The Journal of Immunology

RESEARCH ARTICLE | NOVEMBER 01 2003

Estimating Hypermutation Rates from Clonal Tree Data [1] FREE

Steven H. Kleinstein; ... et. al

J Immunol (2003) 171 (9): 4639–4649.
https://doi.org/10.4049/jimmunol.171.9.4639

Estimating Hypermutation Rates from Clonal Tree Data[1]

Steven H. Kleinstein,[2]* Yoram Louzoun,[†‡] and Mark J. Shlomchik[§]

To understand the mechanisms underlying the varying patterns of mutations that occur during immune and autoimmune responses, estimates of the somatic hypermutation rate are critical. However, despite its significance, precise estimates of the mutation rate do not currently exist. Microdissection studies of mutating B cell clones provide an opportunity to measure this rate more accurately than previously possible. Each microdissection provides a number of clonally related sequences that, through the analysis of shared mutations, can be genealogically related to each other. The shape of these clonal trees is influenced by many processes, including the hypermutation rate. We have developed two different methods to estimate the mutation rate based on these data. These methods are applied to two sets of experimental data, one from an autoimmune response and one from the antihapten response to (4-hydroxy-3-nitrophenyl)acetyl (NP). Comparable mutation rates are estimated for both responses, 0.7–0.9 $\times 10^{-3}$ and 0.9–1.1 $\times 10^{-3}$ bp^{-1} division^{-1} for the autoimmune and NP responses, respectively. In addition to comparing the results of the two procedures, we investigate the effect on our estimate of assumptions, such as the fraction of lethal mutations. ***The Journal of Immunology*****, 2003, 171: 4639–4649.**

The rate of somatic hypermutation can have profound consequences for the fates of mutating B cells. Low rates may fail to produce sufficient genetic variation to enable efficient affinity-based selection that normally leads to high affinity immune responses. Mutation rates that are too high would lead to very poor clonal expansion, due to the high frequency of mutations that inactivate Ig V regions, and thus would lead to death because of inability of the B cell receptor to continue signaling. These consequences of mutation are largely inferred from theoretical considerations and computer simulations of the germinal center (GC)[3] response (1–3). Little is actually known about the true in vivo rates of somatic hypermutation. There are few prior estimates of mutation rates, and most estimates derive from cell lines (4–7), which clearly are not reflective of mutation rates in vivo.

Estimating the mutation rate requires a combination of different experimental observations. Simply counting the number of mutations is not enough without knowing the division rate in vivo. Even with such knowledge, extensive mutation content data on cells that have undergone the same or similar number of divisions are required to obtain an accurate estimate. This task is complicated by the observation that the numbers of mutations accumulated by cells are generally small. Furthermore, mutation may not start right away and, over a longer period of time, the effects of positive and negative selection can confound any simple interpretation of the distribution of the number of mutations (5).

The most recognized in vivo estimates (8, 9) relied on a guess as to the number of cell divisions that transpired over a course of a few weeks and two separate immunizations. These estimates considered a division time of 18 h, which is most likely much longer than the actual division time in GCs (2), estimated to be 6–8 h (10). Over the course of the 21 days of the experiment, differences in division times such as this would introduce a major source of error. Furthermore, most of the mutations were replacements in complementarity-determining regions (CDRs), and so unavoidably the estimates were affected by positive selection of such replacement (R) mutations. Nonetheless, estimated rates in vivo and in vitro have been in the range of 5×10^{-5} bp^{-1} division^{-1} to 10^{-3} bp^{-1} division^{-1}. Computer simulations find that clonal expansion as well as sufficient diversification to account for affinity maturation can occur within this window, with rates $\sim$0.5–1.0 $\times$ 10^{-3} bp^{-1} division^{-1} being optimal (1–3). This has generated a level of comfort in the accepted rate estimates, but this may be no more than coincidence given the major issues with these current estimates.

Mutation data can have multiple dimensions beyond the simple sequence. For example, under certain experimental conditions, sequences come from related cells, and the pattern of sharing of mutations among these demonstrate the genealogical relationships (8, 9, 11–13). Modeling can reconstruct the nature of these relationships and allow for a much deeper understanding of the events that led to the generation of the cells, including the inherent mutation rate.

Three obstacles preventing an accurate estimate of B cell mutation rates are: variation in time of mutation onset in different clones, uncertainty concerning the division rate, and the confounding influence of positive and negative selection. One recently introduced experimental paradigm is particularly suited for solving these issues. This involves the microdissection of small areas of proliferating B cells, for example in GCs (14), followed by PCR amplification, cloning, and sequencing. Because this is done at the DNA level, the frequency of sequences will represent the cells from which they are derived. Moreover, the cells are spatially related, and most experience (12, 13, 15, 16) (see below) indicates

Departments of *Computer Science and [†]Molecular Biology, Princeton University, Princeton, NJ 08544; [‡]Department of Mathematics, Bar-Ilan University, Ramat-Gan, Israel; and [§]Department of Laboratory Medicine and Section of Immunobiology, Yale University School of Medicine, New Haven, CT 06520

Received for publication February 27, 2003. Accepted for publication August 4, 2003.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] The work of S.H.K. was supported in part by a National Science Foundation Integrative Graduate Education and Research Traineeship Program Grant (DGE-9972930) and a Presidential Early Career Awards for Scientists and Engineers Grant (CCR-9702115), and through the Center for Discrete Mathematics and Theoretical Computer Science. Y.L. was partially covered by a Program for Mathematical Molecular Biology Burrough Welcome Grant and by the Horowitz Foundation. M.J.S. was supported by National Institutes of Health Grants R01-AI43603 and P01-AI36529.

[2] Address correspondence and reprint requests to Dr. Steven H. Kleinstein, 35 Olden Street, Princeton, NJ 08544. E-mail address: stevenk@cs.princeton.edu

[3] Abbreviations used in this paper: GC, germinal center; AID, activation-induced cytidine deaminase; CDR, complementarity-determining region; FWR, framework region; NP, (4-hydroxy-3-nitrophenyl)acetyl; R, replacement; S, silent; RF, rheumatoid factor.

0022-1767/03/$02.00

that immediate clonal siblings/progeny remain in close proximity after cell division. Small picks of closely related cells allow us to estimate an upper bound on the number of divisions over which mutations are observed and permit the reasonable assumption of constant mutation rates during the creation of these cells. The limited size of each cell cluster further guarantees that the number of cell divisions is relatively small. Thus, the effect of positive selection is limited, and we can more safely assume that all cells divide an equal number of times. Finally, we can know the number of mutations that are generated, the order in which they were generated, and their relationship as embodied in the shape of the inferred genealogic tree. Under these conditions, the shape of trees is largely determined by the mutation rate and the fraction of framework region (FWR) mutations that are lethal to the cell, as we will show. In this study, we used all of this information to generate an accurate estimate of the in vivo mutation rate.

To generate this estimate, we first determined metrics, or measures, of tree shape that were highly reflective of the underlying mutation rate. We were able to apply both Monte-Carlo-type simulation modeling as well as analytic methods to match tree shape metrics to experimental data and thereby estimate compatible mutation rates. We believe these are the best estimates of rate available in vivo because they overcome many previous problems. Moreover, as the mechanisms of hypermutation are in the process of being determined, it will be important to measure how mutations in proteins putatively involved in the mutator pathway actually affect mutation in vivo. The methods described in this work provide a valid, comparable, and generally accessible way to do this.

Materials and Methods

Clonal tree data

Two in vivo mutating systems were examined. In one case, rheumatoid factor (RF) B cells undergoing an Ag-driven autoimmune response at the T zone-red pulp border in MRL/*lpr* mice were used to obtain extensive sequence data from small numbers of localized cells. This system has recently been described (13). Briefly, RF B cells are identified as a small population of cells (~2%) in RF H chain transgenic mice. The RF B cells are those that express one of two endogenous Vκ8 genes and can be detected with an anti-Id reagent. These cells were observed to undergo clonal expansion in auto-Ag-expressing autoimmune-prone MRL/*lpr* mice. Microdissection, followed by Vκ8 gene sequencing of 5–50 (typically 10–20) cells at local sites of such proliferation, revealed extensive intraclonal V gene variability, indicative of ongoing somatic hypermutation. The number of microdissected cells in a pick (i.e., a single microdissection) was estimated by comparing photomicrographs taken just before and just after the manipulation (13) (J. William and M. Shlomchik, unpublished data). A second source of microdissection-based mutation data comes from the sequences of V_H186.2 genes microdissected from GCs generated by immunization with (4-hydroxy-3-nitrophenyl)acetyl (NP)-chicken gamma globulin. These picks were generally a larger size (100–200 cells per pick) (12, 14). These data were kindly provided by G. Kelsoe (Department of Immunology, Duke University Medical Center, Durham, NC).

By analyzing the pattern of shared mutations in the set of sequences generated from each pick, a genealogical tree is created (13). This clonal tree depicts the ancestral relationships between the sequences all the way back to the germline sequence. This complete tree is not used in this study. Rather, the tree is split into (potentially many) subtrees to ensure that the mutations in each subtree have occurred in situ as the cells proliferate (Fig. 1). In particular, the new roots are those vertices of the tree closest to the germline sequence that fulfills either of the following two criteria: 1) the vertex contains an observed sequence, or 2) the vertex contains two or more branches. This pruning results in the removal of the (sometimes long) stem from many trees (see Fig. 1). In addition to ensuring that all of the sequences in each tree are descended from the same common ancestor, this pruning has another advantage in that the clone size (i.e., the number of progeny directly descended from the cell represented by the root of the tree) can now be reasonably bounded based on the size of the pick.

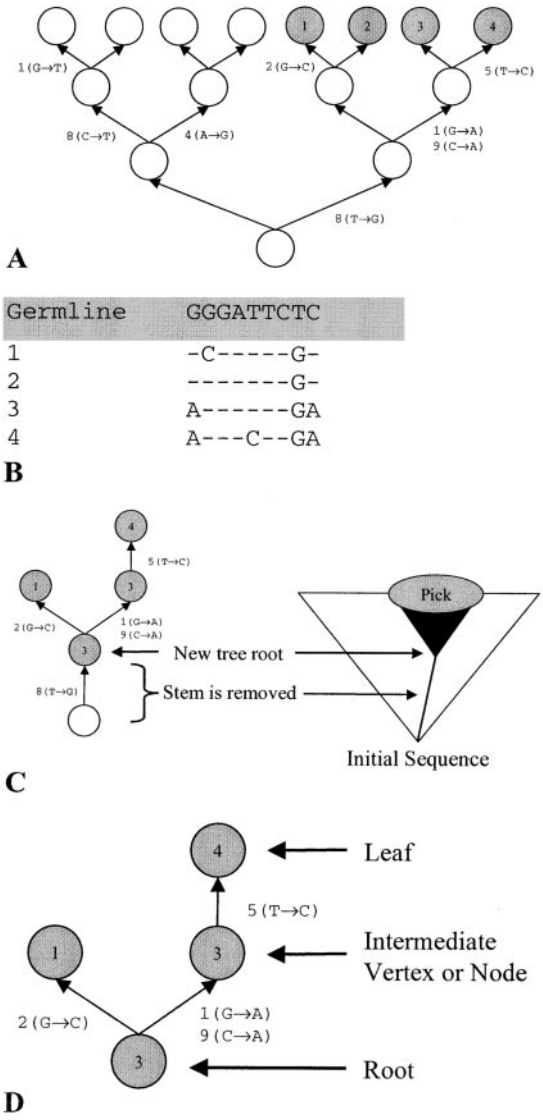

FIGURE 1. Hypothetical example of creating and pruning a clonal tree. *A*, Clonal expansion involving three divisions. The germline sequence occurs at the root of the clonal tree. All divisions are shown, with individual point mutations indicated along the branches. This tree could either be an in vivo expansion, or the result of a simulation. An example pick is indicated by the gray cells (1, 2, 3, and 4). *B*, Individual mutation data for DNA sequences amplified from the pick. In contrast to the simulation in which the relationships between all the cells are known precisely, these sequence data are the only information available from experimental observations. *C*, As was done for the experimental data (Tables II and III), the clonal tree on the *left* is created from the set of DNA sequences by minimizing the number of independent mutation steps required to explain the observed mutations (i.e., the parsimony criteria). In practice, ambiguities in the experimental data were very rare. This tree can also be created from the full simulated tree by collapsing branches that do not contain mutations, as described in *Materials and Methods*. The figure on the *right* is a schematic view of an expanding clone indicating how the tree stem (a single somatic change in this case) could have arisen from the local nature of microdissection. *C*, The clonal tree used to calculate shape measures is obtained after pruning (i.e., removal) of the stem, as described in *Materials and Methods*. Various properties of the clonal tree are specified. The root is the earliest vertex in the tree, leaves are nodes with no child branches, and everything else is called an intermediate vertex (the term node can be used to refer to a vertex containing one or more observed sequences). In the example shown, $S_t = 4$, $U_t = 4$, $N_t = 4$, $M_t = 1.5$, $R_t = 1$, $P_t = 0$, and $I_t = 1$ (definitions of these shape measures are given in Table II).

Underlying assumptions

Both of the computational methods presented in this work for estimating the mutation rate rely on a number of assumptions: 1) mutation rate is a constant, with mutation and/or the fixation of mutations being associated with division; 2) the mutation rate underlying all clonal trees is equivalent; 3) all cells in a particular tree have undergone the same number of divisions; 4) positive selection can be ignored.

Support for the first assumption comes from a variety of studies on cell lines (4–6, 17). In addition, even if the mutation process itself is not associated with cell division, division is required to fix and propagate the mutations. The second assumption arises from the need to precisely define the mutation rate; there is no reason to think that, on average, mutation rates will vary significantly among similar B cells isolated from similar immune responses at similar times. Nonetheless, this assumption is required for any estimate of mutation rate, and the result is that our estimates reflect the average rate of many clones. As for the third assumption, both mutation rate estimation methods presented in this work consider that all of the cells in a given tree have undergone the same number of divisions relative to the root of the tree. The small pick sizes ensure that only a relatively small number of cell divisions could have elapsed between the root of the tree and the leaves. Thus, even if cells divide at slightly different rates, most cells are expected to complete the same number of divisions. For a similar reason, the effects of positive selection during a brief interval will be negligible and can be ignored. Mutations that can be positively selected include only those that substantially increase the affinity of the B cell receptor for the Ag. These are expected to be relatively rare; thus, whether or not they are enriched will a priori have a minimal impact on the mutation rate estimate. The expression of a positively selected phenotype consists of relative enrichment compared with siblings, which in turn means increased division rates and/or increased survival. These differences are likely to be subtle and small over the span of 5–10 divisions, especially compared with the catastrophic loss of Ag binding anticipated for most mutations in FWRs that will disrupt the Ig molecule itself.

We know that ~50% of all R mutations in FWRs will eventually be eliminated from the (auto)Ag-responsive cell population (18). In addition, an unknown, but potentially significant number of R mutations in CDRs will also destroy Ag binding and presumably would also be purged. Because the frequency is high of this category of mutations that would be selected against, it is important to consider them. All mutations will have a lag time between the event in the DNA and expression of the phenotype. It is unknown how long this lag is in the case of mutating and proliferating B cells. However, resting B cells that lose a functional receptor die rapidly (19). The cell division span we are examining in picks of 20–50 cells is on the order of 5–10 divisions. We expect that the phenotype of negatively selectable mutations will be partially and most likely substantially expressed when the entire population is considered. Therefore, it makes sense to consider this in our model, particularly as this is the major category of mutations. To do this, we have estimated rates assuming a range of efficiencies for negative selection, although we consider most accurate rates determined by assuming that negative selection is fairly efficient (i.e., values close to 0.5). To summarize, we ignore the effects of positive selection in our model, but do consider negative selection.

Procedures for estimating the somatic hypermutation rate

Two independent methods are used to estimate the mutation rate: one is numerical (based on a computer simulation of clonal expansion), while the other is analytical. Although both methods are based on the same general ideas outlined below, they were developed separately and rely on somewhat different measures of clonal tree shape to infer the mutation rate. A more detailed description of each method can be found in the *Appendix*.

B cells are assumed to mutate with constant rate μ bp^{-1} division^{-1}. Mutations can be either neutral or lethal. Positive selection is ignored because, as discussed above, it is assumed to operate at a time scale longer than the experimental observations analyzed in this study (pick sizes usually <50 cells suggesting fewer than 10 generations). The effect of each mutation is modeled as a random variable according to the mutation decision tree presented in Fig. 2. The parameter λ gives the fraction of FWR R mutations that are lethal to the cell. Its precise value is uncertain, but an upper bound can be determined based on the finding that 50% of FWR R mutations are negatively selected (18). Thus, the shape of a clonal tree (t) will be determined by the following four parameters: 1) the number of cell divisions in clonal tree t (d_t); 2) the mutation rate (μ); 3) the fraction of mutations that are lethal (λ); and 4) the total number of sequences used to create clonal tree t (S_t).

The first two of these quantities are unknown and are the focus of the estimation procedures presented in this work. Under the assumption that daughter cells remain close to each other spatially, a few times the size of the pick represents a reasonable upper bound on the clone size, and consequently on the number of cell divisions. In fact, we use twice the pick size as an upper bound. We will show later that the precise bound is not important so long as it is large enough. The third quantity will not be estimated, but rather is a required assumption of the methods we present. The final quantity is known.

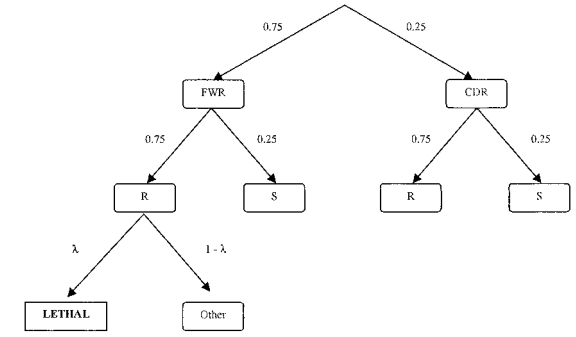

FIGURE 2. The mutation decision tree. The numbers associated with each branch indicate the probability that an individual mutation will fall into the specified category. For example, initially there is a 75% chance the a mutation will fall into the FWR and a 25% chance that it will fall into the CDR. Similar splits are shown for R and silent (S) mutations. Because only the number, and not the type of mutations are important to the procedures we present here, only the overall probability of a mutation being lethal (λ_1) is of any importance. According to the decision tree shown here, $\lambda_1 = 0.75 \times 0.75 \times \lambda$ (e.g., $\lambda_1 \approx 0.28$ when $\lambda = 0.5$).

Given a value for the fraction of mutations that are lethal (λ), the mutation rate is estimated by determining which combination of parameter values for the mutation rate (μ) and the number of divsions (d_t) is expected to produce trees with shapes that are most equivalent to the entire set of observed clonal tree shapes. The mutation rate is assumed to be the same for all of the observed trees in a particular experimental system, as is the fraction of FWR R mutations that are lethal. However, the number of cell divisions and the number of sampled sequences may be different for each tree.

A critical component of this general estimation procedure is deciding which shapes should be used to define the equivalence of clonal trees. We identified tree shape measures that were highly influenced by changes in the mutation rate, and that supplemented the information gained by simple mutation counting. This was accomplished by considering two hypothetical clones with different mutation rates, but which have undergone a varying number of divisions such that they both are expected to carry an equivalent total number of mutations. The simulation was used to evaluate a

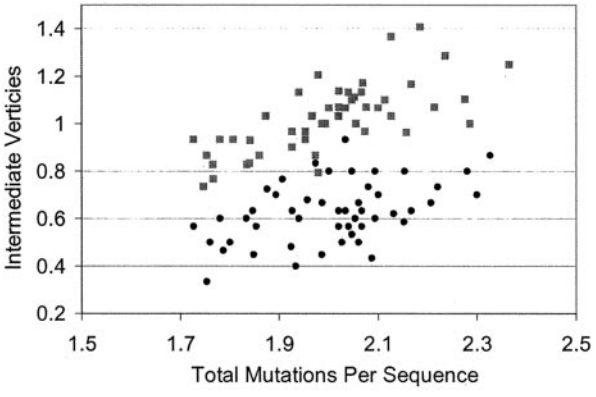

FIGURE 3. Clonal tree shape measures can differentiate between clones with different mutation rates, even when the average number of mutations is equivalent. Red squares show simulated data from clones with a mutation rate of 0.4 per division after 7 divisions. Blue circles show simulated data from clones with a mutation rate of 0.2 per division after 14 divisions. Each point represents a synthetic data set consisting of 30 simulated clonal trees. Notice that both clones have the same average number of total mutations per sequence.

Table I. *Clonal tree 'shape' measures used for defining equivalence between trees*[a]

Shape Measure	Numerical Method	Analytical Method
Branch length (N,M)	Total unique mutations	Average mutations per sequence
Sequences at root (R)	Sequences present at the root of the tree (yes or no)	Average number of sequences present at the root of the tree
Repetitions in nodes (P)	Not used	Total number of sequences in nodes with repeated sequences
Intermediate vertices (I)	Number of vertices that have children in the tree	Not used

[a] Notice that the definitions of the tree shape measures used by the numerical (simulation) method are different than those used by the analytical method.

large number of different tree shape metrics, and we identified a set that could differentiate between these two hypothetical clones (e.g., the number of intermediate vertices, as shown in Fig. 3). Although the expected number of sequences at the root of the tree is correlated with the number of mutations per sequence, it is also included among the shape measures defining equivalence because simulation results suggested that it reduces the number of different clonal trees that correspond to a single set of shape measures. Slightly different criteria for determining equivalence are used by the numerical and analytical methods (Table I).

Given S_t, and assuming some values for the parameters μ and λ, both the simulation and analytical methods predict the distribution of each of the clonal tree measures listed in Table I for various values of d_t. These distributions are compared with experimental observations. For all of the measures used, the simulation code and formulas developed to estimate these measures were cross-validated by comparing the values predicted by both methods with each other. The methods were found to be in good agreement, although minor differences were observed in the number of sequences in nodes with repeated sequences resulting from approximations made in the analytical method (described in the *Appendix*).

Validating the procedures using synthetic data sets

To validate the proposed procedures for estimating the mutation rate, a simulation was used to create sets of synthetic clonal trees with known mutation rates. We then tested whether the estimation procedures proposed in this study could produce correct estimates from these synthetic data sets. This analysis also lets us estimate the precision of our methods, as well as evaluate the sensitivity to any errors in the underlying assumptions.

Consistent with the number of experimental observations described in Table II, each synthetic data set consists of 31 clonal trees. After deciding on values for the mutation rate (μ) and the fraction of FWR R mutations that are lethal (λ), there are four steps in creating each of the clonal trees that comprise the data set (depicted in Fig. 1):

Simulate the clonal expansion. Choose a value for the number of divisions d_t, and simulate a B cell dividing for the specified number of generations. For each tree, d_t is a random number between 2 and 13. The resulting number of cells (up to 8192) spans the range of reasonable clone sizes based on the size of the experimental microdissections. During each division, a Poisson distributed number of mutations occurs with average μ.

Table II. *Summary of clonal tree data from the autoimmune response*[a]

Mouse	Pick/Tree	A_t	S_t	U_t	E_t	M_t	N_t	R_t	P_t	I_t
2205	5a2,3	30	8	7		1.38	11	2	0	0
2205	5a5	20	7	6		1.00	7	2	0	0
2205	5f1, 2.K	100	6	2	37,38,40,46,47,73,74	0.17	1	5	0	0
2205	5f1,2.L	100	5	1	73,74,77,81,82,87,93,95	0.00	0	5	0	0
2205	5f1,2.M	100	2	1	37,38,40,46,47,77,81,82,87,93,95	0.00	0	2	0	0
2540	11f,g	85	18	6	11g1-1,3,4,5,6	2.17	7	0	14	2
2540	11g1	30	5	2	2	1.20	3	0	4	0
4270	10j1	10	6	1		0.00	0	6	0	0
4270	10j2	20	6	3		1.83	6	0	4	1
5281	14c1	30	4	3	107,109,112	6.00	3	0	0	1
5281	14a1,3	20	11	3		0.18	2	9	0	0
5281	14c2	10	5	3		2.00	4	2	2	1
7976	16b3.A	15	6	3	15,20	2.17	6	0	5	1
7976	16b3.B	15	2	2	14,16,18,19,21,22	0.50	1	1	0	0
7976	16c1	50	5	1		0.00	0	5	0	0
7976	16c2.A	20	7	3	21,22,23,24,26,27,28	0.29	2	5	0	0
7976	16c2.B	20	7	1	102,103,105,106,107,108,109	0.00	0	7	0	0
7976	16d1	20	8	3		0.25	2	6	9	0
7976	16d2	20	7	2		0.14	1	6	0	0
7976	16c3.A	10	4	2	34	0.75	1	1	3	0
7976	16c3.B	10	1	1	32,37,39,88	0.00	0	1	0	0
7983	17a4.A	10	3	2	61,64,68,69	0.33	2	3	0	0
7983	17a4.B	10	4	2	62,63,66,68,70	0.25	1	2	0	0
7983	17a4.C	10	1	1	61,62,63,64,66,69,70	0.00	0	1	0	0
7983	17a5.A	50	8	6	77	3.25	15	1	4	1
7983	17a5.B	50	1	1	71,72,73,74,75,76,79,106	0.00	0	1	0	0
4641	12a2,b1[b]	17	12	3	62,63	0.92	2	2	0	1
4641	12c1	10	6	1		0.00	0	6	0	0
4641	12c2	6	7	2		0.14	1	6	0	0
4641	12d2	50	5	4		2.80	9	1	2	1
4641	12e1,2	20	11	2		0.09	1	10	1	0

[a] A_t, Total number of cells used to generate the sequences (i.e., the pick size). S_t, Total number of sequences used to create the tree. U_t, Number of *unique* sequences used to create the tree. E_t, Labels of sequences that were not used to create the tree. M_t, The average number of mutations per sequence. N_t, The number of *unique* mutations in the tree. R_t, The number of sampled sequences at the root of the tree. P_t, The number of sampled sequences in nodes with repeated sequences. I_t, The number of intermediate vertices in the tree.

[b] Pick 12a2,b1 from mouse 4641 was not used in the analytical estimate of the mutation rate.

Mutation can occur either in the CDR or FWR, and they can be either replacements or silent. Some of the framework R mutations are considered to be lethal. The probability assigned to each mutation type is described in the mutation decision tree presented in Fig. 2. Cells with lethal mutations are removed after every generation. If all cells accumulate lethal mutations before generation d_t, then the last available generation is used.

Sample sequences. Randomly choose a subset of cells from the last generation to reflect the limited size of the experimental microdissection and amplification. Remove all other cells from the tree.

Collapse and prune the tree. Collapse any branches that cannot be differentiated by looking at the pattern of accumulated mutations. This is necessary because each branch in the simulated clonal tree represents a cell division, and all branches are known. In contrast, branches in the experimentally observed trees can only be differentiated where mutations have occurred. Next, prune the stem of the tree (i.e., remove it) so that either: 1) the root of the tree contains a sampled sequence, or 2) there are two or more branches leading out from the root of the tree.

Compute and output tree shape measures. Use the resulting clonal tree to calculate the required measures of clonal tree shape (described in Table I). The total size of the clone at the time of sampling is taken to be the equivalent of the pick size of the experimental data. These measures, along with the pick size, constitute our synthetic data set and provide the only input to the estimation procedures.

Results

Estimate of somatic hypermutation rate in an autoimmune response

Sequence data were produced from small groups of cells microdissected from sites of proliferation in autoimmune mice (13). Clonal trees were created from each of these picks and pruned. The resulting data set consists of 31 clonal trees (detailed in Table II). Nine of these trees contain a single unique sequence. The other 22 clonal trees contain 7.0 sequences on average, with 3.2 of these being unique (Table II). The numerical and analytical methods were used to estimate the mutation rate from these data. An important unknown is the fraction of mutations that are lethal. When 50% of FWR R mutations are assumed to be lethal (i.e., $\lambda = 0.5$), both the simulation and analytical methods estimate that the mutation rate is $\sim 0.8 \times 10^{-3}$ bp^{-1} division^{-1} (Fig. 4). This figure is determined by dividing the mutation rate per division (estimated directly by our methods) by the sequence length in base pairs ($\sim$340). Decreasing the assumed fraction of mutations that are lethal leads to only slightly lower estimates for the mutation rate. Even under the most extreme assumption that no mutations are lethal, the mutation rate is estimated to be $\sim 0.7 \times 10^{-3}$ bp^{-1} division^{-1}.

To test the validity of these estimates, both methods were applied to synthetic data sets with known mutation rates. These synthetic data sets were produced in such a way to ensure that they contained an equivalent amount (or less) of information as the experimental data. Each data set consists of 31 independent clonal trees, with each tree containing an equivalent number of total sequences as the corresponding tree in the experimental data set (S_t in Table II). Note that the number of unique sequences in each tree may be different in the synthetic and experimentally observed data sets because the number of sequences that are repeated depends on a number of probabilistic processes.

Fig. 5 shows that both the simulation and analytical methods provide reasonably precise estimates for the mutation rate. When estimating the mutation rate from these synthetic data sets, the same value for the fraction of FWR R mutations considered lethal (λ) used to generate the clonal trees is assumed by the methods. Even with this perfect information, different estimates for the mutation rate were produced from independent synthetic data sets constructed using exactly the same assumptions. However, the SD was small ($<0.1 \times 10^{-3}$ bp^{-1} division^{-1}) (Fig. 5). Because each synthetic data set is equivalent to the entire experimental data set,

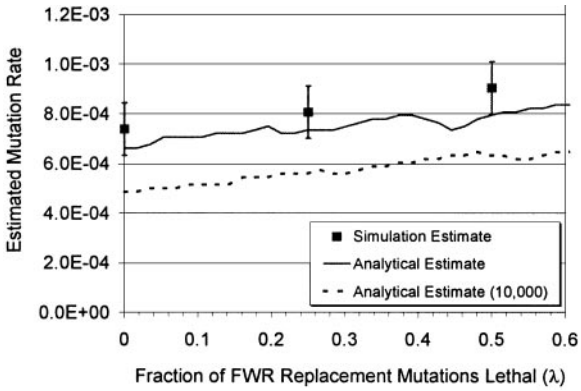

FIGURE 4. The estimated value of the somatic mutation rate (per base pair per division) for the autoimmune response data in Table II. Squares give results using the simulation method, with error bars indicating the maximum SD as determined from Fig. 5. In this case, the maximum clone size is twice the pick size (A_t). The lines indicate results using the analytical method in which the maximum clone size is given by twice the pick size (solid line) or 10,000 cells (dashed line). Our methods directly estimate the mutation rate per division. This value is then divided by 340 (the approximate number of sequenced base pairs) to calculate the mutation rate per base pair per division.

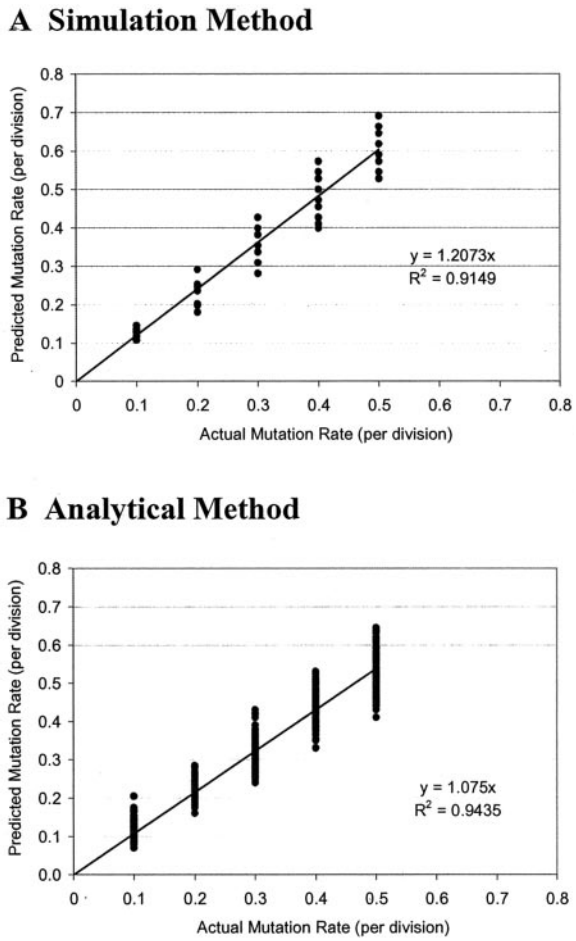

FIGURE 5. Validation of the mutation rate estimation procedures using synthetic data sets. *A*, Results using the simulation procedure. Note that this method has a slight positive bias that is easily divided out of all the mutation rate estimates presented here. This bias is smaller for larger data sets (data not shown). *B*, Results using the analytical procedure.

Table III. *Summary of clonal tree data from the primary NP response*[a]

Germinal Center	Day	A_t	S_t	U_t	E_t	M_t	N_t	R_t	P_t	I_t
61AM40	8	?	8	7		1.50	11	1	2	1
61AM41.A	8	?	6	5	3,5,8,9	1.50	7	0	2	1
61AM41.B	8	?	3	3	1,2,4,6,7,8,10	1.33	3	0	0	1
61AM14	8	?	11	5	7	0.73	8	7	0	0
61AM16	8	?	9	8	2,5,6	1.67	15	2	0	0
61AB08	10	?	2	2		0.50	1	1	0	0
B12	10	?	3	3	3,8,9	1.00	3	1	0	0
B17.A	10	?	3	2	1,2,3,7,8,9	2.00	1	2	0	0
B17.B	10	?	3	3	1,2,4,5,6,7	0.33	2	1	0	0
B17.C	10	?	2	2	2,3,4,5,6,8,9	0.67	4	0	0	0
L1AB01	10	?	9	2		0.11	1	8	0	0
L1AB02	10	?	10	2		0.10	1	9	0	0
L1AB03	10	?	4	2		0.75	1	1	3	0
L1AD01	14	?	7	4	A,C,F,L	1.86	9	4	0	1
L1AD02	14	?	10	7		1.30	9	3	2	3
L1AD03	14	?	8	4	B,J	2.33	9	0	5	2
L1AD05	14	?	9	7		4.10	10	0	3	4
61AD01	14	?	8	7		2.25	9	1	2	3
61AD02	14	?	10	10		4.80	51	0	0	4
61AA02	16	?	6	4	K	1.33	5	0	4	1
GC8	16	?	10	10		5.30	23	0	0	6
GC24	16	?	13	13	II9	4.54	37	0	0	6

[a] Descriptions of shape measures are given in Table II. Note that sequences with stop mutations have been excluded from the clonal trees.

this variance is the precision by which the mutation rate could be computed from the experimental observations, in the absence of other sources of noise. It is reasonable to assume in this analysis that other biasing elements have a weaker effect than this sampling error.

Potential effect of incorrectly estimating the clone size

Up to this point, we have assumed that twice the pick size was a reasonable upper bound for the total size of the clone. As long as this upper bound is large enough, it is not expected to impact the estimate very much because our methods inherently estimate both mutation rate and clone size. Indeed, increasing the maximum clone size to 10,000 cells does not greatly affect the predicted mutation rate (Fig. 4). In addition to shortening the required computation time for the simulation method, the inclusion of a limited clone size prevents undersampled trees from greatly affecting the estimated mutation rate.

The clone size may also be significantly underestimated. There are several reasons that this might happen: either cells may migrate quickly throughout the site of proliferation and/or the pick may include cells from branches that originate far away in the tree. The effect of underestimating the clone size would be to increase the predicted mutation rate above its actual value (data not shown). However, this does not seem to be an important practical factor for our in vivo estimates because increasing the maximum clone size does not greatly impact the estimated mutation rate (Fig. 4).

Estimate of somatic hypermutation rate in the primary NP response

Both procedures for estimating the mutation rate have also been applied to a set of clonal trees obtained during the primary response to the hapten NP. Table III shows the shape measures calculated from these trees. Specific information on the number of cells microdissected from each of the GCs was not available in this case. In our analysis of the autoimmune response, this information was used as a rough upper bound on the clone size. Because we did not have this information in the case of NP responses, estimates for the mutation rate were determined under a number of different assumptions about the maximum clone size (described in Fig. 6).

Another difference with the autoimmune response data is the existence of a source of unobserved mutations (i.e., in the L chain) that can be lethal to the cell. In the autoimmune system, the unsequenced chain was a transgene that does not mutate. The situation for NP is modeled by including a second source of mutations with a rate that is equivalent to the observed rate. The effect of these mutations is governed by the same mutation decision tree shown in Fig. 2. However, the mutations are unobserved so that they are not used in the creation of the clonal tree, nor are they counted when calculating the tree shape measures. Their only potential effect is cell death.

Fig. 6 shows the estimated mutation rate for the primary NP response, which ranges between 0.9×10^{-3} and 1.1×10^{-3} bp^{-1}

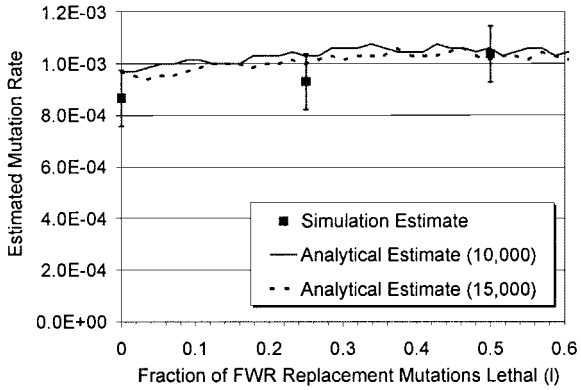

FIGURE 6. The estimated value of the somatic mutation rate (per base pair per division) for the primary NP response data in Table III. Squares give results using the simulation method, with error bars indicating the maximum SD as determined from Fig. 5. In this case, the maximum clone size is 2,000 cells. The lines indicate results using the analytical method in which the maximum clone size is either 10,000 (solid line) or 15,000 (dashed line) cells. Estimates when the maximum clone size is 2,000 (data not shown) fall at $\sim$1.1–1.2 $\times$ 10^{-3} bp^{-1} $division^{-1}$. Both methods directly estimate the mutation rate per division. This value is then divided by 330 (the approximate number of sequenced base pairs) to calculate the mutation rate per base pair per division.

division^{-1} depending on the assumed fraction of FWR R mutations that are lethal. Once again, the particular choice an upper bound for the clone size has little effect on the resulting estimate because, as seen by comparing the two lines in Fig. 6, allowing much larger clone sizes does not have a large effect on the estimated mutation rate.

Unlike the case for the autoimmune response, the methodology for collecting the sequence data used to generate these clonal trees was not optimal for the methods presented in this work. They were produced from larger picks, sometimes from multiple areas within the same GC (12). Thus, it is possible that positive selection has influenced the tree shape and that cells in the same tree have undergone different numbers of divisions. We are currently using the framework presented in this work for validating the methods using synthetic data to test the impact of these processes. Although we have not yet worked out the details, it is clear that positive selection will lead to an overestimate of the mutation rate because it is manifested through the fixing of advantageous mutations in the population. Thus, NP response estimates may be somewhat higher than the actual rate.

Discussion

B cell somatic hypermutation is an important process. It generates mutations that are the substrate for affinity selection. This in turn determines the quality of immune responses to pathogens and vaccines. It also generates diversity that may be important in responses to rapidly mutating pathogens. In pathologic situations, mutation is key in the generation of high affinity autoantibodies (11). Because of the prevalence of deleterious mutations, hypermutation also sharply curtails the clonal expansion rate in mutating cell populations, as demonstrated by computer simulations (2), and by the remarkably large GCs in activation-induced cytidine deaminase (AID) knockout mice that cannot hypermutate (20). The somatic hypermutation process is probably also important in the genesis and progression of various common B cell tumors, including lymphoma, Hodgkin's disease, and myeloma (21–23). The significance of mutation has stirred intense interest in the mechanism of mutation and the control of its onset. The discovery that AID is required for mutation has been a major breakthrough, yet it does not fully explain the control or mechanism of mutation; AID levels alone cannot control mutation because there are cells, e.g., early plasmablasts and T-independent responses, that undergo isotype switch (another AID-dependent process), but do not mutate (12, 24, 25). Perhaps the greatest testimony to the significance of mutation is that it appears in evolution as early as VDJ recombination and adaptive immunity itself, as even cartilagenous fish have populations of B cells that mutate their Ag receptors (26–28).

Despite the importance of mutation rate in determining outcome of immune and autoimmune responses, very little is really known about the rate of mutation in vivo. For example, it is not clear whether the rate can be differentially controlled or whether it is always either off or else on at a single maximal rate. Mutation is best characterized in the GCs of lymphoid follicles, but can occur at other sites, including ectopic GC-like structures in rheumatoid synovia and at the marginal sinus-bridging channels in self IgG-specific B cells proliferating in the marginal sinus-bridging channels of MRL/*lpr* mice (13, 29). Whether the rates at these different sites are the same has not been clear.

As studies progress on AID structure-function and control of its expression, it will be important to have methods to measure the mutation rate in vivo. Furthermore, as putative new sites of mutation are identified, it will be important to measure the rates at these sites. In this study, we demonstrate a new and better approach for measurement of mutation rates in vivo and apply it to mutation at two different sites, the extrafollicular site in MRL/*lpr* mice and GCs formed in response to NP. We used two separate methods to assess the mutation rate using the same data sets. Both methods are based on the same fundamental idea that clonal tree shapes reflect the underlying biology. Although this idea is not new (18, 30, 31), its application to determine specific quantitative features is. We found that mutation at both sites is in the range of $0.7–1.1 \times 10^{-3}$ bp^{-1} division^{-1} and is not significantly different between the responses. These rates are compatible with optimal rates for generation of high affinity along with substantial clones sizes, as determined in our simulation models (2, 3). They are also close to the threshold at which clonal collapse would be common due to the burden of deleterious mutations (2).

The methods presented in this work to estimate the somatic hypermutation rate in vivo are more precise and believable than any previous attempt. Previous estimates of mutation rate have required the cell division rate as an assumption of the estimate (8, 9). A novel component of our methods is the removal of this requirement by considering a combination of multiple clonal tree shape measurements. Moreover, by independently developing and implementing both simulation and analytical procedures, as well as by testing the methods on synthetic data sets, we confirmed the accuracy of our mutation rate estimates. One of the main factors limiting the precision of these estimates is the sampling error due to the small number of available clonal trees. Inclusion of more sample data (i.e., independent trees) would decrease the error of the estimate. However, some error is unavoidable due to the limited number of experimentally observed clonal trees as well as the requirement for a small pick size, which limits number of sequences in each tree. To estimate the maximum precision obtainable by these methods, we used our synthetic data sets to compute the expected variance in the estimate for larger numbers of trees. Increasing the number of trees can reduce the SD from 0.1×10^{-3} bp^{-1} division^{-1} for the current level of experimental data to 0.046×10^{-3} bp^{-1} division^{-1} when 155 trees are present in each data set (data not shown).

A potential error in the mutation rate estimate using the analytical method is the possibility that repeated sequences are the result of sequencing the same template (i.e., a sequence from a single cell) due to amplification by the PCR procedure. However, PCR reamplification (rather than identical independent cells) is unlikely to be a source of identical sequences in our data set because most of the repeated sequences were obtained in the tree roots. If they were indeed due to PCR experimental error, we would have expected them to be distributed throughout the tree. We thus suspect that these repetitions are indeed real. In any case, this concern can be minimized by obtaining a number of sequences that is half or less than the number of cells in the pick, as was typically the case in our autoimmune data set (Table II). This potential error is not a factor when using the simulation method because the clonal tree shape measures used in this case are the same whether or not the repeated sequences represent distinct cells. The fact that the analytical and simulation methods produce similar estimates for these data also argues against this being a real problem. In fact, this observation suggests that none of the theoretical shortcomings unique to either method have a dramatic impact on the mutation rate estimate.

In conclusion, we have demonstrated novel methods for estimating the mutation rate directly from clonal tree data. These methods illustrate how the simple idea that clonal tree shapes reflect underlying biology can be practically applied to provide quantitative estimates of biologically important parameters. In addition to estimating the mutation rate under different conditions, the general framework presented in this work can be adapted to

other important issues. For example, if one were to assume that mutation rates were known and constant, then analysis of tree shape would yield insights into the nature of positive and negative selection. Indeed, data on hypermutation and clonal selection can and should be analyzed using such methods rather than simple statistical tests and analysis of replacement (R)/silent (S) ratios, as we have demonstrated in this work for the case of mutation rate.

Appendix

Methods for estimating the hypermutation rate from clonal tree data

In the following sections, two separate methods for estimating the mutation rate from experimentally derived clonal tree data are described in detail. The first method uses a simulation of B cell clonal expansion to estimate the distribution of clonal tree shapes that will be produced for any given mutation rate. This predicted distribution is compared with experimental observations to estimate the mutation rate with the maximum likelihood of producing these data. The second method follows the same basic strategy for estimating the mutation rate, but uses analytical equations to approximate the expected tree shapes. In both methods, the fraction of FWR R mutations that are lethal (λ) is an assumption.

One significant difference between the simulation and analytical methods concerns their treatment of sequences that are repeated in the experimental data. Due to the experimental techniques used to generate these sequences, it is possible these sequences may result from the repeated amplification of a single cell. This potential error is not a factor when using the simulation method because the clonal tree shape measures used in this case are the same whether or not the repeated sequences represent distinct cells. In contrast, the analytical method assumes that all repeated sequences are derived from individual cells. For example, while the simulation method only looks for whether or not any sequences are present at the root, the analytical method uses the number of sequences at the root of the tree (Table I). As described in the main text, there are a number of reasons to reject the hypothesis that repeated sequences are artifacts making both methods equally valid. In both methods, experimentally observed clonal trees that consist of a single repeated sequence (i.e., $U_t = 1$) are excluded from the data.

Simulation method

In this method, the estimated mutation rate is the value that maximizes the likelihood of producing the observed experimental data. Given any mutation rate μ, the first step in calculating this overall likelihood is to determine $L(t|\mu, \lambda)$, the likelihood of producing the experimentally observed tree t given mutation rate μ per division and assuming that the fraction of FWR R mutations that are lethal is λ. This likelihood is calculated for each of the experimentally observed trees by creating a pair of two-dimensional matrices: O(t,d) and E(t,d) (the observable and equivalent matrices, respectively). As described below, each entry of the observable matrix will be filled in with the number of simulated trees that have undergone d divisions, and which have a clone size consistent with the pick that produced the particular experimentally observed clonal tree t. The equivalent matrix holds the number of observable trees that are precisely equivalent to the experimental tree according to the criteria in Table I. All values in both matrices are initially set to zero.

To fill in these matrices, an expanding B clone is simulated. Details of this simulation were provided when the creation of synthetic data sets was described in *Materials and Methods*. After each division in the simulation, the size of the clone (i.e., the number of live cells) is calculated. This clone size is used to determine whether the simulated clone could possibly have produced each of the observed clonal trees. Specifically, the simulated tree is defined to be observable if the clone size is: 1) greater than the observed number of unique sequences U_t, and 2) less than twice the pick size A_t. These constraints define the sample space as including all trees that could have produced the set of sequences used to create the clonal tree. For each simulation in which the clone is observable according to the constraints for tree t, the matrix entry O(t,d) is incremented, where d is the number of divisions undergone by the clone.

After determining whether the simulated clone is observable, the clonal tree is tested for equivalence with each of the observed trees t. The simulated and observed trees are defined to be equivalent if all of their shape measures (described in Table I) are equal. Shape measures for all of the experimentally derived clonal trees are given in Tables II and III. To calculate analogous shape measures from the simulated clone, an equivalent number of unique sequences (U_t) is randomly sampling from the simulated population. Next, a clonal tree is created from these sequences. This tree is collapsed and pruned, as described for the synthetic data in *Materials and Methods*, so that it can be compared with the experimental data. Shape measures are then calculated on the resulting tree, and if the tree is equivalent, then the matrix entry E(t,d) is incremented. This sampling, collapsing, and pruning process is done after each division separately for each of the observed trees t. In this way, a single simulation run updates all matrix entries.

After running the simulation many times for a given mutation rate μ, the likelihood of producing each of the observed trees is determined. Specifically, the likelihood of producing the experimental tree t is taken to be the probability that the clonal tree is equivalent in shape, given that it is observable (i.e., could have come from the experimental pick A_t):

$$L(t|\mu, \lambda) = \frac{\sum_d E(t, d)}{\sum_d O(t, d)}$$

The overall likelihood for producing the entire experimental data set is the product of the likelihoods for each observed tree:

$$L(\mu) = \prod L(t|\mu,\lambda)$$

By repeating the above procedure for different values of the mutation rate, the value with the maximum likelihood for producing the set of observed clonal trees is determined. This value is the estimated mutation rate. In practice, a slightly modified version of Golden Section Search is used to quickly arrive at the maximum likelihood mutation rate (32). In addition, the optimization is performed many times (i.e., >10) for each data set, and the average value is taken as the final estimate. These independent predictions were always very close to each other (SD $\ll 10^{-4}$).

We used 128,000 independent simulation runs to calculate the likelihood at each value of the mutation rate. Increasing the number of runs beyond this point did not provide additional accuracy (data not shown). Furthermore, we found that the optimization procedure sometimes produced inconsistent estimates when using fewer runs (e.g., the likelihood function $L(\mu)$ appeared to have multiple local maxima). We did not find this to be a problem when using 128,000 runs.

Analytical method

In this approach, formulas are derived for the average (i.e., expected) value of the tree shape measurements M, R, and P (described in Table I). Note that these shape measures are defined

somewhat differently for the simulation and analytical procedures. Using the formulas, these values can be easily calculated under different assumptions about the mutation rate (μ), the fraction of FWR R mutations that are lethal (λ), and the number of divisions undergone by the clone (d). Given an assumption for λ, the difference between the expected and observed tree shapes for each experimental tree t is computed. For each of the trees, we assume that the number of divisions d is the one which minimizes this error. The overall error in the mutation rate, $X(\mu)$, then is the sum of these errors over all the experimentally observed clonal trees:

$$X(\mu) = \sum_t \underset{d}{MIN}\left(\frac{(M_t - M)^2}{VAR(M_t)} + \frac{(R_t - R)^2}{VAR(R_t)} + \frac{(P_t - P)^2}{VAR(P_t)}\right)$$

where the expected tree shape measures M, R, and P depend on the number of divisions d, the fraction of FWR R mutations that are lethal λ, as well as various other aspects of the clonal tree t. $VAR(S_t)$ is the variance of the tree shape measure S among all experimentally observed clonal trees. Potential values for the number of divisions (d) considered in the minimization range from the smallest required to produce (on average) the number of sequenced samples (S_t), to a value one larger than that required to produce (on average) the total number of sampled cells (A_t) (i.e., the maximum clone size is twice the pick size). These upper and lower bounds vary with particular assumptions for μ and λ, because higher mutation rates and lethal frequencies imply a higher death rate, and thus fewer cells on average.

By minimizing the error $X(\mu)$, we arrive at the estimated mutation rate μ. In fact, we do not minimize this function directly, but rather fit a quadratic to the curve to avoid noisy fluctuations around the minimal μ, and then calculate the minimum.

The following sections derive the formulas for each of the tree shape measures: M, R, and P. Predictions from each of the formulas were cross-validated by comparing with estimates using the simulation model. We are currently deriving formulas for other tree shapes in addition to those below. However, as described in the text, these three are sufficient to obtain an accurate estimate of the mutation rate.

The average number of mutations per sequence (M)

Consider a cell that has undergone d divisions and accumulated m mutations. The number of such cells surviving (e.g., accumulating no lethal mutations) is: $\alpha^d(1 - \lambda_1)^m$ where $\lambda_1 = 0.75 \times 0.75 \times \lambda$ is the overall probability that a mutation will be lethal (see Fig. 2), and α indicates the total number of surviving cells surviving independent of the observed mutations. The parameter α is also used to account for the potential presence of unobserved mutations. More specifically, $\alpha = 2 \times (1 - \lambda_2 r\mu)$, where λ_2 indicates the overall probability that an unobserved mutation is lethal, and r indicates the relative mutability of the unobserved chain. Results for the autoimmune response presented in this work have $r = 0$ because there are no unobserved mutations. For the NP response, $\lambda_2 = \lambda_1$ and $r = 1$ so that both chains are assumed to accumulate mutations at the same rate.

To determine the total number of cells that have undergone d divisions and accumulated m mutations, consider that the probability of having m mutations after d generation is a Poisson process with an average of μd. Thus, after d divisions, the expected number of cells with m mutations that are still alive is:

$$\alpha^d (1 - \lambda_1)^m e^{-\mu d} \frac{(\mu d)^m}{m!}$$

By summing over all possible numbers of mutations, we arrive at the total number of live cells after d divisions:

$$\sum_m \alpha^d (1 - \lambda_1)^m e^{-\mu d} \frac{(\mu d)^m}{m!}$$

The average number of mutations per sequence M is then simply the expected number of mutations carried among all surviving cells divided by the total number of live cells:

$$\begin{aligned}
M &= \frac{\sum_m \alpha^d (1 - \lambda_1)^m e^{-\mu d} \frac{(\mu d)^m}{m!} m}{\sum_m \alpha^d (1 - \lambda_1)^m e^{-\mu d} \frac{(\mu d)^m}{m!}} \\
&= \frac{\alpha^d e^{-\mu d} \sum_m \frac{((1 - \lambda_1)\mu d)^m}{m!} m}{\alpha^d e^{-\mu d} \sum_m \frac{((1 - \lambda_1)\mu d)^m}{m!}} \\
&= \frac{\sum_m e^{-(1-\lambda_1)\mu d} \frac{((1 - \lambda_1)\mu d)^m}{m!} m}{\sum_m e^{-(1-\lambda_1)\mu d} \frac{((1 - \lambda_1)\mu d)^m}{m!}} \\
&= \frac{E(\text{Poisson process with mean } (1 - \lambda_1)\mu d)}{1} \\
&= (1 - \lambda_1)\mu d
\end{aligned}$$

After simplifying, the average number of mutations per sequence (M) is simply the expected branch length in the absence of lethal mutations (μd), multiplied by the survival probability ($1 \times \lambda_1$).

The above formula actually has a subtle problem when comparing with experimental data that has been processed, as described in *Materials and Methods*. If all the sequences sampled from the tree share one or more branches (e.g., they all come from one-half of the tree), then these common branches will be removed during the pruning procedure. To estimate how large this error is, first consider the probability that the first branch emanating from the root will be pruned. If we sample S_t cells out of the total population, and S_t is reasonably small compared with the total clone size, the probability of choosing all cells from only half of the tree is: $2 \times (0.5)^{St}$.The initial factor of 2 arises because we can choose from either half of the tree. Similar formulas can be derived for the probability that the second level branches will be erased (i.e., all the sequences will be from one-quarter of the tree), etc. If we sum this potential error, it is: $(0.5)^{St}/(1 - (0.5)^{St-2})$, for large enough S_t, and S_t much less than the total clone size (a reasonable assumption because S_t is usually much smaller than the pick size A_t). This value can be approximated in the same limit as: $(0.5)^{St-1} + (0.5)^{2St-3}$. Although it is easy to see that this error decreases rapidly for increasing values of S_t, we still take into account the full error in our predictions by adding this factor to the observed average number of mutations (M_t) before comparing with the predicted value (M).

The average number of sequences present at the root of the tree (R)

Each sequence in the root of a clonal tree represents a cell that has undergone d divisions without accumulating any mutations. The probability of this occurring is $e^{-\mu d}$. These unmutated cells will be enriched in the population due to the death of cells that accumulate lethal mutations. The fraction of cells in the root is thus

theexpected number of unmutated cells divided by the total number of surviving cells:

$$\frac{\alpha^d e^{-\mu d}}{\alpha^d(1-\lambda_1\mu)^d} = \left(\frac{e^{-\mu}}{(1-\lambda_1\mu)}\right)^d$$

Finally, multiplying this fraction by the number of sampled sequences in any particular clonal tree (t) gives the number of sequences expected to be present at the root:

$$R = S_t \times \left(\frac{e^{-\mu}}{(1-\lambda_1\mu)}\right)^d$$

Total number of sequences in nodes with repeated sequences (P)

Sequences will appear together in a clonal tree if all of their mutations occurred along common branches in the tree (i.e., all of their mutations are shared). We derive a formula for the number of sequences appearing in any of the nodes of a clonal tree (including the root) that contain multiple sequences (P). To estimate this shape measure, we derive the probability that a sampled cell shares a common mutation history with one or more other sampled cells. This estimation is an approximation that only becomes precise for large enough clone sizes and may contain sampling artifacts for small clones. Comparing this approximation with simulation results suggests that the error is relatively small (data not shown).

To exemplify the algorithm for computing P, assume the cell A in Fig. 7 has been sampled. We compute the number of other cells that share the same mutation history as A by summing over generations. As labeled in Fig. 7, the expected number of live cells is $\alpha/2$ among the B's, $\alpha^2/2$ among the C's, $\alpha^3/2$ among the D's, and so on. In order for A and B to appear together, each must have not mutated during the last division cycle. For A and C to appear together, each must have not mutated during the last two division cycles, and analogously for A and D.

The probability that a particular cell A accumulated no mutations during its previous i division cycles is $e^{-\mu i}$. By summing over the categories (A, B, C, etc.), the total number of other cells that are expected to carry the same sequence as A can be calculated as:

$$\sum_{i=1}^{d} \frac{\alpha^i}{2} e^{-2\mu i} = \frac{\alpha}{2} e^{-2\mu} \frac{1-(\alpha e^{-2\mu})^d}{1-(\alpha e^{-2\mu})}$$

In this study, the factor $\alpha^i/2$ accounts for the number of cells that are expected to be alive in each category, and the factor $e^{-2\mu i}$ accounts for the fact that both cells must be unmutated during their final i division cycles to carry equivalent sequences.

Dividing the above formula by the expected total clone size (except for the current cell), $[\alpha(1-\lambda_1\mu)]^d - 1$, we arrive at the probability of a given cell to carry the same sequence as A. We denote this probability of two cells having a common history as p_1:

$$p_1 = \frac{\alpha}{2} \times \frac{e^{-2\mu}}{[(1-\lambda_1\mu)\alpha]^d - 1} \times \frac{1-(\alpha e^{-2\mu})^d}{1-(\alpha e^{-2\mu})}$$

The expected number of sampled sequences (S_t) appearing in nodes containing repeated sequences is:

$$P = \sum_{i=2}^{S_t} i \times p(\textit{node with i copies})$$

The probability of exactly i copies of a sequence appearing within a single node is:

$$p(\textit{node with i copies}) = p_1^{i-1}(1-p_1)^{S_t-i}\binom{S_t}{i}$$

The power of p_1 is i-1, because once we have randomly selected a sequence we have probability p_1 that any other sampled sequence will have a common mutation history and thus appear in the same node. The sum over all values of i yields:

$$\begin{aligned} P &= \sum_{i=2}^{S_t} i p_1^{i-1}(1-p_1)^{S_t-i}\binom{S_t}{i} \\ &= \frac{1}{p_1}\sum_{i=2}^{S_t} i p_1^{i}(1-p_1)^{S_t-i}\binom{S_t}{i} \\ &= \frac{1}{p_1}\left[\sum_{i=0}^{S_t} i p_1^{i}(1-p_1)^{S_t-i}\binom{S_t}{i} - \sum_{i=0}^{S_t} i p_1^{i}(1-p_1)^{S_t-i}\binom{S_t}{i}\right] \\ &= \frac{1}{p_1}[S_t p_1 - S_t p_1(1-p_1)^{S_t-1}] \\ &= S_t[1-(1-p_1)^{S_t-1}] \end{aligned}$$

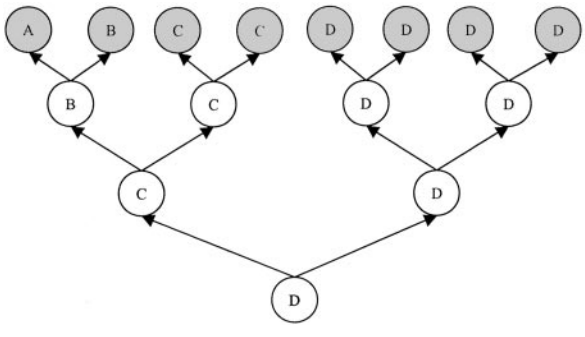

FIGURE 7. Schematic of clonal expansion showing how cells were labeled in the derivation of the analytical formula for the total number of sequences that appear in nodes with repeated sequences (P).

Acknowledgments

We thank Garnett Kelsoe, Jacqueline William, and Chad Euler for sharing their mutation data. We also thank Timothy Hilton for providing some of the early computer code for analyzing the simulated clonal trees.

References

1. Celada, F., and P. E. Seiden. 1996. Affinity maturation and hypermutation in a simulation of the humoral immune response. *Eur. J. Immunol. 26:1350.*
2. Shlomchik, M. J., P. Watts, M. G. Weigert, and S. Litwin. 1998. "Clone": a Monte-Carlo computer simulation of B cell clonal expansion, somatic mutation and antigen-driven selection. *Curr. Top. Microbiol. Immunol. 229:173.*
3. Kleinstein, S. H., and J. P. Singh. 2001. Toward quantitative simulation of germinal center dynamics: biological and modeling insights from experimental validation. *J. Theor. Biol. 211:253.*
4. Wabl, M., P. D. Burrows, A. von Gabain, and C. Steinberg. 1985. Hypermutation at the immunoglobulin heavy chain locus in a pre-B-cell line. *Proc. Natl. Acad. Sci. USA 82:479.*
5. Wabl, M., H. M. Jack, J. Meyer, G. Beck-Engeser, R. C. von Borstel, and C. M. Steinberg. 1987. Measurements of mutation rates in B lymphocytes. *Immunol. Rev. 96:91.*
6. Sale, J. E., and M. S. Neuberger. 1998. TdT-accessible breaks are scattered over the immunoglobulin V domain in a constitutively hypermutating B cell line. *Immunity 9:859.*
7. Martin, A., P. D. Bardwell, C. J. Woo, M. Fan, M. J. Shulman, and M. D. Scharff. 2002. Activation-induced cytidine deaminase turns on somatic hypermutation in hybridomas. *Nature 415:802.*
8. McKean, D., K. Huppi, M. Bell, L. Staudt, W. Gerhard, and M. Weigert. 1984. Generation of antibody diversity in the immune response of BALB/c mice to influenza hemagglutinin. *Proc. Natl. Acad. Sci. USA 81:3180.*
9. Sablitzky, F., G. Wildner, and K. Rajewsky. 1985. Somatic mutation and clonal expansion of B cells in an antigen-driven immune response. *EMBO J. 4:345.*
10. Zhang, J., I. C. M. MacLennan, Y.-J. Liu, and P. Lane. 1988. Is rapid proliferation in B centroblasts linked to somatic mutation in memory B cell clones? *Immunol. Lett. 18:297.*
11. Shlomchik, M. J., A. Marshak-Rothstein, C. B. Wolfowicz, T. L. Rothstein, and M. G. Weigert. 1987. The role of clonal selection and somatic mutation in autoimmunity. *Nature 328:805.*
12. Jacob, J., and G. Kelsoe. 1992. In situ studies of the primary response to (4-hydroxy-3-nitrophenyl)acetyl. II. A common clonal origin for periarteriolar lymphoid sheath-associated foci and germinal centers. *J. Exp. Med. 176:679.*

13. William, J., C. Euler, S. Christensen, and M. J. Shlomchik. 2002. Evolution of autoantibody responses via somatic hypermutation outside of germinal centers. *Science 297:2066*.
14. Jacob, J., G. Kelsoe, K. Rajewsky, and U. Weiss. 1991. Intraclonal generation of antibody mutants in germinal centres. *Nature 354:389*.
15. Schiaffella, E., D. Sehgal, A. O. Anderson, and R. G. Mage. 1999. Gene conversion and hypermutation during diversification of V_H sequences in developing splenic germinal centers of immunuzed rabbits. *J. Immunol. 162:3984*.
16. Vora, K. A., K. Tumas-Brundage, and T. Manser. 1999. Contrasting the in situ behavior of a memory B cell clone during primary and secondary immune responses. *J. Immunol. 163:4315*.
17. Papavasiliou, F. N., and D. G. Schatz. 2000. Cell-cycle-regulated DNA double-stranded breaks in somatic hypermutation of immunoglobulin genes. *Nature 408:216*.
18. Shlomchik, M. J., S. Litwin, and M. Weigert. 1990. The influence of somatic mutation on clonal expansion. In *Proceedings of the Seventh International Congress of Immunology. Prog. Immunol. VII:415*.
19. Lam, K. P., R. Kuhn, and K. Rajewsky. 1997. In vivo ablation of surface immunoglobulin on mature B cells by inducible gene targeting results in rapid cell death. *Cell 90:1073*.
20. Muramatsu, M., K. Kinoshita, S. Fagarasan, S. Yamada, Y. Shinkai, and T. Honjo. 2000. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell 102:553*.
21. Kanzler, H., R. Kuppers, M. L. Hansmann, and K. Rajewsky. 1996. Hodgkin and Reed-Sternberg cells in Hodgkin's disease represent the outgrowth of a dominant tumor clone derived from (crippled) germinal center B cells. *J. Exp. Med. 184:1495*.
22. Goossens, T., U. Klein, and R. Kuppers. 1998. Frequent occurrence of deletions and duplications during somatic hypermutation: implications for oncogene translocations and heavy chain disease. *Proc. Natl. Acad. Sci. USA 95:2463*.
23. Pasqualucci, L., A. Migliazza, N. Fracchiolla, C. William, A. Neri, L. Baldini, R. S. Chaganti, U. Klein, R. Kuppers, K. Rajewsky, and R. Dalla-Favera. 1998. BCL-6 mutations in normal germinal center B cells: evidence of somatic hypermutation acting outside Ig loci. *Proc. Natl. Acad. Sci. USA 95:11816*.
24. Lentz, V. M., and T. Manser. 2001. Cutting edge: germinal centers can be induced in the absence of T cells. *J. Immunol. 167:15*.
25. Toellner, K. M., W. E. Jenkinson, D. R. Taylor, M. Khan, D. M. Sze, D. M. Sansom, C. G. Vinuesa, and I. C. MacLennan. 2002. Low-level hypermutation in T cell-independent germinal centers compared with high mutation rates associated with T cell-dependent germinal centers. *J. Exp. Med. 195:383*.
26. Du Pasquier, L., M. Wilson, A. S. Greenberg, and M. F. Flajnik. 1998. Somatic mutation in ectothermic vertebrates: musings on selection and origins. *Curr. Top. Microbiol. Immunol. 229:199*.
27. Diaz, M., A. S. Greenberg, and M. F. Flajnik. 1998. Somatic hypermutation of the new antigen receptor gene (NAR) in the nurse shark does not generate the repertoire: possible role in antigen-driven reactions in the absence of germinal centers. *Proc. Natl. Acad. Sci. USA 95:14343*.
28. Diaz, M., and M. F. Flajnik. 1998. Evolution of somatic hypermutation and gene conversion in adaptive immunity. *Immunol. Rev. 162:13*.
29. Schroder, A. E., A. Greiner, C. Seyfert, and C. Berek. 1996. Differentiation of B cells in the nonlymphoid tissue of the synovial membrane of patients with rheumatoid arthritis. *Proc. Natl. Acad. Sci. USA 93:221*.
30. Kepler, T. B., and A. S. Perelson. 1993. Somatic hypermutation in B cells: an optimal control treatment. *J. Theor. Biol. 164:37*.
31. Kepler, T. B., and A. S. Perelson. 1995. Modeling and optimization of populations subject to time-dependent mutation. *Proc. Natl. Acad. Sci. USA 92:8219*.
32. Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1992. *Numerical Recipes in C.* Cambridge University Press, Cambridge.