be able to define all possible commonly expressed antibodies and their phylogenies, not just within individuals but across populations.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Austin Hughes
Department of Biological Sciences, University of South Carolina, Columbia, SC, USA

Approved with reservations: 22 April 2013

Referee Report: 22 April 2013
The method presented in this paper depends on the assumption that reconstructing somatic rearrangement events occurring in the ancestors of B cell clones can be improved by using phylogenetic methods to infer the unmutated ancestor (UA) of mature antibodies. The methods used to reconstruct the UA are standard phylogenetic methods, but the statistical approach taken ignores biological complexities. First, the process of affinity maturation of antibodies is a selective process, not just the accumulation of mutations. Certain mutations, which increase affinity, are selectively favoured, while those that decrease affinity are eliminated. Thus, the author's assumption of the independence of nucleotide positions in the sequence seems unjustified, as does the use of the simple Jukes-Cantor model.

In addition, phylogenetic reconstruction in this case (involving short sequences subject to positive selection, resulting in terminal branches that are very long in comparison to internal branches) is likely to be unreliable. The author cites one study suggesting that ancestral sequence reconstruction may not be highly sensitive to tree topology, but the sequences used in that study may not be directly comparable to the present case. It would have been nice to see some quantitative evidence regarding the influence of tree topology on UA reconstruction with these data.

Furthermore, if tree topology doesn't matter for the results, why go through the whole elaborate process of phylogenetic tree reconstruction?

In spite of these reservations, the author is to be congratulated for drawing attention to the potential value of using the information in clonally related sequences for inference of ancestral rearrangement.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Deborah Dunn-Walters
Department of Immunobiology, King's College London School of Medicine, London, UK

Approved: 09 April 2013

Referee Report: 09 April 2013
This is a really useful tool for immunoglobulin affinity maturation studies – particularly now that technology enables us to generate large clonal families.

A minor point, in the file downloads information you state “The program requires a Windows operating system to run.” It would be useful to be more specific – 32 or 64 bit? Which version of Windows?

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.