Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Evolutionary Bioinformatics

SWAMP: Sliding Window Alignment Masker for PAML

Peter W. Harrison[1], Gregory E. Jordan[2] and Stephen H. Montgomery[1]

[1]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom. [2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, United Kingdom.

ABSTRACT: With the greater availability of genetic data, large genome-wide scans for positive selection increasingly incorporate data from a range of sources. These data sets may be derived from different sequencing methods, each of which has potential sources of error. Sequencing errors, compounded by alignment errors, greatly increase the number of false positives in tests for adaptive evolution. Genome-wide analyses often fail to fully address these issues or to provide sufficient detail on postalignment masking/filtering. Here, we introduce a Sliding Window Alignment Masker for Phylogenetic Analysis by Maximum Likelihood (SWAMP) that scans multiple-sequence alignments for short regions enriched with unreasonably high rates of nonsynonymous substitutions caused, for example, by sequence or alignment errors. SWAMP prevents their inclusion in downstream evolutionary analyses and therefore increases the reliability of downstream analyses. It is able to effectively mask short stretches of erroneous sequence, particularly prevalent in low-coverage genomes, which may not be detected by existing methods based on filtering by sitewise conservation or alignment confidence. SWAMP offers a flexible masking approach, and the user can apply different masking regimens to specific branches or sequences in the phylogeny allowing the stringency of masking to vary according to branch length, expected divergence levels, or assembly quality. We exemplify SWAMPs effectiveness on a dataset of 6,379 protein-coding genes from primate species, including data of variable quality. Full reporting of the software parameters will further improve the reproducibility of genome-wide analyses, as well as reduce false-positive rates.

KEYWORDS: sequence analysis, phylogenetics, PAML, molecular evolution, genome evolution, adaptive evolution

AVAILABILITY: SWAMP is freely available, published under GNU GPL v3, including documentation and example test data from http://github.com/peterwharrison/SWAMP

CITATION: Harrison et al. SWAMP: Sliding Window Alignment Masker for PAML. *Evolutionary Bioinformatics* 2014:10 197–204 doi: 10.4137/EBO.S18193.

RECEIVED: June 26, 2014. **RESUBMITTED:** August 21, 2014. **ACCEPTED FOR PUBLICATION:** September 1, 2014.

ACADEMIC EDITOR: Jike Cui, Associate Editor

TYPE: Technical Advance

FUNDING: SHM is grateful to the Royal Commission for the Exhibition of 1851 for financial support. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: p.w.harrison@ucl.ac.uk

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

Introduction

New sequencing technologies and decreasing sequencing costs are leading to a rapid increase in the availability of DNA sequence data. This is true both within and across species, permitting increasing numbers of genome-wide scans for positive selection incorporating a more diverse range of species.[1–4] A general aim of these analyses is to identify genes that evolved adaptively, either across a clade or on specific lineages, and to understand the biological processes targeted during periods of phenotypic change. The *codeml* program in Phylogenetic Analysis by Maximum Likelihood (PAML) is a powerful suite of models routinely used for phylogenetic analyses of protein-coding sequence.[5] These models calculate the ratio of nonsynonymous and synonymous substitution rates (*dN/dS*), a measure of selection pressure, and comparisons between different models permit tests for pervasive or episodic positive selection, acting on a gene-wide or codon-specific manner.

The performance of these models can be strongly influenced by gene misannotation, alignment error, and sequence quality.[6–11] Unfortunately, with the adoption of next-generation sequencing methods,[12] the likelihood of certain errors, including sequencing errors and misalignment

caused by splicing variants, has vastly increased.[13–15] The inclusion of alignments containing nonhomologous data caused by these effects can drastically inflate false-positive rates in PAML, and may also influence false-negative rates.[6,7] Alignment quality is therefore of major importance in the accurate inference of positive selection.

The alignment program chosen can have a significant effect on the reliability of PAML analyses, with some, such as PRANK,[16] outperforming others (ClustalW,[17] MAFFT,[18] ProbCons,[19] and T-coffee[20]).[6,7] Postalignment filtering provides an additional step to improving alignment quality and has been implemented in two main ways. Column-based programs, such as G-Blocks[21] or Noisy,[22] examine the degree of conservation at each position in the alignment, removing contiguous stretches of sequence that are not conserved across species. G-Blocks' original purpose was not to filter alignments for tests of positive selection, but instead to remove unreliable sequence data for phylogenetic studies.[21] As such, although this approach may have some benefits with low-quality alignments, the columnwise nature of the method can remove high proportions of data, greatly reducing power.[7] In addition G-Blocks will fail to remove sequencing errors that affect just one species in a large, multiple-sequence alignment. Where sequencing error is present at a site in one species, Gblocks and other column based filtering methods will not mask the data if that site is conserved across the rest of the alignment. Branch-specific analyses of evolutionary rates will therefore be vulnerable to this source of error.

An alternative approach is to use a measure of alignment confidence to filter the data set. These can be obtained from some alignment programs[16,20] or through additional programs such as GUIDANCE[23] or ALISCORE.[24] These filters can effectively reduce false positives when alignment confidence is low. However, adding these additional filters provides little benefit beyond using the top-performing alignment program for well-supported alignments.[7]

The merit of implementing existing filters is therefore open to debate.[7] This is particularly true when sequence divergence is low, leading to alignments with high confidence. In these cases, short stretches of sequencing errors or longer stretches of nonhomologous sequence caused, for example, by splicing variation or misannotation, can have an overly dominant effect on tests for positive selection. To this end, we have developed a Sliding Window Alignment Masker for PAML (SWAMP). This script provides an additional preprocessing step designed to mask these problematic sections of sequence.

Implementation

SWAMP analyses DNA sequences in a phylogenetic context, identifying regions with a high concentration of nonsynonymous substitutions along a branch, over a short sequence window. The method utilizes the summary of nonsynonymous codon substitutions along branches within a phylogeny obtained by running a one-ratio model (model = 0, NSsites = 0) in *codeml*.[5] These summary data contain details of the codon positions of all predicted substitutions along each branch. SWAMP uses this information to conduct a sliding-window scan across a gene to quantify the number of nonsynonymous substitutions within a user-provided codon window length. If a user-defined threshold is exceeded, the window is masked to exclude it from downstream analyses (Table 1). Masking is achieved by converting the sequence within the window to a sequence of *N* characters. These stretches of *N* characters are removed by PAML and therefore do not influence tests for positive selection or accelerated evolution (Fig. 1A and B).

A "branchnames" file, provided by the user, defines which row(s) of sequences (in most cases this will be species) will be affected by substitution patterns along each branch in the phylogeny. Generally, this relates species data to their ancestral lineages. The use of this "branchnames" file provides a further advantage in that a user may vary the masking parameters across specific branches or sequences, allowing different masking regimens to be applied to different parts of the phylogeny through multiple SWAMP iterations. This is achieved by the user listing only a subset of sequences and/or branches in the "branchnames" file in each of multiple iterations of SWAMP, while supplying different thresholds and window sizes for each run. An example of this is provided in the SWAMP documentation along with step-by-step instructions on how to implement the program. This branch-specific masking may be useful in a number of contexts, for example, if the data for one of a number of species is more likely to contain errors than others, where assembly quality varies, or if significant variation in branch length demands a flexible approach to alignment masking.

In some cases, initial masking can leave small "islands" of sequence data flanked by masked sequence. These potentially problematic stretches of sequence in close proximity to masked sections can be masked with the optional "interscan" function (Table 1). This function masks regions based on their length in comparison to neighboring masked regions (Fig. 1B and C). This ensures longer stretches of nonhomologous data that by chance share some similarity are still masked from the alignment.

Finally, SWAMP also notifies the user if the total length of the sequence falls below a defined minimum. These cases are likely to be incomplete sequences, which may then be excluded from downstream analyses if desired.

SWAMP is not a computationally expensive filtering approach. For example, across a data set of >6,000 four-way 1:1 orthologs, described below, using a threshold of 10 and window size of 15, SWAMP ran in 93.3 seconds on a Mac with a 2.93-GHz i7 and 16-GB RAM.

Results and Discussion

Effects of SWAMP on branch-site tests for adaptive evolution. For developmental purposes, and to provide some guidance on initial parameters, we utilized a primate data set, consisting of 6,379 orthologs from *Homo sapiens*, *Pongo*

Table 1. Description of key SWAMP parameters.

Threshold	-t THRESHOLD or --threshold THRESHOLD	A threshold-positive integer of the number of nonsynonymous substitutions at and above which the window will be masked.
Windowsize	-w WINDOWSIZE or --windowsize WINDOWSIZE	An integer window size for the sliding-window scan, given in numbers of codons.
Minimum sequence length	-m MINSEQLENGTH or --minseqlength MINSEQLENGTH	The required minimum number of informative codons in each of the sequences in the multiple-sequence alignment postmasking. This is a positive integer. The program will print a warning to the user in the standard output if a masked sequence is shorter than this minimum length. The default is 33 codons (99 base pairs).
Interscan masking	-s or --interscan	Activates interscan masking. This will additionally mask regions adjacent to already masked regions based on relative sequence length. This additional masking is performed at the start or end of the sequence alignment if the unmasked sequence region length is shorter than twice the length of the preceding or subsequent masked section. Where a sequence contains multiple masked regions, interscan will also mask internal unmasked regions that are shorter than the combined length of their flanking masked regions. This process occurs repeatedly until no more sections that meet the interscan masking criteria are found. Interscan is useful for removing very short stretches of sequences or sequences at the edge of masked regions that are possibly unreliable, but that do not themselves meet the masking criteria.

pygmaeus, *Colobus angolensis*, and *Papio anubis*. Although small, this data set is comparable in size with the majority of genome-wide scans for positive selection during human or anthropoid evolution performed to date (eg, Nielsen et al.[3], and Scally et al.[4]) and permits the implementation of branch-specific tests for adaptive evolution that are widely used in studies of adaptive evolution (eg, Kosiol et al.[1], and McGowen et al.[2]) but are known to be sensitive to alignment and sequencing errors.[6–9] The data were selected as they incorporate a range of sequencing technologies and coverage levels, therefore encompassing a range of potential errors. This allows for a comparison between high-quality Sanger-sequenced genomes, such as the genome of *H. sapiens*, with genomes and exomes constructed using next-generation sequencing strategies that may have higher error rates, lower assembly quality, and greater difficultly in resolving repetitive regions.[13–15] This data set therefore provides a useful exploration of the effectiveness of the approach taken in SWAMP.

Protein-coding genes for *H. sapiens*, *P. pygmaeus*, and *P. anubis* were obtained from Ensembl v73[25] and *C. angolensis* was obtained from the National Center for Biotechnology Information (NCBI).[26] One-to-one orthologs were aligned with PRANK,[16] the alignment program that results in the lowest false-positive rates in downstream analyses,[6,7] to minimize errors that are attributable to a suboptimal alignment strategy and could otherwise be avoided. We ran the branch-site test for positive selection on the four terminal branches.[27] This model produces significant results when there is an acceleration in *dN/dS* ratio at a subset of sites along a selected set of branches.[6] Without masking, tests on all four branches are significant (likelihood ratio [LR] >2.71 at $P = 0.05$) for a large number of genes (Fig. 2). A number of these show likelihood ratios above 50. Visual inspection of these alignments revealed that the majority of these cases contain short stretches of sequence with low conservation in the focal species, typical of sequencing or alignment error. With increasingly stringent SWAMP masking across a constant-sized window (15 codons) the number of these outliers decreases. Notably, the effect is most modest in the species with the highest quality data, *H. sapiens* (a Sanger-sequenced genome with high coverage) and *C. angolensis* (an exome with high coverage[26]), indicating the effect is largely due to data quality rather than removing true positives. Note the *Homo* and *Pongo*, and *Colobus* and *Papio*, branch lengths are equal in length (time since divergence) so variation in evolutionary distance does not explain the variable effects of masking.

The effect of varying the window size is illustrated in Figure 3. Here we focus on the terminal *Papio* branch and compare results of the branch-site test using unmasked data, a window size of 5 and a window size of 15 with a threshold of 3 and 10, respectively (ie, 60%–66% conservation). We limited the comparison to those genes with significant results at $P = 0.01$ (LR >5.41) with the unmasked alignments. Under both comparisons, a large number of genes have much higher likelihood ratios using the unmasked data (Fig. 3A and B). In many cases previously highly significant results have a LR of zero after masking. The difference in LR before and after masking was used to compare the two masking regimens (Fig. 3C). In a large number of cases, the effect size is similar, with 72/356 (20%) genes losing significance under either masking regimen, but in some cases, they differ. The shorter window size specifically reduces the LR below significance for 119 more genes and may produce

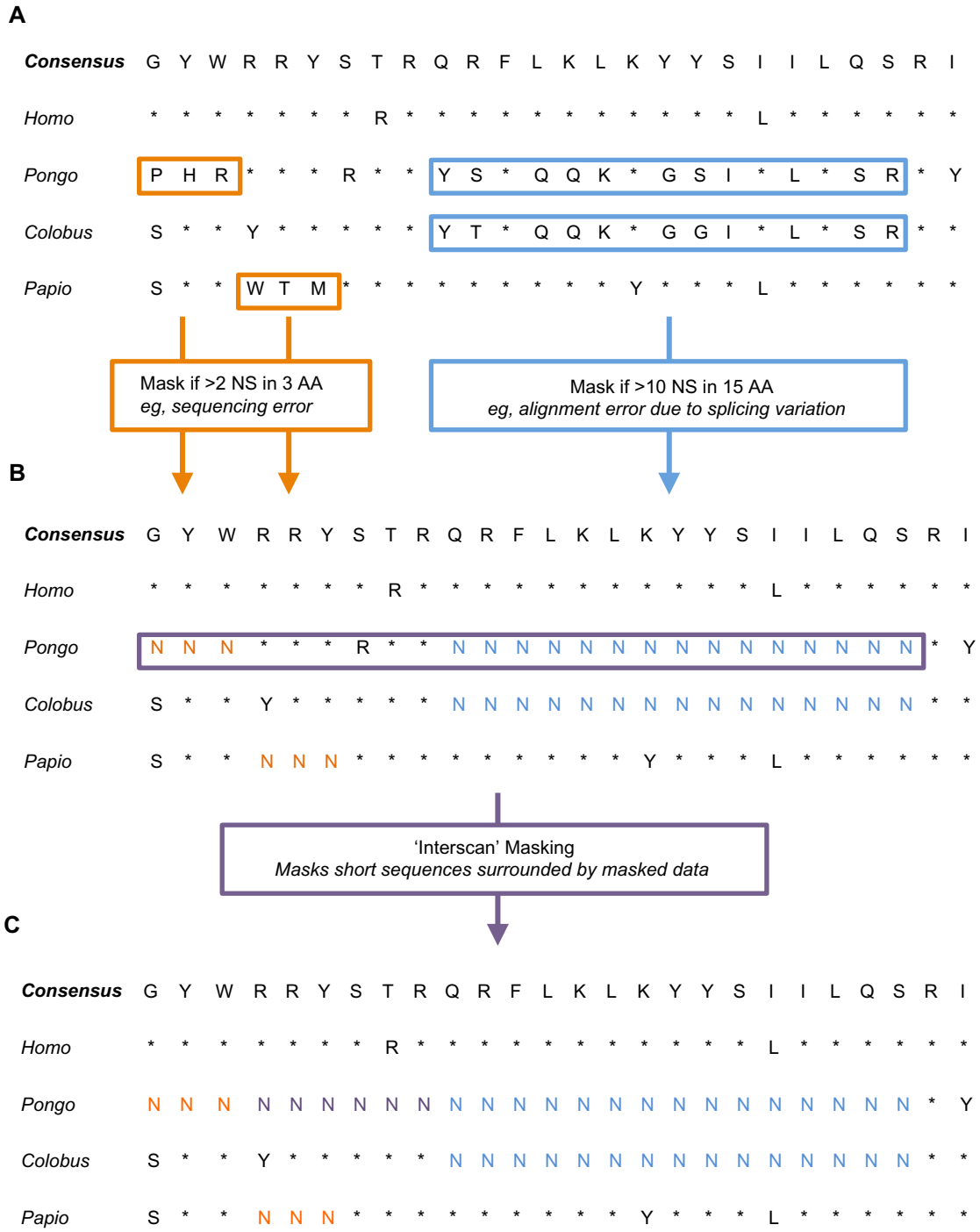

Figure 1. Illustrative schematic of SWAMP masking. (**A**) Unmasked alignment with two potential data errors, short stretches of divergent sequence possibly caused by sequencing errors, and longer stretches possibly caused by exon splicing. (**B**) Application of a two-step filtering masks data errors; a shorter window of 3 AA with a maximum of 2 substitutions effectively masks possible sequence error (orange), and a longer window of 15 AA with a maximum of 10 substitutions effectively masks longer alignment errors (blue). (**C**) The 'Interscan' option further masks the data if short stretches of sequence are surrounded by masked data and the sum of the masked data either side of the sequence exceeds the length of the interceding unmasked data (purple).

more conservative results in this case, while the longer window size specifically reduces the LR of a smaller number of genes.[16] Visual inspection of clear outliers in Figure 3C suggests smaller window sizes are more effective at filtering short clusters of nonsynonymous substitutions, such as those that may be caused by sequencing errors, while larger window sizes may remove alignment errors caused by variation in exon splicing. Therefore, a two-step masking procedure with a large and small window size can effectively remove most of the problematic sequence.

Comparison with the gorilla genome analysis. A similar approach was previously implemented in a genome-wide

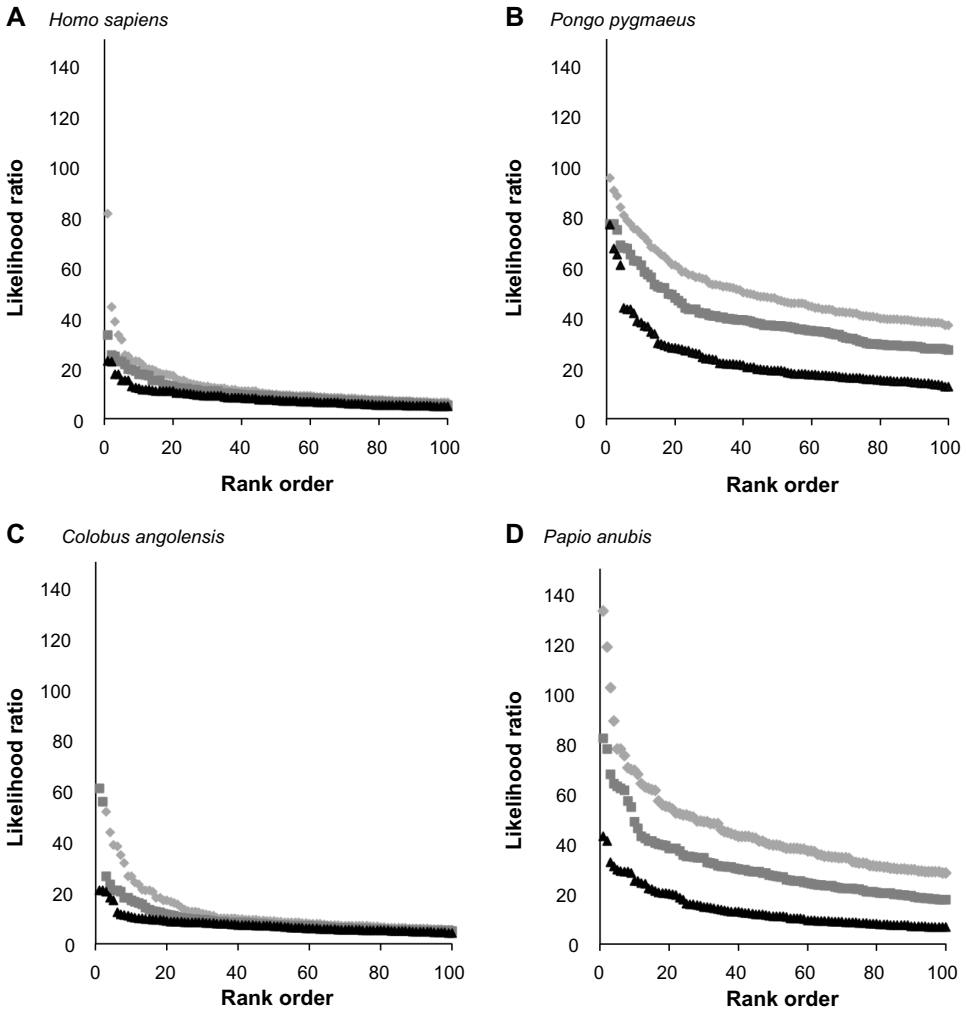

Figure 2. Effects of SWAMP filtering on branch-site tests for positive selection. Branch-site tests were conducted on the terminal *Homo* (**A**), *Pongo* (**B**), *Colobus* (**C**), and *Papio* (**D**) branches in a four-species alignment ((*H. sapiens*, *P. pygmaeus*),(*C. angolensis*, *P. anubis*)). The top 100 LR scores are shown for unfiltered data (light grey diamonds) and filtered for 10 or more nonsynonymous substitutions in 15 codons (medium grey squares) and 5 or more nonsynonymous substitutions in 15 codons (black triangles). As the filtering becomes more stringent, the number of LR values above 50 decreases, and when these alignments are visually inspected, they are often found to contain stretches of poorly aligned or nonhomologous sequence. The effects of genome coverage and assembly quality can also be seen by comparing the results of *Homo*, which has a high-coverage genome, with *Pongo* and *Papio* which have lower coverage genomes. The exome assembly of *Colobus* appears to be of comparatively high-quality consistent with its high sequence coverage.[26]

evolutionary analysis of protein-coding genes across African Great Apes[4] using an unpublished forerunner of SWAMP. This analysis masked a 1:1 orthologous genes set for humans, chimpanzee, gorilla, orangutan, macaque, and marmoset using a window size of 15 codons and a threshold of 10 nonsynonymous substitutions per window. Of 11,538 gene alignments 1,156 (10.1%) were masked in at least one window. This mirrors our results in which of 6,379 alignments 1,022 (16.0%) genes were masked under the threshold of 5 nonsynonymous substitutions in 15 codons and 429 (6.7%) were masked under the threshold of 10 nonsynonymous substitutions in 15 codons. Notably, this analysis also found much lower numbers of genes being partially masked in humans compared to other primates that have lower quality genomes. The masking performed in the analysis of the gorilla genome[4] affected downstream PAML analyses in a similar way to that described above (G. E. Jordon and S. H. Montgomery, personal observation).

Comparison with column-based masking. A major difference between the approach taken in SWAMP and currently available postalignment filtering methods is the orientation of data analysis. Existing methods tend to filter alignments based on conservation within a column of a multiple-sequence alignment (ie, at a codon or nucleotide across species), whereas SWAMP analyzes data within rows. This is advantageous as sequencing or alignment errors may not sufficiently reduce similarity at conserved sites to be filtered by column-based approaches. To demonstrate the effects of this difference, we filtered our alignments using G-Blocks[21] under default parameters. The filtering resulted in the removal of data from 185 (2.9%) of 6,379 alignments. In contrast to

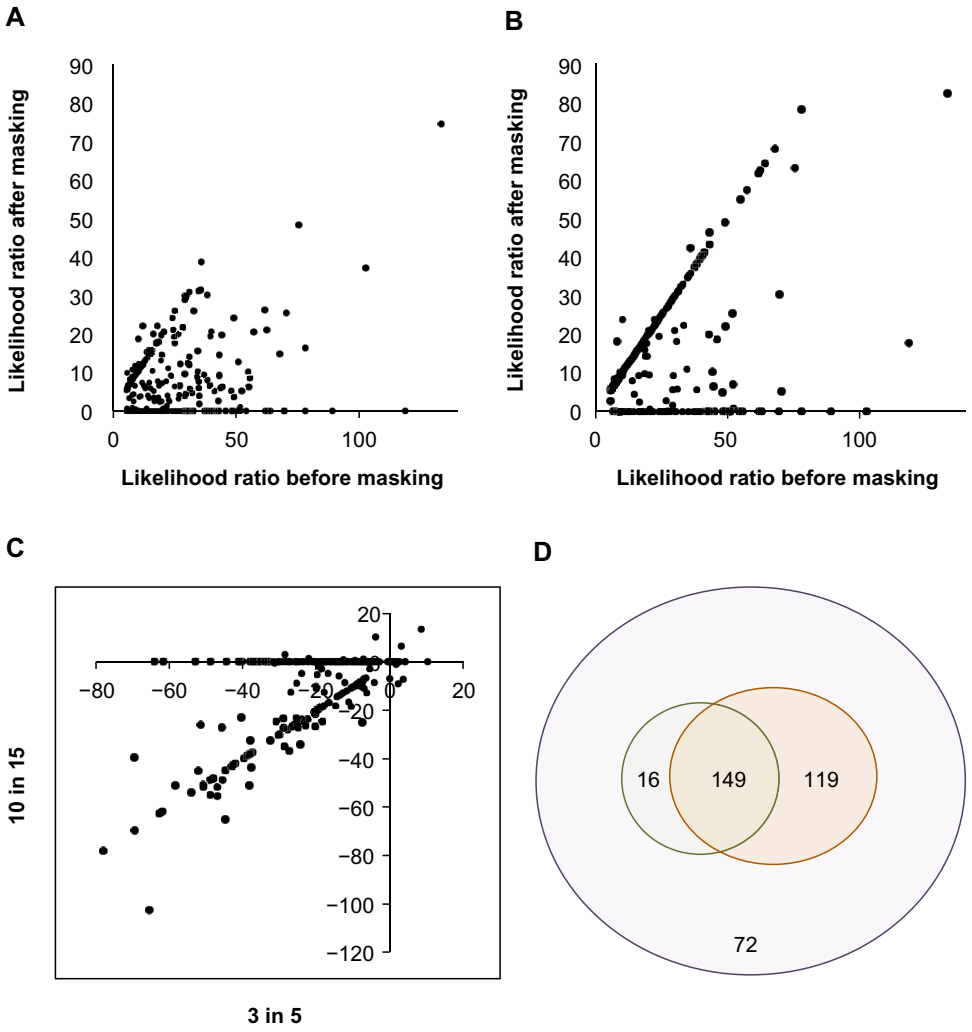

Figure 3. Effects of masking at different window sizes. In (**A**) and (**B**), the LR statistics of genes significant for the branch-site test on the terminal *Papio* branch ($P < 0.01$) using the unfiltered alignments are plotted against the LR for the same test after masking with either (**A**) a window size of 5 codons and a nonsynonymous threshold of 3 or (**B**) a window size of 15 codons and a nonsynonymous threshold of 10. In each case, the percentage of conserved codons required in the window is approximately constant (60% or 67%, respectively). In (**C**) the difference in LR statistic between the unfiltered alignments and the masked alignments for the two masking regimens are plotted against each other. In this case, many genes are equally affected, with 20% losing significance, but a large number appear to provide more conservative results under the smaller window size. This is illustrated in (**D**), where the overlap between genes significant after masking using the 5-codon window size (green), the 15-codon window size (orange), and the unmasked alignments (purple) are shown.

SWAMP masking, the downstream PAML results based on these G-Block-masked alignments are almost identical to those obtained using unmasked data (Supplementary Fig. 1). Of course, across more divergent data sets that include regions where the alignment is problematic, users may find column-based filtering more useful; indeed, this was the intended use of G-Blocks.[21] However, given PAML is optimized for data sets that are unsaturated at synonymous sites, and therefore relatively well conserved, we expect the phylogenetic row-based approach of SWAMP will be preferable in the majority of cases.

Usefulness of SWAMP and potential caveats. These results demonstrate SWAMPs utility on data from genomes of lower quality than those of the gold standard model organisms (Fig. 2). SWAMP provides a flexible framework to mask large data sets, removing stretches of low-quality alignment, probable sequencing errors, and nonhomologous data that could otherwise inflate false-positive rates in tests for adaptive evolution.

Effective masking with the approach taken by SWAMP will most likely produce conservative results as a minority of masked sequence may reflect genuine divergence concentrated in a short stretch of sequence. This could conceivably occur in proteins with key functional domains coded by a contiguous stretch of sequence. While this may result in some false negatives, a conservative approach is preferable in genome-wide studies, particularly when used to generate candidates for functional analyses. However, if this is a concern for a user, we recommend testing for enrichment of protein domain types within the masked genes and extending the window size during filtering.

A further caveat is that users must currently optimize their masking parameters manually. This can be done based on the genome-wide average rate of nonsynonymous substitutions/codons or simply by optimizing the parameters to ensure genes with significant results in downstream analyses do not contain spurious alignments when the most significant genes are inspected manually. The increased confidence in downstream analyses and the reduction in manual filtering of results should offset this investment in time.

It is generally accepted that short sequences and those that contain internal stop codons should be removed from genome-wide scans for positive selection. Sequences that contain repetitive elements could also be masked, for example, with Repeatmasker.[28–30] By implementing SWAMP in conjunction with optimal alignment programs and these established masking steps, researchers can increase their confidence in conclusions drawn from evolutionary and phylogenetic analysis performed in PAML are other analysis suites. SWAMP provides a useful addition to methods of postalignment filtering, improving the reliability and reproducibility of genome-wide analyses using PAML.

Conclusions

SWAMP effectively masks regions with high rates of nonsynonymous substitutions concentrated in short runs of sequence typical of sequence or alignment errors, preventing their inclusion in downstream evolutionary analyses. This removes sequence that violates the assumptions of the phylogenetic model implemented in the software package PAML that could otherwise give a false signal of positive selection. SWAMP effectively masks short stretches of erroneous sequence that may not be detected by existing masking/filtering methods but will be prevalent in low-coverage genomes and the branch- and sequence-specific operation allows different masking regimens to be applied to selected parts of the phylogeny. Although specifically designed for implementation with PAML, SWAMP will be useful as a preprocessing step for any analysis that requires the prevention of the influence of sequence error and misannotation. In addition to the reduction in false-positive rates achieved through SWAMP preprocessing, the inclusion of the implemented SWAMP parameters in future publication methodologies will improve the reproducibility of genome-wide analyses for positive selection.

Availability and Requirements

Project name: SWAMP

Project home page: http://github.com/peterwharrison/SWAMP

Operating systems: Platform independent

Programming language: Python

Other requirements: PAML 4.7 or higher, Python 2.6 or 2.7

License: GNU GPL v3

Acknowledgments

We would like to thank Fabian Zimmer for helpful comments with this article. The authors acknowledge the use of the UCL *Legion* High Performance Computing Facility (Legion@UCL) and associated support services, in the completion of this work.

Author Contributions

Conceived and designed the experiments: PWH, GEJ, SHM. Developed and tested the software: PWH, GEJ. Wrote the first draft of the manuscript: PWH, SHM. Contributed to the writing of the manuscript: PWH, GEJ, SHM. Revised the manuscript: PWH, SHM. All authors reviewed and approved of the final manuscript.

Supplementary Data

Supplementary Figure S1. Comparison of results using row-based SWAMP and column-based G-Blocks filtering. Results show the rank order of the Likelihood ratio test (LRT) for the top 100 genes under a branch-site model for four species as in Figure 2. The dashed grey lines are the results from the unfiltered alignments. Orange circles show the G-Blocks filtered data, which deviates very little from the results using unfiltered alignments. The black triangles show the results after SWAMP filtering with a window size of 15 codons and a threshold of five or more nonsynonymous substitutions.

REFERENCES

1. Kosiol C, Vinař T, da Fonseca RR, et al. Patterns of positive selection in six mammalian genomes. *PLoS Genet*. 2008;4(8):e1000144.
2. McGowen MR, Grossman LI, Wildman DE. Dolphin genome provides evidence for adaptive evolution of nervous system genes and a molecular rate slowdown. *Proc Biol Sci*. 2012;279(1743):3643–51.
3. Nielsen R, Bustamante C, Clark AG, et al. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*. 2005;3(6):e170.
4. Scally A, Dutheil JY, Hillier LW, et al. Insights into hominid evolution from the gorilla genome sequence. *Nature*. 2012;483(7388):169–75.
5. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24(8):1586–91.
6. Fletcher W, Yang Z. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol*. 2010;27(10):2257–67.
7. Jordan G, Goldman N. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol Biol Evol*. 2012;29(4):1125–39.
8. Mallick S, Gnerre S, Muller P, Reich D. The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res*. 2009;19(5):922–33.
9. Privman E, Penn O, Pupko T. Improving the performance of positive selection inference by filtering unreliable alignment regions. *Mol Biol Evol*. 2012;29(1):1–5.
10. Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol Evol*. 2009;1:114–8.
11. Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. *Science*. 2008;319(5862):473–6.
12. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008;26(10):1135–45.
13. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultrashort read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*. 2008;36(16):e105.
14. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res*. 2010;38(12):e131.
15. Huse S, Huber J, Morrison H, Sogin M, Welch D. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol*. 2007;8(7):R143.
16. Löytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*. 2008;320(5883):1632–5.
17. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 1994;22:4673–80.

18. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*. 2005;33:511–8.
19. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res*. 2005;15:330–40.
20. Notredame C, Higgins DG, Heringa J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol*. 2000;302:205–17.
21. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 2000;17(4):540–52.
22. Dress AWM, Flamm C, Fritzsch G, et al. Noisy: identification of problematic columns in multiple sequence alignments. *Algorithms Mol Biol*. 2008;3:7.
23. Penn O, Privman E, Landan G, Graur D, Pupko T. An alignment confidence score capturing robustness to guide tree uncertainty. *Mol Biol Evol*. 2010;27(8):1759–67.
24. Kuck P, Meusemann K, Dambach J, et al. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Front Zool*. 2010;7(1):10.
25. Flicek P, Amode MR, Barrell D, et al. Ensembl 2014. *Nucleic Acids Res*. 2014;42:D749–55.
26. George RD, McVicker G, Diederich R, et al. Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. *Genome Res*. 2011;21(10):1686–94.
27. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*. 2005;22(12):2472–9.
28. Pointer MA, Harrison PW, Wright AE, Mank JE. Masculinization of gene expression is associated with exaggeration of male sexual dimorphism. *PLoS Genet*. 2013;9(8):e1003697.
29. Singer MF. SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes. *Cell*. 1982;28(3):433–4.
30. Smit A, Hubley R, Green P. RepeatMasker Open–3.0. http://www.repeatmasker.org. 1996–2010.