

## Letter to the Editor

### The Power to Detect Recombination Using the Coalescent

Celeste J. Brown,\*† Ethan C. Garner,†<sup>1</sup> A. Keith Dunker,† and Paul Joyce\*‡

\*Division of Statistics, University of Idaho, †School of Molecular Biosciences, Washington State University; and ‡Department of Mathematics, University of Idaho

**Author:** This article has been lightly edited for grammar, style, and usage. Please compare against your original document and make changes on this galley. Please limit yourself to substantive changes that affect meaning. Copyeditor

There are a wide variety of models for estimating the phylogenetic relationships among amino acid and nucleotide sequences sampled from organisms at the population, species, and kingdom levels (reviewed by Swofford et al. 1996). One of the assumptions of these models is that recombination has not occurred in the history of the sampled sequences. When recombination is present, the relationships may be illustrated by a bifurcating graph (fig. 1), which is a network of relationships between parts of all sequences (Griffith and Marjoram 1996). Most commonly used phylogeny methods do not construct such a network, and other measures, such as construction of separate phylogenies, must be taken to find relationships using these methods. Recombination may also be a problem for tandemly repeated genes where intragenic recombination, also known as gene conversion, may occur at a substantial rate (Wiuf 2000).

Before investigators use techniques that assume there is no recombination, they must be convinced that recombination is not present. Even low levels of recombination may have profound effects on phylogeny reconstruction and the conclusions one draws, so using a method that has a good chance of detecting low levels of recombination is important. Statistical power analysis can be used to evaluate various methods to determine the best method. Once a researcher is convinced that a method for detecting recombination has substantial power, that method may be used with confidence. Our purpose is to show that rigorous statistical methods exist for evaluating programs that detect recombination.

Power is the probability that a statistical test will reject the null hypothesis. This probability is calculated using a statistical model that includes both the null and alternative hypotheses. Since the calculation depends on the parameters of the model, power is not a single probability, but rather a series of probabilities, one for each parameter value. For tests of recombination, the parameter of interest is the recombination rate,  $r$  (recombinations per mutation per site per generation), and the null hypothesis is that there is no recombination in the history of the sample ( $r = 0$ ). The power of a test depends

on many factors: the level of significance or type I error, the sample size, and how far the alternative hypothesis is from the null hypothesis. The  $\alpha$ -level, or type I error, that is chosen is the probability that the null hypothesis will be rejected when it is true. This level should not be too high, since looking for recombination that does not exist can be an aggravating task. A power analysis uses data simulated under the null hypothesis and various alternative hypotheses to determine how often the method rejects the null hypothesis for a specific type I error rate. The power should increase toward 100% as  $r$  increases; the method whose power increases the fastest is the most powerful.

Other researchers have investigated the power of various methods for detecting gene conversion. PLATO is a program that detects spatial phylogenetic heterogeneity and can be used to detect recombination (Grassly and Holmes 1997). Grassly and Holmes (1997) tested the power of PLATO to detect recombination by simulating phylogenies with different topologies under a single coalescent model and then manually recombining sequences between phylogenies. They also compared the results from PLATO with a well-documented example of recombination among the *argF* genes of *Neisseria*. Drouin et al. (1999) also used well-documented examples of gene conversion to test various programs for detecting gene conversion. Neither method is optimal for conducting a power analysis in which thousands to tens of thousands of data sets generated under various alternative hypotheses are needed.

The purpose of this note is to increase awareness that coalescent theory can be used to make direct, statistically valid comparisons among recombination detection methods. The coalescent with recombination (Hudson 1983) is a description of ancestral relationships and can be described as follows. As one traces the ancestry of a sample into the past, three possible evolutionary events may arise: a mutation may have occurred in one of the ancestral lines, two individuals may have a common ancestor, or the genetic material in question may have arisen as a result of recombination (see fig. 1). In this last case, the genetic material from a given individual has two common ancestors, where one piece of the DNA came from one ancestor and another piece came from a different ancestor. As the histories of the individuals in the sample are traced back into the past, eventually a common ancestor of all of the individuals in the sample will emerge.

The program TREEEVOLVE, version 1.32 (<http://evolve.zoo.ox.ac.uk/>), by Grassly and Rambaut, generates sequences according to the coalescent model of evolution with recombination (Hudson 1983). This method generates a network of relationships with mutations, recombinations, and coalescences randomly as-

<sup>1</sup> Present address: Department of Biochemistry and Biophysics, University of California at San Francisco.

Abbreviation: LRT, likelihood ratio test.

Key words: recombination, statistical power, gene conversion, hypothesis testing, likelihood ratio.

Address for correspondence and reprints: Celeste J. Brown, School of Molecular Biosciences, Washington State University, Pullman, Washington 99164-4660. E-mail: celesteb@disorder.chem.wsu.edu.

*Mol. Biol. Evol.* 18(7):000–000. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

“ $\alpha$ -level or type I error” changed to “ $\alpha$ -level, or type I error.”  
Correct?

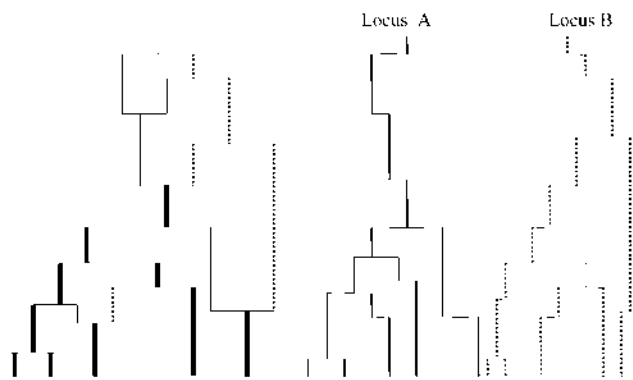


FIG. 1.—Coalescent phylogeny with recombination. The network on the left represents a phylogeny with both loci A and B: thin lines have only A, dotted lines have only B, and thick lines have both and result from recombination events. Trees on the right are for A and B separately.

signed according to parameters determined by the user. A random sequence is then produced and evolved along this network. TREEEVOLVE was used to generate sequences with and without recombination. The substitution model used by TREEEVOLVE was implemented using the F84 option, but with parameter settings that emulated the Kimura (1980) model: a transition/transversion ratio of 2, equal nucleotide frequencies, equal rates of substitution, and a Wright-Fisher model of evolution with no population subdivision.

Other parameter settings for our simulated data were motivated by a recent paper by Cheung et al. (1999), one of many demonstrating the importance of detecting gene conversion. Cheung et al. (1999) show that gene conversion has occurred at least once among the ADH genes of Old World monkeys and at least once among the ADH genes of humans. The total length of their sequences was 1,125 bp, and the number of sequences sampled was 6. RECOMBINE (Kuhner, Yamato, and Felsenstein 2000), a program that estimates population genetics parameters using maximum likelihood (see below), was used on Cheung et al.'s primate ADH sequences to estimate theta ( $\Theta$  = four times the effective population size times the mutation rate per site per generation) and the recombination rate  $r$ . An effective population size of 250,000 and a mutation rate of  $10^{-7}$  mutations per site per generation were used as parameters for TREEEVOLVE based on RECOMBINE's estimate of  $\Theta = 0.1$ . The recombination rate for these data was estimated by RECOMBINE to be 0.035 recombinations per mutation per site per generation. The rates used by TREEEVOLVE were chosen to provide a range of values around 0.035: 0, 0.01, 0.035, 0.06, 0.1, and 0.15 per mutation rate per site per generation, (with the given mutation rate of  $10^{-7}$ , this translates to 0,  $10^{-9}$ ,  $3.5 \times 10^{-9}$ ,  $6 \times 10^{-9}$ ,  $10^{-8}$ , and  $1.5 \times 10^{-8}$  recombinations per site per generation). One thousand data sets of six sequences each were generated for each of these recombination rates, except 0, for which 10,000 data sets were generated.

Theory guarantees that the likelihood ratio test (LRT) will be the most powerful test when the assump-

tions of the statistical model are met (Bain and Engelhardt 1992, pp. 417–428). We used a modified version of RECOMBINE, version 1.0 alpha, by Kuhner, Yamato, and Felsenstein (2000) (<http://evolution.genetics.washington.edu/lamarc/recombine.html>), to calculate the likelihoods used in our likelihood ratio hypothesis test. This program analyzes the data under the same model used to generate the data, such that the assumptions are met, and therefore the best power is guaranteed. The power of other methods for detecting recombination can be compared against this benchmark.

RECOMBINE calculates the log likelihood ratio for  $\hat{r}$  and  $\hat{\Theta}$   $\ln\{[L(\hat{\Theta}, \hat{r})]/[L(\Theta_0, r_0)]\}$ , and the log likelihood ratio at  $r = 0$  and  $\hat{\Theta}$   $\ln\{[L(\hat{\Theta}, r = 0)]/[L(\Theta_0, r_0)]\}$ . The log likelihoods are relative to the starting values of  $\Theta_0$  and  $r_0$  used at the beginning of the final long chain. The likelihood ratio test is a method for testing the likelihood of an alternative hypothesis relative to the null hypothesis:

$$\Lambda = \frac{L(\hat{\Theta}, r = 0)}{L(\Theta_0, r_0)} \bigg/ \frac{L(\hat{\Theta}, \hat{r})}{L(\Theta_0, r_0)} = \frac{L(\hat{\Theta}, r = 0)}{L(\hat{\Theta}, \hat{r})} \quad (1)$$

and

$$-\ln \Lambda = \ln L(\hat{\Theta}, \hat{r}) - \ln L(\hat{\Theta}, r = 0). \quad (2)$$

The program RECOMBINE was implemented using a starting value for  $\Theta$  of 0.1, the same value used to generate the sequences. The starting value for  $r$  was set to the recombination rate at which the sequences were generated, except for  $r = 0$ , which was set to 0.01. An unweighted pair grouping method with arithmetic means tree was estimated for each data set using the program PHYLIP (Felsenstein 1993) and was provided to RECOMBINE as the starting tree. RECOMBINE was implemented using 10 short chain runs with a sampling increment every 40 steps for 400 steps total, followed by a single long chain run with a sampling increment every 10 steps for a total of 20,000 steps. The program provided the log likelihood of the final genealogy relative to the starting genealogy for various values of  $\Theta$  and  $r$ . The log likelihoods at  $\hat{r}$  and  $\hat{\Theta}$  and at  $r = 0$  and  $\hat{\Theta}$  were recorded.

RECOMBINE was started at the true values of  $\Theta$  and  $r$  in order to assure that the correct LRT statistic was obtained. While we admit this is not a fair way to assess the efficiency of RECOMBINE as an algorithm for producing maximum-likelihood estimates, it will produce the correct LRT statistic needed to evaluate the power of other statistical methods relative to the best they are expected to do.

The 10,000 data sets generated by TREEEVOLVE using  $r = 0$  were used to determine the distribution of  $-\ln \Lambda$  under the null hypothesis of no recombination. The 10,000  $-\ln \Lambda$  values were ordered by size, and the value at the 9,580th place (1.03) became the critical value for an  $\alpha$ -level of 0.042. When certain regularity conditions apply, a chi-square distribution with one degree of freedom can be used for likelihood ratio tests. However, these regularity conditions do not appear to be met

Brown et al.

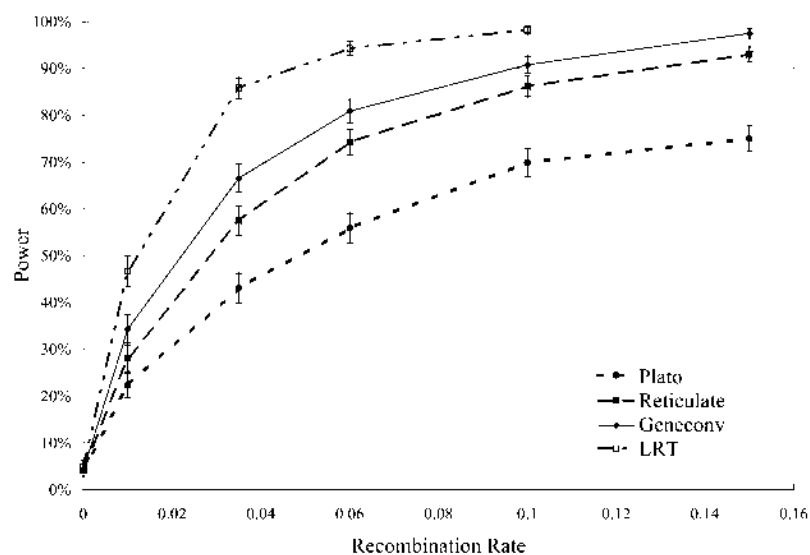


FIG. 2.—Power of the four methods for detecting recombination at  $\alpha = 0.042$ . LRT is the likelihood ratio test. Error bars are two times the standard error.

for coalescent models, so asymptotic chi-square distributions should not be used. Our results indicate that using a chi-square distribution produces an excessively conservative test. The only way to obtain the null distribution for  $-\ln \Lambda$  is through simulation, and this is necessary when any of the parameters ( $\Theta$ , sample size, sequence length) are altered.

In order to determine the power of the LRT, the fraction of data sets having  $-\ln \Lambda$  values greater than or equal to the critical value was found for each recombination rate. These values indicate the power of the test to detect recombination for each of the alternative hypothesis (fig. 2). Note that at the lowest nonzero recombination rate, the power of the LRT was less than 50%. This suggests that no method for detecting recombination will do well at very low rates. The power rose rapidly to almost 90% at the next higher rate, and it was so close to 100% at  $r = 0.1$  that we did not feel it was necessary to use the computing time for the highest recombination rate.

It should be noted that the  $\Lambda$  statistic used here is not exactly the generalized likelihood ratio. The true likelihood ratio maximizes the likelihood of  $\Theta$  at  $r = 0$ , whereas here the likelihood was taken at the joint maximum-likelihood estimate of  $\Theta$  and  $r$  when  $r$  was set to 0. We find that these two values are not very different from each other, so our power estimates are probably only slightly lower than those for the true test.

The website by D. Robertson ([http://grinch.zoo.ox.ac.uk/RAP\\_links.html](http://grinch.zoo.ox.ac.uk/RAP_links.html)) conveniently categorizes various recombination detection programs. We chose one method from each of the categories in his list to perform our statistical power analysis. RECOMBINE is one of the population genetics-based methods described above. We used the phylogenetic tree method PLATO, version 2.11 (<http://evolve.zoo.ox.ac.uk/>), developed by Grassly and Holmes (1997). PLATO identifies lengths of sequences that do not share the phylogenetic relationships of the complete data set. The aim of PLATO is to detect

spatial phylogenetic variation, which might be caused by recombination as well as other factors. We used RETICULATE (<http://jcsmr.anu.edu.au/dmm/humgen/in-grid/reticulate.htm>), a method that analyzes patterns of sites developed by Jakobsen and Easteal (1996). RETICULATE identifies stretches of sequence within a multiple alignment that have phylogenetic relationships that are incompatible with each other and estimates whether these stretches are longer than would be expected by chance. The pairwise comparison method we used was GENECONV, version 1.02 (<http://www.math.wustl.edu/~sawyer/mbprogs>), developed by Sawyer (1989). This method uses pairwise comparisons to find stretches of sequence pairs that are more similar than would be expected by chance. We cannot claim that methods classified as similar according to Robertson have similar performances. Our comparisons are meant to illustrate power analysis.

RETICULATE and PLATO were implemented using their default values throughout. For each data set, PLATO was provided with a maximum-likelihood tree estimated using DNAML from the program PHYLIP (Felsenstein 1993). GENECONV was implemented using the default values except that the *g*scale was set to 1 in order to take into account the probability that mutations had occurred within recombined regions after the recombination event. The number of permutations of the data set for both RETICULATE and GENECONV was 10,000.

The first 1,000 data sets generated by the TREEVOLVE program at  $r = 0$  were used to determine the level of significance,  $\alpha$ . The default  $\alpha$ -levels given in the documentation of the programs are not necessarily based on a coalescent model. Therefore, if coalescent theory is to be used to make comparisons, the tests may need to be adjusted so that the  $\alpha$ -levels under the coalescent model are approximately the same for each method. The type I error for PLATO at  $\alpha = 0.05$  should have been determined by ordering the significant Z-val-

Letter to the Editor

ues reported by PLATO for these data and using the 50th largest Z-value as the critical value for an  $\alpha$ -level of 0.05. Unfortunately, there were only 42 Z-values greater than the built-in and immutable Z-value provided by the program. The  $\alpha$ -level reported for all of the programs was 0.042, instead of the more common  $\alpha = 0.05$ .

As with PLATO, the  $\alpha$ -levels reported in the documentation for GENECONV and RETICULATE were not the same as the levels determined by the simulations. To correct for this phenomenon, the criteria for rejecting the null hypothesis of no recombination was changed for each program tested. As a result of our adjustments, each test mistakenly detected recombination in 42 out of 1,000 data sets generated under the null hypothesis of no recombination. GENECONV and RETICULATE generate  $p$ -values based on a permutation test. Global simulated  $p$ -values were used for detecting recombination using GENECONV as recommended in the documentation. The  $\alpha$ -levels for GENECONV and RETICULATE were determined by ordering the  $p$ -values calculated for the 1,000 data sets generated with no recombination. The 42nd smallest value was taken as the critical value for an  $\alpha$ -level of 0.042. Global simulated  $p$ -values were used for detecting recombination using GENECONV as recommended in the documentation.

The power of each method was determined using the 1,000 data sets generated at each of the five recombination rates. For PLATO, the fraction of data sets out of 1,000 having significant Z-scores were counted. For RETICULATE and GENECONV, the fraction of data sets out of 1,000 having  $p$ -values less than or equal to the critical value were counted. The results for several thousands of hours of CPU time are summarized in figure 2. As expected, the LRT has the greatest power to detect the presence of recombination. GENECONV has the second greatest power, with values above 90% at the two highest recombination rates. RETICULATE has the third greatest power, surpassing 85% at the two highest rates. PLATO has the lowest power to detect recombination across all rates. PLATO's power reaches only 56% at a recombination rate of 0.06; the best this method does is 75% at the highest rate.

It is interesting that GENECONV and RETICULATE have similar powers, considering that one uses pairwise sequence comparisons and the other uses multiple-sequence comparisons in the form of parsimony trees at each informative site. Both analyses identify stretches of similarity or dissimilarity using their respective metrics, and this may be why they have similar powers. PLATO detects changes in the likelihood of the phylogeny along the length of the sequences. This could be caused either by rate heterogeneity, selection, or recombination. Since the program is not specifically de-

signed to detect recombination, it may be more powerful for detecting these other alternatives.

The ability to simulate sequences under the coalescent with recombination makes power analysis much easier for those who wish to develop new recombination detection methods or test one of the many methods that are currently available.

#### Acknowledgments

We thank Dr. Mary Kuhner for her assistance in devising the likelihood ratio test and for providing a modified version of RECOMBINE. We thank all of the individuals who freely provide their programs via the Internet.

#### LITERATURE CITED

- BAIN, L. J., and M. ENGELHARDT. 1992. Introduction to probability and mathematical statistics. Duxbury Press, Belmont, Calif.
- CHEUNG, B., R. S. HOLMES, S. EASTEAL, and I. R. BEACHAM. 1999. Evolution of class I alcohol dehydrogenase genes in catarrhine primates: gene conversion, substitution rates and gene regulation. *Mol. Biol. Evol.* **16**:23–36.
- DROUIN, G., F. PRAT, M. ELL, and G. D. P. CLARKE. 1999. Detecting and characterizing gene conversions between multigene family members. *Mol. Biol. Evol.* **16**:1369–1390.
- FELSENSTEIN, J. 1993. PHYLIP (phylogeny inference package). Version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.
- GRASSLY, N. C., and E. C. HOLMES. 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.* **14**:239–247.
- GRIFFITHS, R. C., and P. MARJORAM. 1996. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3**:479–502.
- HUDSON, R. R. 1983. Properties of the neutral allele model with intergenic recombination. *Theor. Popul. Biol.* **23**:183–201.
- JAKOBSEN, I. B., and S. EASTEAL. 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *CABIOS* **12**:291–295.
- KUHNER, M. K., J. YAMATO, and J. FELSENSTEIN. 2000. Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**:1393–1410.
- SAWYER, S. 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**:526–538.
- SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–514 in D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. *Molecular systematics*. Sinauer, Sunderland, Mass.
- WIUF, C. 2000. A coalescence approach to gene conversion. *Theor. Popul. Biol.* **57**:357–367.

XXXXXXXXXXXXXXXXXX, reviewing editor

Accepted March XX, 2001

Please provide  
name of reviewing  
editor and  
accepted date.