Letter

Limited utility of residue masking for positive-selection inference

Stephanie J. Spielman[1]* and Eric T. Dawson[1] and Claus O. Wilke[1]

Address:
[1]Department of Integrative Biology, Center for Computational Biology and Bioinformatics, and Institute of Cellular and Molecular Biology. The University of Texas at Austin, Austin, TX 78712, USA.

*Corresponding author
Email: stephanie.spielman@utexas.edu

Manuscript type: Letter

Keywords: multiple sequence alignment, alignment filters, positive-selection inference, sequence simulation

Abstract

Multiple sequence alignment (MSA) errors are known to reduce accuracy in positive-selection inference. In response to this issue, some have suggested filtering MSAs before conducting further analyses. One widely-used filter, Guidance, generates site-specific MSA confidence scores, allowing users to remove positions of low confidence. Studies investigating this filter's utility for positive-selection inference have yielded inconsistent results; some have demonstrated that Guidance substantially improved accuracy, but others have found that Guidance affected accuracy minimally. Motivated by these discrepancies, we have conducted a extensive simulation-based study to characterize how Guidance impacts positive-selection inference on protein-coding sequences of realistic divergence levels. We particularly investigated whether novel scoring algorithms, which phylogenetically corrected confidence scores, and a new gap-penalization score-normalization scheme improved upon Guidance's performance. Instead, we found that no filter, including the original Guidance, substantially improved positive-selection inference across multiple inference methods. Although Guidance-based filters were able to improve performance in certain circumstances, filters worsened inference in many scenarios as well. **note that abstract is 7 words too long as of now**

Constructing a multiple sequence alignment (MSA) represents the most fundamental step in most molecular evolution studies, including phylogenetic reconstruction and evolutionary rate inference. Recently, several studies have shown that poor MSA quality can hinder accuracy in positive-selection inference (Schneider et al. 2009; Fletcher and Yang 2010; Markova-Raina and Petrov 2011). As a consequence, some have advocated that users filter MSAs before subsequent analyses to remove putatively-poorly aligned regions (Privman et al. 2012; Jordan and Goldman 2012), thereby reducing noise and maximizing signal in the MSA.

One filter, known as Guidance (Penn et al. 2010), is widely used in positive-selection inference. Guidance derives a confidence score for each MSA position by sampling variants in the guide tree used to construct progressive alignments. Using these confidence scores, users can mask positions that score below a set threshold, effectively removing residues that may produce misleading signal. Unfortunately, studies investigating Guidance's utility in positive-selection inference have produced conflicting findings. While one study by Privman et al. (2012) found that Guidance dramatically improved accuracy and power, a separate study by Jordan and Goldman (2012) found that Guidance affected positive-selection inference modestly, if at all. In particular, both Privman et al. (2012) and Jordan and Goldman (2012) concluded that Guidance filtering improved inference primarily at high insertion/deletion (indel) rates (e.g. 10%) and/or high divergence levels (e.g. mean-path-length of 1.8). However, it is important to recognize that typical positive-selection study rarely contain sequences separated by such high divergences.

In sum, Privman et al. (2012) strongly advocated Guidance's use, while Jordan and Goldman (2012) emphasized relying primarily on robust MSA construction methods. To reconcile these distinct recommendations, we have conducted an extensive simulation-based study to fully elucidate how the Guidance filter affects positive-selection inference. In particular, we examined the potential benefits to modifying the Guidance scoring scheme in several ways. First, we assessed whether two novel algorithms that correct Guidance scores for the sequences' phylogenetic relationships could improve upon the original Guidance algorithm. Briefly, the first phylogenetically-corrected method incorporated a weight, as calculated by BranchManager (Stone and Sidow 2007), for each MSA sequence, and the second method incorporated patristic distances (the sum of branch lengths between two taxa), as calculated using Python phylogenetics package Dendropy (Sukumaran and Holder 2010). We refer to these methods, respectively, BMweights and PDweights. Moreover, we tested a new gap-penalization score-normalization scheme, which scales a given residue's score according to the number of gaps in its column. This strategy naturally assigned lower scores to residues in gappy columns, thereby capturing the inherent unreliability of such regions. We refer to filters using the gap-penalization scheme as GuidanceP, BMweightsP, and PDweightsP. To assess the performance of these novel algorithms, we reimplemented the Guidance software (available at https://github.com/clauswilke/alignment_filtering).

We simulated protein-coding sequences using Indelible (Fletcher and Yang 2009) according to two different selective profiles: H1N1 influenza hemagluttinin (HA), which featured a mean $dN/dS = 0.37$, and HIV-1 envelope protein subunit GP41, which featured a mean $dN/dS = 0.89$. We used these two selective profiles because, while both genes are known to contain positively selected regions (Bush et al. 1999; Frost et al. 2001; Bandawe et al. 2008; Meyer and Wilke 2012), the majority of sites in HA are either under strong purifying or positive selection, whereas the GP41 subunit contains a much larger proportion of sites near neutral, making positive-selection inference more challenging. For each selective profile, we simulated 100 MSA replicates along each of four different gene trees consisting of 11 (Spielman and Wilke 2013), 26 (Spielman and Wilke 2013), 60 (Yang et al. 2011), and 158 (Betancur-R et al. 2013) taxa, yielding a total of 800 simulated MSAs. All sequences were simulated with a 5% indel rate, which has been shown to be typical of mammalian genomes (Cooper et al. 2004).

We processed the unaligned amino-acid sequences with our Guidance reimplementation using the aligner MAFFT L-INS-I (linsi) (Katoh et al. 2002, 2005) and calculated confidence scores for all inferred MSAs using each of the six scoring algorithms detailed above. When filtering MSAs with Guidance-based methods, one must select a specific score cutoff below which to mask residues. We investigated any potential differences resulting from the four masking cutoffs 0.3, 0.5, 0.7, and 0.9 (see Supplementary Materials for details), and found that the 0.9 cutoff, at times, performed much worse than the other cutoffs, which performed similarly to one another (Table S1). Therefore, we masked positions with scores below 0.5, the same threshold as used by Jordan and Goldman (2012). We used two methods to infer positive selection: FUBAR (Murrell et al. 2013), as implemented in HyPhy (Kosakovsky Pond et al. 2005), and the widely-used M8 model in PAML (Yang et al. 2000; Yang 2007). Phylogenies used in positive-selection inference were constructed in RAxMLv7.3.0 using the "PROTGAMMAWAG" model (Stamatakis 2006). Note that while we processed all MSAs with FUBAR, we did not infer positive selection with PAML for MSAs of 158 taxa due to prohibitive runtimes. A detailed description of all methods, including the Guidance software reimplementation, is available in Supplementary Material.

Guidance-based filters have a minimal effect on positive-selection inference

We first compared the resulting true positive rates (TPRs) of positive-selection inference between each filtered MSA and its corresponding unfiltered MSA. We chose to primarily compare the TPRs, as opposed to the false positive rates (FPRs), as the FPRs we recovered were exceedingly small, never surpassing an average of 1% across simulation sets and inference methods. TPRs were calculated using the true evolutionary rates assigned during sequence simulation. For this analysis, we considered sites to be positively selected if the given inference method (i.e. FUBAR or PAML) returned a posterior probability ≥ 0.90. For each simulation set, we fit a series of mixed-effects models using the R package lme4 (Bates et al. 2012). Each model consisted of TPR as the response variable, filtering algorithm (including unfiltered and the six filtering algorithms) as a fixed effect, and simulation count as a random effect, which accounted for the paired structure of our analysis.

Table 1 highlights key findings from these models, as well as some additional information regarding the simulated sequences. We found that, in general, there was no significant mean TPR difference among filters within a given normalization scheme. In other words, Guidance, BMweights, and PDweights performed similarly, and the three gap-penalization filtering algorithms GuidanceP, BMweightsP, and PDweightsP performed similarly. Therefore, Table 1 displays results for only Guidance and GuidanceP, and Table S1 contains results for all filtering algorithms. Table 1 demonstrates that, for the majority of datasets studied here, Guidance-based filtering had no significant effect on the mean TPR, relative to an unfiltered MSA. However, in several cases, MSA filtering did significantly increase mean TPR, but in other cases MSA filtering significantly decreased mean TPR. Importantly, for all cases in which MSA filtering significantly changed mean TPR, the effect magnitude was very small.

Moreover, inference methods did not respond to MSA filtering consistently, as demonstrated in Figure 1, which gives a graphical representation of the linear models' results for the 26- and 60-sequence simulation sets, for both HA and GP41 selective profiles. On one hand, FUBAR yielded consistent mean TPR trends among filters (Guidance mean TPR tended to be higher than both unfiltered and GuidanceP), but the majority of these trends are statistically insignificant. PAML's behavior in response to MSA filtering, on the other hand, varied substantially among simulation conditions. For instance, the largest TPR improvement we recovered in this study was for the HA 26-sequence simulation set; when positive-selection was inferred with PAML on MSAs filtered with GuidanceP, TPR improved by roughly 4%, relative to unfiltered MSAs. However, GuidanceP significantly reduced mean TPR for the GP41 60-sequence simulation set, also as inferred with PAML. Therefore, there does not appear to be a strong universal trend dictating when MSA filtering will be helpful. We do note, however, that the GuidanceP filter was more likely to reduce mean TPR than was the Guidance filter, which only significantly reduced TPR in one case (Table 1). Additionally, for both simulation sets of 158 sequences, all filtering algorithms increased TPR by an average of 2%, relative to unfiltered MSAs. However, as we did not process these sequences with PAML, we caution that these results may not be easily extrapolated to inference methods other than FUBAR.

We additionally used receiver operating characteristic (ROC), which are commonly used to evaluate the performance of binary classification methodscurves to qualitatively assess differences in positive-selection inference for unfiltered versus filtered MSAs. Importantly, this analysis did not bias results to those obtained from a single posterior probability threshold for calling positive-selected sites, but instead considered the overall methodological performance. ROC curves for the HA and GP41 60-sequence simulation sets are shown in Figure 2. In each sub-plot, the top curve gives results from the HA selective profile, and the bottom curve gives results from the GP41 selective profile. The left-hand panels display the entire ROC curves, while the right-hand panels display only the region of the curves with relatively low FPRs.

Several trends emerge from Figure 2. First, while positive-selection inference power was universally greater for the HA simulation sets than for the GP41 simulation sets, Guidance-based filters behaved nearly identically for both selective profiles. That power was higher for the HA simulation sets was largely expected, given that the GP41 selective profile featured a greater proportion of sites with dN/dS near 1, which are more difficult to classify. Second, across the entire span of the ROC curve, there is no dramatic difference in area between curves for the unfiltered and filtered MSAs, although, MSAs processed with gap-penalization algorithms did, at certain FPR levels, perform worse than did both unfiltered and Guidance-filtered MSAs. Third, filtering did confer substantial power boosts at low FPR rates, as seen in the right-hand panels, in particular when PAML was used as the inference method. Even so, these benefits only existed at FPR levels of roughly 1% - 4%, above which any benefits quickly dissipated. Outside of this narrow FPR region, filtered MSAs either yielded results comparable to or worse than those of unfiltered MSAs. Importantly, our linear models which considered positively-selected sites only at a 0.9 posterior probability threshold all yielded mean FPRs under 1%, below the region where MSA filtering increases power. Therefore, while Guidance-based MSA filtering can confer power to detect positive-selection, it does so in an extremely limited region and it therefore not robust to varying FPR levels. ROC curves for all other simulation sets are available in Figures S1 and S2, and yielded broadly consistent results to those described here.

Discussion and Conclusions

The primary goal of using the Guidance, or similar, MSA filters is to remove excessive noise while maintaining informative data. Ideally, our study would have recovered a clear set of circumstances for which Guidance-based filters could consistently achieve this goal. Instead, we were unable to recover any universal trend to ensure that filtering will only remove noise without potentially compromising valuable data. While Guidance-based filters did improve positive-selection inference under certain conditions, they also diminished power to detect positively selected sites under other circumstances. We do, however, emphasize that MSA filtering universally improved positive-selection inference for the largest simulation sets of 158 taxa when analyzed with FUBAR, although the magnitude was very small. Conversely all filters significantly reduced accuracy for the GP41 11-sequence simulation set, as analyzed with PAML. Results for the 26- and 60-sequence simulations sets, unfortunately, did not feature statistically robust trends. Therefore, while Guidance-based filtering certainly had the potential to boost power, there was no apparent way to predict whether a given data set would experience this benefit, or worse, have valuable information removed. These broad inconsistencies highlighted that Guidance is not a particularly robust method when applied to positive-selection inference, and does not need to be a requisite step in sequence analyses.

Our study focused primarily on divergence levels representative of realistic protein-coding data typically used in positive-selection inference. It is possible that we were unable to recover some of the stronger benefits to Guidance filters when applied to highly diverged data. However, as seen in Table 1, our MSAs contained gaps in up to 60% of columns, meaning that constructing MSAs on our datasets was not a trivial task, and portions which were difficult to align certainly existed (Table S2). Again, a given positive-selection study is unlikely to feature a higher proportion of indels than our MSAs, but should such a situation arise, it is possible that Guidance filtering will have a more substantial effect (Privman et al. 2012).

In addition, we note that, for nearly all simulation cases, FUBAR outperformed PAML both in positive-selection TPR and runtime. Each FUBAR inference completed in under 20 minutes, but a single PAML inference took between two hours and a week to complete.

Interestingly, incorporating phylogenetic information into the scoring algorithm generally performed the same as did the original Guidance algorithm. This result indicated the minimal benefits that MSA filtering in this manner produced at all. Were the original Guidance to offer robust improvements in positive-selection detection, one might expect that our more statistically controlled approach would boost the method's performance. However, as we have found that masking individual positions in an MSA only marginally affected positive-selection inference in the first place, the algorithmic changes we implemented might not be expected to have a dramatic effect.

In sum, we have found that, while MSA filtering offered some benefits to positive-selection inference, those improvements were marginal at best. With such a minimal effect, MSA filtering could easily decrease accuracy in a given positive selection study. Indeed, we noted that using a stringent masking cutoff of 0.9 for algorithms normalized with our gap-penalization strategy, or with a large sequence set, resulted in extreme decreases in TPR relative to an unfiltered alignment. Choosing a low filtering threshold was necessary to achieve any improvement in positive-selection inference.

Overall, we cannot unequivocally recommend the use of a Guidance-based MSA filter when inferring positive selection. Once an MSA has been constructed, it does not seem that much can be done to eliminate any misleading information. In sum, we recommend that users employ high-quality alignment and inference methods, in which error can be minimized as much as possible without necessitating post-hoc correction. Moreover, should users decide to filter their MSAs, we recommend using a lenient cutoff (≤ 0.5) to avoid potentially removing signal.

Acknowledgements

This work was supported in part by ARO Grant W911NF-12-1-0390 and in part by the National Institutes of Health grant R01 GM088344 to COW. This material is based in part upon work supported by the National Science Foundation under Cooperative Agreement No. DBI-0939454. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors thank Eyal Privman for MS comments and Sergei L Kosakovsky Pond for helpful comments and FUBAR(?).

References

Bandawe G, Martin D, Treurnicht F, Mlisana K, Abdool Karim S, Williamson C, The CAPRISA 002 Acute Infection Study Team. 2008. Conserved positive selection signals in gp41 across multiple subtypes and difference in selection signals detectable in GP41 sequences sampled during acute and chronic HIV-1 subtype c infection. Virology Journal 5:141.

Bates D, Maechler M, Bolker B. 2012. lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0.

Betancur-R R, Li C, Munroe T A, Ballesteros J A, Orti G. 2013. Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). Systematic Biology 62(5):763–785.

Bush R, Bender C, Subbarao K, Cox N, Fitch W. 1999. Predicting the evolution of human influenza a. Science 286:1921–1925.

Cooper G, Brudno M, Stone E, Dubchak I, Batzoglou S, Sidow A. 2004. Characterization of evolutionary rates and constraints in three mammalian genomes. Genome Research 14:539 – 548.

Fletcher W, Yang Z. 2009. INDELible: A Flexible Simulator of Biological Sequence Evolution. Molecular Biology and Evolution 26(8):1879–1888.

Fletcher W, Yang Z. 2010. The Effect of Insertions, Deletions, and Alignment Errors on the Branch–Site Test of Positive Selection. Molecular Biology and Evolution 27(10):2257–2267.

Frost S, Gunthard H, Wong J, Havlir D, Richman D, Brown A. 2001. Evidence for positive selection driving the evolution of HIV-1 env under potent antiviral therapy. Virology 282:250–258.

Hothorn T, Bretz F, Westfall P. 2008. Simultaneous inference in general parametric models. Biometrical Journal 50(3):346–363.

Jordan G, Goldman N. 2012. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol Biol Evol 29:1125–1139.

Katoh K, Kuma K I, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518.

Katoh K, Misawa K, Kuma K I, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066.

Kosakovsky Pond S L, Frost S D W, Muse S V. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 12:676–679.

Markova-Raina P, Petrov D. 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(6):863–874.

Meyer A G, Wilke C O. 2012. Integrating sequence variation and protein structure to identify sites under selection. Mol Biol Evol 30:36–44.

Murrell B, Moola S, Mabona A, Weighill T, Scheward D, Kosakovsky Pond S L, Scheffler K. 2013. FUBAR: A Fast, Unconstrained Bayesian AppRoximation for inferring selection. Molecular Biology and Evolution 30:1196–1205.

Penn O, Privman E, Landan G, Graur D, Pupko T. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. Mol Biol Evol 27:1759–1767.

Privman E, Penn O, Pupko T. 2012. Improving the performance of positive selection inference by filtering unreliable alignment regions. Mol Biol Evol 29:1–5.

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet G H, Graur D. 2009. Estimates of positive darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biology and Evolution 1(0):114–118.

Spielman S J, Wilke C O. 2013. Membrane environment imposes unique selection pressures on transmembrane domains of G protein–coupled receptors. Journal of Molecular Evolution 76:172–182.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:21:2688–2690.

Stone E, Sidow A. 2007. Constructing a meaningful evolutionary average at the phylogenetic center of mass. BMC Bioinformatics 8:222.

Sukumaran J, Holder M T. 2010. DendroPy: A python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Yang Y, Maruyama S, Sekimoto H, Sakayama H, Nozaki H. 2011. An extended phylogenetic analysis reveals ancient origin of "non-green" phosphoribulokinase genes from two lineages of "green" secondary photosynthetic eukaryotes: Euglenophyta and Chlorarachniophyta. BMC Research Notes 4:330.

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

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

Table 1: Summary statistics for effect of masking.

			Mean True Positive Rate				
Profile	Num	Method	True	Unfiltered	Guidance	GuidanceP	Percent Gaps
HA	11	FUBAR	0.093	0.084	0.085 (1.33%)	0.085 (0.74%)	11.9%
		PAML	0.086	0.082	0.081 (-0.48%)	0.081 (-0.60%)	
	26	FUBAR	0.252	0.227	0.229 (0.84%)	0.226 (-0.38%)	26.0%
		PAML	0.209	0.176	**0.178 (1.36%)****	**0.183 (4.04%)****	
	60	FUBAR	0.551	0.474	0.479 (0.99%)	**0.464 (-2.16%)***	59.4%
		PAML	0.422	0.347	0.342 (-1.68%)	0.337 (-2.92%)	
	158	FUBAR	0.515	0.458	**0.467 (1.89%)***	**0.468 (2.12%)***	50.5%
GP41	11	FUBAR	0.062	0.058	0.057 (-1.55%)	0.057 (-1.21%)	11.6%
		PAML	0.096	0.098	**0.095 (-3.49%)****	**0.095 (-3.80%)****	
	26	FUBAR	0.216	0.196	**0.20 (1.89%)***	0.197 (0.36%)	31.3%
		PAML	0.237	0.216	0.220 (1.54%)	0.217 (0.244%)	
	60	FUBAR	0.359	0.308	**0.313 (1.77%)***	0.304 (-1.16%)	58.1%
		PAML	0.341	0.304	0.302 (-0.77%)	**0.296 (-2.71%)****	
	158	FUBAR	0.348	0.320	**0.325 (1.77%)****	**0.326 (2.02%)****	48.4%

NOTE.— Profile: selective profile used to simulate sequences; Num: number of taxa. Method: positive-selection inference method used. Significance levels: $^{**}P < 0.001$; $^{*}P < 0.01$. Mean TPR values shown in bold represent those which are significantly different from the respective unfiltered MSA mean TPR. Values shown in parentheses refer to the average TPR percent change from the respective unfiltered MSA. All significance levels were corrected for multiple comparisons using the R multcomp package (Hothorn et al. 2008). Note that the true MSAs were not included in the linear models but are shown here for comparative purposes. Percent gaps were calculated from unfiltered alignments as the total number of gaps divided by the total number of MSA positions.

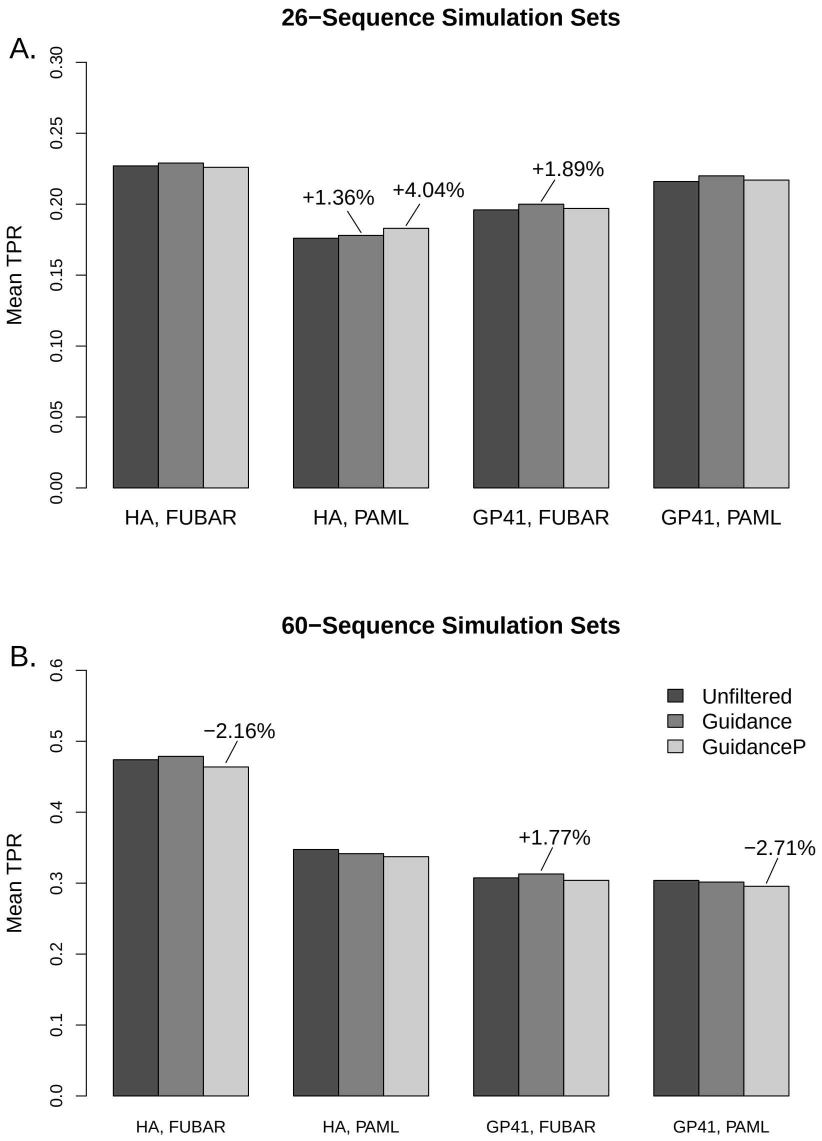

Figure 1: Mean TPR for and 26- and 60-sequence simulation sets. Percentages, which represent the average percent TPR change from the unfiltered MSAs, are shown only for those changes which are significant. Significance levels are the same as those given in Table 1. Dark gray bars represent unfiltered MSAs, medium gray bars represents MSAs filtered with Guidance, and light gray bars represent MSAs filtered with GuidanceP.

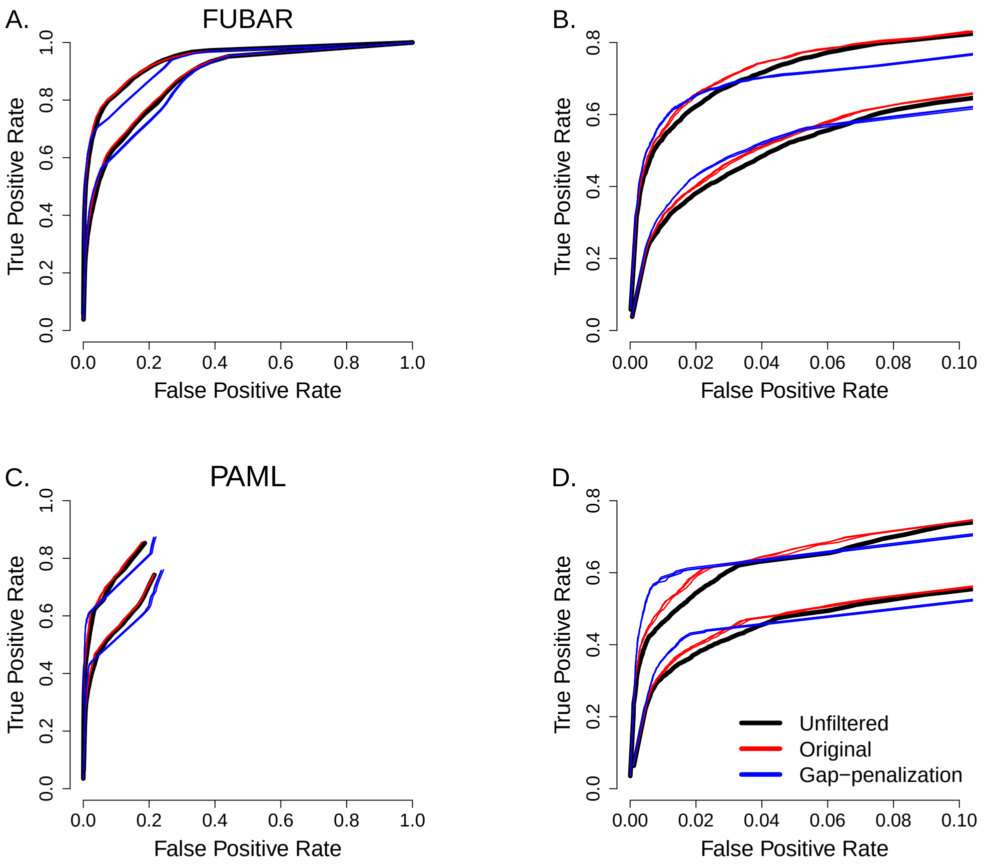

Figure 2: ROC curves as averaged across the two 60 sequence simulation sets. Within each panels, the top curve represents results from the HA selective profile, and the bottom curve represents results from the GP41 selective profile. Full ROC curves are shown in the left-hand panels. Note that, for the full ROC curves, methods only achieved FPR levels shown. The right-hand panels highlight specifically the low FPR regions $(0 - -0.1)$ of the ROC curves. A-B) ROC curves for positive-selection inf8erence by FUBAR. C-D) ROC curves for positive-selection inference by PAML M8.