J Mol Evol (2011) 73:273–286
DOI 10.1007/s00239-011-9473-0

Recombination Detection Under Evolutionary Scenarios Relevant to Functional Divergence

Rachael A. Bay · Joseph P. Bielawski

Received: 22 June 2011 / Accepted: 7 November 2011 / Published online: 1 January 2012

Abstract Recombination can negatively impact methods designed to detect divergent gene function that rely on explicit knowledge of a gene tree. However, we know little about how recombination detection methods perform under evolutionary scenarios encountered in studies of functional molecular divergence. We use simulation to evaluate false positive rates for six recombination detection methods (GENECONV, MaxChi, Chimera, RDP, GARD-SBP, GARD-MBP) under evolutionary scenarios that might increase false positives. Broadly, these scenarios address: (i) asymmetric tree topology and sequence divergence, (ii) non-stationary codon bias and selection pressure, and (iii) positive selection. We also evaluate power to detect recombination under truly recombinant history. As with previous studies, we find that power increases with sequence divergence. However, we also find that accuracy to correctly infer the number of breakpoints is extremely low. When recombination is absent, increased sequence divergence leads to increased false positives. Furthermore, one method (GARD-SBP) is sensitive to tree shape, with higher false positive rates under an asymmetric tree topology. Somewhat surprisingly, all methods are robust to the simulated heterogeneity in codon bias, shifts in selection pressure and presence of positive selection. Based on these findings, we recommend that studies of functional divergence in systems where recombination is plausible can, and should, include a pre-test for recombination. Application of all methods to the core genome of *Prochlorococcus* reveals a substantial lack of concordance among results. Based on analysis of both real and simulated datasets we present some guidelines for the investigation of recombination in genes that may have experienced functional divergence.

Keywords Recombination · Power · Accuracy · Functional divergence · Positive selection · Non-stationary evolution

Introduction

Inter-lineage and inter-species recombination can facilitate the process of functional divergence by allowing organisms to access an "extended gene pool." Through the exchange of genetic material between organisms, or even between species, recombination increases the genetic diversity in a population, which can facilitate population evolution. Examples of recombination-assisted evolution are well documented in bacteria (e.g., Koonin et al. 2001; Boucher et al. 2003; Narra and Ochman 2006). One of the more striking examples concerns the core photosystem genes of cyanobacteria. Core photosystem genes are encoded in the genomes of both the photosynthetic *Synechococcus* and their phages, with the phages mediating recombination among the gene sequences and serving as a reservoir for an

Electronic supplementary material The online version of this article (doi:10.1007/s00239-011-9473-0) contains supplementary material, which is available to authorized users.

R. A. Bay · J. P. Bielawski
Department of Biology, Dalhousie University, Halifax,
NS B3H 4J1, Canada

R. A. Bay (✉)
Department of Biology, Hopkins Marine Station,
Stanford University, 120 Oceanview Blvd.,
Pacific Grove, CA 93950, USA
e-mail: rbay@stanford.edu

J. P. Bielawski
Department of Mathematics and Statistics, Dalhousie University,
Halifax, NS B3H 4J1, Canada

extended pool of sequence diversity. (e.g., Mann et al. 2003; Zeidner et al. 2005; Sullivan et al. 2006). There is considerable interest in functional divergence within systems such as these; however, resolving the relative roles of adaptive substitution and recombination is a challenge, as methods for detecting the signatures of adaptive substitution typically assume that gene sequences are non-recombinant.

In addition to providing an understanding of gene sequence evolution, knowledge of recombination events also helps avoid error in phylogeny-based analysis. Recombination events can cause the phylogeny of a gene or gene fragment to be incongruent with the evolutionary history of the organism (e.g., Ochman et al. 2000; Ragan 2001). While this phylogenetic disagreement may be useful for understanding the recombination process, it sometimes causes errors in analyses that depend on accurate knowledge of the phylogeny. For example, detection of positive selection using models of codon substitution can yield false positives when recombinant segments are present within gene sequences (Anisimova et al. 2003; Shriner et al. 2003; Scheffler et al. 2006). Although the sensitivity of methods formulated at the amino acid level (e.g., Gu 1999; Knudsen and Miyamoto 2001; Susko et al. 2002) has not been systematically evaluated, their explicit reliance on a phylogeny is also expected to make their results sensitive to recombination. For this reason, studies of functional divergence among organisms where recombination is plausible should be accompanied by an analysis of recombination.

A wide variety of methods are available for the detection of recombination, and each has unique advantages and limitations. Here we focus on a subset of methods that we place into two broad groups. The first group, the “substitution-distribution methods,” (e.g., GENECONV: Sawyer 1989) test for significant clustering of substitutions within gene sequences. Methods in the second group are referred to as “phylogenetic methods” because they search for significant variability in tree topologies among adjacent sequence fragments. Earlier methods based on this concept (e.g., RDP: Martin and Rybicki 2000), are not computationally expensive. However, the trend has been to employ increasingly complex models and statistical machinery in an effort to improve power of the phylogenetic methods. The tradeoff is increased computational cost. Examples include Bayesian methods (e.g., DualBrothers: Suchard et al. 2002; BARCE: Husmeier and McGuire 2003), which use posterior probabilities to identify adjacent regions with discordant phylogenies, and the Genetic Algorithm for Recombination Detection (GARD: Kosakovsky Pond et al. 2006), which uses a likelihood-based heuristic algorithm to find the best-fit number and location of recombination breakpoints. More detailed descriptions of the recombination detection methods are presented in the Supplementary Materials. We focus on commonly used methods from both groups because the consensus opinion is that researchers should not be overly reliant on a single approach (e.g., Posada 2002).

There have been several large-scale simulation studies of performance. Posada and Crandall (2001) simulated varying levels of diversity, recombination, and rate variation, finding that recombination detection methods are not generally powerful, although power does increase with diversity. In addition, few false positives were found in their simulations. Wiuf and colleagues (2001) found that certain combinations of branch lengths (e.g., short internal branches and long tips) might cause detection methods to have increased power. Chan and colleagues (2006) found that post-recombination substitutions decrease the ability of several methods to detect breakpoints. While all these studies employed a wide variety of evolutionary scenarios to evaluate recombination detection methods, none included null scenarios (i.e., no recombination) most relevant to the evolution of functionally divergent molecules.

The purpose of this study is to investigate the performance of recombination detection methods under evolutionary scenarios with particular relevance to the process of functional divergence. Three simulation studies are used to evaluate false positive rates for recombination under a range of conditions. In Simulation 1, we investigate the impact of tree shape, as genes experiencing functional divergence have varying levels of tree asymmetry and sequence divergence. In Simulations 2 and 3 we evaluate false positive rates when the mode of evolution changes across the phylogeny, as is often the case with functional divergence. Simulation 2 includes shifts in selection pressure and codon bias characteristic of gene sequences that have experienced a divergence in gene function whereas Simulation 3 focuses on sequences experiencing positive selection in a subset of lineages. In Simulation 4, we use sequences simulated with recombination to investigate the ability of recombination detection methods to accurately predict the presence and amount of recombination under different levels of recombination and divergence (i.e., power analysis). Finally, we apply the same set of recombination detection methods to a set of genes from the core genome of the cyanobacteria *Prochlorococcus*, which plays a significant role in photosynthesis on a global scale and has experienced functional divergence in a number of genes with respect to exploiting high and low light environments.

Methods

Methods for Detecting Recombination

We select three substitution-distribution methods (GENECONV, MaxChi and Chimaera) and three phylogenetic methods (RDP, GARD-SPB and GARD-MBP) for evaluation in series of four different simulation studies. We provide the primary reference for each method in Table 1, and review the analytical basis of each method in the on-line Supplementary Materials. GENECONV, MaxChi, Chimaera and RDP are selected because they are widely used, and because they were found to be relatively powerful in the simulation study of Posada and Crandall (2001). We select two GARD-based methods because they were found to have high accuracy in the simulation study of Kosakovsky Pond et al. (2006). Furthermore, the computational cost of the GARD-based methods is not so large as to preclude the analysis of large numbers of sequences typical of studies directed at functional divergence. MaxChi, Chimaera, RDP, and GENECONV are employed as currently implemented in the RDP3 software package (Martin 2009). Note that the RDP3 implementation of Chimaera, MaxChi, and GENECONV differs from the original implementation of those methods (see Supplementary Materials for additional details). GARD-based methods are employed as implemented in the HyPhy software package (Kosakovsky Pond and Muse 2005).

The GARD-based methods differ by focusing on either a single breakpoint (GARD-SBP) or multiple breakpoints (GARD-MBP). GARD-SPB employs a maximum likelihood framework to conduct rapid screening for a single breakpoint with discordant phylogenies on either side. The SBP method is recommended by Kosakovsky Pond et al. (2006) as a qualitative test of the presence of recombination. Note that we refer to the SBP method as a "GARD" method because it is suggested as a supplement for GARD-MBP, but it does not actually employ a genetic algorithm. GARD-MBP employs a genetic algorithm to identify multiple recombination breakpoints. Additional details about both methods are included in the on-line Supplementary Materials. When recombination is detected under GARD-MBP, a Kishino–Hasegawa (KH) test (Kishino and Hasegawa 1989) is employed with a Bonferroni correction for multiple testing. The purpose of the KH test is to determine whether phylogenies before and after a putative breakpoint are significantly different. Although the correction for multiple tests makes GARD-MBP less powerful as compared to GARD-SBP, it helps to control the rate of false positives (Kosakovsky Pond et al. 2006).

Table 1 Recombination detection methods evaluated by simulation and applied to real data

Method	Type	Reference
GENECONV	Substitution	Sawyer (1989)
MaxChi	Substitution	Maynard Smith (1992)
Chimaera	Substitution	Posada and Crandall (2001)
RDP	Phylogenetic	Martin and Rybicki (2000)
GARD-SBP	Phylogenetic	Kosakovsky Pond et al. (2006)
GARD-MBP	Phylogenetic	Kosakovsky Pond et al. (2006)

Generation and Evaluation of Simulated Codon Sequences

The program INDELible (Fletcher and Yang 2009) is used to simulate codon sequences for Simulations 1–3 (described in the next section), which represent different non-recombinant evolutionary scenarios that might be encountered in a study of functional divergence and could possibly increase false positives in recombination tests. In these studies, we are concerned only with the number of replicates that contain a false signal for recombination rather than the number of breakpoints detected. A replicate is considered to contain false signal for recombination if, for a given method, at least one breakpoint is detected having a p-value ≤ 0.05. In addition, for the phylogenetic methods (RDP and GARD-based methods), a recombination event is only considered significant if there is phylogenetic incongruence on either side of the breakpoint.

Simulation 4 is employed to make a direct comparison of each method's capacity to correctly infer the presence of recombination and the number of breakpoints in truly recombinant sequences. GARD-SBP is not evaluated as it is recommended only as a qualitative test for recombination. Here, we reuse simulated data previously employed by Kosakovsky Pond et al. (2006) to evaluate only GARD-MBP. Power is measured as the percentage of replicates having a significant signal for recombination. Accuracy is measured with respect to the number of inferred breakpoints.

Modeling Selection Pressure in Simulated Codon Sequences

Selection pressure is simulated at the codon level using the parameter ω, which is equal to the ratio of non-synonymous to synonymous substitutions and represents the strength and direction of selection pressure. A value of $\omega > 1$ specifies positive selection while $\omega < 1$ specifies purifying selection. Because selection pressure typically varies among sites within a given sequence, we use a beta distribution to model the distribution of ω among sites with no positive selection ($0 < \omega < 1$). The beta distribution is convenient for this purpose because its range from zero to one is ideal for modeling an ω distribution with no positive selection while employing only two shape parameters (p, q). When positive selection is simulated, an additional discrete category of sites in which $\omega > 1$ is added.

Real Sequence Data

For the analysis of genomic data, we use a set of 585 genes from the *Prochlorococcus* core genome. For the purposes of this study, the "core genome" contains all genes present in all 12 *Prochlorococcus* genomes. These 12 genomes (GenBank accession numbers BX548174, CP000552, CP000576, CP000551, CP000825, CP000111, CP000553, AE017126, CP000878, CP000554, BX548175, CP000435) were previously clustered into orthologous groups by Zhaxybayeva et al. (2009). In order to avoid cases of lateral gene transfer (LGT), which has been well documented in *Prochlorococcus*, only those sequences having a gene tree with the same topology as the organismal tree are analyzed for within-gene homologous recombination. We apply the same six recombination detection methods evaluated via simulation to analyze these real gene sequences (see the on-line Supplementary Materials for additional details regarding the filtering and processing of these genomic data).

Simulations and Analyses

Simulation 1: Tree Topology and Sequence Divergence

Genes that have experienced functional divergence often evolved under asymmetric tree topologies. One well-known example occurs in viral genes, which commonly exhibit rapid evolution in a single part of the phylogeny, resulting in asymmetry (e.g., Duffy and Holmes 2008; Pagán et al. 2010). Previous simulation studies employed either an artificial tree with a symmetric topology or the empirical estimate of a tree from a gene of interest (e.g., Posada and Crandall 2001; Chan et al. 2006). However, because the selected methods for detecting recombination measure parameters from the entire alignment, an asymmetric tree topology may inflate the number of informative, or variable sites, which may ultimately skew the parameter estimates. This issue has not been formally addressed, although the RDP manual (Martin 2009) warns that having both closely related and very divergent sequences in an alignment may result in errors for some methods. Interestingly, some recombination detection methods have been shown to have different performance when branch lengths within a tree are varied (Wiuf et al. 2001). Because divergence is known to impact recombination detection, we employ Simulation 1 to explore the effects of tree shape on false positive rates under a range of tree lengths.

Sequences (200 codons) are simulated based on a 16-taxon phylogeny, which is either asymmetric or symmetric (Fig. 1). In this simulation study the evolutionary model is a process of codon substitution that is homogeneous throughout the phylogeny, with no positive selection (beta distribution parameters: $p = 0.5$, $q = 2$). All non-stop codons have equal frequencies and the transition/transversion ratio (κ) is set to 2. For both phylogenies, all internal branches are of equal length and tips are adjusted to be consistent with a molecular clock. Sequences are generated under 10 levels of sequence divergence, with the total tree length (the sum of all branch lengths in the tree) being an integer from 1 and 10 nucleotide substitutions per codon site. To minimize the computational expense, we simulate $n = 50$ replicates for each level of divergence. This level of sampling is sufficient when the goal is to investigate differences among methods in the relationship between sequence divergence and false positive rate. A small scale-simulation study (tree length = 4) indicates variability due to $n = 50$ is small (Supplementary Table S1) relative to the tree length effect (Fig. 2).

We apply each of the recombination detection methods (Table 1) to the simulated data. Results indicate tree length affects the false positive rate for some methods. MaxChi and Chimaera both suffer increasing false positives as tree length increases (Fig. 2). False positives for MaxChi range from 6% for the smallest tree length to 56% for the largest. For Chimaera, false positives increase from 8 to 60%. RDP also suffers an increase in false positives with tree length, but to a much lesser extent than MaxChi and Chimaera (the greatest false positive rate is 20%). False positives for MaxChi, Chimaera and RDP are not sensitive to tree shape. Our findings contrast with the study of Posada and Crandall (2001), which found the MaxChi and Chimaera did not yield large numbers of false positives, even under their most divergent datasets. However the total divergence simulated by Posada and Crandall was substantially lower than the levels covered by our more divergent datasets. We

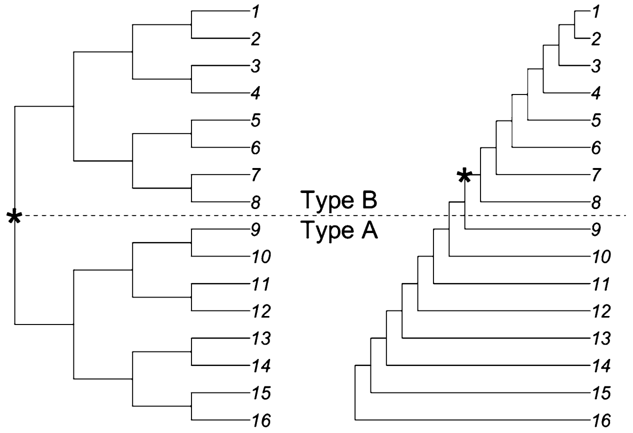

Fig. 1 Symmetric and asymmetric tree topologies used in simulation studies 1, 2 and 3. In Simulations 2 and 3, shifts in selection pressure and codon bias occur at the point represented with an *asterisk* (*), which separates Type A and Type B evolution. Type A evolution differs from Type B due to a shift in the evolutionary process

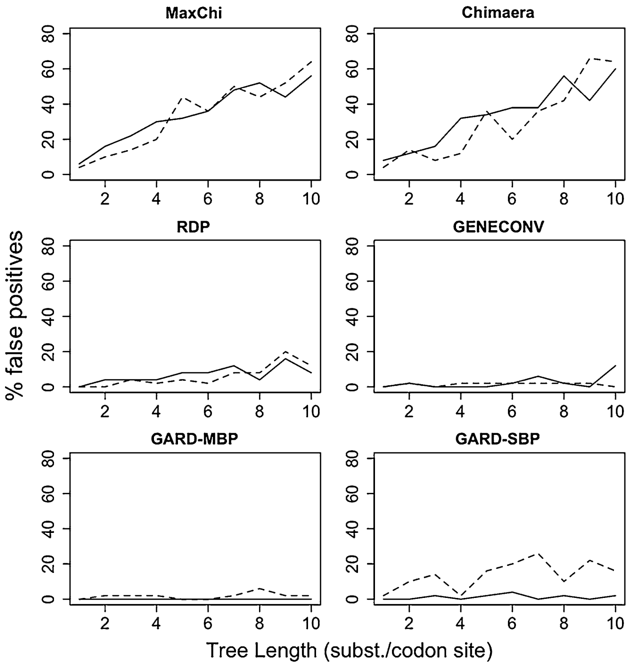

Fig. 2 Results from Simulation 1. *Solid lines* indicate the percentage of false positives when tree topology is symmetric while *dashed lines* indicate results under an asymmetric topology

note that the largest branch lengths for codon evolution in our study (symmetric tree: 0.33; asymmetric tree: 1.11) are still well below saturation, which is 3 substitutions per codon site. Interestingly, previous simulation studies found that power to detect recombination, when it truly exists, increases with sequence divergence (Posada and Crandall 2001; Kosakovsky Pond et al. 2006). We suggest that those findings are related to ours; specifically, when recombination is truly present, the increased variability provides additional signal for recombination, but when it is absent (as in this simulation study) the increased variability provides more opportunities for the detection of false signal.

In addition to divergence, tree topology has a strong influence on the false positive rate for GARD-SBP (Fig. 2). False positive rates for GARD-SBP are larger under the asymmetric topology as compared to the symmetric topology. This method performs well when tree topology is symmetric, with the false positive rate no larger than 4% at any of the simulated tree lengths. However, when topology is asymmetric, false positives increase with tree length, reaching a maximum of 26%. When the topology is asymmetric, the alignment contains a broader range of pairwise divergences between taxa. GARD-SBP relies on a complex model of evolution, measuring substitution parameters from the entire alignment and using them to compute the likelihood of the data under a tree estimated via neighbor-joining. $\mathrm{AIC_C}$ is then used to compare the explanatory power of a single tree and branch lengths to the case of separately estimated trees and branch lengths for either side of a breakpoint. It seems that $\mathrm{AIC_C}$ might not be a good approximation to the Kullback–Leibler divergence in this setting, perhaps due to non-independence among branch length estimates.

The remaining methods, GENECONV and GARD-MBP, yield consistently lower levels of false positives as compared with the other methods we evaluated (Fig. 2). GARD-MBP performs exceptionally well under the symmetric topology, with no false positives. However, there is a very small effect due to asymmetry, with an average (but acceptable) false positive rate of 4% for the asymmetric trees. The average false positive rate for GENECONV is 2.8% for the asymmetric trees and 4.8% for the symmetric trees. The average over symmetric trees is affected heavily by the data simulated under a tree length of 10 (12% false positives); without this condition the average false positive rate drops to 2%. These results are comparable to those from previous work (Posada and Crandall 2001), which also found GENECONV to have low rates of false positives.

To determine the robustness of these results to the data sampling strategy, we conduct additional simulations varying both sequence length (200–600 codons) and number of taxa (16 and 32 taxa). Sequence length affects the number of false positives inferred by GARD-SBP, RDP, MaxChi, and Chimaera (Supplementary Table S2), with number of false positives increasing with sequence length. At all sequence lengths simulated, the negative impact of tree asymmetry is still observed for GARD-SBP. Alternatively, the impact of tree length seems to be diminished by increasing sequence length. However, this is not due to improved performance when divergence is large; rather the equalizing effect among tree lengths is due to false positives increasing with sequence length. Increasing number of taxa decreases the number false positives detected by GARD-SBP, MaxChi, and Chimaera (Supplementary Table S3). An important finding is that denser sampling of taxa from a given level of sequence divergence helps to diminish the negative impact of increased tree length. These results indicate that factors related to sampling strategies can also have a very strong effect on performance.

Taken together, our findings indicate that both tree shape and sequence divergence can negatively affect recombination detection methods. Three methods, MaxChi, Chimaera, and RDP, yield increased false positive rates with increased tree length, possibly because more variability leads to more opportunity for false signal. GARD-SBP, on the other hand, can be strongly affected by tree topology for average-sized genes (200 codons), with much higher levels of false positives under an asymmetric phylogeny. Two methods (GARD-MBP and GENECONV) are recommended under conditions of asymmetric topology or high sequence divergence, as they are more robust to differing tree shapes than GARD-SBP and are more robust to large divergences than MaxChi and Chimaera.

Simulation 2: Non-Stationary Evolution

Recombination detection methods often assume that every aspect of the evolutionary process is homogeneous over time, with average values for parameters such as nucleotide composition and substitution rates assumed to apply equally well to all parts of a phylogeny. However, this assumption is often violated in real data. For example, sequences that have experienced a divergence in gene function often possess sites that exhibit a shift in the substitution rate across the phylogeny (Gaucher et al. 2002). In some cases, divergence of function can be associated with a substantial shift in other aspects of the substitution process such as equilibrium frequencies or amino acid exchangeabilities (e.g., Kettler et al. 2007; Zhang et al. 1998). Simulation 2 is designed to explore the impact of a shift in the substitution process on recombination detection methods. Specifically, we simulate a range of shifts in (i) codon usage bias and (ii) the distribution of selective effects, and compute false positive rates for the set of six detection methods.

As in Simulation 1, datasets are 200 codons in length and generated using a 16-taxon phylogeny that is either symmetric or asymmetric (Fig. 1). Data are simulated to yield a total tree length of 4 substitutions per codon site (average number of substitutions per branch is 0.1333). When a shift in either codon bias or selection pressure is simulated, it occurs at the node shown in Fig. 1 indicated by the asterisk. This node effectively splits the branches of the tree into two types, "Type A" and "Type B," having different evolutionary models.

Codon bias is modeled using the method of Aris-Brosou and Bielawski (2006). This method employs a single parameter, "η," to specify codon frequencies for changing proportions of G and C at the third codon position, or GC3 content (Supplementary Fig. S2). Codons with a G or C in the third position are given a frequency of η/Σ, while codons with an A or T at the third position have a frequency of $(1 - \eta)/\Sigma$, where Σ is a scaling factor. The values of η range from $0 \leq \eta \leq 1$, where a value of $\eta = 0.5$ indicates a GC3 content = 50% and all non-stop codons have equal frequencies. As η approaches zero, GC3 content decreases. Using this system, we can easily calculate and specify separate codon biases for different parts of the phylogeny; e.g., $\eta_A \neq \eta_B$.

The strength and direction of selection pressure varies among sites, and is determined by specifying a distribution for ω separately for each part (A and B) of the phylogeny (e.g., Fig. 3; also see Supplementary Fig. S3). Type A evolution remains constant throughout all simulations while Type B evolution changes among six different scenarios. For the Type A evolutionary model, most codon sites are under strong purifying selection ($\omega < <1$) and very few sites are evolving close to neutrality ($p_A = 0.5$, $q_A = 2$); in general, this beta distribution can be thought of as having an "L-shape" (e.g., Fig. 3). The codon bias parameter for Type A evolution is set to $\eta_A = 0.5$. Type B evolution (p_B, q_B and η_B) varies for each simulation condition. There are two basic cases concerning selection pressure for Type B evolution:

Case 1: Most sites under strong purifying selection, similar to Type A evolution. This is modeled by an L-shaped beta function (Supplementary Fig. S3).

Case 2: A larger proportion of sites evolving close to neutrality. This is modeled with a U-shaped beta function (Fig. 3).

For each case above, we simulate 3 different ω distributions (designated sub-cases "a," "b," and "c"). Case 1a is the "null" for selection pressure. Under the null, Type B evolution is simulated under the same conditions as Type A evolution, so selection pressure is homogeneous across the phylogeny. In all other cases (Cases 1b, 1c, and 2a–c) there is a shift in ω distribution (parameters p and q of the beta distribution differ between Type A and B evolution). In addition, each case for selection pressure is simulated both with a shift in codon bias ($\eta_A = 0.5$ and $\eta_B = 0.1$) and with no shift in codon bias across the phylogeny ($\eta_B = 0.5$). The result is a total of 12 unique evolutionary scenarios: six "sub-cases" for the shift in selection pressure, each with two conditions for a shift in codon usage (condition 1: $\eta_A = \eta_B = 0.5$; and condition 2: $\eta_A = 0.5$, $\eta_B = 0.1$).

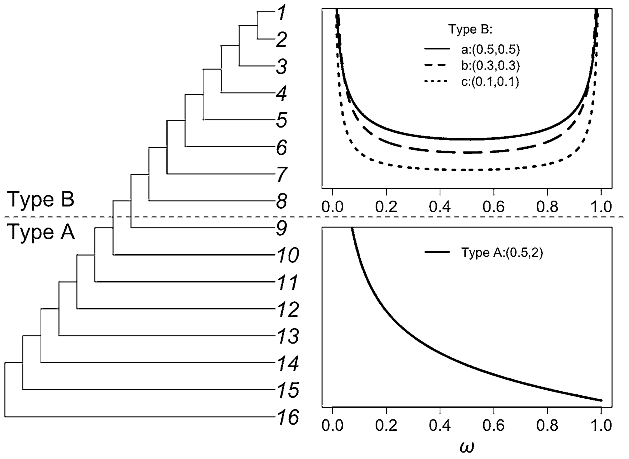

Fig. 3 Example distribution of selection pressure employed in Simulation 2, Case 2. Only the asymmetric tree topology is presented. Shape parameters for beta distribution (*p, q*) for sub-cases a, b and c are given in legends of the beta distribution plot. The distributions for Simulation 2, Case 1 can be found in the Supplementary Materials

Under this design, we generate four different scenarios under the "null" case for selection pressure; i.e., Case 1a is applied to both topologies (symmetric and asymmetric) and both codon usage conditions. For each of these four scenarios, 250 replicate datasets are simulated. Table 2 shows the results for all four Case 1a scenarios, and highlights our finding that, regardless of tree shape, a shift in codon bias alone does not substantially affect the false positive rates.

While a shift in codon bias alone does not negatively impact tests for recombination, it could be the case that those same tests might incorrectly interpret rate shifts at sites within a gene (i.e., arising as a function of a shift in the intensity of selection pressure) as signal for recombination. The tests are surprisingly robust to a shift in selection pressure, and even the combined effect of shifts in both selection pressure and codon bias. In Tables 3 and 4 we generalize the false positive rate over the different scenarios for divergent selection pressure; i.e., we compute the false positive rate for each recombination method as an average over Cases 1b, 1c, and 2a–c (for full results see Supplementary Tables S4 and S5). Table 3 provides results for just a shift in selection pressure and Table 4 provides results for the combined effect. Although false positives for GARD-SBP are again noticeably higher for the asymmetric tree topology, there is no affect for shifts in either codon bias or selection pressure. We also conduct additional simulations to test whether these results are robust at different levels of sequence divergence (see Supplementary

Table 2 Percentage of false positives under homogeneous selection pressure (Case 1a) in Simulation 2 for both symmetric and asymmetric tree topologies with both stationary and non-stationary codon bias

	GARD-MBP	GARD-SBP	RDP	GENECONV	MaxChi	Chimaera
Symmetric						
Stationary	0	3.2	8.8	3.2	28	28.8
Non-stationary	0	2.0	8.8	2.0	33.2	34.4
Asymmetric						
Stationary	1.2	13.2	4.8	2.0	30.8	19.6
Non-stationary	2.8	14.4	5.6	1.6	29.6	23.2

Percentages in each cell are computed from 250 alignments. Stationary codon bias is generated under $\eta_A = \eta_B = 0.5$. Non-stationary codon bias is generated under $\eta_A = 0.5$ and $\eta_B = 0.1$

Table 3 Percentage of false positives under a shift in selection pressure in Simulation 2

	GARD-MBP	GARD-SBP	RDP	GENECONV	MaxChi	Chimaera
Symmetric						
Null (Case 1a)	0	3.2	8.8	3.2	28	28.8
Shift in selection pressure	0	4.0	7.0	2.0	35	30
Asymmetric						
Null (Case 1a)	1.2	13.2	4.8	2.0	30.8	19.6
Shift in selection pressure	2.0	11	1.0	1.0	32	25

Percentages in each cell are computed from 250 alignments. False positive rates are combined across different models for the shift in selection pressure (Cases 1b, 1c, 2a–c)

Table 4 Combined effect of a shift in selection pressure and codon bias on false positives for recombination

	GARD-MBP	GARD-SBP	RDP	GENECONV	MaxChi	Chimaera
Symmetric						
Null (Case 1a)	0	4	7	2	35	30
Shift in selection pressure	0	2	8.8	2	33.2	34.4
Asymmetric						
Null (Case 1a)	2	11	1	1	32	25
Shift in selection pressure	2.8	14.4	5.6	1.6	29.6	23.2

In all cases codon frequencies differ between Type A and Type B evolution ($\eta_A = 0.5$, $\eta_B = 0.1$). Percentages in each cell are computed from 250 alignments. False positive rates are combined across different models for the shift in selection pressure (Cases 1b, 1c, 2a–c)

Tables S6 and S7). The effect is the same as found in Simulation 1; i.e., false positive rates for some methods increase with tree length, but are not negatively impacted by non-stationary selection pressure or codon bias. These data confirm that the major effects on false positive rate are due to the method (e.g., MaxChi and Chimera consistently exhibit the highest false positive rates), sequence divergence, and, to a lesser extent, tree shape.

Simulation 3: Positive Selection

Genetic variability generated by recombination may resemble patterns of molecular adaptation because nonsynonymous substitution rates of recombinant gene fragments appear to differ from the rest of the alignment. Analysis under the standard models of codon evolution can lead to artificially elevated rates of nonsynonymous substitution for the recombinant regions of an alignment, which will be resolved as a fraction of sites having $\omega > 1$ (Anisimova et al. 2003; Shriner et al. 2003; Scheffler et al. 2006). In this simulation study we investigate if the converse might also be true; that is, if the presence of sites in a dataset subject to positive selection in part of a phylogeny might likewise impact statistical tests for recombination. Simulation 3 differs from Simulation 2 by generating an alignment with a fraction of sites having $\omega > 1$ whereas all sites in Simulation 2 had $\omega < 1$.

Simulation 3 is comprised of three “null” scenarios and three positive selection scenarios. The null scenarios have a shift in the intensity of purifying selection pressure; i.e., the shift does not involve positive selection (as in Simulation 2). Sequences are simulated under the same two phylogenies as the previous simulations (symmetric and asymmetric) and each alignment is 200 codons in length. The shift in selection pressure occurs at the same node as in Simulation 2, again splitting the branches of the tree into two types, “Type A” and “Type B,” having different evolutionary models. The ω distributions are specified for each part of the phylogeny using the same models used in Simulation 2. The three positive selection scenarios extend the null scenarios by adding a single discrete category (10% of sites) having $\omega = 2$ in Type B branches.

As in Simulation 2, Type A evolution remains constant throughout all simulations and has most codon sites evolving under strong purifying selection ($p = 0.5$, $q = 2$). Type B evolution, on the other hand, changes for each of six scenarios (Fig. 4). The shift in evolution is determined by a change in shape of the beta distribution, and the presence of sites subject to positive selection in three of the scenarios. The beta distribution for Type B branches is U-shaped, with a large proportion of sites evolving nearly neutrally. We simulate three different sets of shape parameters (hereafter referred to as a, b, and c) for the ω distribution. Adding 10% of sites to the positive selection ($\omega = 2$) scenarios requires that the amount of sites that follow the beta distribution is reduced by 10%.

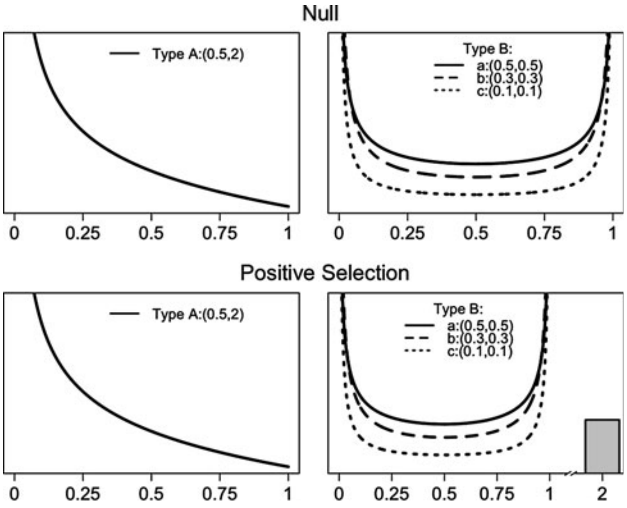

Fig. 4 Omega distributions for Simulation 3. Parameters used for beta distributions (*p, q*) are shown in legends. Each sub-case is simulated with both stationary and non-stationary codon bias

To summarize our findings (Table 5) we combine results for the different tree shapes (symmetric and asymmetric) and beta functions (a, b and c) and compare false positive rates between the positive selection and null scenarios (see Supplementary Table S8 for full results). This comparison reveals that, in general, recombination detection methods are not sensitive to the presence of positively selected sites. While all methods except Chimaera yield some increase in false positives when positive selection is present, this increase does not exceed 4% for any method. False positive rates for MaxChi and Chimera are substantially higher than all other methods in Simulation 3. This finding is consistent with Simulation 1, which shows increased false positives for these methods at this level of sequence divergence (Fig. 2).

Additional simulations are presented in Supplementary Table S9 that cover the same scenarios, but under shallower (tree length: 2 subst./codon site) and deeper divergences (tree length: 10 subst./codon site). Those results indicate that the recombination detection methods are robust to positive selection over a wide range of sequence divergence. Supplementary Table S10 presents a simulation study where 10% of sites are subject to extremely intense positive selection pressure ($\omega = 10$). Here, we also find that the recombination detection methods are robust to positive selection in a portion of the tree.

Lastly, we investigate the impact of positive selection when it is acting throughout the entire phylogeny. Although this mode of evolution is not necessarily connected with the process of functional divergence, it does represent a very strong signal that could be mistaken for recombination.

Table 5 Effect of positive selection on the percentage of false positives for recombination in Simulation 3

	GARD-MBP	GARD-SBP	RDP	GENECONV	MaxChi	Chimaera
No positive selection	0.7	6.7	7.3	1.0	33.3	28.3
Positive selection	1.0	8.7	8.0	3.0	37.0	24.3

Percentages in each cell are computed from 300 alignments. Results are combined across symmetric and asymmetric tree topologies and across sub-cases for selection pressure (beta functions denoted as a, b and c)

To test this hypothesis, we perform a small simulation study in which 10% of sites are subject to positive selection in each branch of the phylogeny and compare results to a "null" scenario having 0% of such sites (see Supplementary Materials for detailed methods and results). Again, we find that none of the recombination detection methods tested here are sensitive to mode of evolution; i.e., false positive rates are similar for data generated under the null and positive selection scenarios. Note that our simulation scenarios with 10% of sites having $\omega > 1$, although realistic, represents the upper range of empirically derived estimates (e.g., Yang et al. 2005). Evolution by positive selection in many to all the branches in the tree is most consistent with diversifying selection, which leaves the strongest signal within real datasets (Bielawski and Yang 2005). Collectively, these results suggest that recombination detection methods are largely robust to the presence of sites subject to this type of positive selection.

Although previous studies have shown that recombination events can lead to false inferences in positive selection analysis (Anisimova et al. 2003; Shriner et al. 2003; Scheffler et al. 2006), our results show that the converse is not true. One possible explanation is that positively selected sites, while often localized in 3D space of the folded protein product of the gene, are typically dispersed along the primary gene sequence. Recall that recombination detection methods search for local variability in adjacent gene fragments; i.e., they search for spatial organization along the gene sequence. However, this pattern is unlikely to result from selection acting on the mature and folded protein product. We expect that sequences generated in an unrealistic way, with continuous blocks of sites subject to positive selection, would have impacted the false positive rates for recombination. As the purpose of the tested recombination detection methods is to identify heterogeneity associated with phylogenetic variability among sites, they appear to be performing as intended in so far as they are not yielding false positives for recombination when evolution by positive selection is distributed in a realistic way.

Simulation 4: Power to Detect Recombination Under Different Levels of Diversity

Because recombination analysis is a key step in phylogeny-based inference, detection methods must also be reliable when recombination has truly impacted the evolution of a set of gene sequences. Methods should be able to both (i) detect whether or not recombination is present, and (ii) estimate the number and location of breakpoints. Several simulation studies have investigated the power of detection methods under different levels of recombination and divergence (Posada and Crandall 2001; Wiuf et al. 2001; Chan et al. 2006; Kosakovsky Pond et al. 2006). However, except for GARD-MBP (Kosakovsky Pond et al. 2006), none of the methods have been evaluated for their ability to determine the correct number of breakpoints when multiple events have occurred. In Simulation 4, we analyze both their power to qualitatively detect the presence of recombination and their accuracy to correctly infer number of breakpoints.

For this simulation, we use datasets from a previous study that, to date, have been analyzed only with GARD-MBP (Kosakovsky Pond et al. 2006). These simulated datasets consist of 8-taxon alignments with different levels of recombination and diversity. Each alignment is 3,000 bp long and has 0, 1, 2, 4, or 8 recombination breakpoints. In addition, for each number of breakpoints, there are datasets with both low (5%) and high (25%) genetic diversity for a total of 10 simulation conditions, each with 100 replicate datasets.

Consistent with previous simulation studies (e.g., Posada and Crandall 2001), our analysis shows that recombination detection methods are not powerful. Power for detecting a single recombination event is low for all methods (Table 6). When just one recombination event is simulated and diversity is low, RDP, GENECONV, MaxChi, and Chimaera have similar performance, detecting just 12–19% of replicates as having been subject to recombination. GARD-MBP has substantially lower power, only detecting recombination in 8% of replicates. This is much lower than previously reported for the same set of simulations (Kosakovsky Pond et al. 2006: 56%) because we chose to apply the KH test for phylogenetic incongruence, which has the desirable effect of controlling the number of false positives (see Simulations 1–3).

For all methods, increasing diversity increases power to detect recombination (Table 6). For example, in sequences with one simulated breakpoint, RDP detects recombination in 38% of alignments with high diversity, compared to just 16% of low diversity replicates. This general pattern of

Table 6 Capacity of five recombination detection methods to correctly infer 0–8 breakpoints

	Low diversity (5%)					High diversity (25%)				
	0	1	2	4	8	0	1	2	4	8
RDP										
0	99	83	73	49	32	98	62	53	19	11
1	1	16	26	39	35	2	37	38	44	30
2		1	1	11	25		1	8	25	30
3				1	3			1	6	16
4					5				4	7
5+									2	6
T	1	17	27	51	68	2	38	47	81	89
GENECONV										
0	99	88	75	59	33	98	76	68	41	29
1	1	12	25	32	47	2	23	27	43	43
2				8	15		1	5	14	17
3				1	2				2	8
4					2					3
5+										
T	1	12	25	41	67	2	24	32	59	71
MaxChi										
0	97	81	60	37	25	94	69	48	18	10
1	3	19	36	42	38	6	28	40	40	28
2			4	19	27		3	11	25	29
3				2	6			1	14	20
4					4				2	7
5+									2	6
T	3	19	40	63	75	6	31	52	83	90
Chimaera										
0	97	84	61	41	26	94	69	48	18	11
1	3	16	36	43	37	6	29	39	44	25
2			3	14	26		2	12	25	31
3				2	6			1	10	19
4					5				3	9
5+										5
T	3	16	39	59	74	6	31	52	82	89
GARD-MBP										
0	99	92	72	74	51	94	70	70	57	47
1	1	8	28	24	43	6	30	25	38	34
2				2	6			5	4	16
3									1	3
4										
5+										
T	1	8	28	26	49	6	30	30	43	53

For each method evaluated, column labels indicate the number of simulated breakpoints while row labels indicate the number of breakpoints inferred by a given method. Values within the table indicate the number of replicates (from $n = 100$) for which the number of breakpoints specified by the row label are inferred. Blank cells indicate that no break points are inferred in any of the replicates. The rows marked "T" indicate the total number of replicates with a signal for recombination

increasing recombination detection at high diversity is present throughout all levels of recombination. In addition, when multiple breakpoints are present, a larger number of replicates have a signal for more than one breakpoint when diversity is high. These results are consistent with previous findings, which suggest the increased information available when genetic diversity is high leads to increased power for detection (Posada and Crandall 2001; Kosakovsky Pond et al. 2006).

For both high and low diversity, as the number of simulated recombination events increases, so does the number of replicates in which recombination can be detected (Table 6). When eight breakpoints are simulated at low diversity, MaxChi and Chimaera detect recombination in 74–75% of replicates. RDP and GENECONV have slightly lower power, detecting recombination in 67–68% of replicates. These results are consistent with previous studies (Posada and Crandall 2001; Kosakovsky Pond et al. 2006), which show that recombination detection methods have higher power when levels of recombination are higher.

Taken over all levels of recombination, GARD-MBP has much lower power, detecting recombination in less than half of replicates even when multiple breakpoints are simulated. Kosakovsky Pond and colleagues (2006) endorse using GARD-MBP without requiring phylogenetic incongruence on either side of a breakpoint, as this yields very good power (recombination is detected in as many as 98% of replicates when both recombination and diversity levels are high). However, it also increases the number of false positives. Without the KH test, Kosokovsky Pond and colleagues (2006) observed a false positive rate of 10%, whereas it falls to 1% (Table 6) in our simulations when the KH test is applied.

Although most methods have increased ability to qualitatively assess recombination with increasing number of breakpoints, accuracy in inferring the correct number of breakpoints when multiple are present is extremely low (<4% for low diversity and <12% for high diversity). In fact, even when eight breakpoints are simulated, no method infers greater than five breakpoints when diversity is low. As we apply it, GARD-MBP is particularly conservative, never inferring more than two breakpoints at low diversity. In general, all methods underestimate the number of recombination events. This decreasing ability to correctly identify multiple breakpoints may be due to decreased amount of information from which to make inferences. Small fragments simply do not provide enough information for accurate estimation of some parameters, including substitution parameters and phylogenies. Because recombination detection methods require parameter estimates to be compared with those in adjacent fragments, users of these methods should be aware that more false negatives may result with decreasing fragment sizes in real data.

Real Data Analysis: Testing the *Prochlorococcus* Core-Genome for Recombination

The cyanobacteria *Prochlorococcus marinus* are responsible for a significant fraction of primary productivity in the world's open oceans (e.g., Suzuki et al. 1995; Liu et al. 1997). Specific lineages have adapted to high-light and low-light conditions, and functional divergence among the genes of those lineages is the subject of considerable research interest (e.g., Ting et al. 2001; Zhao and Qin 2007). As methods for measuring the strength and direction of natural selection pressure in such sequences (e.g., codon models) can be negatively impacted by recombination (Anisimova et al. 2003; Shriner et al. 2003; Scheffler et al. 2006), the capacity to reliably detect recombination is critical to such a research program. Indeed, homologous recombination is well known to have impacted the evolution some of their photosystem genes (e.g., Zeidner et al. 2005; Sullivan et al. 2006). While several studies have explored the role of LGT in the genomic evolution of *Prochlorococcus* (e.g., Zhaxybayeva et al. 2006, 2009), none have attempted to quantify within-gene recombination. Here, we use the knowledge acquired from our simulation studies to test and interpret a set of 585 genes from the core genome of *Prochlorococcus* for recombination.

Each gene is analyzed for within-gene recombination by using all six recombination detection methods evaluated in our simulation studies. The number of genes in which recombination is detected is extremely variable, from just nine genes (1.5% of those analyzed) using GARD-MBP to 534 genes (91.3%) using Chimaera. Interestingly, GARD-based methods infer among the highest and lowest numbers of genes having a history of recombination. The extremely conservative performance of GARD-MBP (nine genes) is consistent with results from our simulation studies, where this method had low power even when multiple breakpoints were present. However, the GARD-SBP method detects recombination in 83.2% of genes, with only Chimaera detecting higher levels. The high rates detected by MaxChi (476 genes: 81.4%) and Chimaera may be due, in part to the level of divergence observed in *Prochlorococcus* genes; the genome tree is ~10 subst./codon site (average branch length ~0.46 subst./codon site), corresponding to the upper end of sequence divergence covered in Simulation 1. In addition, the real gene sequences have an asymmetric tree topology, and given the high false positive rate of GARD-SBP observed under such a tree in Simulation 1, those real-data results are best interpreted as negatively impacted by false positives.

RDP and GENECONV fall between these two extremes and, based on our simulations, they are expected to more reliably identify actual cases of recombination in *Prochlorococcus*. In our simulations, these two methods show reasonable power without yielding a large number of false positives when branch length is increased. In addition, RDP and GENECONV perform well under an asymmetric topology and non-stationary evolution, which are characteristic of these real *Prochlorococcus* gene sequences.

However, the difference between levels of recombination detected with these two methods is still quite large; RDP detects recombination in 209 genes while GENECONV detects recombination in only 50. With such vastly different estimates, it is difficult to determine the actual level of recombination present in the *Prochlorococcus* core genome, but the range (8.5% under GENECONV to and 37.5% under RDP), suggests an important role for within-gene recombination in the evolution of the *Prochlorococcus* core genome.

Prochlorococcus genes exhibit all of the complex evolutionary scenarios simulated in this paper: the topology is largely asymmetric, a shift in codon frequencies occurs within the phylogeny, and genes have undergone a divergence in function. Although our simulations show that false positive rates for many methods tested should be robust to some or all of these conditions, the number of genes predicted to have undergone recombination in this complex setting is still highly variable among detection methods. It may be the case that these evolutionary scenarios impact other aspects of test performance, such as power. This lack of concordance among tests leads to a high amount of uncertainty, not only in measuring the amount of recombination present in the genome, but also in analyzing specific genes for the presence of recombination.

Conclusions

Of the scenarios examined, we find tree length, tree shape, and data sampling strategy to have the most substantial impact on false positive rate. In particular, we find that GARD-SBP can yield an increased number of false positives under an asymmetric topology. For other methods (RDP, MaxChi, and Chimaera), the false positive rate depends more on tree length. The magnitude of these effects depends on the sample of data (number of sequences and their length). Based on these findings we suggest that RDP, MaxChi, and Chimaera should not be used when sequence divergence is too large, and we cannot endorse GARD-SBP due to its sensitivity to tree topology. Somewhat surprisingly, most methods are robust to shifts in codon bias and the distribution of selection pressure. In addition, none of the methods yield a major increase in number of false positives when sites are evolving under positive selection in some, or even all, of the branches of the tree. Two methods (GENECONV and GARD-MBP) stand out for being highly robust to false positives over all of the evolutionary scenarios that we tested.

The power of all methods tested is sensitive to the level of divergence and the amount of recombination. Although power is typically low for all recombination detection methods, the ability to qualitatively detect recombination increases with both diversity and number of breakpoints. However, even when diversity is high, accuracy to determine the number of recombination breakpoints in an alignment is low. It seems that for all methods, there is a tradeoff between power to detect recombination and false positive rate. While the more powerful methods (MaxChi, Chimaera) are more likely to detect the presence of recombination when it truly exists, they are also more likely to detect a false signal for recombination when it is absent. On the other hand, the more conservative methods (GARD-MBP, GENECONV, RDP) have low false positive rates, but are also likely to miss recombination when it is present. Future studies may examine different scenarios related to recombination (e.g., stronger signal or location of breakpoints) in order to further investigate factors that might impact the power of these methods to determine both number and location of recombination breakpoints.

Robustness to shifts in the distribution of selective constraints, and to the presence of positively selected sites, is particularly important because model-based methods that test for such features of molecular evolution are not robust to the presence of recombination; they assume that all sites share a common phylogeny and some have been shown to yield false results in the presence of recombination (Anisimova et al. 2003; Shriner et al. 2003; Scheffler et al. 2006). Therefore, studies of functional divergence in systems where recombination is plausible can, and should, include a pre-analysis for recombination. In this setting, the user might be less concerned with recombination as a process and more concerned with knowledge that will be used to guide subsequent analysis of functional divergence.

When recombination is detected preceding analysis of functional divergence, downstream analysis can proceed either by (i) excluding genes having recombinant sequences from further analysis (e.g., Sullivan et al. 2006) or (ii) using breakpoints to separately analyze sequence fragments (e.g., Scheffler et al. 2006). If recombinant sequences are to be excluded, it is important to choose methods with low false negative rates so that actual recombinant sequences will not cause further error in downstream analysis. In this setting power might be valued as much as, or more than, controlling the false positive rate. However we find that some methods, under some conditions, can have false positive rates that are too high (e.g., >60%), which would unnecessarily deplete the data selected for downstream analyses. Methods should not be selected based solely on the false negative rate. The alternative is to identify breakpoints and run analyses on each fragment. In this case, the user would be most concerned with accurate identification of the number and location of breakpoints. While we did not investigate location accuracy, we did find that inferences about the number of breakpoints are typically very low. The impact of these types of errors on the downstream analysis of separate fragments is not well known and warrants further investigation. In either setting, a small-scale simulation under evolutionary scenarios derived from the data in hand (e.g., branch lengths, tree shape, sample size, rate distribution, etc.) could prove extremely informative about the best tradeoff between power and accuracy for the data set "in hand".

Although using a combination of methods may be the best way to obtain a better understanding of recombination present in a multi-gene dataset, one must be careful to choose methods that are not prone to especially large false positive rates under the evolutionary processes characteristic of the genomes from which the genes were sampled. For example, the fraction of genes for which recombination is detected in the *Prochlorococcus* genome ranges from 1.5 to 91.3%: fundamentally different conclusions would be derived from each extreme if taken alone. Moreover, the average does not seem to be a biologically defendable estimate and a consensus that includes a method with very low power could yield a substantial underestimate of the recombination rate.

For the time being, users of these methods are encouraged to apply the following guidelines:

- When characteristics of a given dataset may violate model assumptions, simulations should be carried out under conditions that are relevant to the data at hand. These simulations should evaluate both power and false positive rates of candidate methods.
- Even if simulations do not show that a given evolutionary scenario impacts false positive rates, users of these methods should be aware that unexamined factors may impact performance and inflate discordance (e.g., the *Prochlorococcus* data).
- Pay particular attention to tree shape, sequence divergence, and sample size, as these characteristics could have a large and negative impact on the performance of recombination detection methods.
- A consensus of several different methods may provide a robust way of detecting recombination within a gene, but methods for use in a consensus should be chosen carefully. For instance, taking into account a strict consensus across a set of methods that includes an extremely conservative one (i.e., very low power) will be unlikely to detect recombination in most cases where it truly exists.
- Overall, this simulation study finds GENECONV to be both reasonably powerful and robust. Therefore, it may be desirable to include GENECONV when evaluating a set of potential methods under other sets of conditions. When power is valued as much as, or more than, controlling the false positive rate, it may be desirable to employ GARD-MBP without the KH correction.

The problem of recombination detection is obviously a complex one and there is much room for improvement in developing these methods before we have a clear understanding of recombination as an evolutionary process.

Acknowledgments This research was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant awarded to JPB. The research utilized computer hardware funded by a grant from the Canadian Foundation for Innovation to JPB. JBP also acknowledges the support of the Centre for Genomics and Evolutionary Bioinformatics (CGEB) which is funded by the Tula Foundation. We thank Olga Zhaxybayeva for providing amino acid alignments for the *Prochlorococcus* genomic data. We thank Joseph Mingrone for assistance with running the GARD analyses. We thank Katherine A. Dunn for helpful discussions, and for valuable guidance and advice on the development of Perl programs and the automation of analyses of both simulated data and real genome-scale datasets. We thank two anonymous referees for their comments, and for several suggestions that substantially improved this paper.

References

Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164:1229–1236

Aris-Brosou S, Bielawski JP (2006) Large-scale analyses of synonymous substitution rates can be sensitive to assumptions about the process of mutation. Gene 378:58–64

Bielawski JP, Yang Z (2005) Maximum likelihood methods for detecting adaptive protein evolution. In: Nielsen R (ed) Statistical methods in molecular evolution. Springer-Verlag, New York

Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau ME, Nesbo CL, Case RJ, Doolittle WF (2003) Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet 37:283–328

Chan CX, Beiko RG, Ragan MA (2006) Detecting recombination in evolving nucleotide sequences. BMC Bioinform 7:412

Duffy S, Holmes EC (2008) Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus *Tomato yellow leaf curl virus*. J Virol 82:957–965

Fletcher W, Yang Z (2009) INDELible: a flexible simulator of biological sequence evolution. Mol Biol Evol 26:1879–1888

Gaucher EA, Gu X, Miyamoto MM, Benner SA (2002) Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci 27:315–321

Gu X (1999) Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16:1664–1674

Husmeier D, McGuire G (2003) Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. Mol Biol Evol 20: 315–337

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J et al (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genet 3:2515–2528

Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mol Evol 29:170–179

Knudsen B, Miyamoto MM (2001) A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc Natl Acad Sci USA 98:14512–14517

Koonin EV, Makarova KS, Aravind L (2001) Horizontal gene transfer in prokaryotes. Annu Rev Microbiol 55:709–742

Kosakovsky Pond SL, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. In: Nielsen R (ed) Statistical methods in molecular evolution. Springer, New York

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW (2006) Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 23:1891–1901

Liu H, Nolla HA, Campbell L (1997) *Prochlorococcus* growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean. Aquat Microb Ecol 12:39–47

Mann NH, Cook A, Millard A, Bailey S, Clokie M (2003) Marine ecosystems: bacterial photosynthesis genes in a virus. Nature 424:741–742

Martin DP (2009) Recombination detection and analysis using RDP3. Methods Mol Biol 537:185–205

Martin D, Rybicki E (2000) RDP: detection of recombination amongst aligned sequences. Bioinformatics 16:562–563

Maynard Smith J (1992) Analyzing the mosaic structure of genes. J Mol Evol 34:126–129

Narra HP, Ochman H (2006) Of what use is sex to bacteria? Curr Biol 16:R705–R710

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304

Pagán I, Firth C, Holmes EC (2010) Phylogenetic analysis reveals rapid evolutionary dynamics in the plant RNA virus genus *Tobamovirus*. J Mol Evol 71:298–307

Posada D (2002) Evaluation of methods for detecting recombination from DNA sequences: empirical data. Mol Biol Evol 19: 708–717

Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci USA 98:13757–13762

Ragan MA (2001) Detection of lateral gene transfer among microbial genomes. Curr Opin Genet Dev 11:620–626

Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6:526–538

Scheffler K, Martin DP, Seoighe C (2006) Robust inference of positive selection from recombining coding sequences. Bioinformatics 22:2493–2499

Shriner D, Nickle DC, Jensen MA, Mullins JI (2003) Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet Res 81:115–121

Suchard MA, Weiss RE, Dorman KS, Sinsheimer JS (2002) Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage. Syst Biol 51:715–728

Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, Chisholm SW (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol 4:1344–1357

Susko E, Inagaki Y, Field C, Holder ME, Roger AJ (2002) Testing for differences in rates-across-sites distributions in phylogenetic subtrees. Mol Biol Evol 19:1514–1523

Suzuki K, Handa N, Kiyosawa H, Ishizaka J (1995) Distribution of the prochlorophyte *Prochlorococcus* in the Central Pacific Ocean as measured by HPLC. Limnol Oceanogr 40:983–989

Ting CS, Rocap G, King J, Chisholm SW (2001) Phycobiliprotein genes of the marine photosynthetic prokaryote *Prochlorococcus*: Evidence for rapid evolution of genetic heterogeneity. Microbiology 147:3171–3182

Wiuf C, Christensen T, Hein J (2001) A simulation study of the reliability of recombination detection methods. Mol Biol Evol 18:1929–1939

Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107–1118

Zeidner G, Bielawski JP, Shmoish M, Scanlan DJ, Sabehi G, Béjà O (2005) Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. Environ Microbiol 7:1505–1513

Zhang J, Rosenburg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. PNAS 95: 3708–3713

Zhao F, Qin S (2007) Comparative molecular population genetics of phycoerythrin locus in *Prochlorococcus*. Genetica 129:291–299

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT (2006) Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res 16:1099–1108

Zhaxybayeva O, Doolittle WF, Papke RT, Gogarten JP (2009) Intertwined evolutionary histories of marine *Synechococcus* and *Prochlorococcus marinus*. Genome Biol Evol 2009:325–339