LFSC 507 Final Project Research Plan

**Streamlining Orthology Inference through Upfront Paralog Exclusion**

Jackson Turner

October 8, 2024

PROJECT DESCRIPTION

The reconstruction of accurate evolutionary relationships is dependent upon the inference of orthologous genes, or orthologs (Emms & Kelly 2019; Lozano-Fernandez 2022; Fleming et al. 2023). Orthologs, single-copy genes evolved from a common ancestor, are effective and well-utilized representatives of a taxonomic group’s evolutionary history as they lack copies which could possess divergent gene histories (Lozano-Fernandez 2022; Fleming et al. 2023). The process of orthology inference compares sequences between organisms to identify single-copy genes (Emms & Kelly 2019; Lozano-Fernandez 2022). Programs that conduct orthology inference such as OrthoFinder form the underpinning of phylogenetic reconstructions, and especially those that lack publicly curated reference orthologs (Lozano-Fernandez 2022; Yang et al. 2023). Orthology inference is often computationally demanding, as the number of comparisons made is proportional to the number of organisms included in the attempted phylogenetic reconstruction and the volume of available genomic content (Emms & Kelly 2019). The genomes of certain taxonomic groups, including plants or fungi, contain extensive repetitive sequences (Montoliu-Nerin et al. 2021; Yang et al. 2023). These sequences may contain paralogs, multi-copy genes from which a group’s accurate evolutionary history cannot be reconstructed (Lozano-Fernandez 2021; Yang et al. 2023). Removing paralogs from genomic data of individual organisms may reduce the computational burden of downstream orthology inference by reducing the number of sequence comparisons.

Filtering paralogs from individual organisms in this way may be accomplished through the alignment of a set of protein sequences against itself through DIAMOND and the exclusion of highly similar sequence matches. DIAMOND allows for efficient protein-to-protein sequence comparisons while remaining sensitive enough to identify divergent matches (Buchfink et al. 2021), resulting in a table of sequence matches. A custom python script, using this table and a protein sequence file from which the table was generated, may then exclude genes with multiple sequence matches. Only the “query accession” column from DIAMOND sequence match tables will be considered for this script. Orthology inference may be conducted upon sequences filtered for paralogs in this way.

The goal of this project is to expedite orthology inference with upfront removal of paralogs within sampled taxa. A dataset of 29 annotated plant exomes will be filtered for orthologs using DIAMOND blastp searches for each taxon. Identified orthologs will be used for all-vs-all orthology inference with OrthoFinder under default parameters and 8 cores. The wall time of this process will be compared against an identical OrthoFinder job with paralogs retained.

A preliminary python script for this purpose has been written and tested upon an annotated exome (254,895 sequences) from a plant (*Henckelia ceratoscyphus*) using a DIAMOND sequence match table (maximum of 10 sequence matches per gene, 50% query cover, and maximum e-value of 1-e25). Using a subset of 1,000 sequence matches identified with DIAMOND revealed 13 orthologs in < 5 seconds. Executing this script with the full set of 9,499,314 sequence matches identified 71,651 orthologs in 13.5 hours. The relatively low CPU and memory footprint for this python script (1 core and < 5 GB memory respectively) allow it to be parallelized. Further work is necessary to adapt this script to better align with the goals of this project and scale it for the remaining 28 taxa.

REFERENCES

Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using \_\_\_\_\_\_DIAMOND. Nat Methods 18, 366–368 (2021).

Emms, David M., and Steven Kelly. “OrthoFinder: Phylogenetic Orthology Inference for Comparative \_\_\_\_\_\_Genomics.” Genome Biology 20, no. 1 (December 2019): 238.

Fleming, James F., Alberto Valero‐Gracia, and Torsten H. Struck. “Identifying and Addressing \_\_\_\_\_\_Methodological Incongruence in Phylogenomics: A Review.” Evolutionary Applications 16, no. 6 \_\_\_\_\_\_(June 2023): 1087–1104.

Lozano-Fernandez, Jesus. “A Practical Guide to Design and Assess a Phylogenomic Study.” Genome \_\_\_\_\_\_Biology and Evolution 14, no. 9 (September 1, 2022): evac129.

Montoliu-Nerin, Merce, Marisol Sánchez-García, Claudia Bergin, Verena Esther Kutschera, Hanna \_\_\_\_\_\_Johannesson, James D. Bever, and Anna Rosling. “In-Depth Phylogenomic Analysis of \_\_\_\_\_\_Arbuscular Mycorrhizal Fungi Based on a Comprehensive Set of de Novo Genome Assemblies.” \_\_\_\_\_\_Frontiers in Fungal Biology 2 (September 29, 2021): 716385.

Yang, Li-Hua, Xi-Zuo Shi, Fang Wen, and Ming Kang. “Phylogenomics Reveals Widespread \_\_\_\_\_\_Hybridization and Polyploidization in Henckelia (Gesneriaceae).” Annals of Botany 131, no. 6 \_\_\_\_\_\_(July 10, 2023): 953–66.

DATA AVAILABILITY

Preliminary data for this project, including the Jupyter notebook containing the python script and data used to initiate it, is available at <https://github.com/jacksonhturner/lfsc_507/tree/main/final_project> or upon request.