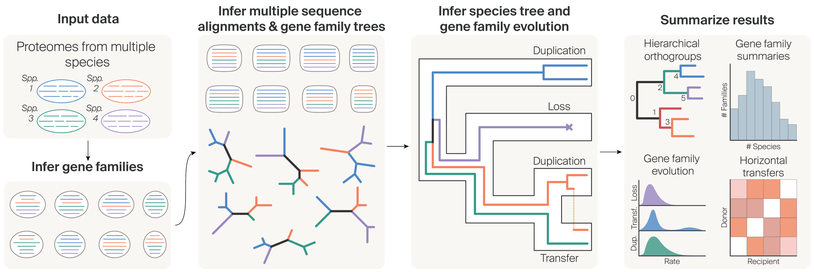
NovelTree: Highly parallelized phylogenomic inference

Presented by  Austin Patton 𝕏 @austinhpatton

ARCADIA

Background

Phylogenomics is a powerful, but burdensome tool



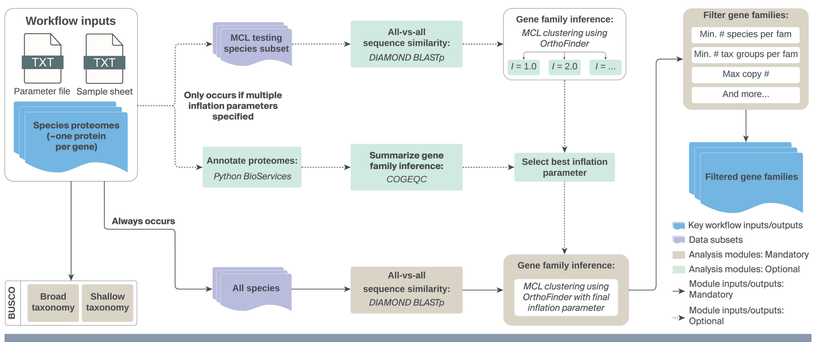
For more, read the full project narrative: **bit.ly/NovelTree**

Key development goals:

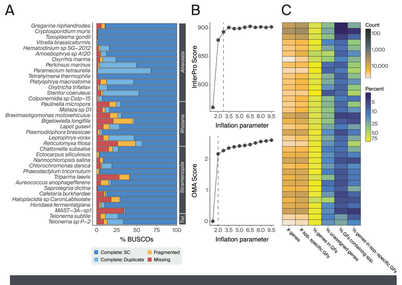
1. Efficient & highly parallelized implementation of core phylogenomic analyses of whole-proteomes
2. Allow for user-determined methodological flexibility at most stages of the workflow
3. Apply a principled, empirically motivated approach to improve gene family inference
4. Implement alignment methods suited to datasets with prevalent sequence-length heterogeneity
5. Automate scaleable inference of gene family evolutionary dynamics (i.e. duplication, transfer, loss)

Approach

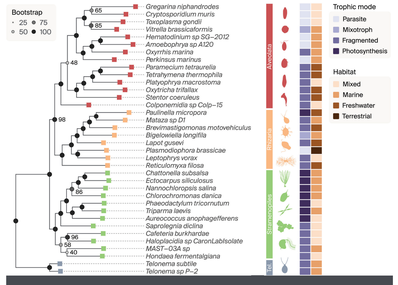
A Swiss army knife solution for phylogenomic analyses at a tree-of-life scale



Gene Family Inference. Detailed workflow diagram outlining the steps involved and implemented in the first stage of the workflow, inferring gene families from species proteomes.

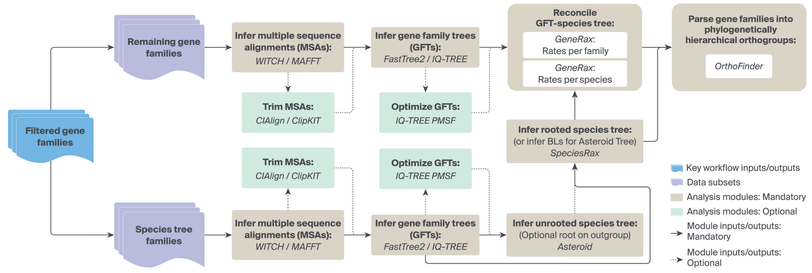


BUSCO proteome quality and summary of gene family inference.



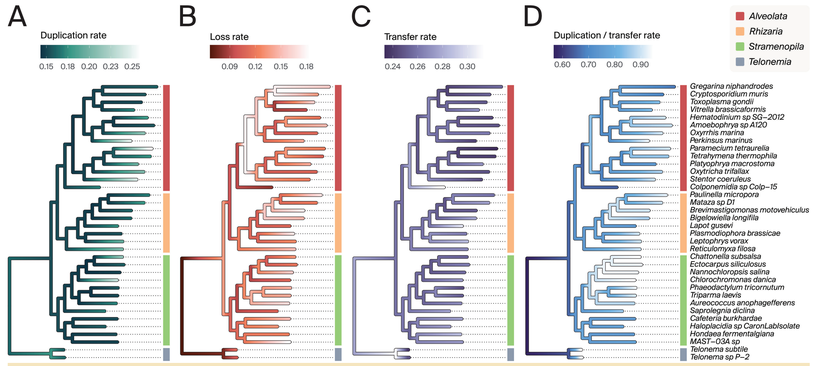
Inferred species tree and ecology of 36 species of TSAR eukaryotes.

Highly modular and customizable phylogenetic analysis

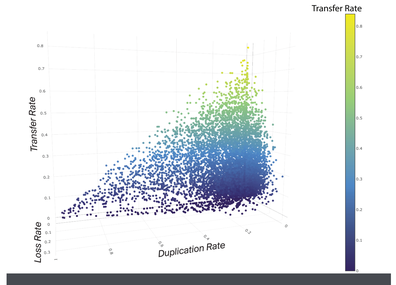


Phylogenetic Inference. Detailed workflow diagram outlining the steps involved and implemented in the second stage of the workflow; phylogenetic inference for all gene families, and inference of gene family evolutiary dynamics (i.e. gene duplication, transfer, and loss).

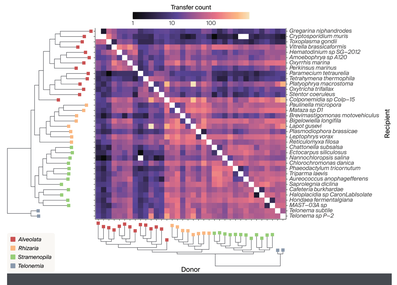
From phylogenetic inference to gene family inference of evolutionary history



Rates of gene duplication, transfer, and loss are inferred for each species and internal node. These rates are inferred for each gene family. Mean rates across all families are plotted above.



DTL Rates inferred across all gene families.



Counts of horizontal transfer events summed across all families, between species.

Next Steps

The upcoming release of NovelTree will implement a number of changes focused on improving efficiency, reliability, and ease of use. That is, we would like to ensure that NovelTree may be more easily applied to all types of datasets, and hope to make specification of different run settings more user-friendly for those with limited experience with NextFlow.

This release will include:

- Dynamic, and Improved handing of large, complex gene families to ensure timely and efficient run completion
- Simple run configurations for “fast” or “exhaustive” analyses
- Standardized tabular and interactive visual summaries of exemplar results for each workflow run
- And more!

Outstanding Questions

We'd appreciate feedback on any part of this work, but we're especially curious about the following:

- How might we improve our gene family inference procedure (e.g. without using protein annotations)?
- How can we more efficiently infer gene family evolutionary dynamics, without loss of accuracy?
- Can we implement protein structural phylogenetics in a similarly scaleable way?
- How to best extend NovelTree to be applicable to the rest of the genome (e.g. leveraging synteny)?

Leave Feedback!

Comment on the pub:

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

𝕏 **Post** with #NovelTree

bit.ly/noveltree-open-questions

All other published work: research.arcadiascience.com

A NOTE ON SHARING WITH US!

Part of our mission is to share as much useful research as we can.

If you choose to share a protocol or other useful information with us after viewing this poster, please understand that we may act upon this knowledge and share it when we publish our work. We publish quickly on an independent platform, so this may happen soon after you share, and we cannot wait for you to publish elsewhere.

If you decide to share anyway, yay! That's what science is all about. If your input is useful, we will include you as a contributor to the publication and explain that your role was in providing "Critical Feedback," likely with an additional description of what you shared.

tl,dr — If you're not ready for everyone to know about something, please refrain from sharing it with us.

Contributors (A–Z)

Prachee Avasthi • Supervision

Feridun Mert Celebi • Software, Validation

Seemay Chou • Supervision

Jonathan A. Eisen (*Advisor: UC Davis*) • Critical Feedback

Megan L. Hochstrasser • Editing, Visualization

Elizabeth A. McDaniel • Critical Feedback, Validation

Erin McGeever • Software

Gilad Mishne • Software

Jasmine Neal • Writing

Austin H. Patton • Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing

Taylor Reiter • Critical Feedback, Validation

Dennis A. Sun • Editing

Ryan York • Conceptualization, Editing, Supervision

References

1. Buchfink B, et al. (2021). https://doi.org/10.1038/s41592-021-01101-x
2. Almeida-Silva F, Van de Peer Y. (2023). https://doi.org/10.1101/2023.04.14.536860
3. Emms DM, Kelly S. (2019). https://doi.org/10.1186/s13059-019-1832-y
4. Shen C, et al. (2022). https://doi.org/10.1089/cmb.2021.0585
5. Katoh K, Standley DM. (2013). https://doi.org/10.1093/molbev/mst010
6. Tumescheit C, et al. (2022). https://doi.org/10.7717/peerj.12983
7. Steenwyk JL, et al. (2020). https://doi.org/10.1371/journal.pbio.3001007
8. Minh BQ, et al. (2020). https://doi.org/10.1093/molbev/msaa015
9. Price MN, et al. (2010). https://doi.org/10.1371/journal.pone.0009490
10. Wang H-C, et al. (2017). https://doi.org/10.1093/sysbio/syx068
11. Morel B, et al. (2022). https://doi.org/10.1093/bioinformatics/btac832
12. Morel B, et al. (2022). https://doi.org/10.1093/molbev/msab365
13. Morel B, et al. (2020). https://doi.org/10.1093/molbev/msaa141