DEPARTMENT OF INTEGRATIVE BIOLOGY
205 W 24TH STREET
THE UNIVERSITY OF TEXAS AT AUSTIN

Austin, Texas 78712 U.S.A. • (512) 232-2459 • fax (512) 471-3878-PAT • fax (512) 232-9529-BIO

Dear Editorial Board,

We would like to submit a manuscript entitled "Limited utility of residue masking for positive-selection inference" for your consideration for publication in *Molecular Biology and Evolution*.

It is well known that errors in multiple sequence alignments (MSAs) can introduce substantial bias into downstream applications, including phylogenetic reconstruction, ortholog identification, and evolutionary rate inference. In particular, low-quality MSAs have been shown to obscure positive-selection inference in protein-coding sequences and lead to erroneous conclusions. To minimize the extent of this issue, recent studies have advocated that researchers remove potentially confounding regions from MSAs to maximize signal. One such method to curate alignments, known as Guidance, has been widely adopted to identify poorly aligned regions in MSAs.

Studies on the efficacy of the Guidance filter, however, have yielded inconsistent results; while some have found that it significantly improves positive-selection inference, others have found that its use only marginally affects results. In order to reconcile these apparent contradictions, we have conducted a large-scale simulation-based study to fully elucidate the effect of applying the Guidance filter when inferring positive-selection. In particular, we have examined the effects of several novel Guidance-based scoring algorithms that incorporate phylogenetic-weighting and penalize highly-gapped regions.

We found that our gap-penalization algorithms improved inferences relative to the original Guidance, but, surprisingly, our phylogenetically-corrected scoring schemes did not behave significantly differently from the original Guidance. Even so, we noted that applying any Guidance-based filter has a minimal effect on positive-selection detection, and in some cases (in particular, when using a stringent filtering threshold) it can even worsen the inference. We therefore conclude that using high-quality alignment and positive-selection inference software is much more beneficial to such analyses than is filtering an already existing alignment.

We believe that results from our research will strongly benefit the molecular evolution community. As MSAs represent the most fundamental step in any sequence analysis, it is crucial that we fully understand how to efficiently and accurately use MSAs in downstream applications, and our findings will help researchers ensure robust results in sequence analyses. We look forward to hearing from you.

Sincerely,

Stephanie J. Spielman, Eric T. Dawson, and Claus O. Wilke