The Journal of Open Source Software

AOC: A Snakemake workflow for the characterization of natural selection in protein-coding genes

Alexander G. Lucaci [1¶] **and Sergei Pond** [2¶]

1 Department of Systems and Computational Biomedicine, Weill Cornell Medicine, Cornell University, New York, NY 10021, United States of America **2** Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, United States of America ¶ Corresponding author

DOI: 10.21105/joss.09872

Software

- Review
- Repository
- Archive

Editor: Sehrish Kanwal

Reviewers:

- @gavinmdouglas
- @juanvillada

Submitted: 01 July 2025
Published: 19 May 2026

Summary

Modern molecular sequence analysis increasingly relies on automated and robust software tools for interpretation, annotation, and biological insight. The Analysis of Orthologous Collections (AOC) application automates the identification of genomic sites and species/lineages influenced by natural selection in coding sequence analysis. AOC quantifies different types of selection: negative, diversifying or directional positive, or differential selection between groups of branches. We include all steps necessary to go from unaligned homologous sequences to complete results and interactive visualizations that are designed to aid in the interpretation and contextualization of inferred selection signals (e.g., site-level dN/dS estimates, lineage-specific selection patterns, and statistical support), enabling users to relate these results to functional, evolutionary, and biological hypotheses. We are motivated by a desire to make evolutionary analyses as simple as possible, and to close the disparity in the literature between genes which draw a significant amount of interest and those that are largely overlooked and underexplored. We believe that such underappreciated and understudied genetic datasets can hold rich biological information and offer substantial insights into the diverse patterns and processes of evolution, especially if domain experts are able to perform the analyses themselves.

1 Introduction

Genomic research is inevitably biased towards certain organisms (humans, model organisms, agriculturally important species, pathogens), and genes (biomedically important, functionally understood) (Stoeger et al., 2018). For example, GeneRif – a database of the reference set of articles describing the function of a gene (*GeneRIF Stats - Gene - NCBI*, n.d., last accessed July 6, 2023), is dominated by 5 species: Humans, Mouse, Rat, Arabidopsis, Drosophila corresponding to about 92% of total coverage; Humans alone represent 63% of all GeneRifs. A highly skewed coverage of protein functional information concentrated in a largely anthropocentric fashion fails to benefit from the potential knowledge gained from studying the diversity of the natural world. The Analysis of Orthologous Collections (AOC) application is designed to be a one-stop shop for molecular sequence evaluation using state of the art methods and techniques. The pipeline is fully automated and incorporates recombination detection, a powerful force in shaping gene evolution which can produce spurious results if not considered. The application is simple to install and use, requiring few dependencies and few input files or configuration. We differentiate ourselves from other approaches in the field (Picard et al., 2020) through both comprehensive data preparation (Figure 1) and the breadth of selection analyses performed. Specifically, AOC integrates lineage-specific and site-level inference to detect pervasive and episodic selection, including negative and positive (diversifying and directional) selection, shifts in amino acid preferences, differential selection between predefined groups of branches, and changes in selection intensity (relaxation or intensification) (Lucaci et al., 2022).

2 Methods

2.1 Implementation

The AOC application is designed for comprehensive protein-coding molecular sequence analysis. AOC (a Snakemake workflow), allows for the inclusion of recombination detection, which is a powerful force in shaping gene evolution and critically important to correctly interpreting analytic results which are vulnerable to changing recombinant topologies. Lineage assignment allows for between-group comparisons of selective pressures using selection analysis. The application accepts CDS FASTA files in which each file corresponds to a single gene (i.e., a set of homologous sequences), typically retrieved from public databases such as NCBI Gene or curated by the user. AOC supports both single-gene analyses (one file) and multi-gene analyses (multiple files) within the same workflow.

2.2 Pre-processing

To generate multiple sequence alignments, we use MACSEv2 (Ranwez et al., 2018) due to its ability to create codon-aware multiple sequence alignment. We also measure the Tamura-Nei 1993 (TN93) genetic distance of alignments using the HyPhy implmentation of TN93. Recombination detection is automatically performed using Genetic Algorithm for Recombination Detection (GARD) (Kosakovsky Pond et al., 2006). A recombination-free set of alignment fragments is placed in the results folder where phylogenetic tree inference and downstream selection analysis are performed. For datasets where recombination is not detected this results in a single file for analysis. In datasets where recombination is detected, we parse out recombinant partitions into multiple files correcting for recombinant breakpoints which occur within a codon. Next, phylogenetic tree inference is done for all the recombination-free FASTA files, we perform phylogenetic inference via FastTree2 (Price et al., 2010). We perform tree labeling via the hyphy-analyses script Label-Trees method and results in one annotated tree with a designation for all lineages:(HyPhy-analyses): Label Trees.

2.3 Selection analysis

All recombination-free alignments and unrooted phylogenetic trees are evaluated using a suite of molecular evolutionary methods, each designed to address specific biological and statistical questions (described in Table 1; (Kosakovsky Pond et al., 2020; Spielman et al., 2019)).

Table 1. Summary of selection analysis methods

Method	Description
FEL	Locates codon sites with evidence of pervasive positive diversifying or negative selection. Answers: Which site(s) in a gene are subject to pervasive diversifying selection? (Kosakovsky Pond & Frost, 2005)
BUSTED[+S+MH]	Tests for gene-wide episodic selection while accounting for synonymous rate variation and multiple instantaneous substitutions. (Lucaci et al., 2023; Wisotsky et al., 2020)

Method	Description
MEME	Detects codon sites under episodic positive diversifying selection. Answers: Which site(s) are subject to episodic or pervasive diversifying selection? (Murrell et al., 2012)
aBSREL	Tests if positive selection has occurred on a proportion of branches. (Smith et al., 2015)
SLAC	Performs substitution mapping to detect pervasive diversifying selection. (Kosakovsky Pond & Frost, 2005)
BGM	Identifies groups of sites that are co-evolving. (Poon et al., 2008)
RELAX	Compares gene-wide selection pressure between a query clade and background lineages to detect relaxation/intensification. (Wertheim et al., 2015)
Contrast-FEL	Compares site-by-site selection pressure between query and background sequences. (Kosakovsky Pond et al., 2021)
FitMultiModel	Tests model fit by allowing multiple instantaneous substitutions. (Lucaci et al., 2021)
FUBAR	Identifies sites under pervasive selection using a fast Bayesian approach. (Murrell et al., 2013)

2.4 Visualizations and Tables

We provide a high-level executive summary and multiple-test correction of the selection analyses and on input files where available for information such as sequence divergence. In addition, we generate figures from all selection analyses along with accompanying summary result tables and figure legends which describe the results. Individual results, specifically output JSON files from HyPhy analyses may also be visualized using Hyphy-Vision or interactive ObservableHQ (Perkel, 2021) notebooks HyPhy: Interactive Observable Notebooks.

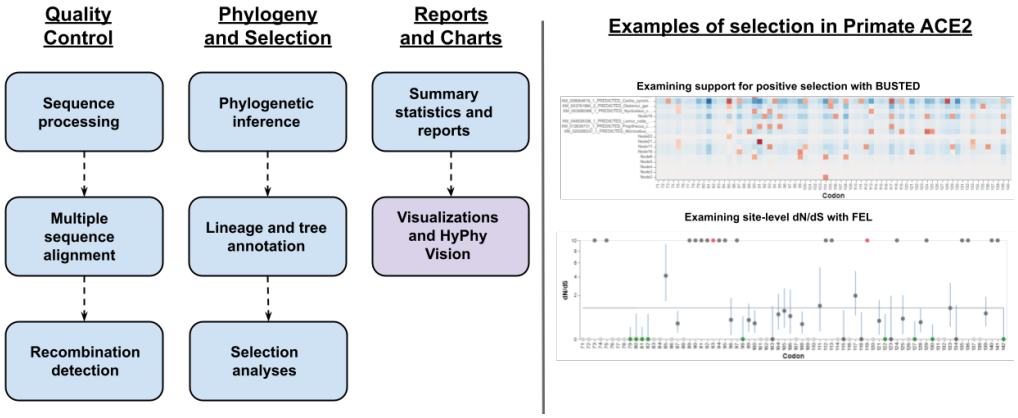

Figure 1: Flowchart diagram of the AOC (Snakemake) workflow and an example using Primate ACE2 data. The workflow consists of three parts, the first of which does quality control, and converts input transcript and protein files from the NCBI ortholog database into codon-aware alignments and checks for phylogenetic evidence of genetic recombination. The second part performs full maximum-likelihood phylogenetic inference and lineage annotation based on NCBI Taxonomy and runs a full suite of selection detection methods using HyPhy. The last part consists of summarizing results into useful tables and visualizations that can be used for post-hoc interpretation and interactions.

2.5 Testing and benchmarking

As an example, using an application of AOC, we were able to report on novel sites of adaptive evolution, broad relationships of coevolution, and independently verify previously reported results on the signatures of purifying selection in the mammalian BDNF (Lucaci et al., 2022) gene, which plays a critical role in brain development. We also explored the evolutionary history of the primate ACE2 protein. Data was accessed from NCBI via the Ortholog database. We downloaded FASTA files from 32 species, with RefSeq Transcripts and RefSeq Proteins (one sequence per species) and metadata in tabular form (CSV). Additional details of our analysis, including all intermediate and HyPhy JSON files are available in our GitHub repository. For more information on how selection analysis scales along with dataset complexity and size, we refer the reader to HyPhy benchmarking results available at HyPhy: Benchmarks and Profiling.

3 State of the field

Codon-based models of molecular evolution are widely implemented in established tools such as PAML(Yang, 2007) and HyPhy(Kosakovsky Pond et al., 2020), which provide statistically rigorous frameworks for detecting selection through site, branch, and branch-site models. While these platforms offer powerful inference engines, they are primarily designed for model execution rather than standardized, large-scale, and reproducible workflow orchestration. Existing wrappers and graphical interfaces, such as the HyPhy graphical user interface (HYPHY Vision), Datamonkey, PAML front-ends, and general workflow platforms like Galaxy, lower the barrier to running individual analyses but do not integrate alignment quality control, phylogenetic reconstruction, coordinated execution of multiple complementary selection tests, structured result aggregation, and publication-ready visualization within a unified, reproducible framework. AOC addresses this gap by building on established inference engines (particularly HyPhy), and embedding them within a modular Snakemake-based system that formalizes best practices for codon-aware evolutionary analysis. Rather than introducing new statistical methodology, AOC contributes a reproducible infrastructure that enables scalable, consistent, and comparative application of state-of-the-art evolutionary models across genes and ortholog collections, addressing a critical workflow-level bottleneck in the field.

4 Statement of need

Comparative evolutionary analyses of protein-coding genes often require coordinating multiple complex steps: alignment quality control, phylogenetic reconstruction, branch labeling, and complementary selection tests across many genes and datasets. While powerful tools such as HyPhy exist, running these analyses reproducibly at scale typically requires substantial scripting, manual intervention, and fragmented result handling. AOC addresses this gap by providing a unified, reproducible workflow that integrates codon-aware alignment processing, tree inference, structured branch labeling, multiple selection analyses, and standardized result aggregation within a single framework. The target audience includes evolutionary biologists, comparative genomicists, molecular evolution researchers, and computational biologists who require scalable, transparent, and publication-ready evolutionary inference across large gene collections or multi-species datasets.

5 Software design

AOC is implemented using Snakemake to ensure reproducibility, scalability, and transparent dependency management across multi-step evolutionary analyses. The workflow adopts a modular, per-gene architecture in which each CDS FASTA file represents a single gene (a set of homologous sequences), enabling straightforward parallelization and fault isolation while supporting both single- and multi-gene analyses. Rather than reimplementing statistical models, AOC integrates established tools in the HyPhy suite (Table 1), prioritizing methodological robustness and community validation over less reliable custom implementations. These design choices allow users to move from raw sequence data to interpretable selection inferences in a reproducible and extensible manner, particularly for large or underexplored gene sets.

6 Research impact statement

AOC is designed to support reproducible and scalable molecular evolutionary analyses and has been validated through application to diverse coding sequence datasets, with fully reproducible workflows and example datasets provided to demonstrate end-to-end functionality. The software integrates established methods in HyPhy within a unified pipeline, lowering the barrier to performing complex selection analyses across genes and lineages. Early indicators of community readiness include its use in internal and collaborative research projects, and interest from external users seeking reproducible evolutionary analysis workflows (Lucaci et al., 2022; Martin et al., 2022; Silva et al., 2023; Zehr et al., 2023). By enabling standardized, multi-gene selection analyses with minimal setup, AOC addresses a clear need for accessible, extensible pipelines in comparative genomics and evolutionary biology.

7 Conclusion

Modern molecular sequence analysis pipelines enable the detection of natural selection and generation of testable biological hypotheses (Martin et al., 2021, 2022; Silva et al., 2023; Tegally et al., 2022; Viana et al., 2022; Zehr et al., 2023). The AOC workflow is designed to play a role in scientific and medical discovery by providing a simple-to-use software application for molecular sequence analysis, especially for insights into unexplored genetic datasets.

Acknowledgements

We would like to thank members of the HyPhy and Datamonkey teams for their contributions to this project, method development, and the maintenance of state-of-the-art molecular sequence analysis software. This work was supported by a NIH grant (GM151683) to SLKP.

AI usage disclosure

Generative AI tools were used to assist with aspects of code development and manuscript preparation. All outputs were critically evaluated, tested, and validated by the authors to ensure correctness and reproducibility. The authors retain full responsibility for the software design, implementation, and scientific conclusions.

References

GeneRIF stats - gene - NCBI. (n.d.). https://www.ncbi.nlm.nih.gov/gene/generif-stats.

Kosakovsky Pond, S. L., & Frost, S. D. W. (2005). Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*, *22*(5), 1208–1222.

Kosakovsky Pond, S. L., Poon, A. F. Y., Velazquez, R., Weaver, S., Hepler, N. L., Murrell, B., & al., et. (2020). HyPhy 2.5—a customizable platform for evolutionary hypothesis testing using phylogenies. *Molecular Biology and Evolution*, *37*(1), 295–299.

Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H., & Frost, S. D. W. (2006). Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular Biology and Evolution*, *23*(10), 1891–1901.

Kosakovsky Pond, S. L., Wisotsky, S. R., Escalante, A., Magalis, B. R., & Weaver, S. (2021). Contrast-FEL-a test for differences in selective pressures at individual sites among clades and sets of branches. *Molecular Biology and Evolution*, *38*(3), 1184–1198.

Lucaci, A. G., Notaras, M. J., Kosakovsky Pond, S. L., & Colak, D. (2022). The evolution of BDNF is defined by strict purifying selection and prodomain spatial coevolution, but what does it mean for human brain disease? *Translational Psychiatry*, *12*(1), 1–17.

Lucaci, A. G., Wisotsky, S. R., Shank, S. D., Weaver, S., & Pond, S. L. K. (2021). Extra base hits: Widespread empirical support for instantaneous multiple-nucleotide changes. *PLOS ONE*, *16*(3), e0248337.

Lucaci, A. G., Zehr, J. D., Enard, D., Thornton, J. W., & Kosakovsky Pond, S. L. (2023). Evolutionary shortcuts via multi-nucleotide substitutions and their impact on natural selection analyses. *Molecular Biology and Evolution*.

Martin, D. P., Lytras, S., Lucaci, A. G., Maier, W., Grüning, B., Shank, S. D., & al., et. (2022). Selection analysis identifies clusters of unusual mutational changes in omicron lineage BA.1 that likely impact spike function. *Molecular Biology and Evolution*, *39*(4), msac061.

Martin, D. P., Weaver, S., Tegally, H., San, J. E., Shank, S. D., Wilkinson, E., & al., et. (2021). The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell*, *184*(20), 5189–5200.

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., & al., et. (2013). FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection. *Molecular Biology and Evolution*, *30*(5), 1196–1205.

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., & Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics*, *8*(7), e1002764.

Perkel, J. M. (2021). Reactive, reproducible, collaborative: Computational notebooks evolve. *Nature*, *593*(7857), 156–157.

Picard, L., Ganivet, Q., Allatif, O., Cimarelli, A., Guéguen, L., & Etienne, L. (2020). DGINN,

an automated and highly-flexible pipeline for the detection of genetic innovations on protein-coding genes. *Nucleic Acids Research*, *48*(18), e103.

Poon, A. F. Y., Lewis, F. I., Frost, S. D. W., & Kosakovsky Pond, S. L. (2008). Spidermonkey: Rapid detection of co-evolving sites using bayesian graphical models. *Bioinformatics*, *24*(17), 1949–1950.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N., & Delsuc, F. (2018). MACSE v2: Toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Molecular Biology and Evolution*, *35*(10), 2582–2584. https://doi.org/10.1093/molbev/msy159

Silva, S. R., Miranda, V. F., Michael, T. P., Plachno, B. J., Matos, R. G., Adamec, L., & al., et. (2023). The phylogenomics and evolutionary dynamics of the organellar genomes in carnivorous utricularia and genlisea species (lentibulariaceae). *Molecular Phylogenetics and Evolution*, *181*, 107711.

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., & Kosakovsky Pond, S. L. (2015). Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular Biology and Evolution*, *32*(5), 1342–1353.

Spielman, S. J., Weaver, S., Shank, S. D., Magalis, B. R., Li, M., & Kosakovsky Pond, S. L. (2019). Evolution of viral genomes: Interplay between selection, recombination, and other forces. In M. Anisimova (Ed.), *Evolutionary genomics: Statistical and computational methods* (pp. 427–468). Springer. https://doi.org/10.1007/978-1-4939-9074-0_14

Stoeger, T., Gerlach, M., Morimoto, R. I., & Amaral, L. A. N. (2018). Large-scale investigation of the reasons why potentially important genes are ignored. *PLOS Biology*, *16*(9), e2006643.

Tegally, H., Moir, M., Everatt, J., Giovanetti, M., Scheepers, C., Wilkinson, E., & al., et. (2022). Emergence of SARS-CoV-2 omicron lineages BA.4 and BA.5 in south africa. *Nature Medicine*, *28*(9), 1785–1790.

Viana, R., Moyo, S., Amoako, D. G., Tegally, H., Scheepers, C., Althaus, C. L., & al., et. (2022). Rapid epidemic expansion of the SARS-CoV-2 omicron variant in southern africa. *Nature*, *603*(7902), 679–686.

Wertheim, J. O., Murrell, B., Smith, M. D., Kosakovsky Pond, S. L., & Scheffler, K. (2015). RELAX: Detecting relaxed selection in a phylogenetic framework. *Molecular Biology and Evolution*, *32*(3), 820–832.

Wisotsky, S. R., Kosakovsky Pond, S. L., Shank, S. D., & Muse, S. V. (2020). Synonymous site-to-site substitution rate variation dramatically inflates false positive rates of selection analyses: Ignore at your own peril. *Molecular Biology and Evolution*, *37*(8), 2430–2439.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Zehr, J. D., Kosakovsky Pond, S. L., Millet, J. K., Olarte-Castillo, X. A., Lucaci, A. G., Shank, S. D., & al., et. (2023). Natural selection differences detected in key protein domains between non-pathogenic and pathogenic feline coronavirus phenotypes. *Virus Evolution*, *9*(1), vead019.