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. As a consequence, 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. 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, consistently improved positive-selection inference across multiple inference methods. Moreover, all improvements detected were of exceedingly small magnitude, and in certain circumstances, MSA filters worsened positive-selection inference.

Multiple sequence alignment (MSA) construction represents the most fundamental step in 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). In response, 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 guide trees when constructing progressive alignments. Users can then mask positions that score below a set threshold, thus 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, a separate study by Jordan and Goldman (2012) found that Guidance affected positive-selection inference only modestly. In particular, both Privman et al. (2012) and Jordan and Goldman (2012) concluded that filtering was primarily beneficial when sequences were highly diverged. However, it is important to recognize that typical positive-selection studies rarely contain sequences separated by such high divergences (e.g. indel rate of 10%), so it remains unclear how useful Guidance filtering is at average divergence levels.

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 elucidate how the Guidance filter affects positive-selection inference, particularly for sequences of realistic divergence levels. We additionally examined the potential benefits to modifying the Guidance scoring scheme in several ways. First, we assessed whether two novel algorithms that corrected 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 through the Python phylogenetics library 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 scaled a given residue's score according to the number of gaps in its column, thus capturing the inherent unreliability of residues in gappy 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 selective profiles because, while both genes 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 a relatively higher proportion of sites in GP41 have dN/dS values closer to 1, 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. We masked all positions with scores below 0.5, the same threshold as used by Jordan and Goldman (2012). We selected this threshold based on results which suggested that higher threshold (e.g., 0.9 as used by Privman et al. (2012)) had the potential to worsen selection inference (see Supplementary Material, Table S1).

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 during positive-selection inference were constructed in RAxMLv7.3.0 using the "PROTGAMMAWAG" model (Stamatakis 2006). While we processed all MSAs with FUBAR, we did not process the MSAs of 158 taxa with PAML 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 contained 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 generally found that all filters within a given normalization scheme performed statistically indistinguishably. In other words, Guidance, BMweights, and PDweights performed similarly to one another, as did the gap-penalization filtering algorithms. Therefore, Table 1 displays results for only Guidance and GuidanceP (refer to Table S2 for results for all filters). Table 1 demonstrates that, for the majority of datasets studied here, Guidance-based fil-

tering did not significant affect TPR. While in several cases, MSA filtering did significantly increase mean TPR, filtering significantly decreased mean TPR in other cases. Even so, all of these effects, which statistically significant, had very small magnitudes. We also found that gap-penalization filters provided both the largest TPR increases and the largest TPR decreases, while the original normalization scheme was much more conservative. Indeed, the Guidance filter only significantly reduced TPR in one case, while GuidanceP significantly reduced TPR in three cases.

Moreover, inference methods responded inconsistently to filtered MSAs, as demonstrated in Figure 1, which gives a graphical representation of the linear models' results for the 26- and 60-sequence simulation sets. On one hand, FUBAR yielded consistent mean TPR trends among filters (Guidance TPR was generally higher than were both unfiltered and GuidanceP), but this trend was mostly 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 was for the HA 26-sequence simulation set; when PAML inferred positive-selection on GuidanceP-filtered MSAs, TPR increased by roughly 4% the TPR of 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.

Additionally, we emphasize that all Guidance-based filters improved positive-selection inference for the largest simulation sets of 158 taxa when analyzed with FUBAR, although the effect magnitude was very small. As we did not analyze these data sets with PAML, we caution this result may not be easily extrapolated to inference methods other than FUBAR. Conversely, all filters significantly reduced accuracy for the GP41 11-sequence simulation set, as analyzed with PAML. Taken together, these results suggested that filtering may, although minimally, be beneficial for relatively large MSAs, but may hinder accuracy for small MSAs.

Guidance-based filters improve power only in a limited range

We additionally used receiver operating characteristic (ROC) curves, commonly used to evaluate the performance of binary classification methods, to qualitatively assess whether MSA filtering influences power in positive-selection inference. 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, power in positive-selection inference for HA simulation sets was universally greater than for GP41 simulation sets. Given that the GP41 sequences featured a greater proportion of sites with dN/dS near 1, this result was largely expected. Even so, Guidance-based filters behaved consistently across all simulation conditions. Second, as all algorithms within a given normalization scheme (original vs. gap-penalization) had nearly identical curves, this analysis confirmed that introducing phylogenetically-weighted scores did not strongly affect Guidance scores. Finally, across the entire span of the ROC curves, no dramatic difference in area between unfiltered and filtered MSA curves existed, although, MSAs filtered with gap-penalization algorithms did, at certain FPR levels (roughly 0.1–0.3), perform worse worse than did both unfiltered and Guidance-filtered MSAs.

However, filtering did substantially increase power at very low FPR rates, as seen in the right-hand panels, in particular when using PAML. These benefits, unfortunately, only existed at FPR

levels of roughly 1% - 4%, above which any improvements 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 with a posterior probability ≥ 0.9 all yielded mean FPRs under 1%, below the region where filtering increased power. Our low recovered FPRs explained why we did not detect substantial benefits to MSA-filtering through our linear models. Taken together, these results demonstrated that Guidance-based filtering was not robust to varying FPR levels. ROC curves for all other simulation sets are available in Figures S1 and S2, and yielded results broadly consistent with 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 recovered few conditions in which filtering strictly removes noise but preserves signal.

Our study focused primarily on divergence levels representative of realistic protein-coding data typically used in positive-selection inference. Therefore, it is possible that Guidance would have provided stronger benefits with highly diverged data Privman et al. (2012). 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).

We additionally found that, for nearly all simulation cases, FUBAR outperformed PAML both in 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. FUBAR, therefore, represents a fast and and accurate alternative to traditional positive-selection inference methods.

In sum, two distinct conclusions may be drawn from our study. First, although Guidance did not universally benefit positive-selection inference, it never entirely precluded detection of positively-selected sites. Therefore, filtering could be used as a conservative method in selection inference. Alternatively, all benefits that filtering conferred were minimal, and filters behaved inconsistently across simulation sets and inference methods. Given these observations, there is no guarantee that MSA filtering will help or harm any given analysis, such that Guidance filtering may inadvertently result in a loss of power.

We conclude that, while potentially beneficial, Guidance-based filtering is not a particularly robust method for positive-selection inference, and therefore does not need to be a necessary component of positive-selection studies. Finally, we recommend that users primarily focus on employing high-quality MSA inference (e.g. linsi (Katoh et al. 2005) or PRANK (Loytynoja and Goldman 2008)) and positive-selection inference (e.g. FUBAR) methods, in which error can be minimized as much as possible preserve informative signal. Should users opt to filter their MSAs, we recommend using a lenient cutoff (≤ 0.5) to preserve informative signal.

Supplementary Material

Supplementary methods, Figures S1 and S2, and Tables S1, S2, and S3, are available at Molecular Biology and Evolution online (http:// www.mbe.oxfordjournals.org/).

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 constructive manuscript discussion and Sergei Kosakovsky Pond for valuable manuscript comments, for providing a GP41 alignment and phylogeny, and for help with using 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.

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

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.

Selective Profile	Number of Taxa	Inference Method	Mean True Positive Rate				Percent Gaps
			True	Unfiltered	Guidance	GuidanceP	
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.— 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 of the respective unfiltered MSA, not the absolute TPR increase or decrease. Significance levels: $^{**}P < 0.001$; $^{*}P < 0.01$. 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, and represents the percentage of columns with at least one gap, averaged across all MSA replicates.

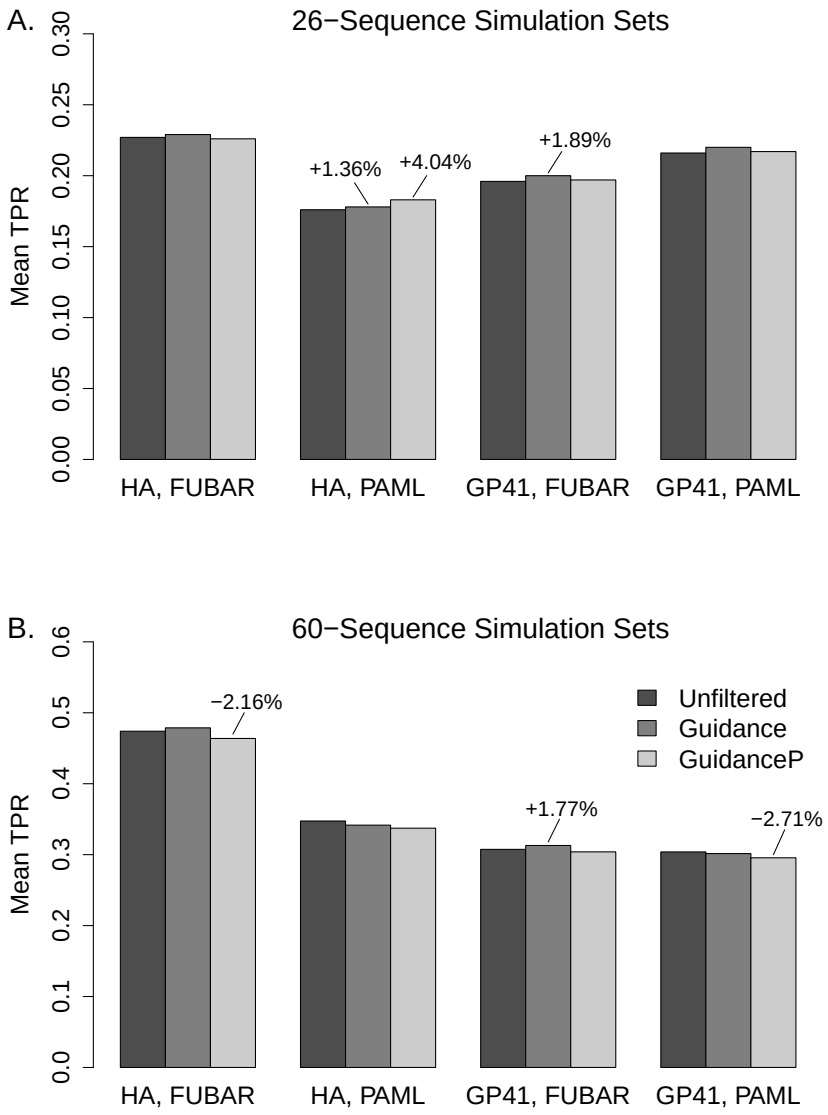

Figure 1: Mean TPR for and 26- and 60-sequence simulation sets. Percentages, which represent the average percent TPR change relative to the unfiltered MSAs, are shown only for those changes which are significant. Significance levels are the same as those given in Table 1. Dark gray bars represent unfiltered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

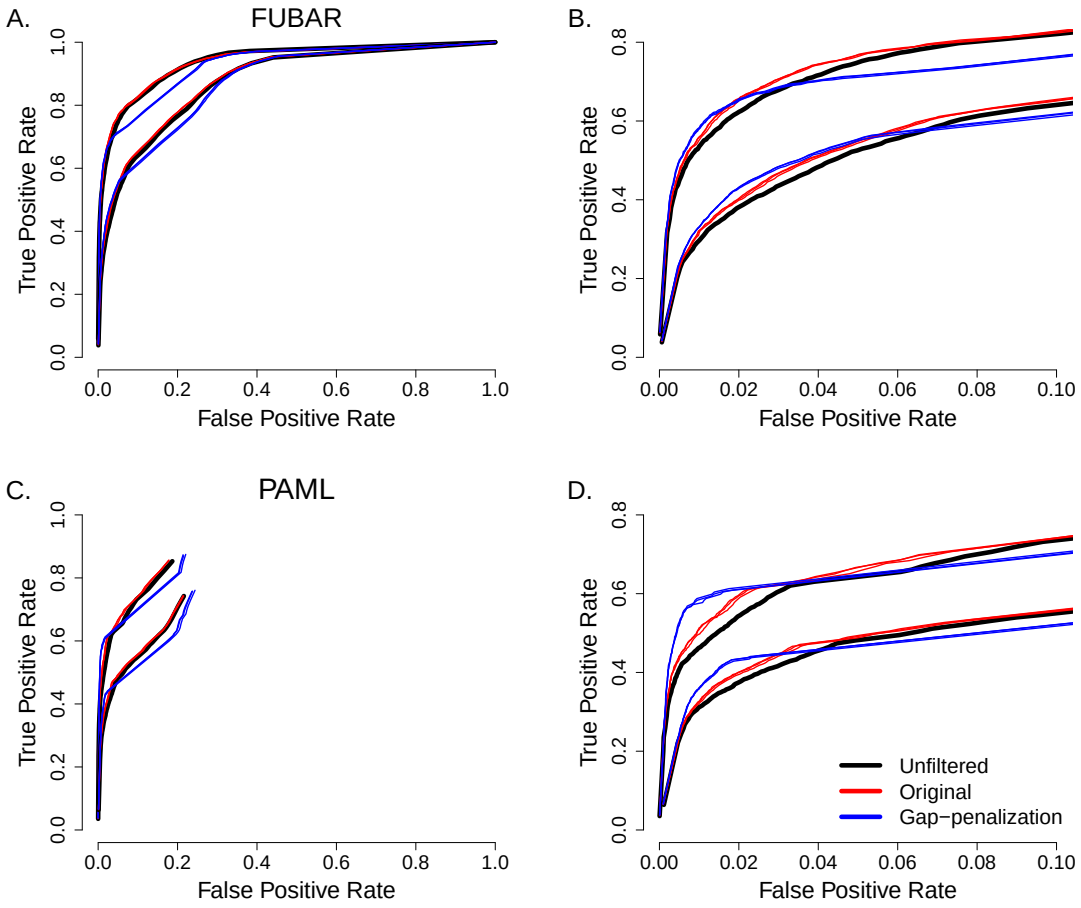

Figure 2: ROC curves as averaged across the two 60 sequence simulation sets. Within each panels, the top curve represents results from the HA selective profile, and the bottom curve represents results from the GP41 selective profile. Full ROC curves are shown in the left-hand panels. Note that, for the full PAML ROC curves, average FPRs higher than shown were not seen. The right-hand panels highlight specifically the low FPR regions (0–0.1) of the ROC curves. All MSA filtering algorithms (Guidance, BMweights, PDweights, GuidanceP, BMweightsP, and PDweightsP) are shown in ROC curves. A-B) ROC curves for positive-selection inference by FUBAR. C-D) ROC curves for positive-selection inference by PAML M8.