A TEST FOR CLONAL RELATEDNESS IN A SET OF LYMPHOCYTES

By SAMUEL LITWIN AND MARK SHLOMCHIK

The basic problem is to decide if a set of immortalized cell lines that have indistinguishable DNA sequences and/or Ig rearrangements could have independently arisen by chance. To establish that any set of cells with the same DNA sequences or rearrangements arose from a single clone, it suffices to show that the set was not likely to have arisen from two clones, since this possibility is more likely than that of three clones, etc.

To assess clonality, we use three different molecular features of B cells that are highly variable among independent clones, and thus can serve as clonal markers: (a) sequences in and around the VDJ junctions (i.e., the “HVR3” region); (b) rearrangements of the nonproductive allele of the Vn locus; and (c) rearrangements of the nonproductive allele at the Vx locus. Members of a single clone will share these features, whereas members of different clones will not unless by coincidence. The confidence with which we can use each criterion to establish clonal relatedness depends on the likelihood of such coincidence. In our analysis, we estimate the frequency of such coincidence by surveying “index sets” of panels of known independent B cells for identity of these molecular features. Our assumption here is that the pool of B cells from which our putative clones are drawn has the same or lower frequency of coincidental identity. As described below, we use these frequencies to generate p values for the null hypothesis that putative members of clones that are identical for those clonal markers were, nonetheless, generated by coincidence.

For estimating the frequency of independent identity at VDJ junctions, we use the data of T. Manser (personal communication), who sequenced multiple antiardsonate V regions that had rearranged the same Vn, Dn, and Jn (and in fact were combined with the same Vx10Jx1 L chain). This is a conservative choice, since in anti-Ars the HVR3s are short, almost always the same length, and encode residues that are important for Ars binding. For estimating the frequency of independent identity at nonproductive allele rearrangements, we use Southern blots of plasmacytommas (1, and our unpublished data). We score comigration (+/- 2 mm) on three sets of single filters each for Jn and Jx. For the VDJ join comparison, the frequency is probably much higher in the index set owing to the restrictions found in the region in anti-Ars antibodies. For the nonproductive allele rearrangements, it is probably the same, as these events take place in pre-B cells.

Table I shows the data from these index sets. The number of distinguishable cell lines, d, for a given characteristic, e.g., VDJ junction (HVR3), in the sample of size

This research was supported by grant CA-06927 from the National Institutes of Health, and by an appropriation from the Commonwealth of Pennsylvania.

293
Appendix

Table I
Tests of Clonality Based on Individual Criteria

| Criterion | r | d | mle(n) | n-low | n-high | n-min | Joint MLE | \( p \) value |
|-----------|---|---|--------|-------|--------|-------|----------|------------|
|           |   |   |        |       |        |       |          | \( m = 2 \)| \( m = 3 \) |
| HVR3      | 28| 26| 180    | 64    | 1,021  | 26    | 180      | 0.0674    | 0.0233    |
| V\(_x\)   | 12| 11| 62     | 17    | 1,290  | 21    | 152      | 0.0698    | 0.0250    |
|           | 22\(^1\)| 21| 224    | 54    | 4,510  |       |          |            |
|           | 6 | 6 | -      | 8     | -      |       |          |            |
| V\(_u\)   | 21\(^1\)| 21| -      | 78    | -      | 32    | 310      | 0.0507    | 0.0167    |
|           | 20| 19| 98     | 45    | 3,710  |       |          |            |
|           | 34| 32| 269    | 94    | 1,526  |       |          |            |

\(^1\) Indicates the data subset used to obtain the \( p \) value for the experiment.

\( r \) can be used to estimate the total number of distinguishable cell lines that make up the sampled population. From any sample, assuming an equal chance of selecting each of the possible cell lines, we can use the method described in reference 2 to determine the number of possible different cell lines in the overall population. With a sample size of \( r \) cell lines with \( d \) distinguishable lines, the number of \( n \) that maximizes:

\[ P(d|r,n) = S(d,r) (n)!/n^r, \]

is the maximum likelihood estimate (MLE) of population size. This relation also yields confidence bounds on \( n \). \( S(d,r) \) are Stirling's numbers of the second kind.

For example, for the \( V_x \) rearrangements, one Southern blot filter had 12 nonproductive rearrangements, of which 11 were different (i.e., one coincidental comigration). The maximum likelihood estimate of the number of possible lines is 62, and a lower (upper) 95% one-sided confidence bound on this number is 17 (1,290). These results, along with another set of similar results, are also given in Table I. A second set of 22 cells taken from another filter yielded 21 different lines. Thus, we are certain of at least 21 possible lines, and 95% confident (see Table I) of at least 54 different lines. A third data set yielded six lines from six cells and, thus, is uninformative as to the upper bound. We wish to refute that even two clones with identical features (e.g., \( V_x \) rearrangements) could have been created independently regardless of the number of isolates in a putative clone. We show that the identity of two independent selections (\( d = 1 \) when \( r = 2 \)) is improbable. In the example above, using the assumption of equally likely selection of all lines, and the 95% lower bound of 54, the chance that two identical lines would be selected is \( 1/54 < 0.05 \) and statistically significant. The absolute lower bound, 21 possible lines, gives a significant result too. Hence, if we accept the assumption of equally likely selection of all possible lines, we can reject the hypothesis that two identical lines were chosen independently. This same conclusion can be drawn from the other experiments in Table I.

Although we have no reason to doubt the equal likelihood assumption, we can dispense with it by supposing that there are \( n \) (an unknown number) possible cell
lines. Let \( p_1, p_2, \ldots, p_n \) be the (also unknown) multinomial probabilities of selecting these lines. The chance of getting the same line given two selections is:

\[
p(d = 1|r = 2) = p_1^2 + p_2^2 + \ldots + p_n^2.
\]

This probability may be made as large as we like, viz., with \( p_1 = 1 \), and the other \( p_i = 0 \), the sum is 1. However, if there is a constraint on how large any one \( p_i \) can be, the sum is smaller than one.

We are interested in distributions that maximize both the chance of making two identical observations on two selections and the chance that a large set of selections yield many different cell lines. These distributions have one or a few large \( p_i \), and the rest very small \( p_i \). Consider distributions with only one large \( p_i \), say \( p_1 \), and the rest small and equal. If there are \( n \) possible cell lines, then:

\[
p(d = 1|r = 2) = p_1^2 + \frac{(1-p_1)^2}{(n-1)} = \text{value 1}.
\]

\( p_1 \) must allow this event \( d = 1|r = 2 \) to occur with reasonable likelihood, say \( p \) value 1 \( \geq 0.05 \). Likewise, to use the \( V_x \) rearrangement example, \( p(d > 20|r = 22) \) \( \Delta \) \text{value 2} \( \geq 0.05 \) must also be true, since this observation was made in one of our index sets. \( p(d > 20|r = 22) \) is determined as follows. The chance of selecting >20 different cells can be maximized by letting \( n \) grow indefinitely. In that case, all selections that are not from population 1 (denoted by \( P_1 \), which has probability \( p_1 \)) are guaranteed to be different. With \( n \) very large, the number of different observations is the number obtained outside \( P_1 \), plus one if any observations are made within \( P_1 \). For example, if 22 selections are made, at most two of them may come from \( P_1 \), else the total number of different selections will be less than the observed 21. Thus, assuming one large \( p_1 \), the rest small yields the inequality:

\[
p(d > 20|r = 22) < b(0,22,p_1) + b(1,22,p_1) + b(2,22,p_1) = \text{value 2},
\]

where \( b(k,n,p) \) is the binomial distribution with parameter \( p \) and \( n \) trials.

Increasing the value of \( p_1 \) increases \( p \) value 1 but decreases \( p \) value 2. A value of \( p_1 \) that makes \( p \) value 1 = \( p \) value 2 gives the largest possible value that can be assigned simultaneously to both events (and exceeds the chance that both occur regardless of possible dependencies). For the above example, we find that if \( p_1 = 0.24293 \), then \( p \) value 1 = \( p \) value 2 = 0.06983. Thus, there exist values of \( p_1 \) that allow both observations to be not wholly improbable. However, 7% is the highest probability that can be simultaneously assigned to the pair under these conditions. Moreover, if \( p_1 < 0.195 \), then \( p \) value 1 < 0.05, and if \( p_1 \) exceeds 0.26, then \( p \) value 2 < 0.05, so there is only a narrow range for \( p_1 \) that makes both observations somewhat likely. To maximize \( p \) value 1, we used the 95% confidence lower bound, namely 54, for \( n \). The results change but little if 100,000 is used instead. In that case, the best value of \( p_1 \) is 0.24885 and the joint \( p \) value declines to 0.062.

Although the above discussion was limited to one large probability population \( (P_1) \), there could be several such higher probability populations. We next show, however, that \( p \) values are maximized by assuming only one high probability population. Suppose instead of only one large probability cell line, there are \( k \). Then:
\[ p(d = 1 | r = 2) = p_1^2 + \ldots + p_k^2 + (1 - \text{sum})^2/(n-k), \]

where \( \text{sum} = p_1 + \ldots + p_k \), and \( n \) is the total number of cell lines.

\[ p(D > d | R = r) \] is upper bounded by a sum of multinomial terms:

\[ \left( \begin{array}{c} r \\ j_1, j_2, \ldots, j_k, j_{k+1} \end{array} \right) \frac{p_{j_1} \cdot p_{j_2} \cdot \ldots \cdot p_{j_k} \cdot r-\text{sum}_j}{p_1 \cdot p_2 \cdot \ldots \cdot p_k \cdot q} \]

where the first term is the multinomial coefficient for \( r \) draws yielding \( j_i \) lines from the first population, \( \ldots j_k \) from population \( k \), and the rest \( (r-\text{sum}_j) \), where \( \text{sum}_j = \sum_{i=1}^{k} j_i \), and \( q = 1 - p_1 \cdot \ldots \cdot p_k \), from the low probability populations with no chance of duplication. The sum is taken over all combinations of \( j_1, \ldots j_{k+1} \), for which \( D > d \). This is determined by defining variables:

\[ h_i = 1 \text{ if } j_i > 0, \text{ zero otherwise, for } i = 1,2, \ldots k \]

\[ h_{k+1} = j_{k+1}. \] Then \( D = h_1 + \ldots + h_{k+1}. \)

This technique was used for two and for three "big" probabilities (i.e., \( k = 2 \) or \( 3 \)). Systematic searches of the spaces \( (p_1,p_2,p_3) \), and also of \( (p_1,p_2,p_3,p_4) \) both showed that the assignment \( p_1 = 0.24293 \) and the other \( p_i \) all near zero gave the largest value of \( \min (p \text{ value } 1, p \text{ value } 2) = 0.06983 \). Thus, the vector \( (0.24293,0,0,\ldots,0) \) is an extreme point for any number of possible high probability cell lines.

We have two other data sets in the same group as \( r = 22, d = 21 \). These, however, are both slightly more compatible with \( r = 2, d = 1 \). Since all the data must be explained, we have chosen the most difficult of the three to stand for the whole set. \( p \) values would be smaller if we were to simultaneously consider all three in relation to the event \( (r = 2 | d = 1) \).

This same technique was used to test the hypothesis of independent clones against the data sets for VDJ junctions and for \( V_n \) nonproductive rearrangements; the results are given in Table I. Since all three experiments are testing the same null hypothesis, namely, that the observed identical cell lines were derived from at least two independent clones, and these experiments are independent, we combine them using Fisher's formula (3):

\[ X^2 = -2 \sum_{j=1}^{s} [\log(p \text{ value } j)], \]

\[ \begin{array}{|c|c|c|c|c|}
\hline
\text{Criteria} & \chi^2 & df & \text{p value}^{*} \\
\hline
\begin{array}{l}
(HVR3,V_e) \\
(HVR3,V_n) \\
(V_e,V_n) \\
(HVR3,V_e,V_n)
\end{array} & \begin{array}{l}
9.7261 \\
10.565 \\
11.287 \\
15.690
\end{array} & \begin{array}{l}
4.896 \\
15.703 \\
15.562 \\
23.081
\end{array} & \begin{array}{l}
4 \\
4 \\
4 \\
6
\end{array} & \begin{array}{l}
0.0453 \\
0.0347 \\
0.0235 \\
0.0155
\end{array} & \begin{array}{l}
0.00492 \\
0.00345 \\
0.00367 \\
0.00077
\end{array} \\
\hline
\end{array} \]

* See Table I.
where the $\chi^2$ statistic has degrees of freedom equal to twice the number of experimental results, $e$, that are combined. All four possible combinations of experiments are presented in Table II. All combinations are statistically significant.

Finally, we consider other possible numbers of clones that might have contributed the original set of hybridomas. If $m$ clones are being considered, then $\rho(d = 1 \mid r = m) = \rho_1^m + \rho_2^m + \cdots + \rho_n^m$. Proceeding as before, we tabulated both $m = 2$ and $m = 3$.

References
1. Coleclough, C., R. P. Perry, K. Karjalainen, and M. Weigert. 1981. Aberrant rearrangements contribute significantly to the allelic exclusion of immunoglobulin expression. Nature (Lond.). 290:372.
2. Weigert, M., and R. Riblet. 1976. Genetic control of antibody variable regions. Cold Spring Harbor Symp. Quant. Biol. 41:837.
3. Fisher, R. A. 1936. Statistical Methods for Research Workers, 6th edition. Oliver & Boyd Ltd., Edinburgh.