The effects of alignment error and alignment filtering on the sitewise detection of positive selection

Gregory Jordan and Nick Goldman

July 22, 2011

Abstract

The effects of alignment error on many types of evolutionary inference are not well understood. Alignment filters are commonly applied to try to reduce errors resulting from misalignment, but neither the prevalence of such errors nor the ability of filters to reduce them has been systematically studied. Focusing on the sitewise detection of positive selection, we performed simulation experiments to quantify the false positives and false negatives introduced by alignment error and the ability of alignment filters to improve performance. We found that some aligners led to many false positive results across a wide range of divergence levels, while others resulted in very few. False negatives were a problem for all aligners, increasing with divergence. Of the aligners tested, PRANK consistently performed the best and ClustalW performed the worst. The best alignment filters showed good ability to reduce the error rates from ClustalW alignments, but they failed to improve the performance of PRANK alignments under most circumstances. Of the filters tested, GUIDANCE performed the best and Gblocks performed the worst, failing to improve even the error-prone ClustalW alignments. Controls using the true alignment and an optimal filtering method suggested that performance improvements could be gained by improving aligners or filters to reduce the prevalence of false negatives, especially at higher divergence levels and indel rates. As no combination of aligner plus filter significantly outperformed the best aligner alone, we conclude that for the sitewise evolutionary analysis of most vertebrate, insect and fungal phylogenies, the use of unfiltered PRANK alignments is recommended.

Introduction

The decreasing cost of DNA sequencing has triggered a striking increase in the number of model and non-model organisms with planned genome sequencing projects, suggesting that the range and scale of comparative genomics applications will continue to expand (Green, 2007; The ENCODE Project Consortium, 2007). These clusters of closely-related genome sequences across a wide taxonomic range have led to a better understanding of which aspects of molecular evolution are variable and which are constant (Wolf et al., 2009) and an increased sampling of species should continue to boost the power and accuracy of individual analyses within a given clade.

The study of protein evolutionary rates and selective pressures in particular has flourished as a result of the growth in comparative genomics datasets. This is especially beneficial for the calculation of spatially precise evolutionary estimates, as additional species sampling has been shown to be an effective means of boosting the accuracy and power of sitewise detection of positive selection and evolutionary constraint (Anisimova et al., 2001; Massingham and Goldman, 2005). Site-specific evolutionary estimates have proved especially valuable when analyzed in conjunction with other protein-based datasets such as protein structural features (Lin et al., 2007; Ramsey et al., 2011), human population diversity (1000 Genomes Project Consortium, 2010) and human disease mutations (Arbiza et al., 2006).

A major concern in the detection of protein positive selection is that the effect of alignment error is not well characterized. Intuitively, one might expect alignment error to result mainly in an increased number of false positives, as the spurious alignment of nonhomologous codons on average would result in a high number of apparent non-synonymous substitutions and a low number of synonymous substitutions (since a randomly chosen pair of codons are more likely to be nonsynonymous than synonymous). However, false negatives may also be introduced, either through over-alignment of synonymous codons (increasing the synonymous substitution rate) or non-alignment of homologous codons (causing reduced power due to less evolutionary information). Since different aligners employ a variety of algorithms, evolutionary models, and heuristic optimizations (Notredame, 2007), each program may be more or less prone to different types of alignment error, causing potentially large variations in the nature and magnitude of its impact on the detection of positive selection.

The protein structure and the evolutionary divergence of a dataset may also contribute to the effects of alignment error. Differently structured protein regions show variable tolerance to biological indels, with indels more common in extracellular and transmembrane proteins than in highly folded enzymes and housekeeping genes (de la Chaux et al., 2007). This suggests that well-folded protein regions will experience fewer biological indels—and will therefore be less susceptible to alignment error—than less structured regions.

Furthermore, the evolutionary divergence of a dataset affects the power of sitewise inference and the prevalence of indels in multiple ways. As maximum-likelihood methods for detecting positive selection require data in the form of observed substitution events, they show little power at low divergence and their highest power at intermediate to high divergence levels (Anisimova et al., 2001). However, alignment error should be greatest at high divergences, which may have the effect of reducing power. These two trends suggest that the overall power will be low at both extremes of divergence, with little inference power at low divergence (due to the scarcity of data in the form of observed substitutions) and an overwhelming amount of alignment error at high divergence (due to the large number of indel events).

Luckily the majority of genes in many biological clades of interest (such as mammals, vertebrates, fruit flies, and yeast) fall within the middle range of divergences where sitewise methods are at their most powerful and where multiple alignment is a difficult—but not hopeless—problem. As such, it is important to seek an understanding of the impact of alignment error on overall error rates within this important range of divergence levels.

A number of empirical analyses have established that errors in gene sequencing, annotation and

alignment can contribute to errors in downstream evolutionary analyses such as phylogeny inference (Wong et al., 2008) and estimates of positive selection (Schneider et al., 2009; Markova-Raina and Petrov, 2011). Most recently, Markova-Raina and Petrov (2011) showed that the detection of positively-selected sites and genes in *Drosophila* genomes is highly sensitive to aligner choice, with PRANK's codon model (Löytynoja and Goldman, 2008) consistently producing alignments with the lowest amount of positive selection. Still, according to the authors' manual inspection of alignments, even positively-selected sites identified with PRANK alignments contained a sizable proportion of apparent false positives.

A limitation of the analysis of error in empirical datasets is the lack of a benchmark set of true alignments and positively-selected sites. Markova-Raina and Petrov (2011) used their expected general effect of alignment error (an increase in false positives due to misalignment of non-homologous codons) as a proxy by which to compare different methods, allowing for the conclusion that PRANK was the least error-prone aligner in their analysis. However, the absolute number of false positives remained uncertain and there was the possibility of conflating multiple sources of error: in addition to alignment error, the authors noted that gene mis-annotation was responsible for many apparent false positives, and there is also an expected error rate from the likelihood inference method itself. This limitation leaves important and interesting questions, regarding the nature of alignment error and its quantitative impact on the detection of positive selection, unanswered by empirical studies.

Controlled simulation experiments provide a natural framework for investigating error rates in detail, allowing one to pinpoint the sources of error in multi-step analyses such as alignment followed by evolutionary inference. This approach has been employed in assessing the robustness of phylogenetic inference methods to misalignment (Dwivedi and Gadagkar, 2009; Ogden and Rosenberg, 2006; Löytynoja and Goldman, 2008) but those results cannot be easily extrapolated to the analysis of sitewise selective pressures. More recently, Fletcher and Yang (2010) performed a series of simulation experiments investigating alignment error in the use of the branch-site test to detect positive selection in genes. Their results showed that most aligners caused false positives by over-aligning codons and that datasets from mammalian and vertebrate gene families contain enough evolutionary divergence to make false positive errors resulting from misalignment a legitimate concern.

Reflecting a widespread awareness of the problem of misalignment, methods for identifying and removing uncertain or unreliable alignment regions have been commonly used in phylogenetic and molecular evolutionary analyses. The popular Gblocks program applies a set of heuristic criteria to identify conserved blocks deemed suitable for phylogenetic or evolutionary analysis (Castresana, 2000) while a number of aligners such as T-Coffee (Notredame et al., 2000) and PRANK (Löytynoja and Goldman, 2008) produce estimates of alignment confidence or reliability. GUIDANCE, which measures the robustness of alignment regions to perturbations in the guide tree used for progressive alignment, has also been proposed as an alignment confidence score (Penn et al., 2010). Unfortunately, despite their widespread use, the impact of the many available alignment scoring and filtering methods on phylogenetic and evolutionary analyses has not been well studied. Even for a single filtering program, Gblocks, results have been contradictory: one simulation-based study found that it improved the phylogenetic signal (Talavera and Castresana, 2007) while an empirical study across a wide range of taxa found that Gblocks-filtered

alignments produced worse phylogenetic trees than unfiltered alignments (Dessimoz and Gil, 2010). With the application of published filtering methods to alignments before testing for positive selection becoming standard practice (Studer et al., 2008; Aguileta et al., 2009), an analysis of the benefits of alignment filtering to the detection of positive selection would seem well-warranted.

This paper aims to use a simulation framework to incorporate alignment error and alignment filtering into estimates of the error rate and power of sitewise evolutionary inference of positive selection. Our approach is similar to that of Anisimova et al. (2002) in that we use simulated protein alignments to evaluate methods for detecting sitewise positive selection, considering each site to be an independent hypothesis test producing a positive or negative result. Our experimental design also shares some features with Fletcher and Yang's (2010) investigation of the performance of branch-site method for detecting genewise positive selection under alignment error. Like Fletcher and Yang (2010), we simulate codon alignments with indels and calculate inferred alignments using a small but diverse sample of aligners. Unlike their study, here we focus on sitewise detection of positive selection occurring throughout a phylogeny (as opposed to genewise detection of selection acting at specific branches), we explore a wider range of plausible tree sizes, indel rates and divergence levels, and we evaluate the impact of a number of alignment filtering methods on the sitewise analysis. Given the recent proliferation of completed and planned vertebrate genomes, we pay special attention to choosing simulation parameters and analysis methods similar to those commonly encountered in the sitewise analysis of vertebrate gene families.

Methods

Alignment Simulations

An overview of the simulation parameters used in this study can be found in Table 1. Three rooted trees were used to guide the simulation of protein-coding DNA alignments: the artificial 6-taxon tree used by Anisimova et al. (2001) and Massingham and Goldman (2005) rooted at its midpoint, the 17-taxon vertebrate β-globin tree from Yang et al. (2000) and the 44-taxon vertebrate tree used by the ENCODE project (The ENCODE Project Consortium, 2007; Nikolaev et al., 2007). Trees, shown with their original branch lengths in Figure 1, were scaled to comparable divergence levels by normalizing their mean path length (MPL), defined as the root-to-tip branch length averaged across all lineages in the tree. We simulated alignments with MPL divergence between 0.05 and 2.0 synonymous substitutions per synonymous site, spanning the range of evolutionary divergences observed in several clades of organisms with fully-sequenced genomes (Table 2).

The INDELible program (Fletcher and Yang, 2009) was used to simulate codon sequences with indels along each phylogenetic tree. The length of the root sequence was set to 500 codons and $\boldsymbol{\kappa}$ (the ratio of transition to transversion substitutions) was fixed at 4. Indel lengths were drawn from a discretized power-law distribution with an exponential decay parameter of 1.8 and a maximum value of 40, yielding a mean indel length of 3.33 codons and standard deviation of 5.51 codons. The power-law model of indel lengths is well-supported by empirical studies (Benner et al., 1993; Cartwright, 2009) and manual inspection of alignments from a range of parameter values identified the chosen model parameters as

resulting in alignments most closely resembling those encountered in vertebrate alignments. The ratio of insertion to deletion events was set to 1, and the rate of indel formation was varied between 0 and 0.2 indel events per substitution per site.

The distribution of sitewise selective pressures (embodied by the parameter $\boldsymbol{\omega}$, the ratio of the rate of non-synonymous substitution to the rate of synonymous substitution) was modeled with a log-normal distribution derived from a maximum-likelihood fit to a large dataset of sitewise selective pressures estimated from mammalian gene trees ((2011); log-normal parameters shown in Table 1). This distribution, with mean $\boldsymbol{\omega}$ of 0.28 and 6% of sites having $\boldsymbol{\omega} > 1$, is consistent with the structure-based expectation of many protein sites under purifying selection and few under neutral selection or positive selection (Smith, 1970; Kimura and Ohta, 1974). INDELible's general discrete model of sitewise $\boldsymbol{\omega}$ variation was used to approximate the log-normal distribution by splitting the probability density into 50 equally-spaced bins between $\boldsymbol{\omega}$ values of 0 and 3, with the highest bin containing the probability density for all values $\boldsymbol{\omega} > 3$.

Branch lengths for each of the simulation trees were scaled before simulation to correct for the difference between our definition of branch lengths as the number of synonymous substitutions per synonymous site (dS) and INDELible's interpretation of branch length as the average number of substitutions per codon (t) (Fletcher and Yang, 2010). They are related approximately by $t = 3(NdN + SdS) = 3dS(\bar{\omega}N + S)$, where N and S are the proportion of nonsynonymous and synonymous sites and $\bar{\omega}$ is the mean $\boldsymbol{\omega}$ across all sites. S is approximately 0.3 when $\boldsymbol{\kappa} = 4$ (Yang and Nielsen, 1998) and the mean $\boldsymbol{\omega}$ ratio for our chosen distribution is 0.277, yielding a dS-to-t conversion factor of 1.48 for all simulations performed.

Sequence Alignment and Filtering

Alignments were inferred using four alignment algorithms chosen for their widespread use or demonstrated accuracy: ClustalW v1.82 (Thompson et al., 1994), MAFFT (Katoh et al., 2005) and two variants of PRANK (Löytynoja and Goldman, 2005) based on an amino acid model (subsequently referred to as PRANK_{AA}) or an empirical codon model (subsequently referred to as PRANK_{C}). Unaligned amino acid sequences were given as input to all alignment programs (except PRANK_{C}, which was provided the unaligned DNA sequences) and all software was run using default parameters with the true phylogenetic tree given as input where possible.

Alignments were filtered by masking out residues based on the output of three alignment scoring methods: Gblocks conserved blocks (Castresana, 2000), T-Coffee consistency scores (Notredame et al., 2000; Notredame and Abergel, 2003), and GUIDANCE alignment confidence scores (Penn et al., 2010). Gblocks, which identifies entire alignment columns as conserved or not conserved, was run using an increased gap tolerance and a reduced minimum block length in order to reduce the amount of each alignment removed (command-line parameters *b5=a* and *b4=3*), and all residues from any columns not within an identified conserved block were masked with Ns.

GUIDANCE and T-Coffee produce individual scores for each residue, allowing individual residues to be masked instead of entire columns. GUIDANCE generates many replicate alignments, each using a slightly perturbed guide tree, with either MAFFT or PRANK_{AA} as the bootstrap aligner. The program

then assigns to each residue from the input alignment a score from 0 to 1 based on how consistently it was placed in the replicate alignments. In order to maximize the similarity between the input aligner and the bootstrap aligner, we ran GUIDANCE with 100 MAFFT replicates when filtering ClustalW alignments and with 30 PRANK_{AA} replicates when filtering PRANK_{C} alignments. T-Coffee calculates the residue-wise consistency between an input multiple alignment and independently calculated pairwise alignments (Notredame and Abergel, 2003), rounding and normalizing residue scores into integers between 0 and 9. T-Coffee was run using its default settings and the *evaluate_mode –output=score_ascii* command-line parameters to output alignment scores.

To filter alignments based on these residue-wise scores, a cutoff threshold was chosen for each method (0.5 for GUIDANCE and 5 for T-Coffee) and residues equal to or below that threshold were masked. On a per-alignment basis, if the default threshold caused greater than 50% of residues to be masked, then the threshold was relaxed to the highest value for which at least 50% of residues remained. We found this adjustment necessary because the scores from GUIDANCE and T-Coffee were strongly affected by the simulation conditions, with much lower average scores at higher indel rates and divergences. Requiring at least 50% of residues to remain unmasked ensured that enough data were available for meaningful evolutionary analysis, mimicking typical treatment of real data sets.

Two unrealistic but informative datasets were produced to serve as controls. First, the true simulated alignment was included in order to evaluate the sitewise performance without any alignment error. Second, an additional filtering method was constructed to represent an unattainable best-case scenario for sequence filtering, using knowledge of the true alignment to assign a score to each residue reflecting how correctly is has been placed in the inferred alignment. The approach taken was to calculate, for each residue, the branch length of the correct sub-tree (defined as the sub-tree connecting all sequences to which the current residue was correctly aligned) divided by the branch length of the total aligned sub-tree (defined as the sub-tree connecting all sequences with non-gap residues at the current alignment column). This residue-wise score ranges from 0 to 1 and reflects the expectation that correctly-aligned evolutionary branch length is the main source of information from which sitewise inference methods derive their power. We refer to this method as the 'optimal' filtering method. Scores were handled in a manner similar to GUIDANCE and T-Coffee, using a score threshold of 0.5.

Sitewise Evolutionary Analysis

Sitewise estimates of selective pressures were calculated using maximum-likelihood methods implemented in the Phylogenetic Analysis by Maximum Likelihood (PAML; Yang 2007) and Sitewise Likelihood Ratio (SLR; Massingham and Goldman 2005) software packages.

The *codeml* program from PAML implements a number of likelihood ratio tests (LRTs) for detecting the presence of positive selection in a gene while allowing the $\boldsymbol{\omega}$ ratio to vary among sites (Yang et al., 2000). These models, known as the sites or random sites models, use a variety of predefined statistical distributions to account for heterogeneous $\boldsymbol{\omega}$ ratios among sites. After the likelihood optimization is performed, Bayesian methods can be used to estimate the posterior probability of each site being drawn from a given site class, where a high posterior probability of a site belonging to a class with $\boldsymbol{\omega} > 1$ can

be considered strong evidence that a site has evolved under positive selection (Yang et al., 2005). We used the two models for which the recommended Bayes Empirical Bayes method are implemented, M2a and M8.

SLR implements a method specifically designed for sitewise estimates which has been shown in simulations to perform as well as or better than PAML's sitewise random sites models (Massingham and Goldman, 2005). SLR models codon evolution as a continuous-time Markov process where substitutions at one site are independent of substitutions at all other sites. No assumptions are made regarding the distribution of $\boldsymbol{\omega}$ ratios within the alignment. The value of $\boldsymbol{\omega}$ is considered to be an independent parameter at each site: after first optimizing shared parameters using the whole alignment, SLR uses the shared parameters and the data at each alignment site to calculate a sitewise statistic for non-neutral evolution. This statistic is based on a likelihood-ratio test where the null model is neutral evolution ($\boldsymbol{\omega} = 1$) and the alternative model is either purifying or positive selection ($\boldsymbol{\omega} < 1$ or $\boldsymbol{\omega} > 1$, respectively). The raw statistic measures the strength of evidence for non-neutral evolution at each site; following Massingham and Goldman (2005) we use a signed version of the SLR statistic (created by negating the statistic for sites with $\boldsymbol{\omega} < 1$) as the test statistic for positive selection.

Measuring Performance

In order to compare sitewise estimates from different alignments, a single sequence from each tree was chosen as the reference (arrows, Figure 1) and all sitewise statistics were mapped from alignment columns to sequence positions in the reference sequence. This approach corresponds to the process of mapping alignment-based evolutionary estimates onto a single member of the alignment for further analysis and integration with other genome-referenced data (as is often done, for example, using mammalian alignments and a human reference). As a result of this reference sequence based mapping, sites which were deleted in the reference sequence or inserted in a lineage not ancestral to the reference were not included in the final performance analysis.

To evaluate the power and error rates that might be achieved in real-world data analysis, the recommended cutoff thresholds for PAML's Bayesian posterior probabilities and the SLR statistic were used to identify positively selected sites. A posterior probability threshold of 0.95 was used for PAML (Yang et al., 2005) and a threshold of 3.84, the 95% critical value of the χ^2 distribution with 1 degree of freedom, was used for SLR (Massingham and Goldman, 2005). Sites were compared to their true simulated state (e.g. positively-selected or non-positively selected) in order to identify correct and incorrect inferences, and from these classifications we calculated the false positive rate (FPR, defined as the proportion of all sites with true $\boldsymbol{\omega} < 1$ falsely identified as positively selected) and true positive rate (TPR, defined as the proportion of all sites with true $\boldsymbol{\omega} > 1$ correctly identified as positively selected).

As the addition of alignment error is expected to affect the power and error rates differently for each combination of simulation condition and aligner, we identified the score thresholds for each dataset that resulted in an actual FPR of 1% and calculated the TPR achieved at this actual error rate (hereafter referred to as $\text{TPR}_{1\%}$ to distinguish it from the TPR described above). Although this estimate of error-controlled power would be impossible to calculate in an empirical analysis where the error rate is unknown,

it is useful in a simulation context for allowing a controlled comparison of the performance of sitewise analysis between different conditions. Specifically, it should be sensitive to changes in the numbers of both false positives and false negatives resulting from alignment error or alignment filtering; in both cases a lowered error-controlled power would result, as fewer true positives are identified at the constant 1% FPR.

We also evaluated the ability of each method to accurately infer the ω value at each site by collecting sitewise ω estimates from the output of each method and calculating the Pearson's correlation coefficient between the true and inferred ω values for each set of simulation conditions.

Results and Discussion

The Performance of Three Methods for Detecting Sitewise Positive Selection

We first evaluated the ability of three sitewise methods, PAML M2a, PAML M8 and SLR, to accurately estimate sitewise ω values and to detect positive selection under a range of tree lengths in the absence of alignment error. Figure 2 shows the TPR, FPR, $TPR_{1\%}$ and sitewise ω correlation over a range of MPLs for each of the three simulation trees.

The detection power and ω correlation were weakest at low divergence levels for all methods and all trees due to the low amount of evolutionary information, as observed in previous simulations (Anisimova et al., 2002). We found a positive correlation between tree size and detection power, with the highest performance in the 44-taxon tree. Power generally increased monotonically with divergence, except for the 6-taxon tree which saw its maximum performance at moderate divergence levels (MPL 0.5–1.0) and began decreasing at higher values. The downward trend in the 6-taxon tree was likely due to the impact of saturation of synonymous sites in the very long branches present in such a sparse tree at high divergence levels. With lower average branch lengths at equivalent MPLs, the two larger trees showed no signs of decreased performance even at a MPL of 2 substitutions per site, which is greater than any of the divergence levels found in groups of commonly analyzed vertebrate, insect and fungal species (Table 2).

Comparing the three methods for detecting positively selected sites, we found that at the recommended cutoff threshold (Figure 2, top row) SLR showed the highest power to detect positive selection in all trees, followed by PAML M8 and PAML M2a. In the smallest tree, the power of the two PAML methods was virtually zero while SLR reached a maximum TPR of 18% (at MPL=0.5). At the same divergence, SLR yielded TPRs of 25% and 45% in the 17-taxon and 44-taxon trees, respectively, with PAML M2a ranging between 50–75% of SLR's power and PAML M8 falling between the two other methods.

The TPR measurements represent the power that might be achieved in real-world analysis using recommended cutoff thresholds, but the higher power from SLR may merely reflect a shifted balance between power and accuracy at the recommended cutoff threshold as opposed to an increased absolute ability to discriminate positive from neutral or purifying selection. The FPR and error-controlled $TPR_{1\%}$ results revealed that this was indeed the case: the FPR from SLR was higher than that from either of the PAML methods for all trees and divergence levels, suggesting that its higher power was the result of a less-conservative cutoff value. This was further verified by evaluating the TPR at a cutoff threshold that

controlled for an actual FPR of 1% for each method ($TPR_{1\%}$, third row in Figure 2). The error-controlled $TPR_{1\%}$ values were virtually identical for all three methods, providing strong evidence that the three methods' sitewise statistics were nearly equally sensitive to positive selection under our chosen simulation conditions.

The conservative nature of the default thresholds for PAML and SLR has been previously noted (Anisimova et al., 2002; Yang et al., 2005; Massingham and Goldman, 2005), but the extremely low false positive rates in our simulations showed that in the absence of alignment error all three methods would yield very few false positives when analyzing genes with a typical mammalian-like distribution of ω values. The low FPRs were likely due to the large proportion of sites under moderately strong purifying selection in our ω distribution used for simulation. Such sites are less likely to yield false positives than sites under neutral evolution (ω=1), the null model against which tests for positive selection are traditionally controlled.

For the purposes of our indel experiments, the observed similarity in error-controlled power levels indicated that the behavior of PAML M2a, PAML M8, and SLR was similar enough not to warrant separately evaluating all three methods in the subsequent indel simulation experiments. As the runtime for SLR was significantly lower than that of either PAML model, all subsequent results are presented only based on the SLR test.

The Effect of Alignment Error on Sitewise Power

When the indel rate was greater than zero, performance levels varied significantly for different tree sizes, alignment algorithms, and evolutionary divergences. Figure 3 shows the same performance measurements as Figure 2 for simulations without indels (gray lines, Figure 3) and with indels (black and textured lines, Figure 3) analyzed using three different aligners (ClustalW, MAFFT, and $PRANK_C$) and the true alignment. (We omitted results for $PRANK_{AA}$ alignments from Figure 3 in order to reduce visual clutter; a full comparison of all aligners tested is shown in Figure 4C and discussed in the next section.) For the indel simulations, the indel rate here was held constant at 0.1 indel event per substitution.

Comparing the results without indels to those with indels under the true alignment we found a slight decrease in power and ω correlation and no noticeable increase in FPR. The decreased power was expected, since even in the absence of alignment error alignment columns containing gaps harbor less evolutionary information than columns with complete sequence data. The lack of increased FPR showed that SLR retained its conservative statistical performance even when analyzing gapped alignments. Surprisingly, at higher divergences (MPL $>$ 1.0) under the six-taxon tree, the FPR with indels was lower than the FPR without indels. This unexpected result may be attributed to the large number of alignment columns under such conditions that contained only a single non-gap sequence, as those columns were never inferred as positively-selected by SLR due to the complete lack of information. The two larger trees did not show a similar trend at high divergence levels, suggesting that this effect was indeed due to the highly sparse nature of the alignments in the 6-taxon tree. All other results from the 6-taxon tree at high divergences were similarly anomalous in this respect; we surmised that the sparseness of the true alignment, combined with the extreme difficulty of accurately aligning sequences along very long branches, made sitewise

analysis with indels very unreliable at high divergences in the smallest tree.

When alignments were inferred using one of the three aligners tested, the TPR, $TPR_{1\%}$ and $\boldsymbol{\omega}$ correlation were all reduced relative to the true alignment (dashed and dotted lines, Figure 3). The degree of reduction varied depending on the aligner, simulation conditions, and performance measurement being analyzed. At low divergences (e.g. MPL < 0.2) the inferred alignments generally showed only a small decrease in performance. As divergence levels increased, so did the difference between the performance of the true alignment and the inferred alignments. The three aligners tested could be consistently and unambiguously ranked by all of the measured performance characteristics, with $PRANK_C$ always performing best and ClustalW performing worst. The same ranking of aligners with respect to detecting positive selection was found by Fletcher and Yang (2010); our results corroborate their finding and provide evidence that this ranking may be applicable to a wider array of evolutionary analyses.

Looking at the TPR results for inferred alignments, we observed that in the 6-taxon tree the three aligners formed a cluster of lines well below the true alignment value, indicating similar tendencies among the different aligners to produce false negatives in the smaller tree. In larger trees the different aligners showed a wider spread of TPR values, but even $PRANK_C$ showed a 5–10% reduction compared to the true alignment at MPL=1.0. These results show that the introduction of false negatives is a significant and seemingly unavoidable result of alignment error at medium to high divergence levels (MPL > 0.5), with even the most successful aligner producing a marked reduction in TPR compared to the true alignment. The $TPR_{1\%}$ and $\boldsymbol{\omega}$ results in the larger two trees were qualitatively similar to the TPR results, showing that the aligners tested led to different levels of sitewise performance even when controlling for actual error rates or assessing the sitewise $\boldsymbol{\omega}$ correlation.

The FPRs for infered alignments exhibited a very different trend from the other performance measures, with generally higher FPRs than the true alignment and the widest range of values occurring in the 6-taxon tree. In this tree at medium divergence levels (e.g. MPL 0.2–0.6) ClustalW showed up to a fourfold increase, and $PRANK_C$ a nearly twofold increase, in FPR over the true alignment. As previously noted, the 6-taxon tree showed an anomalous FPR pattern at higher divergences, with lower FPRs for inferred alignments than the true alignment, likely due to the highly sparse true alignment under those conditions. In the two larger trees, FPRs from inferred alignments were less elevated compared to the true alignment, less variable between aligners, and relatively constant across the range of divergences. ClustalW's FPR ranged between 0.001 to 0.005, while $PRANK_C$'s FPR was virtually identical to that of the true alignment in the 17-taxon and 44-taxon trees.

We found it useful to combine divergence estimates from Table 2 with the results from Figure 3 to characterize the combined effects of alignment error at different commonly analyzed levels of divergence. For example, at a human-mouse divergence level (MPL=0.2) misalignment had little impact on the TPR regardless of what aligner was used. However, ClustalW yielded a notably higher FPR than MAFFT or $PRANK_C$, and the error-controlled $TPR_{1\%}$ was correspondingly lower for ClustalW in all three trees. Thus, at low divergences we found that false positives were the main source of error from misalignment, and different aligners had highly variable tendencies to produce false positive results. At higher vertebrate and *Drosophila* divergence levels (MPL=0.8–1.0) false negatives became much more prevalent. The TPR

for all inferred alignments was virtually zero in the 6-taxon tree, underscoring the necessity of including many species in the analysis of highly diverged sequences. In the two larger trees, PRANK_C resulted in very few additional false positives, but it suffered a 5–10% reduction in TPR relative to the true alignment. Meanwhile, ClustalW showed a 50% TPR reduction and maintained a strongly elevated FPR. At higher divergences and in larger trees, false negatives were thus the most persistent effect of alignment error, causing a marked reduction in sitewise power even with the best-performing aligner. Overall, the ranking of aligners was clear, with PRANK_C performing better than MAFFT and MAFFT performing better than ClustalW across all performance measures and MPL divergence levels.

Sitewise Power Under a Range of Indel Rates and Divergences

To explore the effects of alignment error across a wider range of simulation conditions, we extended the simulations of Figure 3 across multiple indel rates. Figure 4 shows heatmaps of the TPR and FPR for ClustalW, PRANK_C and the true alignment (Figure 4A,B) and a heatmap of the error-controlled $\text{TPR}_{1\%}$ for all aligners tested (Figure 4C). (MAFFT and PRANK_AA were omitted from Figure 4A–B to save space, but their performance fell between that of ClustalW and PRANK_C for all performance measures, with PRANK_AA slightly outperforming MAFFT; a comprehensive set of TPR, FPR, and $\text{TPR}_{1\%}$ results can be found in Supplementary Figure S1.) The results from Figure 3, which were simulated with an indel rate of 0.1, correspond to the middle row of each panel in Figure 4; rows above and below the middle row represent higher and lower indel rates, respectively. Similarly, the bottom row of each panel in Figure 4 was simulated with an indel rate of zero and corresponds to the 'No Indels' data in Figure 3.

The TPR values (Figure 4A) show a consistent pattern across the range of indel rates, with power decreasing as either the indel rate or the divergence level increases (except at the lowest divergence levels, where the lack of evolutionary information yielded slightly lower TPRs in the larger two trees). PRANK_C showed a greater ability than ClustalW to maintain a high TPR at higher indel rates, especially in the 17-taxon and 44-taxon trees. At lower indel rates, the TPR performance of both aligners and the true alignment were qualitatively similar.

PRANK_C and ClustalW both showed a qualitatively similar pattern of elevated FPRs in the 6-taxon tree (Figure 4B), but their behavior diverged significantly in the 17-taxon and 44-taxon trees. In the 17-taxon tree, PRANK_C only showed an elevated FPR compared to the true alignment at very high indel rates and divergence levels, but the ClustalW FPR increased steadily with the indel rate, quadrupling in value from the lowest to highest indel rate. Interestingly, for any given indel rate, the ClustalW FPR showed little variation across the range of divergence levels. This result was counter-intuitive, as we expected alignment errors to become more common as divergence increased and the number of observed indel events grew. Furthermore, PRANK_C behaved according to our expectations, showing increased FPRs only at the highest divergences and indel rates in the 17-taxon tree. The FPR results in the 44-taxon tree confirmed the strange effect of ClustalW's alignments on the sitewise FPR: at the highest indel rates, ClustalW showed a negative relationship between FPR and divergence—exactly opposite to the trend we expected. PRANK_C's FPR in the 44-taxon tree was equal to or below that of the true alignment under almost all conditions.

The error-controlled $TPR_{1\%}$ results (Figure 4C) provide a comprehensive picture of the effect of alignment error on the detection of sitewise positive selection. The two aligners not shown in the two other panels (MAFFT and $PRANK_{AA}$) exhibited $TPR_{1\%}$ values intermediate to those from ClustalW and $PRANK_C$ across the range of parameters tested, with $PRANK_{AA}$ performing better than MAFFT. As expected, performance was very similar between aligners at very low indel rates. At higher indel rates, most aligners yielded similar patterns of low $TPR_{1\%}$ in the 6-taxon tree, but in the larger two trees ClustalW and MAFFT alignments were unable to achieve high $TPR_{1\%}$ values, presumably due largely to their elevated FPR in those trees.

It is worth noting the exceptional ability of $PRANK_C$ to maintain a very low level of false positive sites even under extremely difficult alignment conditions. Although $PRANK_C$ showed slightly elevated FPRs at high indel rates in the 17-taxon tree, FPRs were nearly identical to the true alignment across all simulated conditions in the 44-taxon tree. This impressive performance suggests that, given a large enough number of taxa, $PRANK_C$ alignments would yield very few erroneous false positives in scans for positive selection in sequences with even very high divergence levels. Furthermore, these results showed that false negatives contributed more than false positives to $PRANK_C$'s reduction in sitewise performance—a novel observation which provides insight into the nature of $PRANK_C$ alignments and their application to sitewise evolutionary analysis.

Effect of Alignment Filtering on Sitewise Error Rates

Having established that alignment error can lead to reduced sitewise performance through the introduction of false negatives and false positives, we tested whether alignment filtering methods could reduce error rates and improve the power of sitewise detection of positive selection. Using sequences simulated from the 17-taxon tree and a range of indel rates and divergence levels, we calculated inferred alignments using ClustalW and $PRANK_C$ and applied four filtering methods before performing the sitewise analysis. Since we wished to determine whether alignment filters either improved or worsened the error rates and power of sitewise analysis, we measured the ratio of each performance measure to the value obtained from the equivalent unfiltered alignments. These relative values are presented in Figure 5.

As alignment filters act through the removal of alignment residues or columns, a certain amount of reduction in the FPR and TPR was expected purely from the decreased amount of information available. For example, a filter that randomly removes a fraction of residues of each alignment would be expected to produce equal reductions in FPR and TPR. A more effective filter may also yield a reduced TPR, but the FPR reduction would be larger in magnitude, making the detection of positive selection more powerful for a given error rate. Thus, a reduced FPR is not necessarily indicative of good filtering performance, nor is a reduced TPR necessarily indicative of poor filtering performance. Additionally, the prevalence of false negatives resulting from misalignment suggested that alignment filters may also improve power by removing false negatives, perhaps by masking out residues that were preventing positive sites from being identified. The removal of false negatives would result in an increased TPR, further complicating the assessment of filtering results based on FPR or TPR alone. As a result, we focused on the change in error-controlled $TPR_{1\%}$ as the best single measure of whether a filter had successfully improved the

sitewise power of a dataset since this value is sensitive to changes in both the FPR and TPR. Note that the $TPR_{1\%}$ controls the FPR post-filtering, accounting for the tendency of filtering to reduce the FPR at a given cutoff threshold.

We first examined the two controls, the unfiltered true alignment and the inferred alignments filtered with the optimal filter (top two rows, Figure 5). The true alignment nearly always showed smaller FPR (red cells) and greater TPR and $TPR_{1\%}$ (blue cells) compared to the inferred alignments, with a greater magnitude of change relative to the ClustalW alignments than to the $PRANK_C$ alignments (darker shades cf. lighter shades). These scores represented the direction and an upper limit on the magnitude of change that might be achieved by a perfect alignment filter. One exception to the general trend was the observation of two simulation conditions with slightly elevated FPRs in the true alignment compared to $PRANK_C$ alignments (at an indel rate of 0.04 and MPL of 0.8 and 2.0). This small inconsistency may be explained by stochastic variation in false positive counts, as the absolute value of the FPR was very low in both datasets under those conditions (on the order of 5×10^{-4}, Figure 4B). A similar slight FPR elevation was also observed at the same indel rate for the optimal, GUIDANCE and T-Coffee filters.

Our hypothesis was that the optimal filter would show the same direction of change in FPR and TPR as the true alignment, but with slightly lower magnitudes. Indeed, improved sitewise performance was achieved in nearly all simulation conditions by the optimal filter, with the magnitude of $TPR_{1\%}$ change slightly lower than for true alignment. For ClustalW alignments the amount of improvement was quite large, with >70% increase in $TPR_{1\%}$ for nearly all conditions with an indel rate above 0.1. The improvement was more modest for $PRANK_C$ alignments with a maximum of 15–35% $TPR_{1\%}$ increase.

Looking at the reduction in the number of non-masked residues remaining after filtering, we found that the optimal filter reached the maximum of 50% filtered residues for all ClustalW alignments with MPL>1 and an indel rate>0.1. This meant that more than 50% of residues were correctly aligned across less than 50% of the tree in those alignments. By contrast, the optimal filter applied to $PRANK_C$ alignments only reached the maximum of 50% filtered residues at the highest tested divergence level and indel rate combination.

The TPR improvements achieved by the optimal filter provided some insight into the nature of sitewise false negatives resulting from alignment error. Two different types of alignment error might cause a false negative at a positively-selected site: either misalignment of one or more nonhomologous codons causing the positive signal to be masked, or non-alignment of homologous codons causing the amount of evolutionary information to be reduced. The former type of error would be recoverable by alignment filtering (through removal of the codon(s) masking the positive signal), but the latter would not. Thus, the ability of the optimal filter to improve TPR levels across the board provided evidence that a sizeable portion of false negatives from both ClustalW and $PRANK_C$ alignments were due to misaligned codons and thus amenable to recovery by filtering. Although the optimal filter was unrealistic in that it was based on perfect knowledge of which codons were misaligned, this result provided hope that one of the other filters might show a similar ability to recover false negative errors from $PRANK_C$ alignments.

Turning to the three filters under investigation, we found T-Coffee and GUIDANCE both to be highly effective at improving ClustalW alignments, with magnitudes of improvement near those of the optimal

filter. When applied to PRANK_C alignments, however, the two filters' behavior diverged: T-Coffee only showed unchanged or reduced $\text{TPR}_{1\%}$, but GUIDANCE yielded slightly improved $\text{TPR}_{1\%}$ at high divergence levels and indel rates, with values 5–15% greater than the unfiltered PRANK_C alignments. Both filters removed similar amounts of sequence information and resulted in similarly reduced FPR levels, but GUIDANCE showed a unique ability to recover false negatives from PRANK_C alignments at the highest divergence levels and indel rates, and the resulting TPR elevation appears to have been responsible for the increased $\text{TPR}_{1\%}$ performance.

Gblocks behaved very differently from the other filters tested, resulting in reduced FPR, TPR, and $\text{TPR}_{1\%}$ under nearly all simulation conditions. Only at high indel rates and low divergence levels in the ClustalW alignments did Gblocks show increased $\text{TPR}_{1\%}$ relative to the unfiltered alignments. This poor performance was likely due to overly-aggressive removal of alignment columns. We could not limit the amount of sequence masked by Gblocks, so many alignments saw more than 70% of residues removed, resulting in the loss of a large number of correctly-aligned true positive sites. Dessimoz and Gil (2010) found Gblocks filtering to have a negative effetc on the accuracy of phylogenetic inference; our results provide additional evidence in support of their finding, suggesting that Gblocks filtering tends to reduce, rather than increase, the power and accuracy of alignments when applied to a number of evolutionary analyses.

There was some evidence that the column-wise nature of Gblocks filtering was partly responsible for its poor performance in our experiments. Since false negative errors cannot be recovered through the removal of entire alignment columns, it made sense that the PRANK_C alignments—which resulted in varying numbers of false negatives but always very few false positives—would not see improved performance after column-wise filtering. On the other hand, ClustalW alignments showed a relatively constant level of false positives and an increasing number of false negatives as divergence levels increased. The application of a column-wise filter like Gblocks would thus be expected to show good improvement at low divergences where false positives dominated, but less improvement at higher divergences where false negatives became more prominent. Indeed, this was the pattern observed when applying Gblocks to ClustalW alignments.

Overall, our alignment filtering simulations found that Gblocks rarely improves alignments for site-wise detection of positive selection, but filtering methods based on GUIDANCE and T-Coffee scores have a good ability to mask out misaligned residues that cause false positives and false negatives in site-wise inference. This beneficial effect is highly dependent on simulation conditions and the input aligner. For ClustalW alignments (which, left unfiltered, led to many false positives and false negatives) both GUIDANCE and T-Coffee showed good ability to improve sitewise performance, behaving qualitatively similarly to the 'optimal' filter. In order to compare the TPR, FPR and $\text{TPR}_{1\%}$ values obtained with filtered ClustalW alignments to those obtained with unfiltered alignments, we ran ClustalW with GUIDANCE filtering using all three trees and the same range of indel rates and MPLs as used for Figure 4; these results are included in Supplementary Figure S1. GUIDANCE filtering generally improved the performance of ClustalW alignments, resulting in sitewise performance comparable to (but not better than) unfiltered MAFFT alignments.

Filtering was less beneficial when applied to the more accurate PRANK_C alignments, with T-Coffee

reducing performance and GUIDANCE yielding only mild $TPR_{1\%}$ improvements. Importantly, GUIDANCE only showed improved performance at high divergence levels (e.g., MPL>1.6), well above those found in commonly-analyzed groups of species. Thus, the use of unfiltered $PRANK_C$ alignments would yield the best performance for detecting sitewise positive selection when analyzing protein-coding sequences across a wide range of indel rates and commonly encountered sequence divergence levels. The performance of the 'optimal' filter on $PRANK_C$ alignments suggests that mild further improvements to filtering strategies may be possible, but the potential for improvement is small and may be of little value.

Conclusions

In this paper, we investigated the performance of sitewise detection of positive selection under a range of tree sizes, indel rates, and divergence levels, using simulation parameters designed to approximate the analysis of mammalian and other typical protein-coding genes. We evaluated the ability of four alignment methods and three alignment filtering methods to produce alignments for detecting positively selected sites, using the FPR, TPR, and error-controlled $TPR_{1\%}$ of the sitewise detection of positive selection as our main measures of performance.

The simulation results showed that alignment error can have a measurable impact on the error rates and power of the sitewise detection of positive selection under all but the least difficult alignment conditions. We confirmed and extended the findings of Fletcher and Yang (2010) regarding the relative accuracy of different aligners, showing that $PRANK_C$ had the best performance and ClustalW had the worst performance for subsequent sitewise analysis. Notably, our simulations found that ClustalW produced more sitewise false positives than MAFFT or $PRANK_C$ even at low MPLs, suggesting that its use should be avoided even when analyzing closely-related sequences. $PRANK_C$, on the other hand, resulted in very low FPRs even at higher MPLs. In particular, when the number of sequences in the tree was large, $PRANK_C$'s sitewise FPRs were virtually indistinguishable from those of the true alignment.

An important observation regarding the size of tree analyzed was that the 6-taxon tree caused qualitatively similar problems (e.g., elevated FPRs and reduced TPRs) for all aligners, suggesting that poor performance is inevitable when analyzing a small number of moderately divergent sequences. The small amount of evolutionary information combined with the longer branch lengths makes alignment difficult and increases the tendency of misalignment to cause sitewise false positives. Thus, we reiterate the well-established recommendation to use large numbers of sequences when inferring sitewise positive selection (Anisimova et al., 2001, 2002). We additionally point out that when analyzing sequences with indels, the shape of the tree may matter as well: trees with long internal branches may be especially prone to false positives, as longer branches are more difficult to align.

The very low FPRs observed for $PRANK_C$ alignments conflicted somewhat with the results of Fletcher and Yang (2010), who found that the FPRs for the branch-site test were not under control even with $PRANK_C$ alignments. This apparent discrepancy can be explained by different sensitivities to alignment error: the branch-site test would yield false positives when misalignment causes apparent positive selection along only the foreground branch, while the SLR sitewise test would produce false positives

only when misalignment causes a signal of positive selection strong enough to overpower the non-positive signal throughout the tree. This effect stems from the different biological hypotheses tested by the two methods; their differential sensitivity to misalignment underscores the necessity of considering the biological sensitivity and robustness to alignment error when applying either of these tests to detect positive selection within an alignment.

Despite producing very low FPRs, PRANK_C alignments still resulted in an increased number of false negatives compared to the true alignment. We showed that some of these false negatives were possible to recover as true positives through alignment filtering, and we found that both the 'optimal' filter and GUIDANCE were able to successfully recover these false negatives at high divergence levels, resulting in small but measurable performance improvements over the unfiltered PRANK_C alignments.

The manual or automated adjustment of alignments has been thought by many to be an important step in evolutionary analyses due to fear of a high prevalence of misalignment-induced false positives. While this is true for some aligners, we find that more accurate alignment algorithms result in significantly fewer false positives in the subsequent detection of sitewise positive selection. This strongly reduces the beneficial effect of alignment filtering, so much so that current filtering methods are essentially unable to improve the performance of PRANK_C alignments when analyzed with SLR.

As a result, among the aligners and filters tested here, we recommend the use of unfiltered PRANK_C alignments for the detection of sitewise positive selection in all but the most divergent phylogenies. GUIDANCE showed some ability to improve sitewise power under difficult alignment conditions at high divergence levels, but this improvement was achieved largely through the recovery of false negatives as opposed to the elimination of false positives. For an analysis where the control of false positives is the primary concern, the added computational expense of running many bootstrap alignment replicates (as performed by GUIDANCE) may not be worth the possibility of a slight increase in power. The development of efficient and effective alignment filters, particularly with an eye towards recovering false negatives as identified in this study, will be an interesting area for future research.

Our simulations did not include fully biologically realistic models of spatial or temporal variability in the rate of indel formation or in the distribution of selective pressures (Whelan, 2008), but we do not expect that such heterogeneity would affect our main conclusions regarding the relative performance of different aligners and filters. The trends we observed were consistent across a wide range of parameter values, suggesting that they reflect fundamental differences in each method's ability to align or filter sequences as opposed to artifacts due to the relative simplicity of our simulations.

However, the apropriateness of our simulation scheme should be critically considered when evaluating our specific power and error rate estimates in the context of real-world data analysis. Functional differences between genes may influence the $\boldsymbol{\omega}$ distribution and indel rate—with some genes or domains showing lower or higher tolerance to indels and fewer or more neutrally-evolving sites than modeled in our simulations—making false positive results more or less likely. Second, species-level differences may also be important, as the efficacy of natural selection is highly dependent on effective population size (Ellegren, 2009). For example, proteins evolving in *Drosophila* species with a high effective population size should experience stronger positive and purifying selection than in mammals, potentially leading to

increased power and reduced error rates when compared to our simulations based on a mammalian-like $\boldsymbol{\omega}$ distribution.

A tangential but interesting observation from our simulations was that all three methods we tested for the sitewise detection of positive selection were highly conservative when analyzing alignments with a mammalian-like $\boldsymbol{\omega}$ distribution. SLR, for example, yielded FPRs well below the nominal 5% level. This was due to a mismatch between SLR's null model of neutral evolution ($\boldsymbol{\omega} = 1$) and our more realistic distribution of non-positive sites (where $\bar{\boldsymbol{\omega}} = 0.277$). We note that while the ability of SLR and PAML's sitewise models to distinguish between neutral evolution and positive selection in a well-controlled manner is important, the majority of protein sites evolve under moderate purifying selection. Future work on methods to adaptively adjust the cutoff thresholds to achieve better statistical control under non-neutral (and possibly unknown) $\boldsymbol{\omega}$ distributions could yield much greater power to detect positive selection while maintaining good control of error rates.

We showed here that even relatively simple evolutionary simulation experiments could sensitively assess the performance characteristics of different aligners, provide quantitative insight into the practical effects of alignment error, and suggest areas for future development of alignment and filtering methods. In the future, we expect the development of more realistic simulations for protein evolution—perhaps incorporating structurally-motivated and empirically validated models of mutation, indel formation and constraint—to further increase the applicability and accuracy of such experiments, and we believe that flexible and accessible simulation programs such as INDELible (Fletcher and Yang, 2009) and PhyloSim (Sipos et al., 2011) will play an important role in the quantitative assessment of alignment algorithms and alignment-dependent comparative analyses.

As genomes rapidly accumulate in the databases and large-scale analyses become the norm, we hope that the development and application of alignment methods, which are arguably the most important step in any evolutionary analysis, will be based on a rigorous understanding of their behavior and performance when applied to a wide variety of evolutionary analyses.

Acknowledgments

GJ was funded by a Gates Cambridge Trust Scholarship and is a member of Darwin College, University of Cambridge.

Literature Cited

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature **467**:1061–1073.

Aguileta, G., G. Refrégier, R. Yockteng, E. Fournier and T. Giraud. 2009. Rapidly evolving genes in pathogens: Methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. Infection, Genetics and Evolution **9**:656–670.

Anisimova, M., J. P. Bielawski and Z. Yang. 2001. Accuracy and power of the likelihood ratio test in detecting Adaptive molecular evolution. Molecular Biology and Evolution **18**:1585–1592.

Anisimova, M., J. P. Bielawski and Z. Yang. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. Molecular Biology and Evolution **19**:950–958.

Arbiza, L., S. Duchi, D. Montaner, J. Burguet, D. P. Uceda, A. P. Lucena, J. Dopazo and H. Dopazo. 2006. Selective pressures at a codon-level predict deleterious mutations in human disease genes. Journal of Molecular Biology **19**:1390–1404.

Benner, S. A., M. A. Cohen and G. H. Gonnet. 1993. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. Journal of Molecular Biology **229**:1065–1082.

Cartwright, R. A. 2009. Problems and solutions for estimating indel rates and length distributions. Molecular Biology and Evolution **26**:473–480.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution **17**:540–552.

de la Chaux, N., P. W. Messer and P. F. Arndt. 2007. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. BMC Evolutionary Biology **7**:191.

Dessimoz, C. and M. Gil. 2010. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biology **11**:R37.

Dwivedi, B. and S. R. Gadagkar. 2009. Phylogenetic inference under varying proportions of indel-induced alignment gaps. BMC Evolutionary Biology **9**:211.

Ellegren, H. 2009. A selection model of molecular evolution incorporating the effective population size. Evolution **63**:301–305.

Fletcher, W. and Z. Yang. 2009. INDELible: a flexible simulator of biological sequence evolution. Molecular Biology and Evolution **26**:1879–1888.

Fletcher, W. and Z. Yang. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Molecular Biology and Evolution **27**:2257–2267.

Green, P. 2007. 2x genomes: does depth matter? Genome Research **17**:1547–1549.

Hillier, L., W. Miller, E. Birney et al. (178 co-authors). 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432**:695–716.

Katoh, K., K.-i. Kuma, H. Toh and T. Miyata. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research **33**:511–518.

Kimura, M. and T. Ohta. 1974. On some principles governing molecular evolution. Proceedings of the National Academy of Sciences **71**:2848–2852.

Lin, Y.-S., W.-L. Hsu, J.-K. Hwang and W.-H. Li. 2007. Proportion of solvent-exposed amino acids in a protein and Rate of protein evolution. Molecular Biology and Evolution **24**:1005–1011.

Lindblad-Toh, K., M. Garber, O. Zuk et al. (64 co-authors). 2011. A high-resolution map of evolutionary constraint in the human genome based on 29 eutherian mammals (submitted) .

Löytynoja, A. and N. Goldman. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proceedings of the National Academy of Sciences **102**:10557–10562.

Löytynoja, A. and N. Goldman. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science **320**:1632–1635.

Markova-Raina, P. and D. Petrov. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 drosophila genomes. Genome Research **21**:863–874.

Massingham, T. and N. Goldman. 2005. Detecting amino acid sites under positive selection and purifying Selection. Genetics **169**:1753–1762.

Nei, M., Y. Suzuki and M. Nozawa. 2010. The neutral theory of molecular evolution in the genomic era. Annual Review of Genomics and Human Genetics **11**:265–289.

Nikolaev, S., J. I. Montoya-Burgos, E. H. Margulies, N. C. Program, J. Rougemont, B. Nyffeler and S. E. Antonarakis. 2007. Early history of mammals is elucidated with the ENCODE multiple species sequencing data. PLoS Genetics **3**:e2.

Notredame, C. 2007. Recent evolutions of multiple sequence alignment algorithms. PLoS Computational Biology **3**:e123.

Notredame, C. and C. Abergel. 2003. Using multiple alignment methods to assess the quality of genomic data analysis. Bioinformatics and Genomes: Current Perspectives. Horizon Scientific Press, Wymondham, UK 30–55.

Notredame, C., D. G. Higgins and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology **302**:205–217.

Ogden, T. H. and M. S. Rosenberg. 2006. Multiple sequence alignment accuracy and phylogenetic inference. Systematic Biology **55**:314–328.

Ogurtsov, A. Y., S. Sunyaev and A. S. Kondrashov. 2004. Indel-based evolutionary distance and mouse-human divergence. Genome Research **14**:1610–1616.

Penn, O., E. Privman, G. Landan, D. Graur and T. Pupko. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. Molecular Biology and Evolution **27**:1759–1767.

Ramsey, D. C., M. P. Scherrer, T. Zhou and C. O. Wilke. 2011. The relationship between relative solvent accessibility and evolutionary Rate in protein evolution. Genetics **188**:479–488.

Schneider, A., A. Souvorov, N. Sabath, G. Landan, G. H. Gonnet and D. Graur. 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biology and Evolution **1**:114–118.

Siepel, A., G. Bejerano, J. S. Pedersen et al. (16 co-authors). 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research **15**:1034–1050.

Sipos, B., T. Massingham, G. Jordan and N. Goldman. 2011. Phylosim–monte carlo simulation of sequence evolution in the r statistical computing environment. BMC Bioinformatics **12**:104.

Smith, J. M. 1970. Natural selection and the concept of a protein space. Nature **225**:563–564.

Studer, R., L. Duret, S. Penel and M. R. Rechavi. 2008. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. Genome Research **18**:1393–1402.

Talavera, G. and J. Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology **56**:564–577.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature **447**:799–816.

Thompson, J. D., D. G. Higgins and T. J. Gibson. 1994. CLUSTAL w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research **22**:4673–4680.

Whelan, S. 2008. Spatial and temporal heterogeneity in nucleotide sequence evolution. Molecular Biology and Evolution **25**:1683–1694.

Wolf, J. B., A. Künstner, K. Nam, M. Jakobsson and H. Ellegren. 2009. Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection. Genome Biology and Evolution **1**:308–319.

Wong, K. M., M. A. Suchard and J. P. Huelsenbeck. 2008. Alignment uncertainty and genomic analysis. Science **319**:473–476.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution **24**:1586–1591.

Yang, Z. and R. Nielsen. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. Journal of Molecular Evolution **46**:409–418.

Yang, Z., R. Nielsen, N. Goldman and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

Yang, Z., W. S. W. Wong and R. Nielsen. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Molecular Biology and Evolution **22**:1107–1118.

Table 1: Parameter Values Used in Simulations

Taxa	Tree	MPL	Indel Size Distribution	Mean Indel Length (Std. Dev.)	Indel Rate	$\boldsymbol{\omega}$ Distribution	Mean $\boldsymbol{\omega}$	$p(\boldsymbol{\omega} > 1)$
6	Artificial		power law			lognormal		
17	β-globin	0.05–2.0	decay: 1.8	3.33 (5.51)	0–0.2	log mean: -1.864	0.277	0.06
44	Vertebrates		max length: 40			log SD: 1.201		

NOTE.—MPL is the mean path length of the tree in units of substitutions per synonymous site. Indel lengths are measured in units of codons, and the indel rate is defined as the number of insertion & deletion events per substitution.

Table 2: Genome-wide Divergence Estimates for Commonly Analyzed Eukaryotes

Species	Pairwise dS	Root-to-tip dS	Reference
Human-Chimp	0.01	(0.005)	Nei et al., 2010
Human-Mouse	0.43	(0.215)	Nei et al., 2010
Human-Mouse	0.5 - 0.8	(0.25 - 0.4)	Ogurtsov et al., 2004
Human-Chicken	0.9	(0.45)	Nei et al., 2010
Human-Chicken	1.66	(0.83)	Hillier et al., 2004
Human-Zebrafish	1.38	(0.69)	Nei et al., 2010
Vertebrates	–	0.75	Siepel et al., 2005
Drosophila	–	1.0	Siepel et al., 2005
Yeasts	–	1.25	Siepel et al., 2005

NOTE.—The root-to-tip dS is equivalent to the MPL (mean path length) used in our simulations. For two-species comparisons where the pairwise dS was given, the root-to-tip dS was calculated as half of the pairwise dS and is included in parentheses.

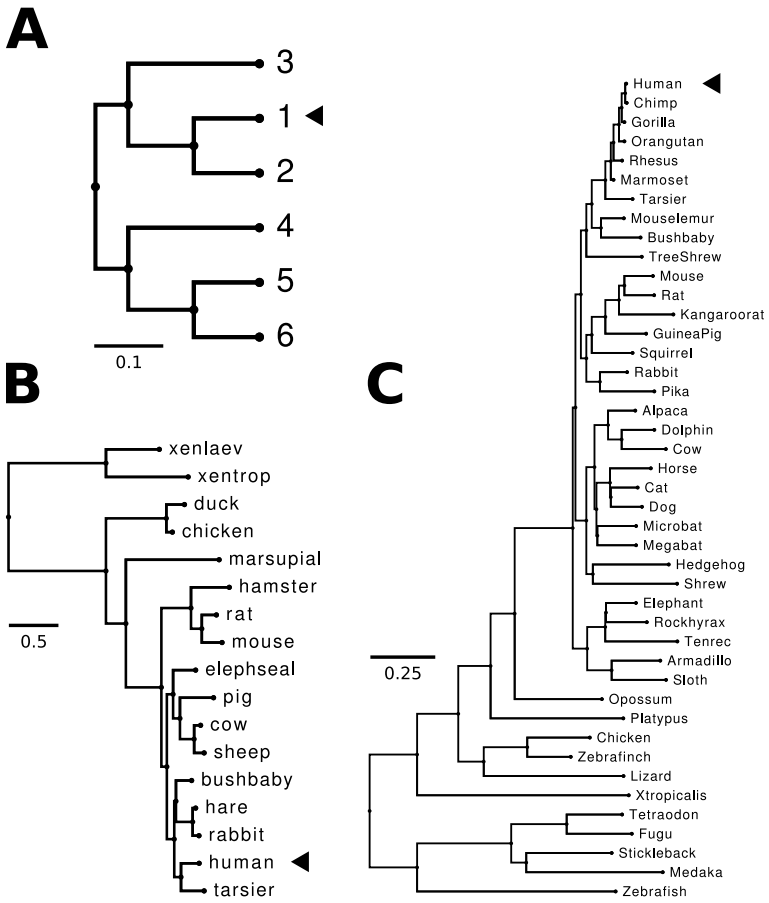

Figure 1: Phylogenetic trees used for simulation and analysis. The original scale for each tree is indicated by a scale bar, but trees were scaled to equal mean path length (MPL) divergence levels for simulation. (a) A 6-taxon artificial tree used in previous simulations (Anisimova et al., 2001; Massingham and Goldman, 2005). (b) A tree estimated from β-globin genes of 17 vertebrates and used in previous empirical analyses and simulation studies (Anisimova et al., 2001, 2002). (c) The 44-species tree used by the ENCODE project (The ENCODE Project Consortium, 2007; Nikolaev et al., 2007). The nodes indicated by arrows were used as the reference species when comparing the true and inferred alignment (see Methods).

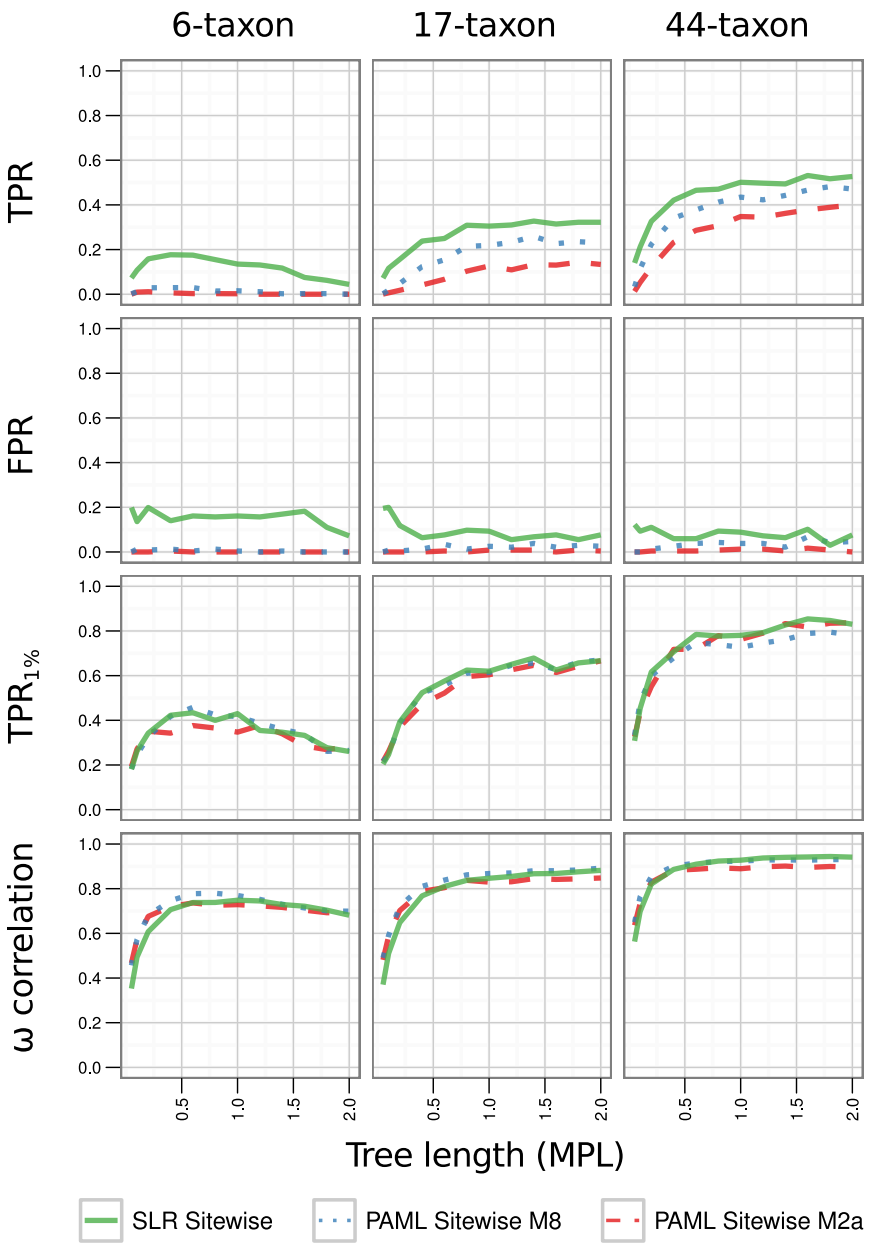

Figure 2: Alignments were simulated without indels for three tree shapes and analyzed with SLR, PAML M8, or PAML M2a. Fifty replicate alignments were simulated for each data point. The performance of each analysis method, as measured by four summary statistics, is plotted as a function of the mean path length (MPL) divergence. From top to bottom: true positive rate (TPR) at the recommended cutoff threshold (0.95 for PAML and 3.84 for SLR); false positive rate (FPR) at the recommended cutoff threshold; TPR at a 1% FPR threshold; Pearson's correlation coefficient between the true and inferred sitewise ω.

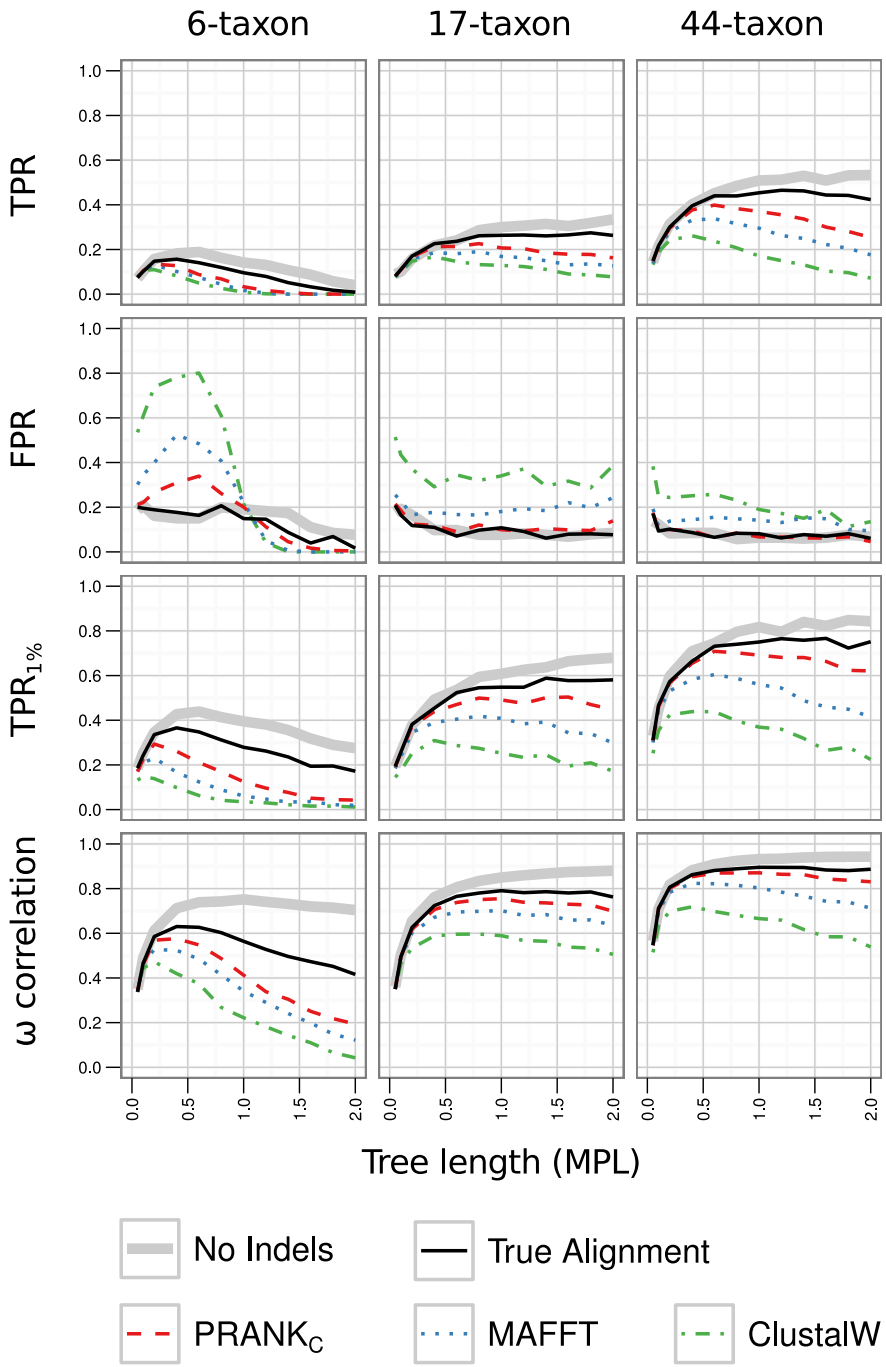

Figure 3: Sequences were simulated without indels (solid gray lines) or with indels (solid black and textured lines) using one of three tree shapes, aligned with one of three aligners, and analyzed with SLR; true alignments were separately analyzed with SLR (solid black lines). One hundred replicate alignments were simulated for each data point. The performance of each dataset, as measured by four summary statistics, is plotted as a function of the mean path length (MPL) divergence. From top to bottom: true positive rate (TPR) at the recommended cutoff threshold; false positive rate (FPR) at the recommended cutoff threshold; TPR at a 1% FPR threshold; Pearson's correlation coefficient between the true and inferred ω.

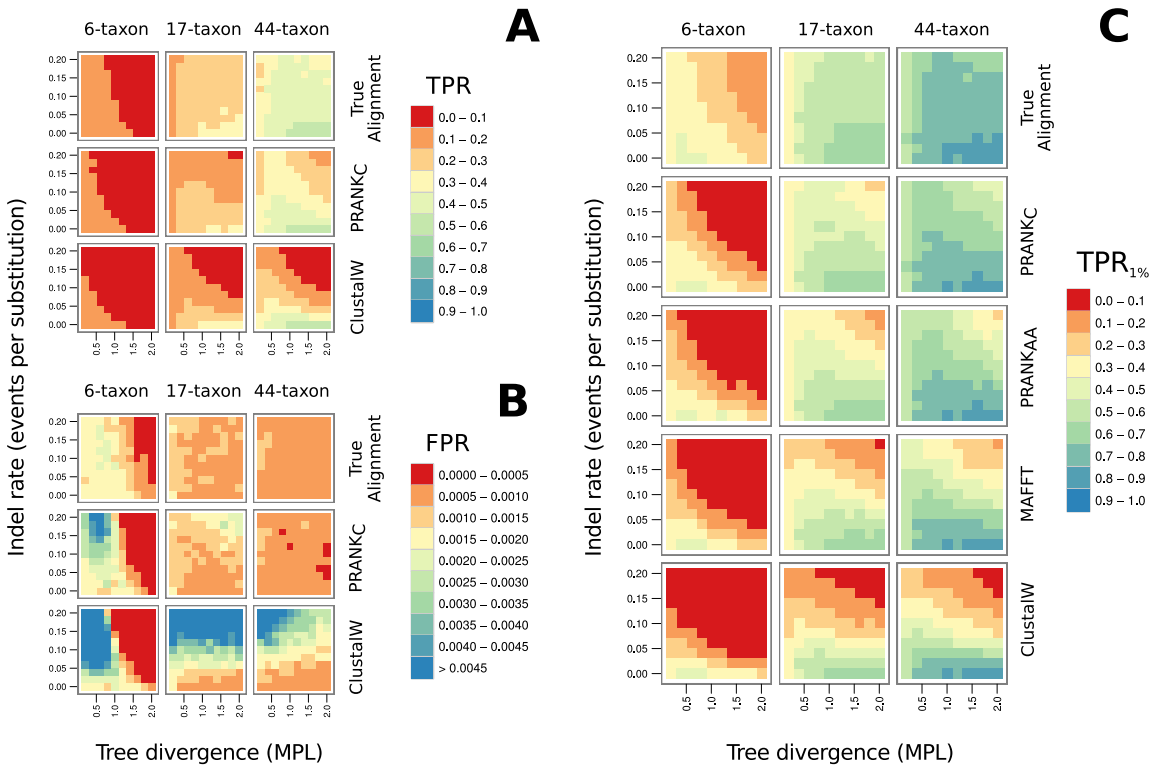

Figure 4: Sequences were simulated with indels using one of three tree shapes (t-, 17- or 44-taxon) and a range of indel rates and mean path length (MPL) divergence levels. Alignments were inferred with one of four aligners (ClustalW, MAFFT, PRANK_{AA}, PRANK_{C}) and analyzed with SLR; true alignments were separately analyzed with SLR. One hundred replicates were simulated for each set of conditions. Each cell is colored according to the performance at a given (indel rate, MPL) pair as measured by one of three summary statistics: (a) the true positive rate (TPR) at the recommended cutoff threshold, (b) the false positive rate (FPR) at the recommended cutoff threshold, or (c) the TPR at a 1% FPR threshold. Results for MAFFT and PRANK_{AA} are omitted from (a) and (b); as in (c) they show characteristics intermediate between ClustalW and PRANK_{C}.

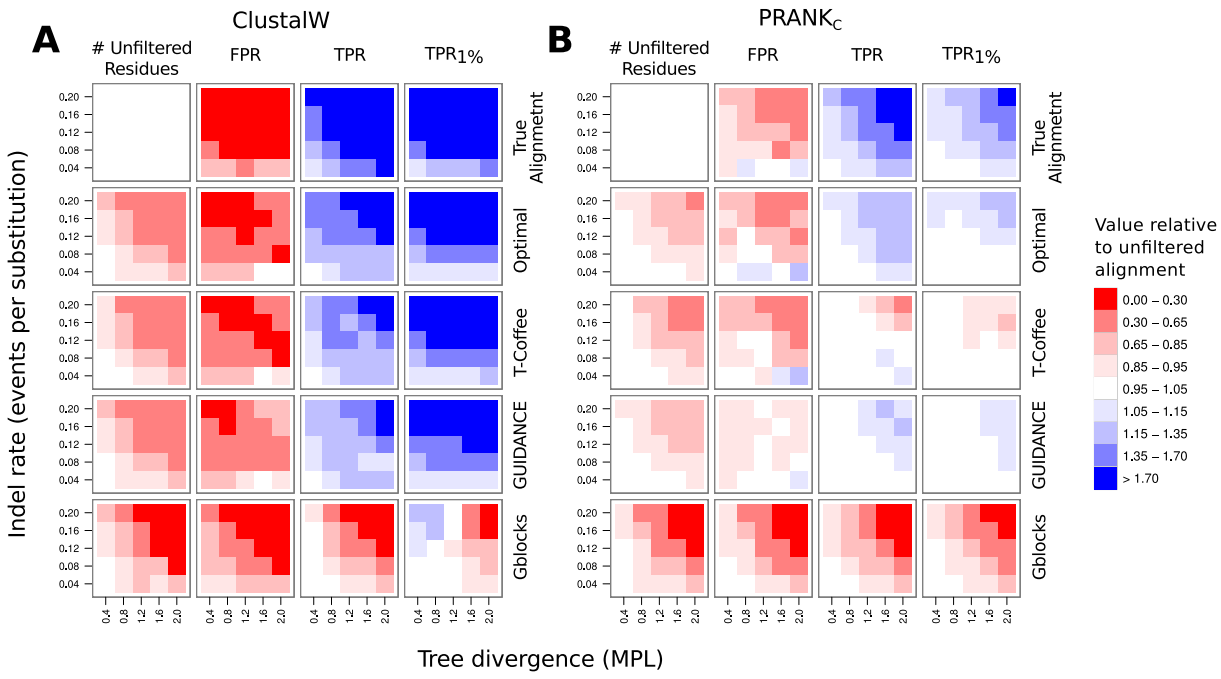

Figure 5: Sequences were simulated using the 17-taxon tree and a range of indel rates and mean path length (MPL) divergence levels. Alignments were inferred using (a) ClustalW or (b) PRANK$_C$, either left unfiltered or filtered with one of four alignment filters (Optimal, T-Coffee, GUIDANCE, Gblocks), and analyzed with SLR; true alignments were left unfiltered and separately analyzed with SLR. One hundred and fifty replicates were simulated for each set of conditions. Cells are colored according to the ratio of the performance of the indicated filter to the performance of the unfiltered ClustalW or PRANK$_C$ alignment as measured by one of four summary statistics. In columns from left to right: the number of unfiltered (i.e., non-N) residues remaining in the alignment; the false positive rate (FPR) at the recommended cutoff threshold; the true positive rate (TPR) at the recommended cutoff threshold; the TPR at a 1% FPR threshold (TPR$_{1\%}$). Note that the maximum percentage of residues removed by filtering was capped at 50% for all methods except Gblocks.