bioRxiv, 2024, 1–13
doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

Poplar: A Phylogenetics Pipeline

Elizabeth Koning[1,2] and Raga Krishnakumar[1,*]

[1]Biosecurity and Bioassurance Department, Sandia National Laboratories, 7011 East Avenue, Livermore, CA, 94550, USA and [2]Department of Computer Science, University of Illinois, Urbana-Champaign, 201 North Goodwin Avenue, Urbana, IL, 61801, USA

[*]Corresponding author. rkrishn@sandia.gov

Abstract

Motivation: Generating phylogenetic trees from genomic data is essential in understanding biological systems. Each step of this complex process has received extensive attention in the literature, and has been significantly streamlined over the years. Given the volume of publicly available genetic data, obtaining genomes for a wide selection of known species is straightforward. However, analyzing that same data in order to generate a phylogenetic tree is a multi-step process with legitimate scientific and technical challenges, and often requires a significant input from a domain-area scientist.
Results: We present Poplar, a new, streamlined computational pipeline, to address the computational logistical issues that arise when constructing phylogenetic trees. It provides a framework that runs state-of-the-art software for essential steps in the phylogenetic pipeline, beginning from a genome with or without an annotation, and resulting in a species tree. Running Poplar requires no external databases. In the execution, it enables parallelism for execution for clusters and cloud computing. The trees generated by Poplar match closely with state-of-the-art published trees. The usage and performance of Poplar is far simpler and quicker than manually running a phylogenetic pipeline.
Availability and Implementation: Freely available on GitHub at `https://github.com/sandialabs/poplar`. Implemented using Python and supported on Linux.
Supplementary Information: Newick versions of the reference and generated trees.

1. Introduction

Creating phylogenetic trees from sequenced genomes is core to most areas of bioscience and biotechnology, as an understanding of the genomic relationships between organisms is critical for prevention, surveillance, intervention and optimization of biological processes (Kapli et al., 2020). In particular, understanding microorganismal diversity is critical both for defending against pathogens and for optimizing organisms to support the bioeconomy, with examples including bacteria, fungi and even members of the plant kingdom such as algae (Spatafora et al., 2017; Nakano et al., 2023; Navarro, 2021; Azouri et al., 2021). In addition, phylogenetics play a critical role in viral, bacterial and fungal pathogen and pandemic monitoring, as exemplified recently by SARS CoV2 and fungal plant pathogens (De Maio et al., 2023; Turakhia et al., 2022; Ki and Terhorst, 2022; Dort et al., 2023).

To date, there are numerous strategies and methods associated with building these trees. During this process, it is essential to combine biological meaning with computational efficiency and flexibility. While early approaches relied on observable physical and metabolic characteristics, the expansion of genetics and genomics over the decades has added a significant amount of resolution to these methods (Hug et al., 2016b).

Building a species tree requires genomic data from each included species to be sequenced, analyzed, compared, and then constructed into a tree. The specific steps may differ, but a typical pipeline includes sequencing, identifying homologs, aligning sequences, inferring gene trees, and finally constructing a species tree (Young and Gillung, 2020). Each step requires one or more software tools requiring various inputs and data formats. In addition, software that offers sequence-to-tree analysis has its own limitations.

For example, MEGA is a comprehensive tool for tree-building (Tamura et al., 2021). However, MEGA requires either a genome annotation or whole genome multiple genome alignment, and while this can be extremely helpful for some applications, it can have limitations in other situations, for example, if users want to focus on specific gene subsets without requiring significant manual intervention, which might be challenging for non-subject matter experts. Another challenge a user may face if attempting to use whole genomes is the requirement of a single sequence for each species, which excludes assemblies of scaffolds. If there is an annotation, the user must

© The Author 2024. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

use another tool to group gene sequences before alignment. Furthermore, MEGA is not open source, so users cannot adapt it or run on incompatible operating systems. It also does not support parallelism.

EasyCGTree is another tool for inferring phylogenetic trees Zhang et al. (2023). It includes options for multiple inference methods and does include a step to group sequences to form gene trees. However, it does not include parallelism, which is a significant roadblock when building increasingly larger trees.

A third software tool for tree inference is Read2Tree (Dylus et al., 2023). While the authors of Read2Tree report highly accurate output trees, Read2Tree requires an internet connection during its run as well as a download from the author's Orthologous Matrix (OMA) database before running. For some applications, the required remote connection would be a security concern or a practical restriction. As for the data required from the external database, the database does not include sequences for all species. While we have not done a comprehensive examination of the included species, in our comparative testing, we found OMA includes very few fungal sequences. Read2Tree also uses reads, rather than assemblies. In some cases, this may be very useful, but if genomes are selected from the NCBI database, they will typically have assemblies available, which allows for downloading and managing a much smaller amount of data. In addition, reads can be assembled into genomes or transcriptomes, but the reverse is not true.

To address the issues of ease of use, parallelism and flexibility outlined above, we present Poplar, a novel software tool for the pipeline from assembled genome to phylogenetic tree. As a tree species, "poplar" connects the concept of biological trees with the phylogenetic trees the software infers. Specifically, poplar trees have a reputation for fast growth, and Poplar, the software tool, is designed to quickly grow a phylogenetic tree (Britannica, 2023). Tasks as seemingly simple as converting file types or as time consuming as genome annotation can bog down the development of a phylogenetic inference pipeline, and Poplar addresses the issue by smoothing the connections between the numerous steps in tree inference and using annotated or unannotated genomes.

2. Background

To obtain a species tree from a collection of assembled genomes typically requires a variety of software packages. While each step in the pipeline has well established and ever improving tools, those tools have unique file formats and input requirements that are frequently incompatible with other tools. Many steps have entirely independent computation, but executing in parallel requires manually initiating parallel jobs. Poplar addresses these concerns and manages converting file types, renaming sequences, and running tasks in parallel.

The key steps are:

1. Assemble genome from reads (if using next-generation sequencing data as opposed to a pre-assembled genome)
2. Identify gene sequences
3. Group sequences into genes across species
4. Align gene sequences
5. Create gene trees
6. Merge gene trees into species tree

There may be variation in these steps, many of which are described in (Kapli et al., 2020). Some approaches will require whole genome alignment (WGA) rather than perform alignments for gene sequences, for example. WGA is done through specialized tools such as Cactus (Armstrong et al., 2020) or more general alignment tools such as RAxML-NG and Muscle (Kozlov et al., 2019; Edgar, 2022). WGA is a significantly costly computation and therefore larger alignments are often not feasible on computational resources that are readily available to most researchers. Kmer-based distance methods can also substitute for alignment, and they bypass using full gene sequences at all, but rather use dictionaries of kmers. These methods are extremely helpful for fast examination of relationships between genomes, particularly large numbers of large genomes Nasko et al. (2018); Manekar and Sathe (2018); Moeckel et al. (2024). However, the caveat of using kmer-based methods is that the gain in speed sacrifices some accuracy, and while this loss in accuracy can be mitigated by having more individual genomes to analyze (i.e. more pairwise distance comparisons can make for a more accurate overall tree), we focused Poplar to use methods (discussed below) that are more resilient and accurate even with low genome/sequence numbers.

Within gene-driven trees are a specific subset that are frequently used - alignment based on ribosomal genes. While there are numerous examples of this in the literature, a significant one is the recent novel view of the tree of life, where multiple ribosomal genes were used (Hug et al., 2016a). While focusing on ribosomal genes can standardize comparisons, it can be limiting in cases needing discernment of similar organisms, or strains within a species, where the ribosomal genes are likely to be very similar and may not provide the desired resolution.

Most often, phylogenetic trees are created through ad-hoc pipelines, such as in the 1000 Plants Initiative and the Avian Phylogenomics Project Wickett et al. (2014); Jarvis et al. (2015). This has the benefit of allowing the scientists to use tools they expect to be most suitable for their data in each step, but it also requires implementing many elements of the application, such as the tools to connect the main steps. Ad hoc pipelines also face challenges in presentation. The methods to construct published trees are often not specific enough for other researchers to adapt or reproduce. Since code for the pipelines are often not included with publications, each research group must implement their own pipeline, which may or may not share characteristics with other pipelines, leading to wasted energy and a lack of clarity in the origins of various phylogenetic trees. In designing Poplar, our goal is to implement a tried-and-tested combination of the algorithms required to complete the key steps outlined below. We acknowledge that there are alternatives, and ultimately the long-term goal of Poplar is to be modular and allow for flexibility in algorithm usage, but we describe here a pipeline that uses established tools in an efficient and accessible way. Specifically we focus on optimizing a few processes outlined below.

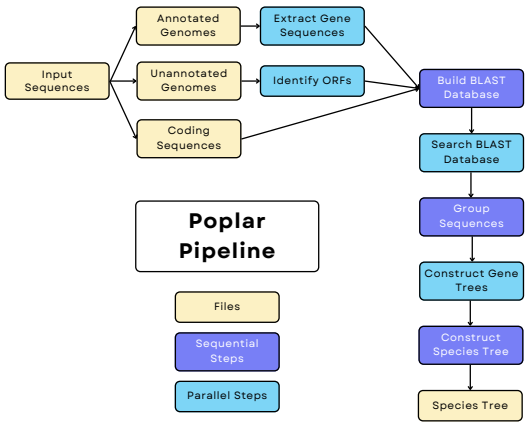

Fig. 1: Illustration of the Poplar Pipeline

Inferring gene trees requires collections of gene sequences. Identifying genes as genes that occur across multiple species requires a genome annotation. De novo gene prediction identifies probable genes based on patterns within the genomic sequence, while alignment based methods rely on a reference database in order to compare sequences in a genome with known genes (Vallender, 2009). Either approach is a significant task in terms of both labor and computation. On the other hand, identifying homologs that do not have an identified function is a simpler task but provides far less information. They may resemble one another due to shared ancestry or horizontal gene transfer, or the sequences may have no relation. Occasionally, this task is done with grouping algorithms such as DBSCAN (Edla and Jana, 2012). If an annotation is available for some genomes of interest but not others, that typically means that the unannotated genomes will be used for all included species.

As described by (Kapli et al., 2020), gene tree inference may be done through techniques including distance methods, such Neighbor Joining (NJ), or character-based methods, such as Maximum Likelihood (ML). Deciding which method is most practical for a particular situation is dependent on factors such as how closely related the sequences are.

Given a set of gene trees, various supertree methods also exist to infer the species trees. A popular tool for species tree inference is ASTRAL-Pro3 (Zhang and Mirarab, 2022), which optimizes for the number of quartets occurring in the gene trees in the resulting species tree. Poplar incorporates ASTRAL-Pro3 as a method for fusing the gene trees generated in the previous step.

3. Algorithm

Poplar provides a structure to connect established tools for each step of phylogenetic tree inference. Beginning from gene or genome data, it works through identifying relevant sequence, building possible gene groupings, inferring gene trees, and constructing a species tree. Figure 1 illustrates the pipeline.

Poplar does not have a next-generation sequencing analysis or genome assembly module, and assumes you are providing either genomes, scaffolds or genes. We will describe the other key steps identified in the background section which are implemented by Poplar.

3.1. Identification of gene sequences

The first step in Poplar is identifying gene sequences for each input species. For some species, these will be input as sequences, and for others, the sequences must be extracted from the genome scaffold sequences. If coding sequences are provided for a species, those sequences will be prioritized. Otherwise, if an annotated genome with identified genes is provided, then the sequences are extracted and used as potential genes. If not, then Open Reading Frames (ORFs) are extracted from the genome using orfipy (Singh and Wurtele, 2021). In a single run of Poplar, each type of data may be used for different genomes. Currently, the implementation requires nucleotide sequences.

3.2. Sequence Grouping into Genes Across Species, and Alignment of Gene Sequnces

Following the identification of genes and ORFs, Poplar groups the sequences in clusters.

First, Poplar uses Nucleotide BLAST to identify similar sequences across genomes. All the sequences are collected in a Nucleotide BLAST database Altschul et al. (1990). Poplar then queries the database for each sequence. The default similarity threshold is an expectation value (evalue) of 1^{-20}, but this threshold may be increased or decreased by the user if the input organisms are less or more closely related, respectively.

The results of the searches are used to create a distance matrix, which is then fed into an implementation of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for grouping (Ester et al., 1996; Schubert et al., 2017). DBSCAN was selected for this

step because it does not force a fixed number of groups, a fixed size of groups, or all sequences to be placed into a group. Instead, it identifies dense clusters as sequences that it declares groups. After using DBSCAN to create these groups, Poplar omits groups that are too small or too large, based on user input. By default, the minimum group size is 3, and the maximum group size is 100.

Using a grouping algorithm like DBSCAN to identify homologs across identified gene sequences is not the most popular way that gene clusters are identified. Often, orthologs will be identified and stored in a database such as OMA for future reference (Vallender, 2009). Using the BLAST search followed by DBSCAN only takes into account the similarity between two sequences, not any contextual information about their respective genomes. Yet, DBSCAN has been used for identifying gene clusters, including as an aspect of the tool GeneGrouper (McFarland et al., 2021). The results of Poplar also demonstrate the effectiveness of using a distance-based clustering for sequences.

Rather than rely on an external database or an intensive annotation process, Poplar considers a large set of sequences, and identifies those that are most similar. The BLAST database is stored locally, allowing Poplar to execute offline. The pairing of BLAST and DBSCAN allows Poplar to identify similarities between sequences, including sequences from genomes that have not been annotated.

The output of the sequence grouping step is the sets of sequences identified as similar.

3.3. Creation of Gene Trees

Once Poplar has a set of sequence groups, it selects a subset of the groups to build gene trees. By default, it will randomly select 50 groups, and a user may specify some other maximum number of trees to build.

Poplar infers gene trees from the sequence groups via MAFFT (Katoh and Standley, 2013) for sequence alignment and RAxML-NG (Kozlov et al., 2019) for gene tree construction. By default, Poplar also uses the default settings for both tools. The user may select the maximum number of gene trees to create, as the number of groups available to create trees is likely to be high, due to the nature of the sequence grouping process. MAFFT and RAxML-NG are both established tools in the field. RAxML-NG outputs the best tree for each set of sequences, which is then the input for creating the species tree.

3.4. Merging of Gene Trees into Species Trees

In order to infer a species tree from a set of gene trees, Poplar uses ASTRAL-Pro3 (Zhang and Mirarab, 2022). Like MAFFT and RAxML-NG, ASTRAL is an established and popular tool. The output of ASTRAL is a single tree including each of the input species.

4. Implementation

Poplar works as a framework or backbone to connect various well established tools. It runs as a Python script that manages inputs and refers to other tools, allowing the user to select options such as the number of gene trees, the maximum number of sequences to include in a gene group, and the evalue threshold for BLAST. In order to manage parallelism, it uses Parsl (Babuji et al., 2019), a Python workflow library. Parsl allows Poplar to fluctuate the hardware allocated to the tasks as needed.

Our tests were performed on a system using Slurm, but Parsl also supports running a workflow locally, or via other schedulers or cloud services. Changing the type of system used is done in a configuration file, and a Parsl "provider" is selected. The configuration file must be customized for the system being used in order for Parsl use the correct Slurm queue or other method of accessing the hardware.

Some steps required new code within Poplar. Most of the new code involves changing file formats and data labels, and are clearly outlined in our user manual. The one more substantial element is in identifying homolog groups using DBSCAN, which are then used for the "gene" trees. The algorithm cannot confirm that the groups represent a single gene across the species. Rather, it merely groups similar sequences together. However, as we will demonstrate in the results, and based on prior literature, this is an effective method for finding groups of sequences that can be annotated as individual genes for tree-building (McFarland et al., 2021).

Not using new, untested algorithms is a strength of the current iteration of Poplar –because it does not utilize new ways of generating gene trees or species trees, the end result when using Poplar is the same as when using RAxML or ASTRAL-Pro, thereby giving expected and trusted results in a more streamlined and efficient way. When new software is introduced or different tools and models are needed for particular uses, Poplar can also be updated to use the most accurate algorithms. The accuracy is that of its component tools, and has the same theoretical and applied accuracy as its components. A user experienced with the component tools may want to modify the calls to these tools to use alternative options from the default configuration of each piece of software.

The input to Poplar is FASTA files. As discussed earlier, these may be coding sequences, annotated genomes, or unannotated genomes. The intermediate data from each of these steps may be preserved for user reference or be deleted to minimize stored files. Poplar also supports check-pointing through Parsl. The clustering algorithm outputs lists of sequence identifier for all groups it creates. MAFFT outputs an alignment file in FASTA format, and RAxML outputs gene trees in Newick format. The output of Poplar is a Newick tree, including branch lengths.

5. Results

5.1. Testing Data

The testing data came from NCBI, JGI's MycoCosm, and JGI's Phycocosm. Specifically, fungal genomes were chosen due to the fact the fungal kingdom is both highly diverse and rapidly growing, and the use of phlogenetic trees to characterize new and emerging fungal species is key and widespread (Frac et al., 2018; Shumskaya et al., 2023; Naranjo-Ortiz and Gabaldón, 2019). The test datasets are:

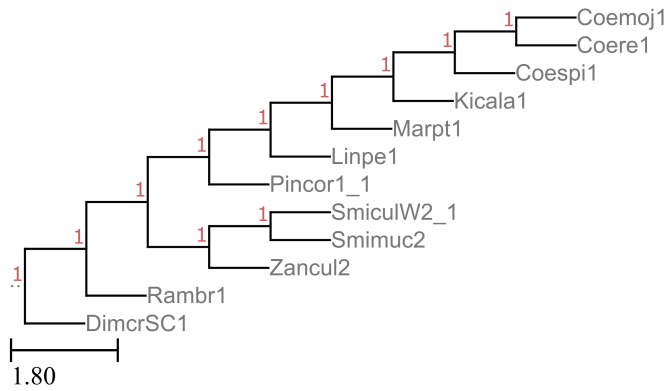

Fig. 2: Tree of Kickxellomycotina genomes, according to JGI's MycoCosm

Input Type	Number of Gene Trees	Number of Runs	Average RF Error Rate	Minimum RF Error Rate	Maximum RF Error Rate
Unannotated genome	50	5	0.20	0.11	0.33
Unannotated genome	100	5	0.11	0	0.22
Unannotated genome	500	5	0.07	0	0.22

Table 1. Poplar Performance results on Kickxellomycotina unannotated genomes

1. The 13 Kickxellomycotina genomes available through MycoCosm (Reynolds et al., 2023; Chang et al., 2015; Amses et al., 2022; Mondo et al., 2017; Wang et al., 2016; Ahrendt et al., 2018)
2. The 88 Mortierellomycotina genomes available through MycoCosm (Vandepol et al., 2020; Mondo et al., 2017; Mathieu et al., 2024; Chang et al., 2022; Mesny et al., 2021; Uehling et al., 2017)
3. The 26 Rhodophyla coding region sets available through PhycoCosm (Schönknecht et al., 2013; Rossoni et al., 2019; Cho et al., 2023; Nozaki et al., 2007; Lee et al., 2018; Nakamura-Gouvea et al., 2022; Collén et al., 2013; Brawley et al., 2017; Nakamura et al., 2013; Bhattacharya et al., 2013)
4. NCBI's 20 commonly used organisms for molecular research projects (Assembly, 1988a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t)
5. The 5 genomes of Puccinia graminis available through MycoCosm Duplessis et al. (2011); Li et al. (2019)

For the datasets from JGI, the reference trees came from the JGI webpage. For the NCBI dataset, the reference tree is from the species classifications. The reference trees do not include branch lengths.

A key criteria in selecting the datasets was the availability of a fully resolved reference tree from clear and well-grounded methodology. JGI publicly describes their Annotation Pipeline, and all included genomes are presented in the species tree, including where there are multiple genomes from a single species. NCBI presents only the classification of species, which means the tree has many polytomies. Polytomies are not an issue with the 20 common species as they are distantly related but are an issue with any closely related set of species or any species with multiple variants.

5.2. Metrics

We report the accuracy of trees by the Robinson-Foulds distance. The Robinson-Foulds metric works by identifying all bifurcations of the trees, and comparing the bifurcations in one tree to those of the other. If a split is present in only one of the two trees, the distance score increases by one. The reported distances are normalized to be out of 100%, where 0% is an identical tree and 100% shares no bifurcations.

Where the reference tree is not a binary tree, we do not report the additional splits in the generated trees as inaccuracies. MycoCosm presents binary trees, while NCBI provides the classifications of the species, which are not binary trees but also do not assert that there are not additional relationships that are not captured by the classification schema. Poplar reports binary trees.

5.3. Kickxellomycotina

The Kickxellomycotina classification within fungi has 13 genomes in JGI's MycoCosm, with the tree shown in Figure 2. Some species within Kickxellomycotina have multiple genomes represented. For this test, we used the genome assemblies rather than the annotated genomes, even when an annotation was available. When run five times with 50 random gene trees generated, the average Robinson-Foulds error rate is 0.20. The results of one run with a 11% Robinson-Foulds error rate is shown in Figure 3. As shown in Table 1, the average error with 100 gene trees is 11% and with 500 gene trees is 7%.

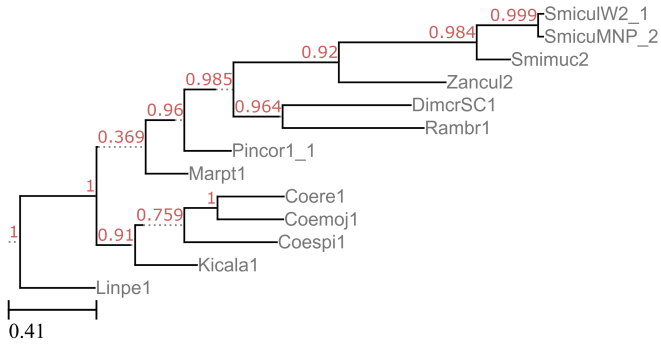

Fig. 3: Tree of Kickxellomycotina genomes, according to one run of Poplar with 50 gene trees

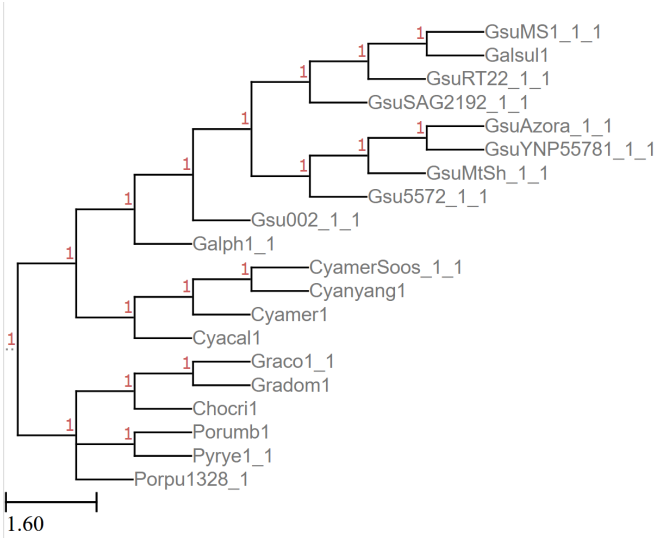

Fig. 4: JGI's Rhodophyla Tree

5.4. Mortierellomycotina

Mortierellomycotina is another fungal classification, with 88 genomes in JGI's MycoCosm. As with Kickxellomycotina, some species have multiple genomes in the database, and the tree was generated from genome assemblies, omitting annotations. In a run with 200 gene trees, Poplar's tree differed from the MycoCosm tree by 21 out of 86 splits, having a Robinson-Foulds distance of 24%. The Newick versions of the trees are included in the supplementary materials.

5.5. Rhodophyla

Rhodophyla is a algal classification, with 26 genomes available through JGI's PhycoCosm, 20 of which are included in their tree. In a single run with 200 gene trees, Poplar had a distance of 15% from PhycoCosm's tree. The two trees are shown in Figure 4 and Figure 5.

5.6. NCBI Commonly Studied Organisms

These 20 organisms from NCBI's database were included in order to demonstrate the usage of Poplar on genomes outside of fungi and algae, as well as on a tree with a vast spread of species.

Two runs of Poplar, both with 200 gene trees, resulted in slightly different trees. Because of the random selection of the groups, Poplar is not deterministic. Given enough gene trees, the accuracy impact of the randomness will typically be small, while the runtime impact may be significant.

The output of Poplar did not include *Mycoplasmoides pneumoniae* or *Escherichia coli*. This exclusion is likely due to their distance from the other species. The other difference was in the placement of *Plasmodium falciparum*. In the first run, it was placed among the plants, leading to a Robinson-Foulds distance of 20%. In the second run, it was placed accurately next to *Dictyostelium discoideum*, and therefore the Robinson-Foulds distance was 0%. The misplacement of *P. falciparum* was most likely due to the large spread of the included species. If a species does not share as much genetic material with the other species in the tree, it will be included in fewer groups and gene trees, and a misplacement in one tree that might otherwise be noise can be perpetuated into the species tree.

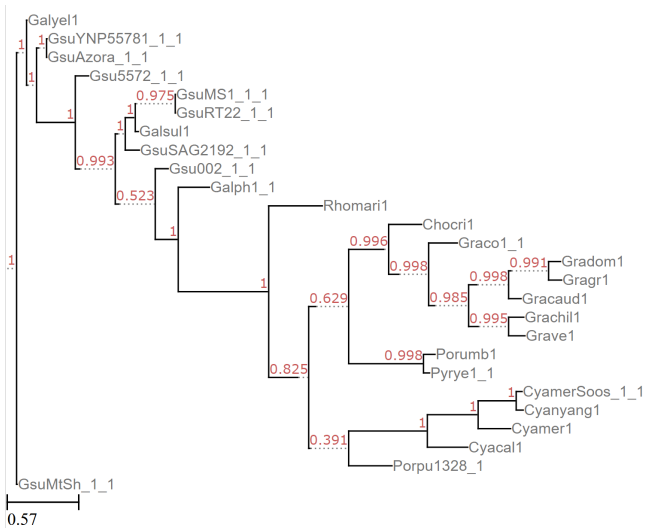

Fig. 5: Rhodophyla tree generated by Poplar with 200 gene trees

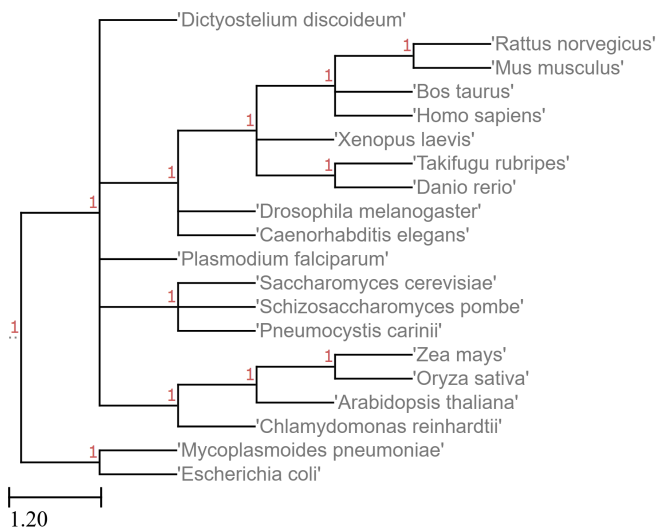

Fig. 6: True tree of species from NCBI

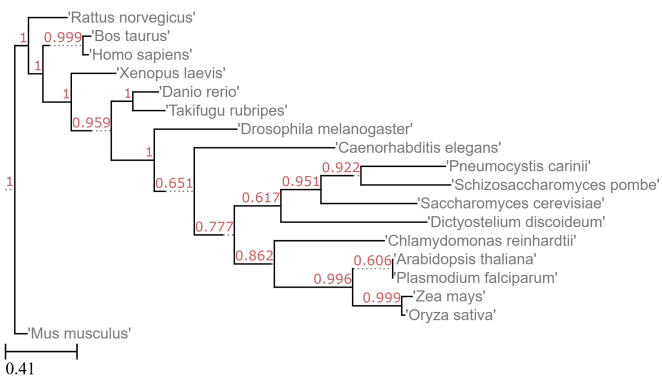

Fig. 7: Tree generated by Poplar from the NCBI genomes

Input Type	Number of Gene Trees	Number of Runs	Average RF Error Rate	Minimum RF Error Rate	Maximum RF Error Rate
CDS	50	5	0.50	0.50	0.50
CDS	100	5	0.50	0.50	0.50
CDS	500	5	0.50	0.50	0.50

Table 2. Poplar Performance results on Puccinia graminis coding sequences

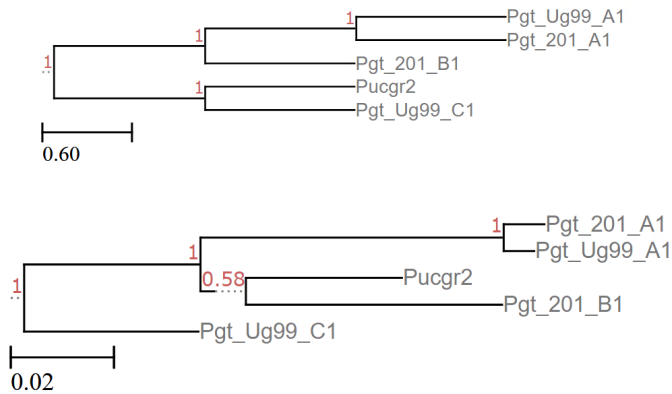

Fig. 8: Puccinia graminis trees from MycoCosm (top) and Poplar with 100 or 500 gene trees (bottom)

Task	Number of Invocations	Average Time	Total Time (hh:mm:ss)
Find ORFs	14	00:00:11	0:02:39
Build BLAST Database	1		0:00:09
Search BLAST	14	0:02:33	0:35:52
Make groups	1		0:05:30
Align Sequences and Construct Gene Trees	100	0:06:36	11:00:44
Other	6	0:01:51	0:11:08
ASTRAL-Pro	1		0:00:00.37
Total			11:56:04

Table 3. Time per task in building the Kickxellomycotina tree with 100 gene trees

5.7. Puccinia graminis

The five genomes of the species Puccinia graminis are included to contrast the diversity of the NCBI Commonly Studied Organisms, and demonstrate the effectiveness of Poplar on genomes that are closely related. Error rates are shown in Table 2. Runs with 50 gene trees varied between each other, though the distance from the MycoCosm tree was consistent for all trees. Using 100 or 500 gene trees, all five trees were the same as each other, though they varied from MycoCosm's tree. As shown in Figure 8, the two trees vary in therelationship between Pucgr2 and Pgt_201_B1. The node in MycoCosm's tree connecting (Pgt_Ug99_A1, Pgt_201_A1) to Pgt_201_B1 has much lower support than the other nodes in the tree. While all other nodes have the maximum support of 1.0, this split has support of only 0.2850. It is plausible that the difference between the MycoCosm and Poplar trees is supported by the available data.

5.8. Performance

We assessed performance on the 14 Kickxellomycotina genomes, using the full genomes. The tests were run on CPU nodes on a High Performance Data Analytics Research Cluster with Dual Socket Intel E5-2683v3 2.00GHz CPUs.

In these runs, each worker uses 2 cores, so each node is able to run 28 tasks at a time. Two nodes were used per block, and Parsl was permitted a maximum of 4 blocks.

Table 3 shows the performance of Poplar when running 100 trees. The majority of the time is spent in building the gene trees. While the total time is nearly 12 hours, the runtime from start to end was actually only 23 minutes. The division of each node into 28 workers also served to reduce the node usage.

Task	Number of Invocations	Average Time	Total Time (hh:mm:ss)
Find ORFs	14	0:00:04	0:00:57
Build BLAST Database	1		0:00:11
Search BLAST	14	0:05:17	0:35:16
Make groups	1		0:05:17
Align Sequences and Construct Gene Trees	500	0:08:49	3 days, 1:35:25
Other	6	0:01:45	0:10:32
ASTRAL-Pro	1		0:00:01
Total			3 days, 2:27:43

Table 4. Time per task in building the Kickxellomycotina tree with 500 gene trees

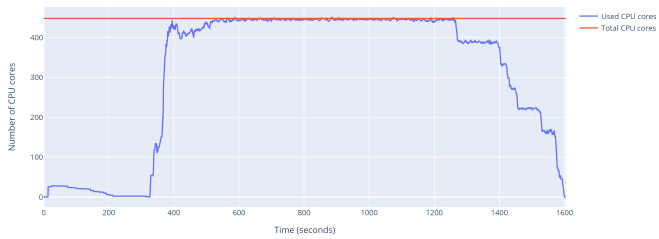

Fig. 9: Graph of CPU usage during execution of Poplar on 14 Kickxellomycotina genomes using 500 gene trees

In the same way, Table 4 shows the performance of Poplar when running 500 trees for Kickxellomycotina. Again, generating the gene trees was where the majority of time was spent.

Figure 9 shows the CPU usage throughout the run with 500 trees. During significant portion of the run, the parallelism is able to use all available CPUs.

6. Discussion

Poplar offers a readily usable alternative to building custom phylogenetic tree pipelines, which often require subject-matter expertise, and the ability to string together multiple pieces of software that do not necessarily have compatible inputs and outputs. In addition, Poplar can utilize genomes with or without annotation, and advanced users can vary the number of gene trees that get constructed, thereby increasing the resolution of the tree. This higher-resolution tree will require more resources, giving users the option to choose between high resolution and speed/resource utilization.

The implementation of Poplar allows flexibility across hardware. While testing was only performed on clusters using Slurm, Parsl supports cloud computing and other cluster management systems. Shifting between systems across workflows or within a single workflow can be done in a configuration file, with Parsl offering examples and support for a wide variety of architectures. While Poplar is limited by the operating systems and hardware where the other software is depends on is able to run, some steps are portable to non-Unix machines and could be ported more widely with minimal user intervention.

Poplar's results match closely with reference trees from accepted sources. Each step in Poplar is not new. Instead, the algorithms and software it incorporates are widely used and based on substantial theoretical and practical study. The novel aspect of Poplar is in the ways it allows a user to more easily run the entire pipeline with minimal interference while still obtaining results from state-of-the-art algorithms. Gathering the phylogenetic inference pipeline into one tool allows for coherency of use and communication.

6.1. Limitations and Future Work

Poplar is designed in a modular manner, so a user may replace a step with an alternative tool. Using an alternative tool will also require providing scripts to convert input or output files if they differ from the default. Future work will include options for different methods for each step, allowing the change to be done by simply changing an argument. Adding additional options for each step will allow users who want to use particular tools or models to simply select the suitable tools via the command line or a configuration file.

Most runs of Poplar are dominated by searching the custom BLAST database for similar sequences. The time spent in this step is especially pronounced when the included genomes do not have provided coding regions or annotations. When sequence selection is not narrowed to only the coding regions or annotated genes, all ORFs in the genome are included, ballooning both the search space and the number of queries. Future work will look into both narrowing the number of queries used as well as more efficiently searching the database, given the goal of obtaining clusters.

Another area of future development of Poplar improving results with distantly related species. When the sequence groups are selected, the currently algorithm selects a random set of groups. When some species are distantly related from the main group, they may not have

sequences occur in any of the groups. In that case, they will be omitted from the species tree. In the future, the selection of sequence groups will include a optimization for including all species.

7. Availability and requirements

Poplar is available at `https://github.com/sandialabs/poplar` under the GPLv3 license.

8. Acknowledgments

We would like to thank Ellis Torrance and Sharon Nademanee for their review of the manuscript. This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mision laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

References

S. R. Ahrendt, C. A. Quandt, D. Ciobanu, A. Clum, A. Salamov, B. Andreopoulos, J.-F. Cheng, T. Woyke, A. Pelin, B. Henrissat, N. K. Reynolds, G. L. Benny, M. E. Smith, T. Y. James, and I. V. Grigoriev. Leveraging single-cell genomics to expand the fungal tree of life. *Nat. Microbiol.*, 3(12):1417–1428, Dec. 2018.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, Oct. 1990.

K. R. Amses, D. R. Simmons, J. E. Longcore, S. J. Mondo, K. Seto, G. H. Jerônimo, A. E. Bonds, C. A. Quandt, W. J. Davis, Y. Chang, B. A. Federici, A. Kuo, K. LaButti, J. Pangilinan, W. Andreopoulos, A. Tritt, R. Riley, H. Hundley, J. Johnson, A. Lipzen, K. Barry, B. F. Lang, C. A. Cuomo, N. E. Buchler, I. V. Grigoriev, J. W. Spatafora, J. E. Stajich, and T. Y. James. Diploid-dominant life cycles characterize the early evolution of fungi. *Proc. Natl. Acad. Sci. U. S. A.*, 119(36):e2116841119, Sept. 2022.

J. Armstrong, G. Hickey, M. Diekhans, I. T. Fiddes, A. M. Novak, A. Deran, Q. Fang, D. Xie, S. Feng, J. Stiller, D. Genereux, J. Johnson, V. D. Marinescu, J. Alföldi, R. S. Harris, K. Lindblad-Toh, D. Haussler, E. Karlsson, E. D. Jarvis, G. Zhang, and B. Paten. Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, 587(7833):246–251, Nov 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2871-y. URL `https://doi.org/10.1038/s41586-020-2871-y`.

Assembly. Accession No. GCF_000001735.4, Arabidopsis thalianaos taurus. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988a. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001735.4`.

Assembly. Accession No. GCF_900660465.1, Mycoplasmoides pneumoniae. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988b. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_900660465.1`.

Assembly. Accession No. GCF_034140825.1, Oryza sativa. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988c. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_034140825.1`.

Assembly. Accession No. GCF_000002765.6, Plasmodium falciparum. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988d. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002765.6`.

Assembly. Accession No. GCF_001477545.1, Pneumocystis carinii. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988e. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_001477545.1`.

Assembly. Accession No. GCF_036323735.1, Rattus norvegicus. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988f. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_036323735.1`.

Assembly. Accession No. GCF_000146045.2, Saccharomyces cerevisiae. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988g. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000146045.2`.

Assembly. Accession No. GCF_000002945.2, Schizosaccharomyces pombe. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988h. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002945.2`.

Assembly. Accession No. GCF_901000725.2, Takifugu rubripes. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988i. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_901000725.2`.

Assembly. Accession No. GCF_017654675.1, Xenopus laevis. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988j. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_017654675.1`.

Assembly. Accession No. GCF_902167145.1, Zea mays. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988k. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_902167145.1`.

Assembly. Accession No. GCF_002263795.3, Bos taurus. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988l. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_002263795.3`.

Assembly. Accession No. GCF_000002035.6, Danio rerio. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988m. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002035.6`.

Assembly. Accession No. GCF_000002985.6, Caenorhabditis elegans. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988n. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002985.6`.

Assembly. Accession No. GCF_000002595.2, Chlamydomonas reinhardtii. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988o. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002595.2`.

Assembly. Accession No. GCF_000004695.1, Dictyostelium discoideum. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988p. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000004695.1`.

Assembly. Accession No. GCF_000001215.4, Drosophila melanogaster. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988q. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001215.4`.

Assembly. Accession No. GCF_000005845.2, Escherichia coli. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988r. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000005845.2`.

Assembly. Accession No. GCF_000001405.40, Homo sapiens. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988s. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40`.

Assembly. Accession No. GCF_000001635.27, Mus musculus. *National Library of Medicine (US), National Center for Biotechnology Information*, 1988t. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001635.27`.

D. Azouri, S. Abadi, Y. Mansour, I. Mayrose, and T. Pupko. Harnessing machine learning to guide phylogenetic-tree search algorithms. *Nat. Comm.*, Mar. 2021.

Y. Babuji, A. Woodard, Z. Li, D. S. Katz, B. Clifford, R. Kumar, L. Lacinski, R. Chard, J. Wozniak, I. Foster, M. Wilde, and K. Chard. Parsl: Pervasive parallel programming in python. In *28th ACM International Symposium on High-Performance Parallel and Distributed Computing (HPDC)*, 2019. doi: 10.1145/3307681.3325400. URL `https://doi.org/10.1145/3307681.3325400`.

D. Bhattacharya, D. C. Price, C. X. Chan, H. Qiu, N. Rose, S. Ball, A. P. M. Weber, M. C. Arias, B. Henrissat, P. M. Coutinho, A. Krishnan, S. Zäuner, S. Morath, F. Hilliou, A. Egizi, M.-M. Perrineau, and H. S. Yoon. Genome of the red alga porphyridium purpureum. *Nat. Commun.*, 4(1):1941, 2013.

S. H. Brawley, N. A. Blouin, E. Ficko-Blean, G. L. Wheeler, M. Lohr, H. V. Goodson, J. W. Jenkins, C. E. Blaby-Haas, K. E. Helliwell, C. X. Chan, T. N. Marriage, D. Bhattacharya, A. S. Klein, Y. Badis, J. Brodie, Y. Cao, J. Collén, S. M. Dittami, C. M. M. Gachon, B. R. Green, S. J. Karpowicz, J. W. Kim, U. J. Kudahl, S. Lin, G. Michel, M. Mittag, B. J. S. C. Olson, J. L. Pangilinan, Y. Peng, H. Qiu, S. Shu, J. T. Singer, A. G. Smith, B. N. Sprecher, V. Wagner, W. Wang, Z.-Y. Wang, J. Yan, C. Yarish, S. Zäuner-Riek, Y. Zhuang, Y. Zou, E. A. Lindquist, J. Grimwood, K. W. Barry, D. S. Rokhsar, J. Schmutz, J. W. Stiller, A. R. Grossman, and S. E. Prochnik. Insights into the red algae and eukaryotic evolution from the genome of *Porphyra umbilicalis* (bangiophyceae, rhodophyta). *Proc. Natl. Acad. Sci. U. S. A.*, 114(31):E6361–E6370, Aug. 2017.

T. E. o. E. Britannica. poplar, 2023. URL `https://www.britannica.com/plant/poplar`.

Y. Chang, S. Wang, S. Sekimoto, A. L. Aerts, C. Choi, A. Clum, K. M. LaButti, E. A. Lindquist, C. Yee Ngan, R. A. Ohm, A. A. Salamov, I. V. Grigoriev, J. W. Spatafora, and M. L. Berbee. Phylogenomic analyses indicate that early fungi evolved digesting cell walls of algal ancestors of land plants. *Genome Biol. Evol.*, 7(6):1590–1601, May 2015.

Y. Chang, Y. Wang, S. Mondo, S. Ahrendt, W. Andreopoulos, K. Barry, J. Beard, G. L. Benny, S. Blankenship, G. Bonito, C. Cuomo, A. Desiro, K. A. Gervers, H. Hundley, A. Kuo, K. LaButti, B. F. Lang, A. Lipzen, K. O'Donnell, J. Pangilinan, N. Reynolds, L. Sandor, M. E. Smith, A. Tsang, I. V. Grigoriev, J. E. Stajich, and J. W. Spatafora. Evolution of zygomycete secretomes and the origins of terrestrial fungal ecologies. *iScience*, 25(8):104840, Aug. 2022.

C. H. Cho, S. I. Park, T.-Y. Huang, Y. Lee, C. Ciniglia, H. C. Yadavalli, S. W. Yang, D. Bhattacharya, and H. S. Yoon. Genome-wide signatures of adaptation to extreme environments in red algae. *Nat. Commun.*, 14(1):10, Jan. 2023.

J. Collén, B. Porcel, W. Carré, S. G. Ball, C. Chaparro, T. Tonon, T. Barbeyron, G. Michel, B. Noel, K. Valentin, M. Elias, F. Artiguenave, A. Arun, J.-M. Aury, J. F. Barbosa-Neto, J. H. Bothwell, F.-Y. Bouget, L. Brillet, F. Cabello-Hurtado, S. Capella-Gutiérrez, B. Charrier, L. Cladière, J. M. Cock, S. M. Coelho, C. Colleoni, M. Czjzek, C. Da Silva, L. Delage, F. Denoeud, P. Deschamps, S. M. Dittami, T. Gabaldón, C. M. M. Gachon, A. Groisillier, C. Hervé, K. Jabbari, M. Katinka, B. Kloareg, N. Kowalczyk, K. Labadie, C. Leblanc, P. J. Lopez, D. H. McLachlan, L. Meslet-Cladiere, A. Moustafa, Z. Nehr, P. Nyvall Collén, O. Panaud, F. Partensky, J. Poulain, S. A. Rensing, S. Rousvoal, G. Samson, A. Symeonidi, J. Weissenbach, A. Zambounis, P. Wincker, and C. Boyen. Genome structure and metabolic features in the red seaweed chondrus crispus shed light on evolution of the archaeplastida. *Proc. Natl. Acad. Sci. U. S. A.*, 110(13):5247–5252, Mar. 2013.

N. De Maio, P. Kalaghatgi, Y. Turakhia, R. Corbett-Detig, B. Q. Minh, and N. Goldman. Maximum likelihood pandemic-scale phylogenetics. *Nat. Genetics*, Apr. 2023.

E. Dort, E. Layne, N. Feau, A. Buyaev, B. Henrissat, F. Martin, S. Haridas, A. Salamov, I. Grigoriev, M. Blancehtte, and R. Hamelin. Large-scale genomic analyses with machine learning uncover predictive patterns associated with fungal phytopathogenic lifestyles and traits. *Sci. Rep.*, Oct. 2023.

S. Duplessis, C. A. Cuomo, Y.-C. Lin, A. Aerts, E. Tisserant, C. Veneault-Fourrey, D. L. Joly, S. Hacquard, J. Amselem, B. L. Cantarel, R. Chiu, P. M. Coutinho, N. Feau, M. Field, P. Frey, E. Gelhaye, J. Goldberg, M. G. Grabherr, C. D. Kodira, A. Kohler, U. Kües, E. A. Lindquist, S. M. Lucas, R. Mago, E. Mauceli, E. Morin, C. Murat, J. L. Pangilinan, R. Park, M. Pearson, H. Quesneville, N. Rouhier, S. Sakthikumar, A. A. Salamov, J. Schmutz, B. Selles, H. Shapiro, P. Tanguay, G. A. Tuskan, B. Henrissat, Y. Van de Peer, P. Rouzé, J. G. Ellis, P. N. Dodds, J. E. Schein, S. Zhong, R. C. Hamelin, I. V. Grigoriev, L. J. Szabo, and F. Martin. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U. S. A.*, 108(22):9166–9171, May 2011.

D. Dylus, A. Altenhoff, S. Majidian, F. J. Sedlazeck, and C. Dessimoz. Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree. *Nat. Biotechnol.*, Apr. 2023.

R. C. Edgar. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.*, 13(1):6968, Nov. 2022.

D. R. Edla and P. K. Jana. A prototype-based modified dbscan for gene clustering. *Procedia Technology*, 6:485–492, 2012. ISSN 2212-0173. doi: https://doi.org/10.1016/j.protcy.2012.10.058. URL `https://www.sciencedirect.com/science/article/pii/S2212017312006032`. 2nd International Conference on Communication, Computing & Security [ICCCS-2012].

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press,

1996.

M. Frac, S. E. Hannula, M. Bełka, and M. Jedryczka. Fungal biodiversity and their role in soil health. *Frontiers in Microbiology*, 9, 2018. ISSN 1664-302X. doi: 10.3389/fmicb.2018.00707. URL https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2018.00707.

L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, and J. F. Banfield. A new view of the tree of life. *Nat. Microbiol.*, 1(5), Apr. 2016a.

L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, and J. F. Banfield. A new view of the tree of life. *Nat. Microbioloy*, Apr. 2016b. doi: 10.1038/nmicrobiol.2016.48.

E. D. Jarvis, S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. W. Ho, B. C. Faircloth, B. Nabholz, J. T. Howard, A. Suh, C. C. Weber, R. R. da Fonseca, A. Alfaro-Núñez, N. Narula, L. Liu, D. Burt, H. Ellegren, S. V. Edwards, A. Stamatakis, D. P. Mindell, J. Cracraft, E. L. Braun, T. Warnow, W. Jun, M. T. P. Gilbert, G. Zhang, and T. A. P. Consortium. Phylogenomic analyses data of the avian phylogenomics project. *GigaScience*, 4(1):s13742–014–0038–1, 02 2015. ISSN 2047-217X. doi: 10.1186/s13742-014-0038-1. URL https://doi.org/10.1186/s13742-014-0038-1.

P. Kapli, Z. Yang, and M. J. Telford. Phylogenetic tree building in the genomic age. *Nat. Rev. Genetics*, May 2020.

K. Katoh and D. M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780, 01 2013. ISSN 0737-4038. doi: 10.1093/molbev/mst010. URL https://doi.org/10.1093/molbev/mst010.

C. Ki and J. Terhorst. Variational phylodynamic inference using pandemic-scale data. *Mol. Biol. and Evol.*, July 2022.

A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, and A. Stamatakis. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455, 05 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz305. URL https://doi.org/10.1093/bioinformatics/btz305.

J. Lee, E. C. Yang, L. Graf, J. H. Yang, H. Qiu, U. Zelzion, C. X. Chan, T. G. Stephens, A. P. M. Weber, G. H. Boo, S. M. Boo, K. M. Kim, Y. Shin, M. Jung, S. J. Lee, H.-S. Yim, J.-H. Lee, D. Bhattacharya, and H. S. Yoon. Analysis of the draft genome of the red seaweed gracilariopsis chorda provides insights into genome size evolution in rhodophyta. *Mol. Biol. Evol.*, 35(8):1869–1886, Aug. 2018.

F. Li, N. M. Upadhyaya, J. Sperschneider, O. Matny, H. Nguyen-Phuc, R. Mago, C. Raley, M. E. Miller, K. A. T. Silverstein, E. Henningsen, C. D. Hirsch, B. Visser, Z. A. Pretorius, B. J. Steffenson, B. Schwessinger, P. N. Dodds, and M. Figueroa. Emergence of the ug99 lineage of the wheat stem rust pathogen through somatic hybridisation. *Nat. Commun.*, 10(1):5068, Nov. 2019.

S. C. Manekar and S. R. Sathe. A benchmark study of k-mer counting methods for high-throughput sequencing. *Gigascience*, 7(12), Dec. 2018.

D. Mathieu, A. E. Bryson, B. Hamberger, V. Singan, K. Keymanesh, M. Wang, K. Barry, S. Mondo, J. Pangilinan, M. Koriabine, I. V. Grigoriev, G. Bonito, and B. Hamberger. Multilevel analysis between physcomitrium patens and mortierellaceae endophytes explores potential long-standing interaction among land plants and fungi. *Plant J.*, 118(2):304–323, Apr. 2024.

A. G. McFarland, N. W. Kennedy, C. E. Mills, D. Tullman-Ercek, C. Huttenhower, and E. M. Hartmann. Density-based binning of gene clusters to infer function or evolutionary history using GeneGrouper. *Bioinformatics*, 38(3):612–620, 11 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab752. URL https://doi.org/10.1093/bioinformatics/btab752.

F. Mesny, S. Miyauchi, T. Thiergart, B. Pickel, L. Atanasova, M. Karlsson, B. Hüttel, K. W. Barry, S. Haridas, C. Chen, D. Bauer, W. Andreopoulos, J. Pangilinan, K. LaButti, R. Riley, A. Lipzen, A. Clum, E. Drula, B. Henrissat, A. Kohler, I. V. Grigoriev, F. M. Martin, and S. Hacquard. Genetic determinants of endophytism in the arabidopsis root mycobiome. *Nat. Commun.*, 12(1):7227, Dec. 2021.

C. Moeckel, M. Mareboina, M. A. Konnaris, C. S. Y. Chan, I. Mouratidis, A. Montgomery, N. Chantzi, G. A. Pavlopoulos, and I. Georgakopoulos-Soares. A survey of k-mer methods and applications in bioinformatics. *Comput. Struct. Biotechnol. J.*, 23: 2289–2303, Dec. 2024.

S. J. Mondo, R. O. Dannebaum, R. C. Kuo, K. B. Louie, A. J. Bewick, K. LaButti, S. Haridas, A. Kuo, A. Salamov, S. R. Ahrendt, R. Lau, B. P. Bowen, A. Lipzen, W. Sullivan, B. B. Andreopoulos, A. Clum, E. Lindquist, C. Daum, T. R. Northen, G. Kunde-Ramamoorthy, R. J. Schmitz, A. Gryganskyi, D. Culley, J. Magnuson, T. Y. James, M. A. O'Malley, J. E. Stajich, J. W. Spatafora, A. Visel, and I. V. Grigoriev. Widespread adenine n6-methylation of active genes in fungi. *Nat. Genet.*, 49(6):964–968, June 2017.

Y. Nakamura, N. Sasaki, M. Kobayashi, N. Ojima, M. Yasuike, Y. Shigenobu, M. Satomi, Y. Fukuma, K. Shiwaku, A. Tsujimoto, T. Kobayashi, I. Nakayama, F. Ito, K. Nakajima, M. Sano, T. Wada, S. Kuhara, K. Inouye, T. Gojobori, and K. Ikeo. The first symbiont-free genome sequence of marine red alga, susabi-nori (pyropia yezoensis). *PLoS One*, 8(3):e57122, Mar. 2013.

N. Nakamura-Gouvea, C. Alves-Lima, L. F. Benites, C. Iha, V. Maracaja-Coutinho, V. Aliaga-Tobar, M. Araujo Amaral Carneiro, N. S. Yokoya, E. Marinho-Soriano, M. A. S. Graminha, J. Collén, M. C. Oliveira, J. C. Setubal, and P. Colepicolo. Insights into agar and secondary metabolite pathways from the genome of the red alga gracilaria domingensis (rhodophyta, gracilariales). *J. Phycol.*, 58(3): 406–423, June 2022.

Y. Nakano, Y. Domon, and K. Yamagashi. Phylogenetic trees of closely related bacterial species and subspecies based on frequencies of short nucleotide sequences. *PLoS one*, Apr. 2023.

M. A. Naranjo-Ortiz and T. Gabaldón. Fungal evolution: diversity, taxonomy and phylogeny of the fungi. *Biological Reviews*, 94(6): 2101–2137, 2019. doi: https://doi.org/10.1111/brv.12550. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.12550.

D. J. Nasko, S. Koren, A. M. Phillippy, and T. J. Treangen. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol.*, 19(1):165, Oct. 2018.

D. e. a. Navarro. Large-scale phenotyping of 1,000 fungal strains for the degradation of non-natural, industrial compounds. *Comm. Biol.*, 7 2021.

H. Nozaki, H. Takano, O. Misumi, K. Terasawa, M. Matsuzaki, S. Maruyama, K. Nishida, F. Yagisawa, Y. Yoshida, T. Fujiwara, S. Takio, K. Tamura, S. J. Chung, S. Nakamura, H. Kuroiwa, K. Tanaka, N. Sato, and T. Kuroiwa. A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga cyanidioschyzon merolae. *BMC Biol.*, 5(1):28, July 2007.

N. K. Reynolds, J. E. Stajich, G. L. Benny, K. Barry, S. Mondo, K. LaButti, A. Lipzen, C. Daum, I. V. Grigoriev, H.-M. Ho, P. W. Crous, J. W. Spatafora, and M. E. Smith. Mycoparasites, gut dwellers, and saprotrophs: Phylogenomic reconstructions and comparative analyses of kickxellomycotina fungi. *Genome Biol. Evol.*, 15(1), Jan. 2023.

A. W. Rossoni, D. C. Price, M. Seger, D. Lyska, P. Lammers, D. Bhattacharya, and A. P. Weber. The genomes of polyextremophilic cyanidiales contain 1% horizontally transferred genes with diverse adaptive functions. *Elife*, 8, May 2019.

G. Schönknecht, W.-H. Chen, C. M. Ternes, G. G. Barbier, R. P. Shrestha, M. Stanke, A. Bräutigam, B. J. Baker, J. F. Banfield, R. M. Garavito, K. Carr, C. Wilkerson, S. A. Rensing, D. Gagneul, N. E. Dickenson, C. Oesterhelt, M. J. Lercher, and A. P. M. Weber. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science*, 339(6124):1207–1210, Mar. 2013.

E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu. Dbscan revisited, revisited: Why and how you should (still) use dbscan. *ACM Trans. Database Syst.*, 42(3), jul 2017. ISSN 0362-5915. doi: 10.1145/3068335. URL `https://doi.org/10.1145/3068335`.

M. Shumskaya, N. Filippova, L. Lorentzen, S. Blue, C. Andrew, and N. S. Lorusso. Citizen science helps in the study of fungal diversity in new jersey. *Scientific Data*, 10, 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01916-z. URL `https://www.nature.com/articles/s41597-022-01916-z#citeas`.

U. Singh and E. S. Wurtele. orfipy: a fast and flexible tool for extracting ORFs. *Bioinformatics*, 37(18):3019–3020, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab090. URL `https://doi.org/10.1093/bioinformatics/btab090`.

J. W. Spatafora, M. C. Aime, I. V. Grigoriev, F. Martin, J. E. Stajich, and M. Blackwell. The fungal tree of life: from molecular systematics to genome-scale phylogenies. *Microbiol. Spectrum*, 9 2017.

K. Tamura, G. Stecher, and S. Kumar. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution*, 38(7):3022–3027, 04 2021. ISSN 1537-1719. doi: 10.1093/molbev/msab120. URL `https://doi.org/10.1093/molbev/msab120`.

Y. Turakhia, B. Thornlow, A. Hinrichs, J. McBroome, N. Ayala, C. Ye, K. Smith, N. De Maio, D. Haussler, R. Lanfear, and R. Corbett-Detig. Pandemic-scale phylogenomics reveals the sars-cov-2 recombination landscape. *Nature*, Aug. 2022.

J. Uehling, A. Gryganskyi, K. Hameed, T. Tschaplinski, P. K. Misztal, S. Wu, A. Desirò, N. Vande Pol, Z. Du, A. Zienkiewicz, K. Zienkiewicz, E. Morin, E. Tisserant, R. Splivallo, M. Hainaut, B. Henrissat, R. Ohm, A. Kuo, J. Yan, A. Lipzen, M. Nolan, K. LaButti, K. Barry, A. H. Goldstein, J. Labbé, C. Schadt, G. Tuskan, I. Grigoriev, F. Martin, R. Vilgalys, and G. Bonito. Comparative genomics of *Mortierella elongata* and its bacterial endosymbiont *Mycoavidus cysteinexigens*. *Environ. Microbiol.*, 19(8): 2964–2983, Aug. 2017.

E. J. Vallender. Bioinformatic approaches to identifying orthologs and assessing evolutionary relationships. *Methods*, 49(1):50–55, May 2009. doi: 10.1016/j.ymeth.2009.05.010.

N. Vandepol, J. Liber, A. Desirò, H. Na, M. Kennedy, K. Barry, I. V. Grigoriev, A. N. Miller, K. O'Donnell, J. E. Stajich, and G. Bonito. Resolving the mortierellaceae phylogeny through synthesis of multi-gene phylogenetics and phylogenomics. *Fungal Divers.*, 104(1): 267–289, Sept. 2020.

Y. Wang, M. M. White, S. Kvist, and J.-M. Moncalvo. Genome-wide survey of gut fungi (harpellales) reveals the first horizontally transferred ubiquitin gene from a mosquito host. *Mol. Biol. Evol.*, 33(10):2544–2554, Oct. 2016.

N. J. Wickett, S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam, M. S. Barker, J. G. Burleigh, M. A. Gitzendanner, B. R. Ruhfel, E. Wafula, J. P. Der, S. W. Graham, S. Mathews, M. Melkonian, D. E. Soltis, P. S. Soltis, N. W. Miles, C. J. Rothfels, L. Pokorny, A. J. Shaw, L. DeGironimo, D. W. Stevenson, B. Surek, J. C. Villarreal, B. Roure, H. Philippe, C. W. dePamphilis, T. Chen, M. K. Deyholos, R. S. Baucom, T. M. Kutchan, M. M. Augustin, J. Wang, Y. Zhang, Z. Tian, Z. Yan, X. Wu, X. Sun, G. K.-S. Wong, and J. Leebens-Mack. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111(45):E4859–E4868, 2014. doi: 10.1073/pnas.1323926111. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1323926111`.

A. D. Young and J. P. Gillung. Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology*, 45(2):225–247, 2020. doi: https://doi.org/10.1111/syen.12406. URL `https://resjournals.onlinelibrary.wiley.com/doi/abs/10.1111/syen.12406`.

C. Zhang and S. Mirarab. ASTRAL-Pro 2: ultrafast species tree reconstruction from multi-copy gene family trees. *Bioinformatics*, 38(21): 4949–4950, 09 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac620. URL `https://doi.org/10.1093/bioinformatics/btac620`.

D.-F. Zhang, W. He, Z. Shao, I. Ahmed, Y. Zhang, W.-J. Li, and Z. Zhao. EasyCGTree: a pipeline for prokaryotic phylogenomic analysis based on core gene sets. *BMC Bioinformatics*, 24(1):390, Oct. 2023.