Gene Tree Affects Inference of Sites Under Selection by the Branch-Site Test of Positive Selection

Yoan Diekmann[1,2,*] and José B. Pereira-Leal[1]

[1]Instituto Gulbenkian de Ciência, Oeiras, Portugal. [2]Programa de Doutoramento em Biologia Computational (PDBC), Instituto Gulbenkian de Ciência, Oeiras, Portugal. *Present address: MACE Lab, Research Department of Genetics, Evolution and Environment, University College London, London, UK.

Supplementary Issue: Evolutionary Genomics

ABSTRACT: The branch-site test of positive selection is a standard approach to detect past episodic positive selection in a priori-specified branches of a gene phylogeny. Here, we ask if differences in the topology of the gene tree have any influence on the ability to infer positively selected sites. Using simulated sequences, we compare the results obtained for true and rearranged topologies. We find a strong relationship between "conflicting branch length," which occurs when the set of sequences that experiences selection for a given topology and foreground is changed, and the ability to predict positively selected sites. Moreover, by reanalyzing a previously published data set, we show that the choice of a gene tree also affects the results obtained for real-world sequences. This is the first study to demonstrate that tree topology has a clear effect on the inference of positive selection. We conclude that the choice of a gene tree is an important factor for the branch-site analysis of positive selection.

KEYWORDS: branch-site test, PAML, dN/dS, positive selection, molecular evolution

SUPPLEMENT: Evolutionary Genomics

CITATION: Diekmann and Pereira-Leal. Gene Tree Affects Inference of Sites Under Selection by the Branch-Site Test of Positive Selection. *Evolutionary Bioinformatics* 2015:11(S2) 11–17 doi: 10.4137/EBO.S30902.

TYPE: Original Research

RECEIVED: June 23, 2015. **RESUBMITTED:** October 27, 2015. **ACCEPTED FOR PUBLICATION:** November 01, 2015.

ACADEMIC EDITOR: Jike Cui, Associate Editor

PEER REVIEW: Seven peer reviewers contributed to the peer review report. Reviewers' reports totaled 1419 words, excluding any confidential comments to the academic editor.

FUNDING: This work was funded by the Fundação para a Ciência e Technologia (FCT) under grant PCDC/EBB-BIO/119006/2010. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: y.diekmann@ucl.ac.uk

Paper subject to independent expert blind peer review. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

The branch-site test of positive selection (BSPS)[1,2] is a standard approach to detect sites that evolve under episodic positive selection, ie, in a subset of branches in a phylogeny. It is based on a codon model of sequence evolution[3] with an explicit parameter ω defined as the nonsynonymous to synonymous substitution rate ratio (dN/dS), which is commonly interpreted as evidence for positive selection when greater than one. Given a multiple sequence alignment (MSA), a gene tree relating these sequences, and a partition of the branches into the so called "foreground" and "background," the parameters of two models are estimated by maximum likelihood (ML). The null model corresponds to the hypothesis that no sites evolve under positive selection in the foreground. It has three free parameters: two specifying proportions of site classes and an ω constrained to values below one (refer the study by Zhang et al.[2] for the detailed mathematical description). The alternative model has an additional free parameter ω constrained to values above one and corresponds to the alternative hypothesis that some sites evolve under positive selection in the foreground branches. Given the model likelihoods corresponding to the ML parameter estimates, a likelihood ratio test (LRT) determines if the alternative model fits the data significantly better than the null model.

If the null model is rejected, the actual sites most likely evolving under positive selection in the foreground branches can be determined in a second step by an Empirical Bayes procedure.[1,4] This is necessary as the previously estimated proportions of site classes and their corresponding ω values do not answer the question directly if any specific site belongs to a class with ω greater than one. In a first implementation, this was achieved by a naive empirical Bayes procedure.[1] Posterior probabilities for the site class of each site were calculated based on the ML estimates of the model parameters, which failed to account for sampling error in these estimates and was found to produce unreliable results on small data sets.[5] This motivated an improved Bayes empirical Bayes (BEB) procedure,[4] which is still the recommended approach to determine the actual sites under selection in the foreground.[6] BEB accounts for the uncertainty in the estimated ML parameters by defining uniform priors and numerically integrating over them. Thus, it is possible to consider not only the ML model parameter values but a whole range of values that are weighted by their likelihood. The output of BEB is a posterior probability for each site that it evolved under selection in the foreground, with probabilities above 0.95 generally considered significant.

The performance of the BSPS, usually defined by the type I and type II errors of the LRT, has been assessed mostly using simulations,[7–9] as in general the true selection history cannot be known. The most common approach is to generate sequences under different simulation parameters and compare the performance of the BSPS on each simulated data set. Examples of parameters that have been varied are sequence length, strength of positive selection, proportion of sites under positive selection,[8] indels and alignment errors,[7] synonymous substitution saturation, and variations in GC-content.[9] In contrast, to the best of our knowledge, the effect of the gene tree on BSPS performance has not been quantitatively evaluated. At least for site models, the gene tree appears not to be of great importance as long as it is "reasonably good,"[10] for example, inferred from the data by ML.

Here, we ask if and how the gene tree affects the performance of the BSPS. We define performance as the ability to retrieve the actual sites under positive selection by BEB and not as the errors committed by the LRT. Except for a short paragraph by Fletcher and Yang,[7] BEB performance has so far been measured only in the context of site models.[4,11–13] However, although "[i]dentifying amino acid residues under positive selection along the lineages of interest is clearly much more difficult than testing for the presence of such sites,"[1] the actual sites are often most useful to molecular biologists (see Yang[14] and the references therein). Therefore, our performance metric is relevant in practice and novel in the context of the BSPS.

Results and Discussion

We first measure the performance of BEB on simulated sequences given the true topology. This establishes the baseline against which the results on rearranged topologies can be compared.

We simulate sequences along the Ensembl Compara[15] species tree for a sample of mammals and chicken shown in Figure 1 in order to ensure a real-world tree topology. We generate eight independent replicas for 21 different foreground branches (each contiguous group of labeled branches in Figure 1, ie, {fg1},…, {fg6}, {fg1, fg2},…, {fg5, fg6},…, {fg1, fg2, fg3, fg4, fg5, fg6}), resulting in 168 sets of 32 sequences in total. Two batches of simulations are performed with different selection schemes in the foreground: a previously published scheme with 20% of the sites that average a dN/dS of three (scheme V in the studies by Zhang et al.[2] and Fletcher and Yang[7]) and a stronger one simulating 30% of the sites at a dN/dS of four (subsequently referred to as scheme W). We do not simulate insertions or deletions (indels), despite them having been shown to be of critical importance for the BSPS[7] as we do not want to confound the effect of tree topology alone. More details on the simulation procedure are given in the "Materials and methods" section. For each set of sequences, we infer the sites under positive selection in the foreground branch by BEB[4] using the program Phylogenetic Analysis by Maximum Likelihood (PAML).[6] All results in the main text are shown for site-specific posterior probability >0.95; however,

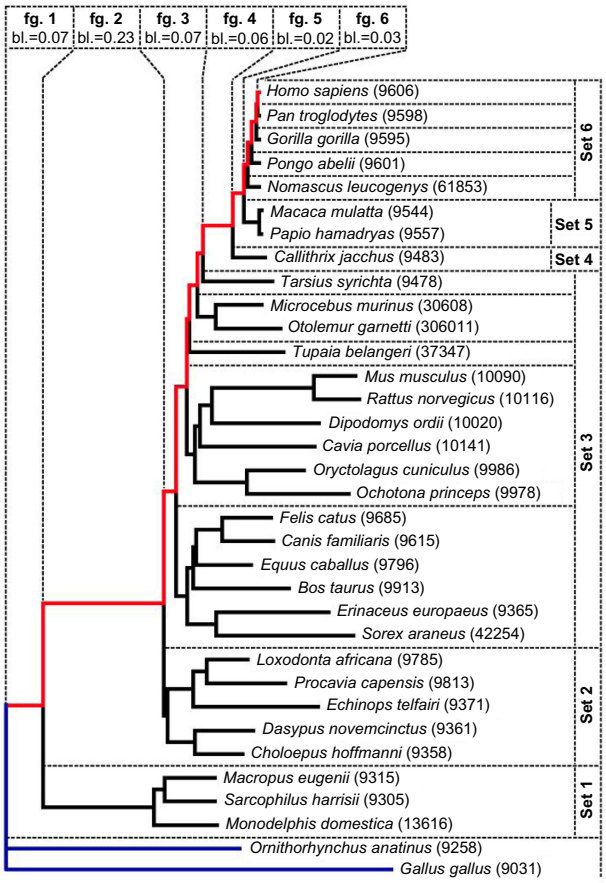

Figure 1. Gene tree underlying sequence simulations – the sequences are simulated along the tree using INDELible.[21] The tree is a subset of the Ensembl Compara[15] species tree; however, species names and NCBI taxon IDs are only for orientation as the simulations start from a random sequences. Every contiguous subset of branches labeled as foreground and highlighted in red, ie, {fg1},…, {fg6}, {fg1, fg2},…, {fg5, fg6},…, {fg1, fg2, fg3, fg4, fg5, fg6}, is simulated under foreground selection schemes described in the Materials and methods section. The remaining branches are simulated using the background scheme. Most basal branches serving as outgroups are shown in blue. Dashed lines are solely for clarity and labeling of sets of leaves for reference in the main text.
Abbreviations: fg, foreground; bl, branch length.

we obtain equivalent results for a more stringent threshold of >0.99 (corresponding plots given as part of Supplementary File 1). We compare them to the true sites and summarize the performance by computing sensitivity (defined as number of true positives [TP] divided by the sum of TP and false negatives [FN]) and specificity (defined as true negatives [TN] divided by the sum of TN and false positives [FP], see Materials and methods section for further information). By default, we average derivations of the confusion matrix over the eight replicates to mask the variation across all replicates.

The overall distribution of sensitivity and specificity (Supplementary Fig. 1) reveals that specificity is generally high, ie, very few sites are wrongly inferred to have evolved under selection in the foreground branches. Hence, in the following, we focus on sensitivity as a measure of BEB performance,

which is the fraction of all sites under selection that has been correctly found.

In all cases, foreground selection scheme V attains only very limited sensitivity, never exceeding 2%, which is consistent with a previous report of less than 1% for simulations including indels (page 2264 in Fletcher and Yang[7]). This low sensitivity makes it hard to systematically analyze the effect of gene tree topology. In the following, we therefore focus exclusively on foreground scheme W and include the corresponding figures for scheme V in Supplementary File 1.

In Figure 2, we represent the sensitivity for each of the 21 different foreground branches. Clear differences are apparent, with poor sensitivity for branches six and one (referring to the labels from Fig. 1) and higher sensitivity for foregrounds stretching (nearly) the entire path from root to leaf.

These performance differences can be explained in terms of properties of the foreground branch. For the power of the LRT, two aspects have previously been shown to be important: the foreground branch length and, to a lesser extent, its age, loosely formalized as the distance to the root.[7,9] Our test corroborates the major influence of the length of the foreground also on the sensitivity of the BEB procedure (simple linear regression $r^2 = 0.85$, $P < 1.95 \times 10^{-9}$). Moreover, adding the age of the foreground (see Materials and methods section for the definition used here) as a second explanatory variable leads to a better model fit (multiple linear regression $r^2 = 0.90$, $P < 5.26 \times 10^{-10}$, the resulting regression plane is shown in Fig. 3). Although the gain is modest, it is favored by model selection (Bayesian information criterion [BIC], ΔBIC = 6.1, see Materials and methods section for more details on BIC).

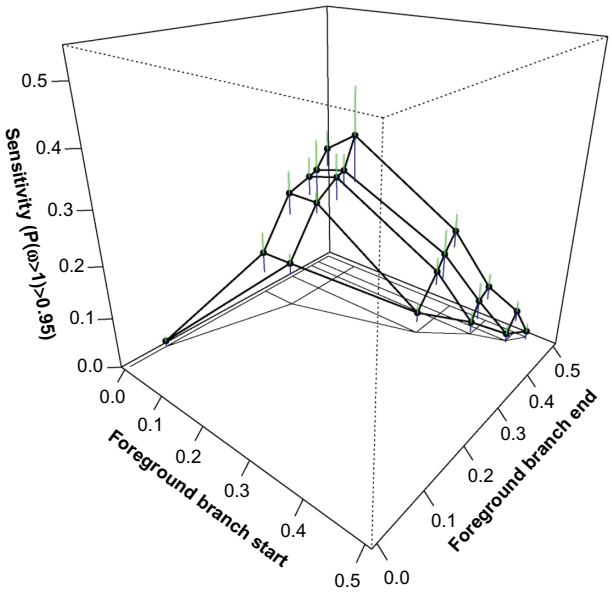

Figure 2. Sensitivity of the BEB procedure on the true gene tree – the mean sensitivity across the eight replicas is shown for every foreground branch, specified here by the pair of distances from the basal trifurcation (Fig. 1) to the start and end points of the branch. Green and blue lines are standard deviations.
Abbreviation: BEB, Bayes empirical Bayes.

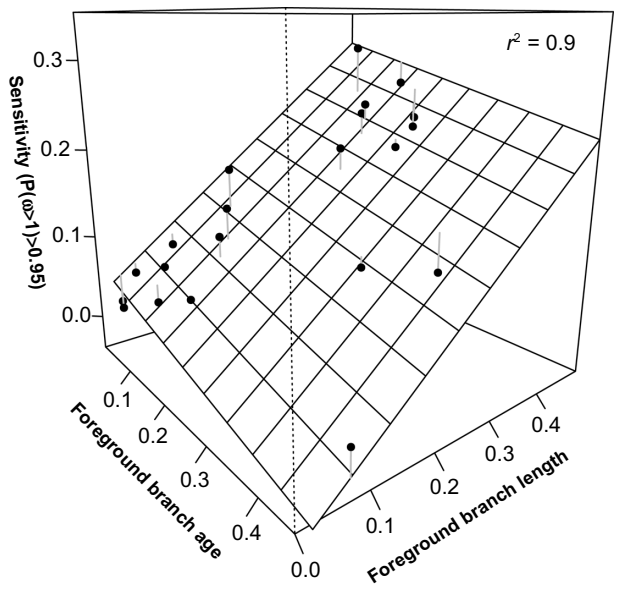

Figure 3. Sensitivity of the BEB procedure on the true gene tree – multiple linear regression of the sensitivity (averaged over the eight replicas) of BEB to infer the sites simulated under positive selection in the foreground in true tree. The explanatory variables are foreground branch length and age (see Materials and methods section for the definition of age employed here).
Abbreviation: BEB, Bayes empirical Bayes.

Hence, we show that the foreground branch length and, to a lesser extent, the age of the foreground are major factors for the performance of both LRT and BEB in the context of the BSPS. We quantify the effect and observe that these factors together account for roughly 90% of the variation in mean sensitivity in our simulations.

Next, we turn to our main question and ask if errors in the topology of the gene tree have any influence on the performance of BEB. Our approach is to introduce topological errors and pass these trees as input to PAML. We quantify the effect on the BEB procedure using the previously obtained sensitivities as a baseline for comparison.

We use 30 different trees given in Supplementary Table 1 and Supplementary File 2 that are generated by swapping or reattaching leaves. This approach of introducing topological errors ensures that the foreground branches remain unaffected and therefore permits us to compare the results of BEB across the 30 trees. Hence, the trees are not inferred from the simulated sequences, for example, by ML. While this may generate very improbable gene trees, the advantage of this approach is that it allows to freely manipulate the topology independently of its likelihood. We reason that any potential effect of tree topology is more easily understood including extremes. In a second step at the end of this section, we show that our findings also have practical relevance on real-world sequences.

Table 1. Results of the branch-site test of positive selection on data from Voordeckers et al.[16] obtained with original and reconciled gene tree – branch names refer to the labels given in Figure 6. H_1 and H_0 correspond to the branch-site model A with and without selection, respectively.

BRANCH	TREE	H_0	H_1	LRT	SIGNIFICANT AT
A	Original	–25025.9	–24994.3	63.2	1% (>5.41)
	reconciled	–25130.1	–25116	28.1	1% (>5.41)
B	Original	–25026.7	–25024.4	4.4	5% (>2.71)
	reconciled	–25147.1	–25144.9	4.3	5% (>2.71)
C	Original	–25023.7	–25005.3	36.9	1% (>5.41)
	reconciled	–25139.1	–25125.8	26.6	1% (>5.41)

First, we establish that rearranged trees do indeed have an effect. The distribution of effects on sensitivity (Supplementary Fig. 2) indicates that sensitivity can drop by more than 0.15 in the most extreme cases, which represents more than 50% of the greatest observed sensitivity (Fig. 2).

Next, we seek to explain this effect as a function of the tree and foreground. We hypothesize that the greatest impact occurs when the altered topology affects the length of time sequences evolve under selection. To quantify this phenomenon, Figure 4 defines "conflicting branch length" (CBL). CBL is introduced when the set of sequences that evolve under selection changes as a consequence of a changing topology. In that case, the difference in length of the branches where selection is acting in the true and in a topologically rearranged tree is defined as CBL. CBL depends both on the gene tree and the position and lengths of the foreground branch simulated to be under selection, meaning that the same tree can result in different CBL values.

We validate our hypothesis by demonstrating a linear relationship between the loss in sensitivity of the BEB procedure and CBL (simple linear regression $r^2 = 0.87$, $P < 2.2 \times 10^{-16}$; Fig. 5). The intercept represents the effect of all rearranged topologies that do not introduce CBL and clusters tightly around zero with very few trees showing differences in sensitivity above 0.02. We also observe a small (here below 0.02) yet significant increase of specificity with CBL (simple linear regression $r^2 = 0.45$, $P < 2.2 \times 10^{-16}$, data not shown), resulting in two opposite effects of CBL on the overall accuracy.

Model violations introduced by rearranged tree topologies can have a strong detrimental effect in the context of the BSPS. We define and single out one parameter – CBL – as an explanatory variable for a strong linear loss in sensitivity. This also implies that the results of BEB are robust against erroneous tree topologies as long as these do not introduce CBLs. Furthermore, it appears that the overall tree quality, which has previously been suggested to be important for site models, is only indirectly related to the loss of sensitivity observed here (see inlay in Fig. 5), as it has less explanatory power (simple linear regression $r^2 = 0.50$, $P < 2.2 \times 10^{-16}$).

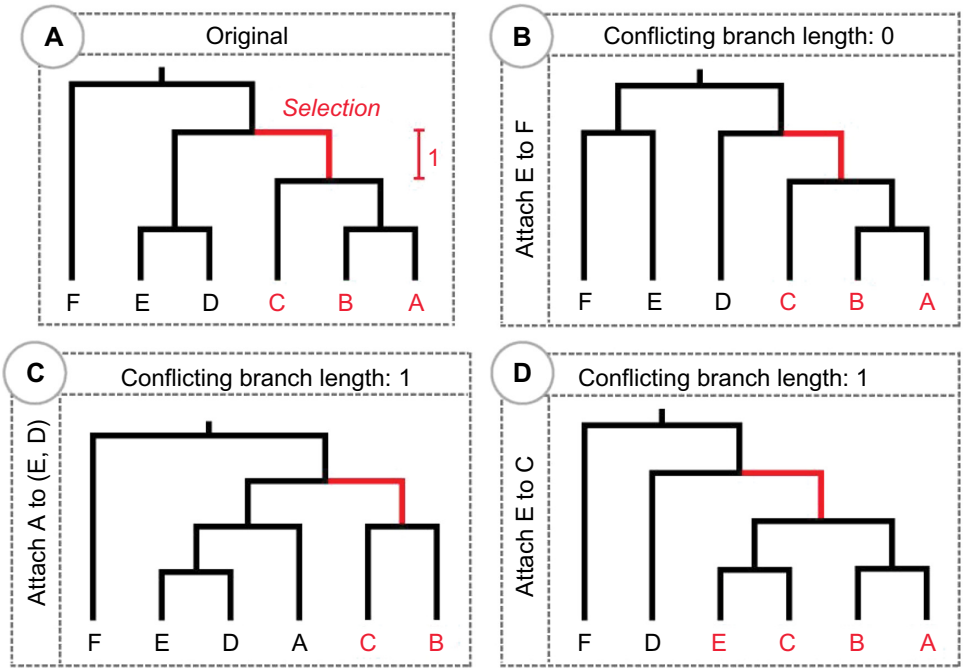

Figure 4. Definition of conflicting branch length – "conflicting branch length" occurs when the set of sequences that evolve under selection changes as a consequence of a changing topology. Below, the original set {**A**, **B**, **C**} (shown in red) from panel A is not altered by the tree in panel B; however, panel C reduces the set to {**B**, **C**}, and panel D increases it to {**A**, **B**, **C**, **E**}. The corresponding difference in branch length that a sequence evolves under positive selection is then defined as "conflicting branch length" (here both times 1).

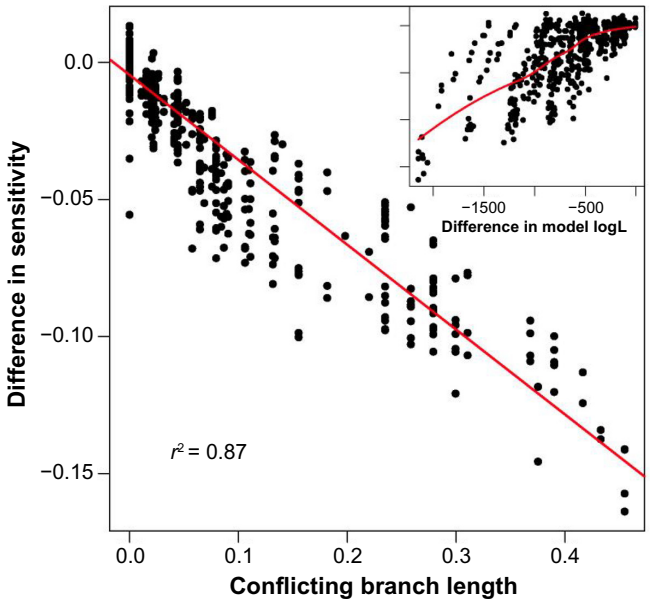

Figure 5. Conflicting branch length explains the loss in sensitivity of the BEB procedure – simple linear regression of change in sensitivity (averaged over eight replicas) over conflicting branch length. The inlay depicts the former quantity over the difference in log-likelihood of the codon model fitted by PAML[6] on the true versus the rearranged topology. The line is a nonparametric local regression (LOESS).
Abbreviation: BEB, Bayes empirical Bayes.

Finally, we ask if the choice of a gene tree also affects the results in real sequences. Real sequences usually evolve in more complex manners than is possible and desirable to simulate, in particular, compared to the rudimentary scheme without the indels used here. Hence, although we lose the certainty about the right tree topology and selective regimes, only these conditions can show if the effect we described remains detectable or if it is minor, and therefore, without a consequence for real sequences.

To answer this question, we reanalyze a data set of fungal glucosidase genes that have been studied by the authors with respect to their functional specialization after gene duplication.[16] We choose this data set as it exemplifies a situation in which alternative tree topologies commonly arise, namely, when the gene tree of orthologs and the species tree are incongruent or alternatively when a gene tree/species tree reconciliation yields a competing gene tree in the presence of paralogs. There seems to be no consensus on which tree to use (refer the studies by Mukherjee et al.[17] and Dasmeh et al.[18], for examples of a gene and a reconciled gene tree, respectively), which suggests that both generally represent plausible choices.

Voordeckers et al.[16] analyzed the sequences using a gene tree inferred by MrBayes[19] testing for sites under positive selection in three different foreground branches. Based on the authors' species tree of yeasts, gene tree/species tree

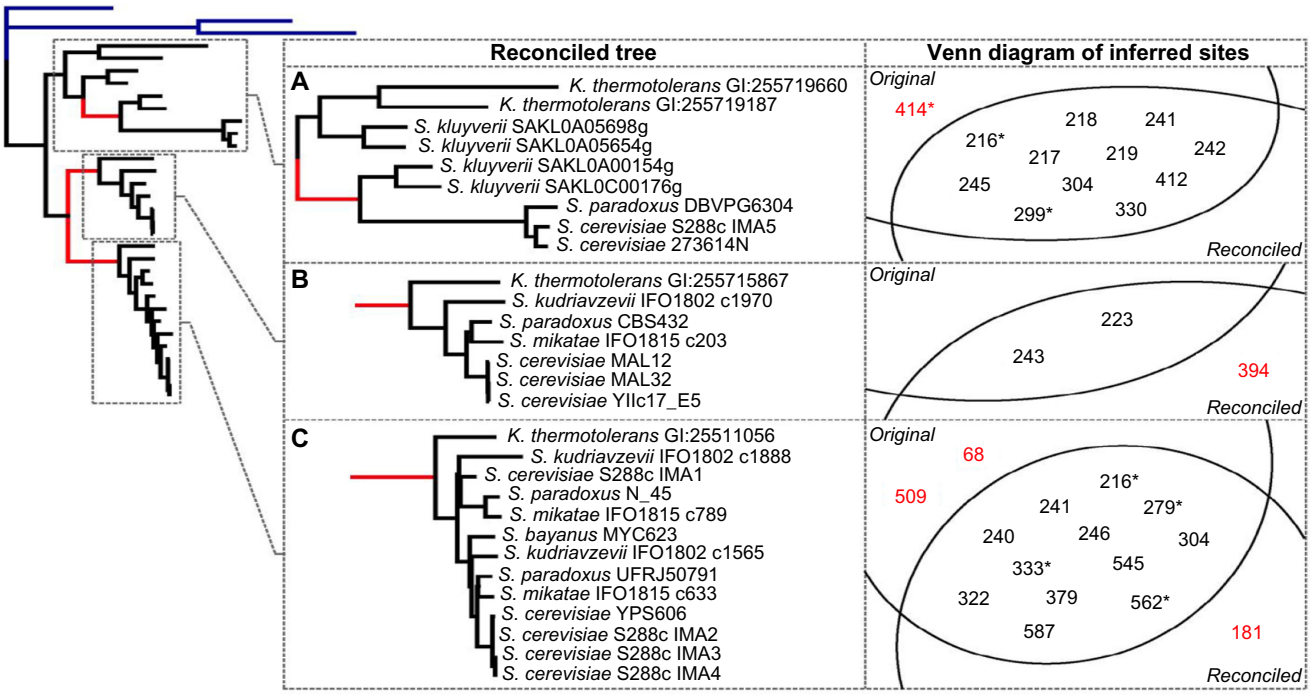

Figure 6. Different gene trees can lead to different BEB results on real sequences – the upper left panel represents the gene tree published as Figure 4 in the study by Voordeckers et al.[16], with the analyzed foreground branches highlighted in red and basal branches in blue. The three boxed areas (labeled **A**, **B**, and **C**) correspond to the subtrees, which we manually reconciled in the middle column using the species tree from Figure 1 in the study by Voordeckers et al.[16] The sites inferred to be under positive selection based on the two different trees (original and reconciled) are compared by Venn diagrams in the right column. Note that we are not able to reproduce the original results (here indicated by stars after the common sites), despite using the same sequences, alignments, trees, and program version. None of the differences we report change the authors' conclusions (all our input and output files are provided as Supplementary File 3).
Abbreviation: BEB, Bayes empirical Bayes.

reconciliation alters the subtree under each of these foreground branches. When the BEB results obtained with the original tree and our reconciled tree are compared, we observe that all three lists of sites change as summarized in Figure 6 (refer Table 1 for the results of the LRT).

This demonstrates that the gene tree influences the results on real sequences also, including alignments with indels. Yet, the experimental setting does not permit identification of the better inference of sites and tree, because outside of simulations, the truth is generally unknown.

Conclusion

This is the first study to systematically assess the performance of the BEB procedure in the context of branch-site models. We show that the length and age of the foreground determine not only the power of the LRT as reported before but also that of the BEB procedure. Most importantly, we find evidence for an effect of the gene tree on the inference of sites under selection in both simulated and real-word sequences. In the simulations, we are able to explain this effect by virtue of a single parameter, which we coin as CBL (Fig. 4).

We conclude that the gene tree is an important factor for the branch-site analysis of positive selection, so far unrecognized. However, unlike in the case of simulations, the true complement of branches and sites under positive selection cannot be usually known when analyzing real-world sequences. Therefore, it is not clear if alternative topologies introduce CBL and which tree to choose to minimize this effect. A simple strategy is to try all competing trees and test if sites are consistently inferred to evolve under positive selection in the foreground branches. If so, the robustness of the results against alternative topologies can be interpreted as evidence for a true statistical signal that increases the confidence in the inferred sites.

In summary, developing guidelines for the choice of a gene tree remains an important problem. An especially interesting case is when both gene and species trees or reconciliated gene tree are available, as for example in the study reanalyzed in Figure 6. Further investigations are also needed to understand the interplay of the tree with other known factors affecting the BSPS. Finally, while we focused solely on the BSPS, it has to be pointed out that alternative approaches for the detection of sites under episodic positive selection exist[20] that would be interesting to test and compare. These do not require an a priori distinction of foreground and background branches, making them well suited to assess the consequence of drastic rearrangements of the gene tree as for example resulting from long-branch attraction artifacts.

Materials and Methods

We simulate MSAs given a phylogenetic tree using INDELible (version 1.03),[21] which is a flexible simulation tool implementing a variety of different substitutions and indel models. It uses a Markov chain approach that allows to deal with the dependency among sites introduced when simulating indels (refer the book by Yang,[22] pp. 302–304). Here, we simulate genes consisting of 522 codons with no indels along the tree depicted in Figure 1, starting from a random sequence at the root. Simulation parameters are the transition/transversion ratio $\kappa = 2.1$, chosen to match the average reported for the human genome (see DePristo et al.[23]) and a background scheme of dN/dS ratios (1, 1, 0.8, 0.8, 0.5, 0.5, 0.2, 0.2, 0, 0) with every class making up 10% of the sites (the same as background scheme X from the studies by Zhang et al.[2] and Fletcher and Yang[7]). Furthermore, we use two foreground selection schemes (0.5, 1, 4, 0.8, 4, 0.5, 4, 0.2, 0.8, 0.5) (referred to as W) and (1.0, 0.7, 4.0, 0.8, 2.0, 0.5, 0.3, 0.2, 0.1, 0.0) (the same as foreground selection scheme V in the references mentioned above). The simulated MSAs and the control file with all parameters are attached as Supplementary File 4.

The sequences simulated in the previous step are analyzed with PAML (version 4.6),[6] which is a package with various programs for the phylogenetic analysis of molecular sequences in an ML statistical framework. It provides a rich repertoire of evolutionary models allowing to test biological hypotheses, for example, of positive Darwinian selection as does the BSPS. We label the branches as foreground that were simulated as such. Branch lengths are estimated by PAML ("runmode = 0," refer Supplementary File 2 for the basic control files with and without selection we used for all simulations). The sites under selection in the foreground branches are obtained by BEB at site-specific posterior probabilities >0.95 and >0.99.

We define the age of a foreground branch spanning nodes n_1 to n_m (ie, for $m > 2$, additional internal nodes are present) as the average distance of the nodes n_1 to n_m to human (the leaf at the end of fg6 in Fig. 1).

Simple and multiple linear regressions are compared using the BIC, which allows to select among a set of models. It compares models based on their likelihood while penalizing for the number of model parameters. Models with the lowest BIC are preferred, with a difference $\Delta BIC = BIC(H_0) - BIC(H_1)$ above 6 indicating strong evidence against the null model H_0.[24]

We summarize the elements of the confusion matrix (ie, TP, FP, TN, and FN), computing sensitivity and specificity according to their standard definitions TP/(TP + FN), TN/(FP + TN), respectively. Tree manipulations are done in Python using Biopython[25] and the ETE library.[26]

The codon MSA for the reanalysis of the sequences from the study by Voordeckers et al.[16] is generated based on the protein MSA provided in their Supporting Information. After retrieving the corresponding cDNA sequences from the NCBI and Sanger Institute, we use Pal2Nal (version 14)[27] to convert the protein alignment into a codon alignment. Pal2Nal automates this conversion, providing robustness against the presence of mismatches, UTRs and polyA tails in the input DNA sequences, frame shifts, and inframe stop codons in the input alignment.[27] Results shown in Table 1 and Figure 6 are generated with PAML version 4.4.[28] We use PAML version 4.4

to exclude different versions of PAML as a reason for different sets of sites inferred to be under positive selection.

Acknowledgments

The authors wish to thank Karin Voordeckers for providing help in retrieving the DNA sequences reanalyzed here. The authors thank the list of anonymous referees for their constructive comments that greatly improved the manuscript and Adrian Timpson for help with the written English. Finally, Pascale Gerbault is gratefully acknowledged for encouraging us to contribute to this issue.

Author Contributions

Conceived and designed the experiments: YD. Analyzed the data: YD. Wrote the first draft of the manuscript: YD. Contributed to the writing of the manuscript: YD. Agree with manuscript results and conclusions: YD, JPL. Jointly developed the structure and arguments for the paper: YD, JPL. Made critical revisions and approved final version: YD, JPL. Both authors reviewed and approved of the final manuscript.

Supplementary Materials

Supplementary File 1. Figures illustrating the results for foreground selection scheme V at site-specific posterior probability >0.95 and scheme W at site-specific posterior probability >0.99.

Supplementary File 2. Graphical representation and Newick files of the trees listed in Supplementary Table 1. Additionally, the basic control files with and without selection that were used for all simulations are given.

Supplementary File 3. All PAML input and output files corresponding to the re-analysis of the data from Voordeckers et al. summarised in Figure 6.

Supplementary File 4. The simulated INDELIble multiple sequence alignments and corresponding the control files.

Supplementary Figure 1. Overall distribution of derivations from the confusion matrix – the boxplots show the distributions of mean sensitivity, specificity, and Matthew's correlation coefficient (MCC; defined as $[TP \times TN - FP \times FN)/(P \times N \times P' \times N']^2$) across eight replicates for all foreground branches.

Supplementary Figure 2. Differences in sensitivity observed with rearranged topologies – histogram of mean differences in sensitivity across the eight replicas observed across all rearranged topologies (listed in Supplementary Table 1) and foreground branches (foreground scheme W).

Supplementary Table 1. List of rearranged topologies tested for their effect on the performance of the Bayes empirical Bayes procedure – the leaf names refer to the NCBI taxon IDs also listed in the species tree in Figure 1.

REFERENCES

1. Yang Z, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*. 2002;19:908–17.
2. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*. 2005;22:2472–9.
3. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*. 1994;11:725–36.
4. Yang Z, Wong WS, Nielsen R. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol*. 2005;22:1107–18.
5. Anisimova M, Bielawski JP, Yang Z. Accuracy and power of bayes prediction of amino acid sites under positive selection. *Mol Biol Evol*. 2002;19:950–8.
6. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24:1586–91.
7. Fletcher W, Yang Z. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol*. 2010;27:2257–67.
8. Yang Z, Reis dos M. Statistical properties of the branch-site test of positive selection. *Mol Biol Evol*. 2011;28:1217–28.
9. Gharib WH, Robinson-Rechavi M. The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Mol Biol Evol*. 2013;30:1675–86.
10. Yang Z, Nielsen R, Goldman N, Pedersen AM. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*. 2000;155:431–49.
11. Privman E, Penn O, Pupko T. Improving the performance of positive selection inference by filtering unreliable alignment regions. *Mol Biol Evol*. 2012;29:1–5.
12. Blackburne BP, Whelan S. Class of multiple sequence alignment algorithm affects genomic analysis. *Mol Biol Evol*. 2013;30:642–53.
13. Jordan G, Goldman N. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol Biol Evol*. 2012;29:1125–39.
14. Yang Z. The power of phylogenetic comparison in revealing protein function. *Proc Natl Acad Sci U S A*. 2005;102:3179–80.
15. Vilella AJ, Severin J, Ureta-Vidal A, Lii H, Durbin RM, Birney E. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*. 2009;19:327–35.
16. Voordeckers K, Brown CA, Vanneste K, et al. Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biol*. 2012;10:e1001446.
17. Mukherjee K, Campos H, Kolaczkowski B. Evolution of animal and plant dicers: early parallel duplications and recurrent adaptation of antiviral RNA binding in plants. *Mol Biol Evol*. 2013;30:627–41.
18. Dasmeh P, Serohijos AW, Kepp KP, Shakhnovich EI. Positively selected sites in cetacean myoglobins contribute to protein stability. *PLoS Comput Biol*. 2013;9:e1002929.
19. Ronquist F, Teslenko M, van der Mark P, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 2012;61:539–42.
20. Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K. A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol*. 2011;28:3033–43.
21. Fletcher W, Yang Z. INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol*. 2009;26:1879–88.
22. Yang Z. *Computational Molecular Evolution*. Oxford University Press; 2006:1–374. http://ukcatalogue.oup.com/product/9780198567028.do
23. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43:491–8.
24. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc*. 1995;90:773–95.
25. Cock PJ, Antao T, Chang JT, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25:1422–3.
26. Huerta-Cepas J, Dopazo J, Gabaldón T. ETE: a python Environment for Tree Exploration. *BMC Bioinformatics*. 2010;11:24.
27. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 2006;34:W609–12.
28. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 2003;52:696–704.