Published on Mar 05, 2024 by Arcadia Science **DOI**: 10.57844/arcadia-5b41-637f

How can we improve upon and expand the scope of our phylogenomic inferences?

We're seeking feedback on NovelTree, our modular phylogenomic workflow. We'd appreciate your insights into how we can improve gene family inference, incorporate protein structure predictions, and expand to whole-genome data as input.

Contributors (A-Z)

Prachee Avasthi, Megan L. Hochstrasser, Jasmine Neal, Austin H. Patton, Ryan York

Version 2 · Mar 31, 2025

Purpose

We recently released NovelTree — a modular Nextflow workflow that takes proteomes from diverse organisms as input and conducts phylogenetic inference for thousands of genes **[1]**. Since its release, we've started to consider alternative methodologies and input data to facilitate a range of new use cases for the method.

We're seeking feedback from the community on how we might improve our approach to inferring gene families, perform protein structural phylogenetics, and conduct phylogenomic inference of not only coding sequences but genome-wide synteny.

Whether you're a phylogeny lover, develop phylogenetic methods, or apply them as a systematist or comparative evolutionary biologist, we'd love to hear from you!

- This is a follow-up to work described in a **prior pub**, "NovelTree: Highly parallelized phylogenomic inference." Visit that pub for complete background info and context.

Background on the original pub

By performing phylogenetic inference for thousands of genes, we can both develop and test hypotheses about the role of proteins and their constituent gene families over evolutionary time scales, teaching us about the tempo and mode of evolution across the tree of life. To do this, we developed NovelTree — a modular Nextflow workflow that takes proteomes from diverse organisms as input and infers orthology, gene family trees, species trees, and gene family evolutionary dynamics **[1]**. The pipeline is a helpful tool to, for instance, associate gene family expansion or contraction with specific organismal traits and generate stronger hypotheses for the function of uncharacterized proteins.

Our latest questions

NovelTree has proven useful for many tasks — generating phylogenomic datasets, mapping genome-wide evolutionary patterns across broad portions of the tree of life, and more. However, there are still multiple areas wherein the framework could be optimized, improved, or expanded upon. Simultaneously, the tools, data types, and theoretical frameworks used in computational and comparative evolutionary research are rapidly changing. We'd therefore love to elicit feedback on the following questions. Rather than leading with our own thoughts on the matter, we'd like to hear your ideas, independent of what we've already begun to consider. We hope this can help us understand what may be useful to the community as a whole.

How can I weigh in?

We hope you'll respond publicly to our questions below by selecting/highlighting the question you'd like to answer, clicking the comment icon, and typing in your thoughts (as shown in the GIF below)! You'll need a PubPub account to do this, but it's free and quick to make one. Here's a quick tutorial on how to comment.

What's the best strategy for gene family inference?

Nearly all gene-based phylogenomic analyses rely on the accurate inference of gene families. Despite this, the methodology underlying gene family inference has historically received relatively little scrutiny compared to that for multiple sequence alignment or inference of phylogenetic trees. Right now, NovelTree uses a procedure based on OrthoFinder's **[2]**, clustering protein sequences into gene families (orthogroups) based on their sequence similarity. We extend this procedure by assessing the impact of the MCL clustering algorithm's inflation parameter using the COGEQC functional annotation metric, which quantifies the extent to which biologically informative protein annotations are distributed within vs. split among gene families **[3]**. However, much can still be done to improve, add to, or extend how we

assess the accuracy of orthogroup inference or how we infer gene families entirely. We're particularly interested in how we can perform gene family inference without the use of protein functional annotations that are frequently unavailable for non-model organisms. Similar to how we've implemented various methods for multiple sequence alignment and gene/species tree inference, we'd like to have multiple gene family inference methods available for NovelTree users.

> Other than the COGEQC functional annotation metric, how might we assess the quality of gene family inference?

> How might we improve our gene family inference procedure (e.g., using alternative methodology), and what types of data would be most suited to doing so?

> Going a step further, how might we infer gene family evolutionary dynamics more efficiently, without loss of accuracy?

How can we level up our phylogenomic inferences using newly abundant protein structure predictions?

Our phylogenetic analyses are based on protein (i.e. amino acid) sequences, as these are readily available in various public databases. Yet given the recent advances in predicting protein structure with tools such as AlphaFold **[4]** and ESMFold **[5]**, there are many opportunities to develop novel statistics accounting for both sequence and structural evolutionary patterns. One small example: it might be possible to infer shifts in the adaptive evolution of certain proteins by investigating discontinuities between sequence and structural similarity. Generating a theoretical and applied framework for this would open up new possibilities for identifying interesting evolutionary patterns.

How could we best incorporate protein structural predictions into phylogenomic analyses?

How can we move beyond just proteins and use whole genomes for phylogenomic analysis?

One of the biggest limitations of NovelTree and other related phylogenomic pipelines is their exclusive applicability to protein-coding sequences. Expanding the framework to accommodate genome-wide sequence data (e.g. whole-genome assemblies from multiple species) would empower us in several ways. Some example benefits include: 1) Expanding our scope beyond identifying orthologous genes and proteins to inferring synteny across the genome. 2) More exhaustively studying patterns of adaptive and non-adaptive molecular evolution of coding sequences (e.g., the ratio of non-synonymous to synonymous substitutions, *dN*/*dS*). 3) Developing tools to investigate the evolution of non-coding and regulatory regions of the genome.

Given these possibilities and numerous others, we'd love to explore means for incorporating whole-genome sequence data into our phylogenomic analyses.

How might we extend NovelTree to conduct truly whole-genome phylogenomics, reaching beyond the scope of coding sequences alone?

Let us know what you think!

We've outlined several outstanding questions and potential development opportunities that should inform our next steps in enhancing NovelTree and, hopefully, future phylogenomics tools from others. That said, this is not an exhaustive list of questions related to NovelTree or related applications of the evolutionary datasets it generates. We encourage public responses to the questions posed above — but we'd love to hear about anything else that came to mind while reading the pub!

References

1. Celebi FM, Chou S, McGeever E, Patton AH, York R. (2024). NovelTree: Highly parallelized phylogenomic inference. https://doi.org/10.57844/ARCADIA-Z08X-V798

2. Emms DM, Kelly S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. https://doi.org/10.1186/s13059-019-1832-y

3. Almeida-Silva F, Van de Peer Y. (2023). Assessing the quality of comparative genomics data and results with the*cogeqc*R/Bioconductor package. https://doi.org/10.1101/2023.04.14.536860

4. Barrio-Hernandez I, Yeo J, Jänes J, Mirdita M, Gilchrist CLM, Wein T, Varadi M, Velankar S, Beltrao P, Steinegger M. (2023). Clustering predicted structures at the scale of the known protein universe. https://doi.org/10.1038/s41586-023-06510-w

5. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. https://doi.org/10.1126/science.ade2574