

UnFATE: A Comprehensive Probe Set and Bioinformatics Pipeline for Phylogeny Reconstruction and Multilocus Barcoding of Filamentous Ascomycetes (Ascomycota, Pezizomycotina)

CLAUDIO G. AMETRANO^{1,2,*}, JACOB JENSEN¹, H. THORSTEN LUMBSCH¹, AND FELIX GREWE¹

¹Grainger Bioinformatics Center and Negaunee Integrative Research Center, Science and Education, Field Museum of Natural History, 1400 S DuSable Lake Shore Drive, Chicago, IL 60605, USA

²Department of Life Sciences, University of Trieste, via Giorgieri 10, 34127 Trieste, Italy

*Correspondence to be sent to: Department of Life Sciences, University of Trieste, via Giorgieri 10, 34127 Trieste, Italy; E-mail: claudioametrano@units.it.

Associate Editor: Ryan Folk

Received 10 May 2022; reviews returned 4 February 2025; accepted 11 February 2025

Abstract.—The subphylum Pezizomycotina (filamentous ascomycetes) is the largest clade within Ascomycota. Despite the importance of this group of fungi, our understanding of their evolution is still limited due to insufficient taxon sampling. Although next-generation sequencing technology allows us to obtain complete genomes for phylogenetic analyses, generating complete genomes of fungal species can be challenging, especially when fungi occur in symbiotic relationships or when the DNA of rare herbarium specimens is degraded or contaminated. Additionally, assembly, annotation, and gene extraction of whole-genome sequencing (WGS) data require bioinformatics skills and computational power, resulting in a substantial data burden. To overcome these obstacles, we designed a universal target enrichment probe set to reconstruct the phylogenetic relationships of filamentous ascomycetes at different phylogenetic levels. From a pool of single-copy orthologous genes extracted from available Pezizomycotina genomes, we identified the smallest subset of genetic markers that can reliably reconstruct a robust phylogeny. We used a clustering approach to identify a sequence set that could provide an optimal trade-off between potential missing data and probe set cost. We incorporated this probe set into a user-friendly wrapper script named Universal Filamentous Ascomycete Target Enrichment (UnFATE) (<https://github.com/claudioametrano/UnFATE>) that allows phylogenomic inferences without requiring expert bioinformatics knowledge. In addition to phylogenetic results, the software provides a powerful multilocus alternative to Internal transcribed spacer (ITS) based barcoding. Phylogeny and barcoding approaches can be complemented by an integrated, pre-processed, and periodically updated database of all publicly available Pezizomycotina genomes. The UnFATE pipeline, using the 195 selected marker genes, consistently performed well across various phylogenetic depths, generating trees consistent with the reference phylogenomic inferences. The topological distance between the reference trees from literature and the best tree produced by UnFATE ranged between 0.10 and 0.14 (normalized Robinson-Foulds) for phylogenies from family to subphylum level. We also tested the *in vitro* success of the universal baits set in a target capture approach on 25 herbarium specimens from 10 representative classes in Pezizomycotina, which recovered a topology congruent with recent phylogenomic inferences for this group of fungi. The discriminating power of our gene set was also assessed by the multilocus barcoding approach, which outperformed the barcoding approach based on ITS. With these tools, we aim to provide a framework for a collaborative approach to build robust, conclusive phylogenies of this important fungal clade. [keyword: Bait set; targeted capture; phylogenomic; multilocus barcoding; next-generation sequencing.]

Filamentous ascomycetes (subphylum Pezizomycotina) represent most of the diversity in the largest phylum of Fungi Ascomycota and is comprised of species that adopt one of the widest ranges of lifestyles in the fungal kingdom. Many of these species have relevance in medicine, biotechnology, and food production, while many others have pathogenic potential. Given the importance of these species, it is surprising that the Ascomycota tree still needs to be fully resolved. Although many major conflicts were solved recently using genome-scale data (Shen et al. 2020), phylogenomic analyses often suffer from incomplete taxon sampling. For instance, 50% of the described Ascomycota orders have no representative genome assembly (Hill et al. 2021). Even when the phylogenetic inference is conducted using a few well-established genetic markers, the accuracy of the results can be compromised by incomplete taxon sampling. For instance, the estimated number of taxa for which the commonly used rDNA large subunit locus (LSU or

28S) is available on National Center for Biotechnology Information (NCBI) Genbank is about 20,000 (data not shown), which roughly corresponds to the number of species that were accepted for the Dothideomycetes class alone 10 years ago (Kirk et al. 2008; Schoch et al. 2009). Within the last 10 years, many more species have been described, since a relevant portion of the fungal diversity is still unknown (Hawksworth and Lücking 2017).

Shotgun genome sequencing (Venter et al. 1998) based on high-throughput sequencing (Ronaghi 2001) technologies provided the necessary impulse for a framework shift in phylogenetic studies. It is now a common practice to consider the evolutionary information carried by hundreds to thousands of genomic loci. Given the large amount of genomic data that is used for phylogenetic work, studies are often coherently labeled as phylogenomic studies (Philippe et al. 2005). However, the term was first introduced by Eisen (1998) to highlight the

need to include evolutionary relationships to infer gene function better. Following this trend, new bioinformatic tools to reconstruct phylogenomic trees emerged and former phylogenetic programs were updated to handle phylogenomic-level data sets (Zhang et al. 2018; Kozlov et al. 2019; Minh et al. 2020). However, even with the possibility of obtaining high-quality fungal genomes from high-throughput sequencing (Rhoads and Au 2015; Jain et al. 2016), sequencing and assembling a draft fungal genome in good quality still presents some challenges. For example, symbiotic organisms and microfungi from complex matrices, such as soil, need to be either isolated or sequenced as metagenomes, and herbarium samples often have poor DNA quality, which is unsuitable for WGS applications.

The economic burden and technical challenges in identifying fungal samples or producing a robust phylogeny of the clade of interest using WGS can be avoided using reduced representation methods. Reduced representation methods require less sequencing and can offer pre-designed pipelines to analyze the sequence data, which can otherwise be overwhelming when hundreds of samples are sequenced (Good 2012). These methods can be based on restriction enzyme-selected loci (e.g., Restriction site associated DNA sequencing, RADseq) or are based on specific loci to be targeted and captured (e.g., ultra-conserved elements, Faircloth et al. 2012). Restriction-based approaches are only effective if restriction sites are conserved among taxa (e.g., RADseq, Davey and Blaxter 2010). Therefore, methods, such as RADseq are better suited to infer shallow evolutionary relationships and are often used for population genomic approaches (Grewe et al. 2017, 2018; Salas-Lizana and Oono 2018). However, target capture methods can obtain complete coding sequences (Schott et al. 2017), which allows the selection of specific marker genes of interest based on their function or their phylogenetic performance. Hence, targeted capture allows the reconstruction of phylogenies within deeper time scales. In addition, it can access an additional source of data from off-target reads, which have been exploited to assemble organellar markers and nuclear rDNA (Weitemier et al. 2014; Allio et al. 2020). Among the possible approaches to enrich a particular set of loci of interest (target enrichment), hybrid capture using nano beads has recently become the technical solution of choice (Mamanova et al. 2010; Mertes et al. 2011). Target capturing and the enrichment of specific loci for a specific taxon has especially proven beneficial when working with older herbarium specimens with degraded DNA, or when symbiotic organisms are targeted (Brewer et al. 2019; Forrest et al. 2019; Grewe et al. 2020). Although phylogenetic studies using this technique have the additional cost of the enrichment step, advanced multiplexing techniques and other strategies (Glenn et al. 2019; Hale et al. 2020) can help dilute this cost. In addition, the marker selection and the probe design process can be both time-consuming and bioinformatically challenging. Bait sets usually are

designed from 1 or a few reference sequences (Widhalm et al. 2019), and the performance of these baits rapidly decreases when the reference sequence diverges too far from the targeted DNA (Liu et al. 2019). To overcome these limitations, universal bait sets have been developed, such as for angiosperms (Johnson et al. 2019) and flagellate plants (Breinholt et al. 2021).

The opportunity to sequence multiple genes with target capture allows a multilocus barcoding approach. Multilocus barcoding (Blaxter 2004) has already proven to be a powerful tool, overcoming the limitations of single-locus barcode approaches based on rDNA (Mallo and Posada 2016). In addition, Liu et al. (2017) showed that a multilocus approach based on tens to hundreds of nuclear protein-coding loci outperformed the Cytochrome c Oxidase I (COI) barcode routinely used for fish when sister species were tested. The standard barcode for Fungi is rDNA ITS (Beimforde et al. 2014). However, this marker lacks the needed resolving power in several groups of Fungi (Roe et al. 2010; Pino-Bodas et al. 2013; Lücking et al. 2020). Thus, we established a multilocus approach to barcoding, to enhance the possibility of reliably describing fungal diversity.

The recent increase in the popularity of target enrichment sequencing led to the development of many bioinformatics tools. Existing bioinformatics pipelines allow users to both identify suitable loci and to design probes (Mayer et al. 2016; Campana 2018; Chafin et al. 2018); other pipelines allow the processing of data from target enrichment sequencing (Johnson et al. 2016; Andermann et al. 2018), whereas others allow reconstructing phylogenomic trees from hundreds to thousands of loci applying various approaches to data filtering and tree reconstruction (Tice et al. 2021; Resl et al. 2022). However, no existing solution can simultaneously accomplish all these steps, for all filamentous ascomycetes.

We aimed to fill this gap by building a comprehensive and user-friendly pipeline that integrates pre-designed universal baits for filamentous ascomycetes and automates all required steps of bioinformatic processing from the raw data to the phylogenetic tree. The pre-designed universal bait sequences will be accessible to the public as well as the computational pipeline Universal Filamentous Ascomycete Target Enrichment (UnFATE) bait set and wrapper script which provides a 1-click straightforward phylogenomic workflow. The universal baits set and UnFATE pipeline, together with the existing genomic resources, will help to populate a database of marker sequences for filamentous ascomycetes. This database will provide a framework for a collaborative approach to build robust phylogenies of filamentous ascomycetes, similar to what the Angiosperms353 (Johnson et al. 2019) and GoFlag 451 (Breinholt et al. 2021) projects are accomplishing in flowering and the flagellate plants tree of life, respectively (Baker et al. 2021; Maurin et al. 2021; McDonnell et al. 2021). This database of filamentous ascomycete genes allowed the implementation of a multigene

barcoding approach. Consequently, a barcoding option that takes advantage of all NCBI publicly available filamentous ascomycete genomes is already implemented in the first version of UnFATE.

MATERIALS AND METHODS

Gene Selection and Bait Set Design

Initially, we determined the minimum number of genes that can produce a reliable phylogeny of Pezizomycotina. We used a pilot data set of 162 genome assemblies from NCBI assembly (11 December 2019; [Supplementary Table S1](#)), picking 1 assembly per genus and excluding assemblies with <90% of BUSCO completeness, >10% of duplicated BUSCOs (BUSCO 5.4.2). From these assemblies, we selected the widely used universal single-copy orthologs (BUSCO) ([Simão et al. 2015](#)) as a starting pool gene set for Ascomycota (1706 protein-coding genes, Orthodb v10, [Waterhouse et al. 2013](#)). The ortholog genes that were retrieved were then aligned with mafft using parameters -maxiterate 1000 and -localpair ([Katoh and Standley 2013](#)) and subsequently block-filtered with Gblocks using relaxed parameters ([Castresana 2000](#)). Genes with aligned exons shorter than 600 bp were discarded. To identify the minimum number of genes that were required to receive a well-supported, stable phylogeny, a set of 20 genes were stepwise added to a concatenated matrix until the full gene set was used. Of each matrix with 20 to all genes, we calculated a maximum likelihood tree with IQTREE 2.1.2 ([Minh et al. 2020](#)). This procedure was repeated a second time, but the genes were ordered by increasing substitution rates as done previously ([Grewe et al. 2020](#)). The rates were calculated using PAML 4.9 baseml ([Yang 2007](#)) using a calibration point of 449 Mya for the Pezizomycotina root (average value among [Beimforde et al. 2014](#); [Nelsen et al. 2020](#) and; [Prieto and Wedin 2013](#)). For the constrained tree for the rates analyses, we used the single-locus tree inferred by reconstruct single-locus ML trees (RaxML) 8.2 ([Stamatakis 2014](#)), rooted with the Orbiliomycetes-Pezizomycetes clade. In addition to calculating species trees from concatenated matrices, we used a coalescent-based approach with Accurate Species TRee ALgorithm (ASTRAL) ([Mirarab et al. 2014](#)) to reconstruct species trees from a stepwise increasing number of single loci RAXML trees. Differences in the topology between the produced trees were measured by normalized Robinson-Foulds distance (nRF) implemented in the ETE toolkit 3.1.3 ([Huerta-Cepas et al. 2016](#)). A nRF distance matrix was built to assess the topological distances between all reconstructed phylogenies and identify at which gene number the phylogenetic trees were not further improving. In addition to nRF, all phylogenetic trees were used to calculate the average bootstrap support values as another parameter to compare the quality of the resulting phylogenetic trees.

After assessing the topological stability and average support values variation produced by a progressively larger number of loci using the pilot data set, a bigger data set was built to identify suitable genes in an amount consistent with the findings of the preliminary analyses. All sequenced draft genomes of Pezizomycotina available at NCBI on 6 May 2020 were processed with BUSCO5.4.2 (Benchmarking single-copy orthologs, [Simão et al. 2015](#)). The most complete assembly for each species was then retained in a comprehensive data set. BUSCO genes from this data set were aligned with mafft 7.52 and block-filtered with Gblocks 0.91. The genes still retaining at least 600 bp of aligned exons were used for the k-medoids clustering to identify the best representative sequences for each gene, that is, the lowest number of sequences in an alignment that represented all known Pezizomycotina species (throughout the remaining text, these alignments of reference sequences from few representative species are referred to “medoid alignments.”). K-medoid clustering works like the k-means approach, except it picks an actual value (i.e., a sequence) from the alignment instead of using the mean value. The aim of this method is to select the sequences (medoids) that minimize the distance to the surrounding values. This approach is described in [Johnson et al. \(2019\)](#) and was used to build a universal bait set for angiosperms. We tested maximum allowed distances from medoids of 0.20, 0.25, 0.30, and a maximum number of medoids of up to 70. We used 100 iterations for each medoid number tested since the clustering method starts from a random sequence in the alignment on each iteration. The genes that successfully converged, were then represented by a number of representative sequences smaller or equal to the maximum, which allowed at least 95% of the sequences in the alignment to be within the maximum distance selected. The clustering parameters (maximum distance from the medoid, medoid number allowed) and the minimum length of exons were adjusted to obtain at least the minimum number of genes that were optimal for topological and nodal support of phylogenetic trees (see above).

The alignments of the selected genes were used to RAXML. Single-locus topologies were used to build a nRF distance matrix, which was used to calculate the average distance of any single-locus tree from the others and the species tree. This distribution of average distances was used to discard outlier topologies using the Thompson-tau test. The single-locus trees distance matrix was then used as an input for a 2D Multi-Dimensional Scaling and DBscan clustering (scikit-learn 1.0, [Pedregosa et al. 2011](#)). The clustering method was used to evaluate the presence of topology clusters corresponding to possible alternative discordant evolutionary histories.

Unfiltered “medoid alignments” were realigned in a codon-aware fashion with MACSE2 ([Eddy 2009](#); [Ranwez et al. 2018](#)). From these alignments, HMMER 3.3.2 profiles were created with hmmbuild and these

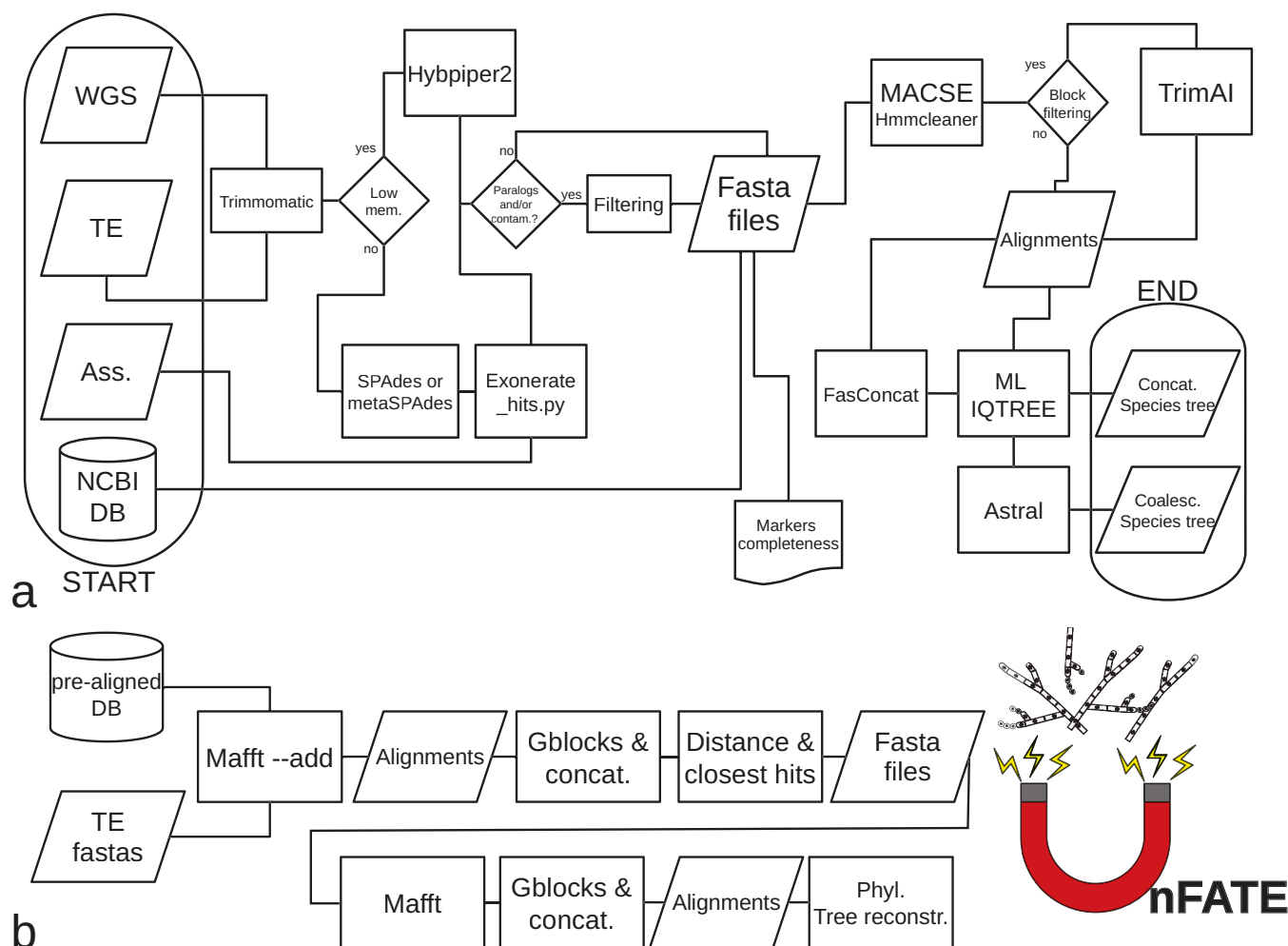


FIGURE 1. (a) Workflow of the UnFATE pipeline; (b) Workflow of UnFATE in its multilocus barcode version. Parallelograms are used for inputs/outputs, rectangles for processes, diamonds for decisions, cylinders for databases. TE: Target Enrichment; WGS: Whole-Genome Sequencing; Ass.: Assembly; DB: Database; ML: Maximum likelihood.

profiles were used to retrieve the target genes from selected transcriptomes and *Aspergillus nidulans* annotated genome (Supplementary Table S2) using *hmmsearch* (Finn et al. 2011). These alignments were then used to manually curate the “medoid alignments” that were then used to design the bait set. This additional step allowed us to delete introns when possible, and to check the ends of the coding sequences predicted by Augustus (Stanke et al. 2006).

The manually curated “medoid alignments” were sent to Arbor Biosciences (Ann Arbor, MI) to produce the RNA baits set. The complete baits set comprises of 80 bp baits and is designed in a 2× tiling fashion. Ns in consecutive runs of 1–10 were changed to Ts. As a control, all baits were BLAST (2.14.1) searched in 13 reference genomes. All baits that produced multiple hits, those with more than 35% repeats, or those in a region with more than 35% repeats were discarded. Baits with extreme GC content were also discarded (25% < GC < 67%). These baits sequences will be made available to the public for reproduction.

Pipeline Workflow

We implemented a user-friendly wrapper script called UnFATE (github.com/claudioametrano/UnFATE). The software allows the processing of filamentous ascomycete target enrichment data and the processing of data from other sources, such as genome assemblies or raw sequence data. As shown in the workflow (Fig. 1a), target enrichment data are trimmed by Trimmomatic (Bolger et al. 2014) before they are assembled with metaSPAdes 3.15.5 (default; Nurk et al. 2017) or Hybpiper 2.1.8 (-y option in UnFATE: mapping to reference, then SPAdes; Johnson et al. 2016). Raw WGS data are trimmed by Trimmomatic 0.39 before they are either assembled with SPAdes 3.15.5 (default; Bankevich et al. 2012) or Hybpiper (see above). SPAdes and MetaSPAdes assemblies are then processed using a script from the Hybpiper workflow (*exonerate_hits.py*), which exploits Exonerate 2.4.0 (Saari et al. 2017) for appropriate gene extraction. After gene retrieval from the

data sources, it is highly recommended to apply the filtering step (`-s` argument), especially for samples from unknown targets. When this filtering is applied, genes that are retrieved in multiple copies (2 or more) in a sample and do not meet the coverage depth condition are removed. By default, a copy must be at least 10 times deeper covered than any other copy obtained. If the optional filtering step is not used, the copy that satisfies the depth condition or, if this fails, the most similar copy to the reference sequence, according to the Exonerate alignment, is retained. It is worth noting that the Hybpiper2 default approach, while is suitable in many cases where contamination is unlikely (e.g., plant contaminants for a plant sample), is not ideal for fresh or herbarium fungal samples, which may contain a significant amount of other fungi efficiently captured—especially when a universal bait set is used. The completeness of the extracted genes in fasta file format is evaluated by comparing them with the reference sequence length. A heatmap and a.csv table are produced to indicate the level of completeness of each gene (provided in the “fastas” output folder). In the next step of the pipeline, nucleotide, and amino-acid translations of the extracted genes are codon-aware aligned with the OMM-MACSE 10.02 pipeline (Ranwez et al. 2018; Di Franco et al. 2019). This pipeline includes a segment-based filtering step based on Hmmercleaner, which is performed in addition to an optional blocks filtering approach using TrimAl 1.4 (Capella-Gutiérrez et al. 2009). Amino-acid and nucleotide alignments are then used both for concatenated (FASconCAT-g 1.04, Kück and Longo 2014) ML species tree, and for single-locus ML tree inferences (both using IQTREE 2.1.1 and selecting the best substitution model with Model Finder Plus) followed by species tree inference using ASTRAL 5.7.7.

Universal Bait Set and Pipeline Testing

Including the data specifically generated to test the universal baits set (see below), here are the data sets that were used to evaluate the performance of the pipeline, the selected genes, and the bait set:

- 1) Twenty-five herbarium specimens of variable age (2004–2023, see NCBI SRA metadata, Data Availability section) from 10 major classes in Pezizomycotina, specifically selected, DNA extracted, captured with the universal baits, and sequenced as 2×150 bp reads on the Illumina Novaseq6000, to test the 195 genes and the universal bait set; 2) High and low ($<15\times$) coverage WGS data from Ametrano et al. (2022) and Pizarro et al. (2018), respectively (195 UnFATE genes); 3) high-quality target enrichment data capturing a wide phylogenetic range of Lobariaceae (96 samples) and other Peltigerales (35 samples) by baits (400 genes) designed from *Lobaria pulmonaria* and *Evernia prunastri* (Widhalm et al. 2019, 2021); and 4) several phylogenomic data sets from literature; from shallower to deeper: 93 genomes in Aspergillaceae (Steenwyk et al. 2019), 48 lichen

metagenomes in Parmeliaceae (Pizarro et al. 2018), 101 genomes in Dothideomycetes (Haridas et al. 2020) and the Pezizomycotina subset of the fungal kingdom phylogeny (Li et al. 2021), containing 758 genomes (195 UnFATE genes).

The performance of the bait set and the pipeline was assessed as follows:

- 1) The completeness of the loci retrieved by UnFATE was estimated both in terms of loci completeness (in comparison to the length of the reference gene) and of percentage of loci retrieved (presence/absence) using WGS data (Pizarro et al. 2018; Ametrano et al. 2022), Peltigerales target enrichment data (Widhalm et al. 2019, 2021) and the newly sequenced Pezizomycotina specimens using the universal bait set; 2) The impact of mapping and coverage threshold parameters in Hybpiper on the completeness of the retrieved genes were also assessed with the Peltigerales target enrichment data; we relaxed the default BLAST e-value (10^{-10}) to 10^{-7} and 10^{-3} and the default coverage threshold ($8\times$) to $3\times$ and $1\times$. We kept the sequence identity threshold for the Exonerate alignment to a default value of 55% for both Hybpiper and “assembly first” methods; and 3) the performance of our 195 phylogenetic markers and the pipeline was assessed in terms of topological consistency using the 4 data sets mentioned earlier (Pizarro et al. 2018; Steenwyk et al. 2019; Haridas et al. 2020; Li et al. 2021). Phylogenies produced by the UnFATE wrapper and the originally published trees were compared both in terms of overall topological distance (nRF) and in detail assessing the topological differences and the nodal support. In addition to the ultrafast bootstrap nodal support provided by IQTREE2 and the local posterior probability from ASTRAL, the level of conflict of single gene topologies was summarized by Phyparts (Smith et al. 2015) and phypartspiecharts.py (github.com/mossmatters/phyloscripts/tree/master/phypartspiecharts).

The Ascomycota BUSCO gene set we used as the starting point for marker selection contains genes that are rarely duplicated. However, duplication, or other events, such as ploidy or contamination (especially when herbarium or fresh environmental specimens instead of cultures are used), can cause more than 1 copy to be obtained from target enrichment, or (meta) genomic data. Therefore the amount of gene copies was assessed using the `paralog_retriever.py` script from Hybpiper, both in 400 genes from Lobariaceae and in the newly generated sequences of the 195 genes in Pezizomycotina.

Pre-computed Pezizomycotina Database

To reduce the computational burden, a pre-computed database containing the selected gene set from any assembly available on NCBI assembly portal (29 July 2021) can be queried within UnFATE (NCBI DB, Fig. 1, option `-n`). The user can choose any main taxonomic rank from NCBI taxonomy belonging to the subphylum Pezizomycotina. The samples belonging to the selected

rank will then be automatically added to the alignment produced by the UnFATE pipeline without the need for the time-consuming processing of assemblies or raw reads archives.

The database was built with the following steps: 1) the assembly reports for the assemblies belonging to kingdom Fungi were downloaded with *ncbi-genome-download* (<https://github.com/kblin/ncbi-genome-download>) and filtered to retain only accession numbers, 2) the taxonomy was associated to each accession (https://github.com/claudioametrano/get_taxonomy_with_edirect), fixed to retain only the main ranks (class, order, family, genus, and species), and only samples classified at least to the genus level, 3) the accessions belonging to Pezizomycotina were downloaded with *ncbi-genome-download*, and 4) UnFATE was used to mine for the Pezizomycotina gene sets from the genomes and to build comprehensive fasta files. The database composed of the taxonomy file and the alignments can be queried through the UnFATE command line. The nucleotide version of the database was also prealigned with *mafft* (Katoh and Standley 2013) which speeds up the multilocus barcode approach of UnFATE.

Multilocus Barcode Mode and “-n AUTO” Mode

multilocus barcoding was added to the pipeline as an additional mode in UnFATE to provide fast multi-gene barcoding identification results, without running a complete phylogenetic analysis. The method follows the outline of Liu et al. (2017) that exploits the speed of a distance method coupled with the accuracy of a phylogenetic reconstruction method to infer the sister clade taxa of the query sample. The multilocus barcode mode (*barcode_wrap.py*, Fig. 1b) starts from a pre-aligned version of the Pezizomycotina database in its nucleotide version. The newly sequenced fasta file produced by UnFATE is added to these alignments using *mafft* (*--add*). The alignments are then block-filtered and concatenated. The polished alignments are used to calculate the pairwise distance (Biopython Bio.Phylo.TreeConstruction module, *DistanceCalculator*, “*blastn*” scores) between the query sequence and each sequence in the database to determine the closest sequences. The twenty most similar sequences are selected (with up to 4 sequences belonging to the same species). The query sequence and the most similar sequences are then realigned with *mafft*, block-filtered, and concatenated to produce a more specific alignment. Finally, a ML phylogenetic tree (IQTREE2) is generated from the concatenated DNA alignment which resolves the sister taxon/a relationship of the queried sequence. This distance-based method used for the barcode mode can also be used in the UnFATE *main_wrap.py* automatic mode (*-n AUTO*) to automatically select closely related sequences to the samples to be included in the phylogeny. The final tree includes the samples to be investigated, plus at most twenty additional tips for each sample. The barcode mode runs 1 sample at a time and does not allow to filter out multiple-copy genes; the

user can use the “-n AUTO” mode of *main_wrap.py* together with the “-s” argument, to both barcode multiple samples simultaneously and to apply the filtering.

The effectiveness of the distance method used for the barcoding mode to discriminate between species using an increasing amount of genetic data was assessed by random sampling (5 replicates) an increasing number of loci (3–150), and comparing the intraspecific and interspecific distances of multiple individuals (3 to 5 individuals of each species; https://github.com/jacob-njensen1/UnFATE_paper_scripts). The effectiveness of the barcoding method, and the barcoding potential of our genes, were assessed by comparing the identification results from our approach to the result obtained by ITS barcoding using BLAST (Altschul et al. 1990). For this approach, the complete assembly database was mined for the ITS region by ITSx (Bengtsson-Palme et al. 2013) after fragmenting the longest contigs to fragments of 100 Kb (ITSx maximum sequence length). An ITS BLAST database containing the same species as our genome-based database was then built. When no ITS region was found in the genome assembly, the ITS database was manually completed by sequences of the same species (where possible, the same strain) from the NCBI nucleotide database to ensure a fair comparison of the 2 barcoding methods.

RESULTS

Gene Selection and Bait Design

The topology and support values of phylogenetic trees calculated from data

sets of increasing gene numbers plateaued rapidly, indicating that about 200 genes are suitable to reconstruct the Pezizomycotina phylogeny with a resolution similar to all BUSCO genes. For randomly sampled 200 genes, the nRF distance between sampled genes and the reference was 0.06 for concatenated and 0.01 for coalescence-based species tree inferences. For 200 genes that were ordered by substitution rate, the nRF distance was 0.06 for concatenated and 0.04 for coalescence-based tree inferences (Supplementary Figs. S1–S4). Moreover, the average support value (bootstrap) for the concatenated ML trees produced with 200 genes was 99.8 for randomly sampled genes and 95.5 when slow to fast-evolving genes were used. The average support value (posterior probability) for coalescence-based trees was 0.981 for randomly sampled genes and 0.977 when slow to fast-evolving genes were used (Supplementary Fig. S5 and S6). Therefore, about 200 genes were set as the minimum number of genes that can construct reliable phylogenies. This number of genes achieved the best balance between the required phylogenetic accuracy (phylogenetic signal increases with the number of genes) and bait set size (more genes will result in a bigger bait set) to represent the entire filamentous ascomycetes phylogeny effectively.

We used the complete set of 1706 semi-universal BUSCO orthologs as a starting point to select about 200 genes. According to the original publication, these universal single-copy orthologs should be present in 90% of the species and they are rarely duplicated (Simão et al. 2015). The alignments used to retrieve the representative sequences by k-medoids clustering were built from 3124 Pezizomycotina draft genome assemblies. Of these genomes, 1163 remained (Supplementary Table S3) after filtering the most complete for each species based on the BUSCO completeness benchmark. The k-medoids clustering converged for 210 genes, represented by up to 50 medoids per gene, with 95% of the sequences within a distance of 0.25 from their medoid (the representative sequence). Of these 210 genes, ten genes were discarded because they were identified as outliers in the distribution of single-locus topology average nRF distance; 2 additional genes were removed by the same method, as they performed consistently poorly in test runs using phylogenomic data sets from the literature. An additional 3 genes were removed because they were rich in repetitive regions (RGG boxes, Thandapani et al. 2013). The DBscan clustering approach performed on the nRF distance matrix from single-locus topology cluster did not identify any dominant alternative topology cluster (Supplementary Fig. S7). The final gene set consisted of 195 genes represented by 3109 individual sequences for a total length of 4,748,098 bp. The bait set for target capture approaches was designed from these sequences in a 2× tiling design. The baits comprised of 100,000 80 bp RNA baits, which survived the filtering steps (see methods).

Pipeline Testing: Average Length, Fraction of Genes and Genes in Multiple Copies

Under the UnFATE pipeline, WGS data can be processed either with Hybpiper, which uses reads mapping or by assembling the WGS data with SPAdes, which does not require read mapping (“assembly first,” default). The pipeline tests indicated that the 2 methods provide similar completeness for high coverage WGS DATA (Hybpiper: 97.1%, “Assembly first”: 96.7%; Supplementary Table S4). Hybpiper had the advantage of reduced memory footprint since it only assembles reads that map onto the reference sequences. This is due to Hybpiper using a mapping approach (e-value 10^{-10} using BLASTx, if an amino-acid reference file is provided) before assembling, and a coverage threshold for the assembled contigs (by default equal to 8). However, the “assembly first” method, on average, reduced the amount of missing data (average marker length completeness: Hybpiper 70.4%, “Assembly first” 96.6%; Supplementary Table S4), when low coverage WGS data was used (Grewe et al. 2020) or, when reference sequences are divergent enough from target DNA (see below).

We also assessed the performance of the pipeline using high quality 96 samples in Lobariaceae captured by baits designed from 2 taxa (*Lobaria pulmonaria* and

Evernia prunastri) (Widhalm et al. 2019), and testing the universal bait set we designed, sequencing 25 herbarium specimens from ten major classes in Pezizomycotina (Supplementary Table S5). In the Peltigerales data set, we determined that using Hybpiper2 or “assembly first” retrieved loci which were on average 0.86 (st. dev. 0.19) and 0.93 (st. dev. 0.16) the length of the respective reference sequences. The “assembly first” approach provided, on average, 8% more loci than Hybpiper2. Relaxing the Hybpiper2 parameters (e-value, coverage) provided an improvement of 4% more loci. It is worth noting that the reference sequence can strongly influence the completeness of the gene retrieved. The difference between the 2 methods, in terms of loci retrieved, became indeed larger when the sample diverged from the 1 used to design the bait set; Supplementary Table S6 shows, using the Widhalm et. al (2021) Peltigerales data, that just 5%–6% loci retrieved increase for Lobarioideae (*Lobaria* was used as reference sequences indeed), but up to 36% more loci are retrieved, using “assembly first,” in more distantly related clades, such as Coccocarpiaceae. In the Pezizomycotina data set, obtained using the universal bait set, we determined that using Hybpiper2 or “assembly first” retrieved loci which were on average 0.92 (st. dev. 0.06) and 0.97 (st. dev. 0.04) the length of the respective reference sequences. The “assembly first” approach provided on average 1% more loci than Hybpiper2. In terms of the number of copies, capture using a bait set specifically designed on 1 to few taxa (Widhalm et al. 2019) resulted in a low fraction of multiple-copy genes (~4%) mostly in 2 copies (average 2.07–2.11 copies in multiple-copy genes), while the universal bait set retrieved a relevant portion of multiple-copy genes (31% with on average 3.4–3.6 copies; Supplementary Table S5).

Pipeline Testing: Comparison Between the 195 UnFATE Genes Phylogenies and Previously Published Phylogenomics Trees

We compared phylogenies of different depths produced with UnFATE based on 195 universal filamentous ascomycete genes with the respective reference phylogenies from the literature. The published maximum likelihood (ML) genus-level phylogeny of *Aspergillus* and *Penicillium* (DNA, concatenation, 1668 loci) (Steenwyk et al. 2019) compared to the ML phylogeny by UnFATE with a nRF distance of 0.11. The distances produced by all other topologies from the original study to the UnFATE topologies had a nRF range of 0.10–0.18. (Supplementary Fig. S8a). The original phylogenies were only marginally influenced by the data set used (nucleotide or AA) and the different reconstruction methods (ML or coalescent-based), while UnFATE topologies were slightly less consistent. The comparison (Supplementary Fig. S9) highlighted only a minor and 2 major incongruences: The sister relationship of *Aspergillus steynii*, the placement of *Aspergillus ochraceoroseus* within the genus *Aspergillus*, and the placement of the genera *Monascus* and *Xeromyces*.

The comparison of the *Parmeliaceae* family-level phylogeny from Pizarro et al. (2018) (2256 loci) showed mostly consistent topologies with 1 major incongruence. The ML phylogenies based on a concatenated data set had a nRF distance of 0.13 in relation to the published tree. The 2D-MDS summarizes the distances between the 4 topologies produced by UnFATE and the 2 topologies from the original study (Supplementary Fig. S8b, nRF range 0.13–0.27). The plot highlights that the UnFATE topologies cluster according to the reconstruction method used, with the UnFATE coalescent-based species trees as having the furthest topologies. The comparison of topologies (Supplementary Fig. S10) highlighted 2 minor and 1 major incongruence: the placement of the *Usnea-Cornicularia* clade, the sister relationship of *Arctoparmelia centrifuga* and *Pseudevernia furfuracea*, and the placement of *Relicina intertexta*, which is not supported in the UnFATE phylogeny.

The comparison between the Dothideomycetes class-level phylogeny from Haridas et al. (2020) (aa, concatenated, ML, 738 loci) and the UnFATE tree (DNA, concatenated, ML, 195 loci) had a nRF distance of 0.14 (the nRF range between the UnFATE and the original phylogenies was 0.09–0.20) (Supplementary Fig. S8c). The comparison of topologies (Supplementary Fig. S11) shows few incongruences, mostly in the placement of samples belonging to small orders, such as Patellariales, Acrospemales, Lineolatales and *incertae sedis* taxa such as *Coniosporium*, or within Pleosporales (*Aaosphaeria*, *Melanomma*, *Pleomassaria*, and *Massariosphaeria*).

The comparison at the highest taxonomic level measured the distance between the phylogenies produced by Li et al. (2021), pruned to the taxa that belong to Pezizomycotina (from a fungal kingdom tree), and the UnFATE phylogenies. The nRF distance between the published tree (290 loci aa concatenated, ML) and the UnFATE ML concatenated DNA tree (195 loci) was 0.11. The nRF distance between the UnFATE topologies and original phylogenies ranged between 0.08 and 0.20 with the UnFATE topologies mostly clustering by data type (DNA, aa) and the Li et al. (2021) one mostly clustering by reconstruction method (concatenation, coalescent-based) (Supplementary Fig. S8d). In this case, the most distantly related UnFATE topology was built from aa alignments and a coalescent-based approach. The comparison of topologies (Supplementary Fig. S12) shows some major incongruences in various lineages involving: *Cryomyces*, *Venturiales*, *Thermoascaceae*, *Aspergillus* (few species), *Sordariales-Coniochaetales*, *Erysiphaceae*, *Diaporthales*, *Glomerellales*, *Sarocladiaceae*, *Stachybotryaceae*, *Bionectriaceae*, *Trichothecium*, *Ophiocordycipitaceae*, and *Cordycipitaceae*.

The ASTRAL species trees reconstructed from single-locus ML trees of the universal target genes highlighted a high degree of conflict between gene trees and species trees at every level of phylogenetic depth (Supplementary Figs. S13–S16). The pie charts show a high number of nodes having a prevalence of

topologies discordant from the ASTRAL species tree. The nodes with a high level of conflict are characterized by multiple conflicting topologies—none of the nodes in any of the 4 test phylogenies was characterized by a prevalence of 2 leading topologies. Most nodes show full support because the published phylogenomic trees are based on many more loci than the reduced representation method we implemented with UnFATE (except for the full fungal kingdom phylogeny from Li et al. 2021). The UnFATE phylogenies present a few more nodes than the references, which are not fully supported, considering the phylogenies up to the class level. However, all the incongruences highlighted by the tanglegrams (Supplementary Figs. S9–S12), except the *Relicina intertexta*, were fully supported. The UnFATE Pezizomycotina tree, being a deeper and larger phylogeny, presents the highest number of unsupported bipartitions, but often these clades also corresponded to incongruences with the reference phylogeny. Relevant examples include: *Venturiaceae* (bootstrap value = 4), *Thermoascaceae* (82), *Erysiphaceae* (4), *Sordariales-Coniochaetales* (84) *Sarocladiaceae* (2), *Stachybotryaceae* (2), *Ophiocordycipitaceae* (2), and *Cordycipitaceae* (2).

Pipeline Testing: A Proof of Concept Pezizomycotina Phylogeny Using the Universal Bait Set

We extracted DNA from 25 herbarium samples, followed by target capturing and sequencing using our universal bait set to assess its efficacy. An average of $\sim 4.8 \times 10^6$ (st. dev. 2.4×10^6) 2×150 bp high quality reads per sample were produced (phred score of raw reads > 28); of these an average of 37% (st. dev 16%) were on target.

The resulting data set was complemented by 1 sample per class from publicly available genomes. As the fraction of multiple-copy genes was higher using our universal bait set in comparison to more specific ones (Supplementary Table S5), and the herbarium samples being of heterogeneous quality, biomass and age, we ran the UnFATE pipeline both with and without the “strict filtering” argument (-s), which deletes every gene in multiple copies, except those for which 1 copy has at least *n* times the coverage of the others (default 10 $\times$).

The heatmap (Fig. 2) shows that this strict filtering increased the number of missing loci of the target captured herbarium samples from nearly zero (only 1 gene in 1 sample completely missing; Fig. 2a) to about 12% of missing loci (Fig. 2b). This increase in missing data due to strict filtering, along with the average number of copies for multiple-copy genes being higher than 2 (Supplementary Table S5), highlights that the herbarium samples might have more contamination from other fungi rather than an abundance of paralogous genes (which should be in 2 copies characterized by similar coverage). Filtering also increased the number of missing loci in the ten high-quality assembly samples, but to a much lesser extent. Since these assemblies

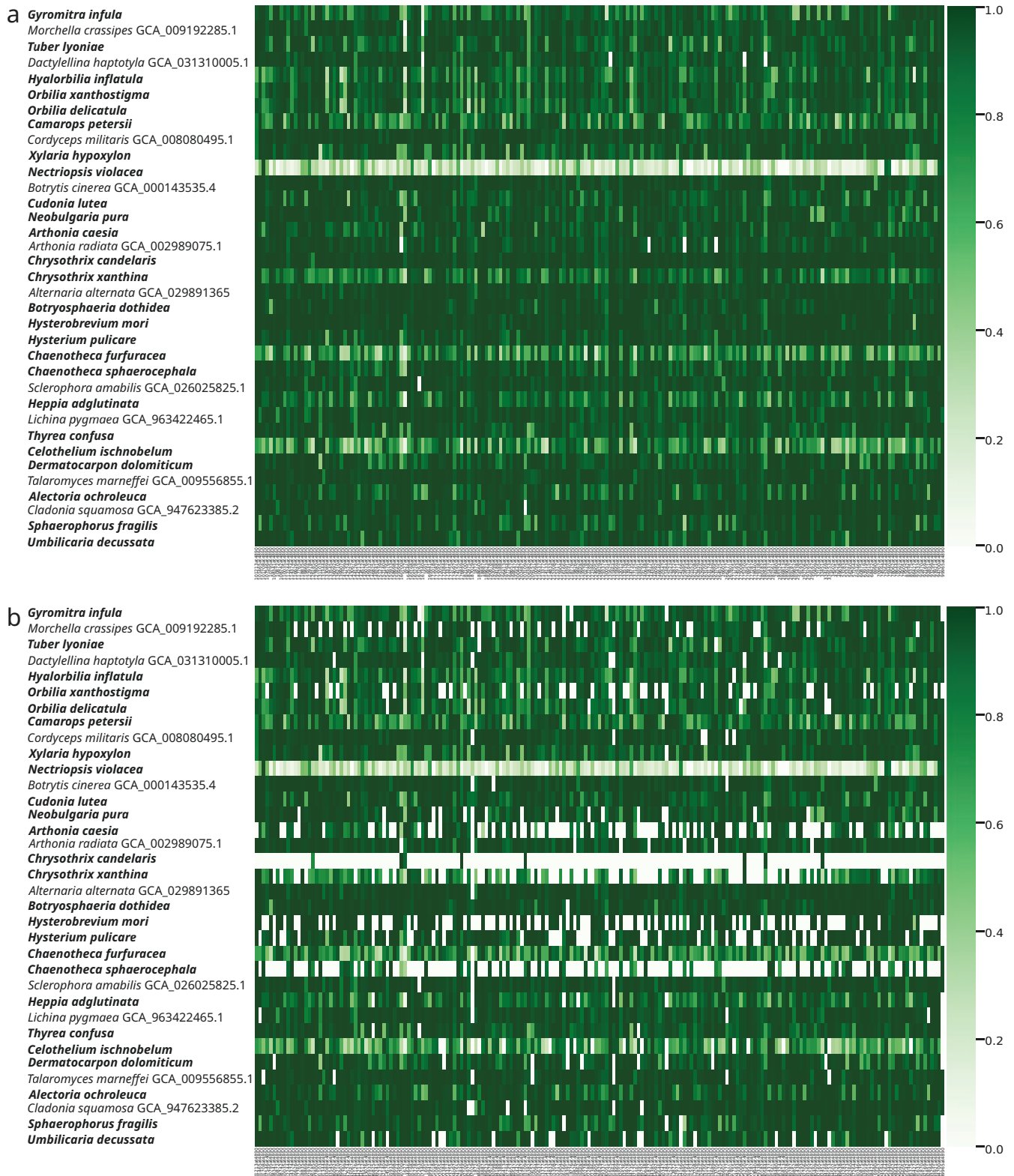


FIGURE 2. Completeness of the 195 UnFATE genes (horizontal axis) as a fraction of the reference sequence lengths for each of the 25 target captured herbarium samples (in bold) and the 10 reference genome (vertical axis) when: (a) all the copies are considered, or (b) only the genes found in single copy, or in multiple copy with a copy having at least ten times the average coverage of any other copy (-s argument). Samples are ordered according to the phylogenetic placement (at the class level) and in alphabetical order within the class.

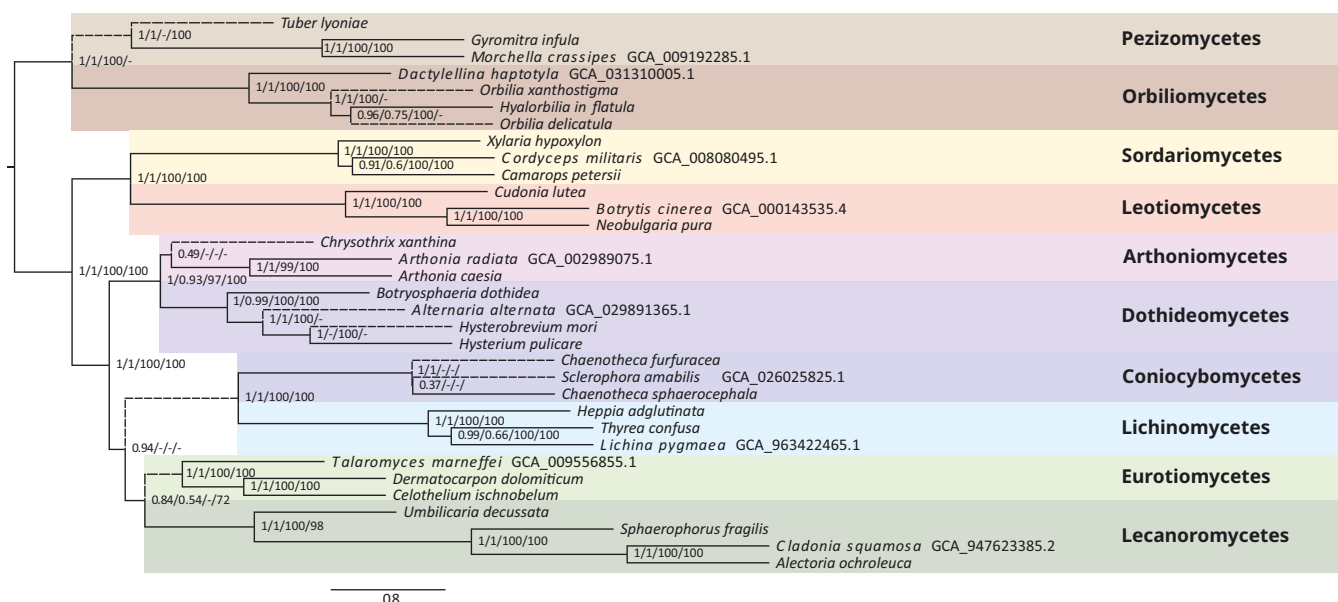


FIGURE 3. Coalescence-based (ASTRAL) species tree of ML single-locus trees (nucleotides). Support values are near branches (posterior probability from this phylogeny/bootstraps of the coalescence-based species tree of amino-acid ML trees/ultrafast bootstraps from the ML tree from concatenated nucleotide alignments/bootstraps from the ML tree from concatenated amino-acid alignments). No support value (“-”) and dashed branch indicate an incongruent topology between this phylogeny and at least one of the other 3 inferred phylogenies (coalescence aa, concatenated nt, concatenated aa). Tips with NCBI accession numbers reported are genome assemblies, while the rest of the tips are taxa from targeted capture.

were derived from whole-genome sequencing projects, it is likely that duplicated gene copies represent paralogs, as they were mostly found in 2 copies with similar coverage. Consequently, due to 31% of loci being in multiple copies, 12% of loci were missing in target enrichment herbarium samples. Similarly, for the 4% of genes found in multiple copies, a comparable proportion of missing loci was observed in the assemblies. Two herbarium samples were discarded: *Chrysothrix candelaris*, for the very high fraction of multiple-copy genes not solved by coverage, which left the sample with only 8 genes in single copy, and *Nectriopsis violacea*, which mostly produced short markers due to poor DNA quality. Other samples with a relevant amount of missing genes were retained to assess how robust the method was to missing data. After alignment and block filtering, the DNA data matrix consisted of ~256,000 bp, on average ~1300 bp per locus, and ~86,000 amino acids, on average ~441 amino per locus.

The phylogenies inferred by default by the UnFATE pipeline are 4: 2 summary species trees (DNA and AA) and 2 ML phylogenies based on concatenated alignments (see methods section). The most consistent phylogenies were obtained by the summary method (Fig. 3) which produced 2 phylogenies (DNA and AA) of 0.17 nRF (normalized Robinson-Foulds distance). The same input data used for the concatenation approach produced 2 phylogenies 0.33 nRF far apart, similar to the distance of the phylogenies obtained with both a different inference method (summary/concatenation) and data (DNA/AA) (0.30 nRF). The phylogenies were

mostly consistent in basal nodes with few incongruences. The most relevant was the placement of the sister clades Lichinomycetes/Coniocybomycetes (recently merged under the name Lichinomycetes, together with other classes not considered here, Díaz-Escandón et al. 2022) in relationship with the Eurotiomycetes/Lecanoromycetes clade. The number of genes used (up to 195), intermediate between a few-loci and a phylogenomic analysis, allowed the support values not to saturate, both in the case of the summary method and the concatenation, highlighting few critical nodes of the presented phylogeny, which did not receive full support.

Multilocus Barcoding

Multilocus barcoding was implemented in the UnFATE pipeline and provides a fully automated feature that allows barcoding of selected taxa based on the database integrated into UnFATE. The comparison of intra- and inter-species distances using randomly sampled loci highlighted a generalized increase in the resolution power of the method with an increasing number of loci (Fig. 4). The UnFATE barcoding mode uses a distance method as a fast-skimming approach to select the closest sequences from the database. These sequences are subsequently used to reconstruct a small phylogeny (20 samples, up to 4 samples per species, in default mode) and to infer the most reliable sister clade of the queried sample. The species we selected to test the power of the distance method exemplify 4 possible

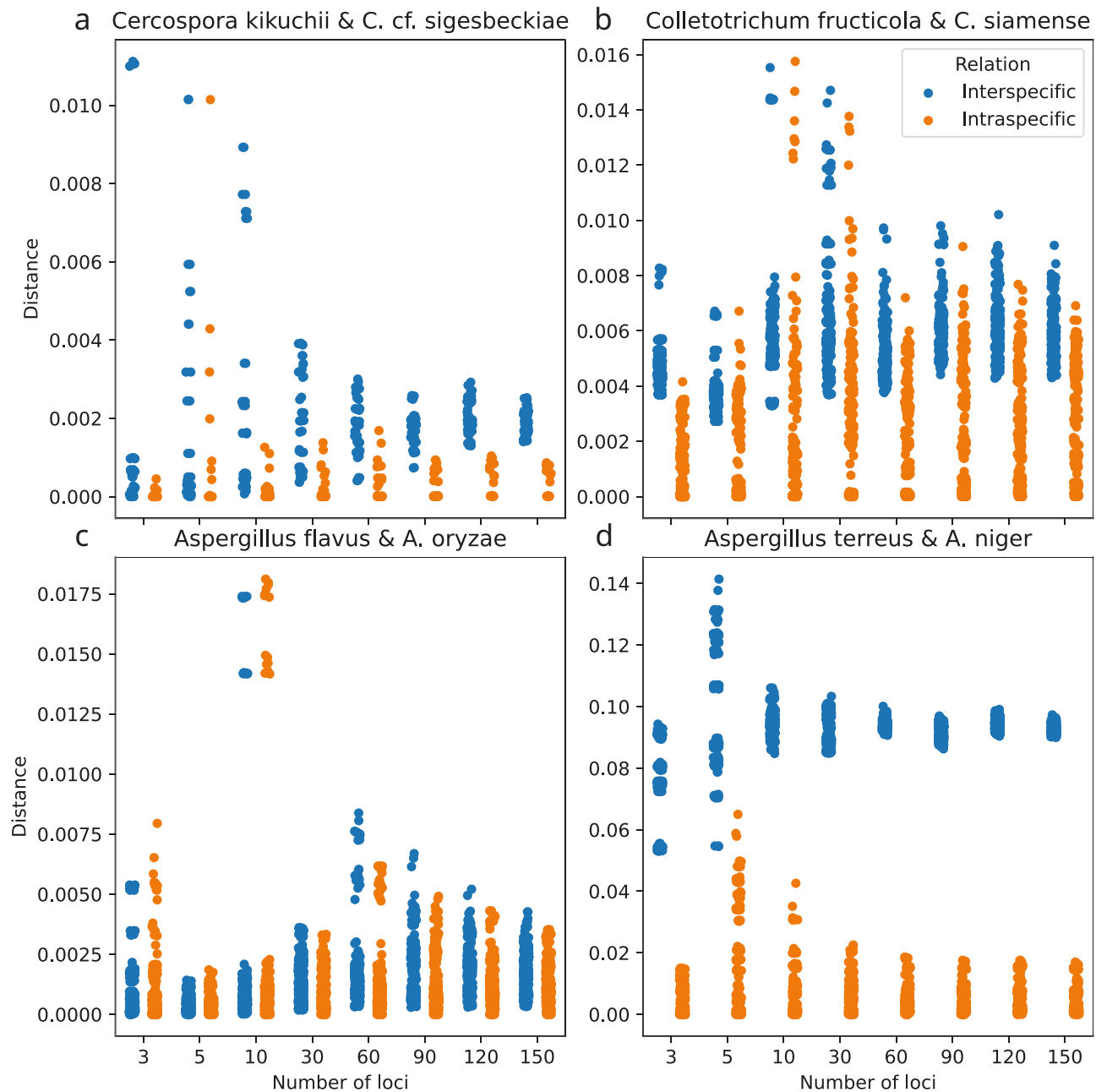


FIGURE 4. Intra- (orange) and inter- (blue) specific distance between multiple samples of 4 pairs of closely related species. Distances were calculated using alignments of an increasing number of randomly sampled markers among the UnFATE genes. Normalized distance is shown on vertical axis, number of sampled loci on horizontal axis.

scenarios: 1) Few genes can lack the resolving power to discriminate closely related species, such as *Cercospora kikuchii* and *C. cf. sigesbeckiae*. However, a higher number of genes both increased the precision (smaller variance among replicates and individuals) and the discrimination accuracy (interspecific distance > intraspecific distance) of the barcoding (Fig. 4a). 2) Another possible scenario is when distributions were not completely disjointed, but different (Fig. 4b). 3) In a third scenario, species that share most of the genomic sequence, such

as *Aspergillus flavus* and *A. oryzae*, and which could be reliably discriminated only by complete genome SNP analysis (Chang 2019) were not resolvable (Fig. 4c). 4) Taxa which are not considered sister species were resolved even with few genes, but a higher number of loci still positively impacted the precision of the analysis (Fig. 4d).

A comparison of the UnFATE multilocus barcode and the ITS BLASTn barcode approach was also applied (Table 1), using the same 4 species couples' scenarios

TABLE 1. Comparison between ITS BLASTn and multilocus UnFATE barcode, a tick is assigned if only the species that corresponds to the query sequence obtained the best alignment score (for ITS BLASTn barcode) or if the query species is delimited in a monophyletic clade (for UnFATE multilocus barcoding), if not a cross is assigned

	ITS BLASTn	UnFATE barcoding
<i>Aspergillus flavus</i> F13	x	✓
<i>Aspergillus terreus</i> ATCC20542	✓	✓
<i>Colletotrichum acutatum</i> C71	x	x
<i>Cercospora</i> cf. <i>sigesbeckiae</i>	x	✓
<i>Colletotrichum fructicola</i> Nara gc5	✓	✓
<i>Fusarium fujikuroi</i> IMI58289	x	✓

from above. In addition, we used 2 more species, that is, *Colletotrichum acutatum* C71 and *Fusarium fujikuroi* IMI 58289, which belong to taxa that were difficult to barcode previously (Lücking et al. 2020). ITS BLAST barcode on *Aspergillus flavus* AF31 (Supplementary Table S6) provided a large number of equally good matches, belonging to multiple species names: *A. minisclerotigenes*, *A. tritici*, *A. flavus*, *A. luchuensis*, *A. oryzae*, and *A. parasiticus*. In comparison, the UnFATE multilocus barcoding (Supplementary Fig. S14) detected the queried samples clearly as *A. flavus* by placing it in a monophyletic clade. Furthermore, *Aspergillus terreus* ATCC20542 ITS sequence blasted with high similarity only on *A. terreus* sequences (Supplementary Table S7), and the multilocus approach placed the query sample in a monophyletic *A. terreus* clade (Supplementary Fig. S15). These results are in accordance with the high interspecific distance of the species to closely related species (Fig. 4). *Colletotrichum acutatum* C71 ITS blasted with high similarity to sequences from *C. scovillei*, *C. acutatum*, and *C. fioriniae* which made an identification ambiguous (Supplementary Table S8). The UnFATE approach clustered the sample with itself (Supplementary Fig. S16). Still, the other selected sample of *C. acutatum* was instead spread within the phylogeny, confirming the difficulty in distinguishing well circumscribed species in the *C. acutatum* species complex. *Cercospora* cf. *sigesbeckiae* PP 2012 003 ITS blasted with similarity to sequences belonging to *C. kikuchii*, *C. cf. sigesbeckiae* and *C. cf. flagellaris* which complicates an identification with a single barcode (Supplementary Table S9). However, identification was possible in the multilocus tree which placed the queried sequence in a monophyletic clade of *C. cf. sigesbeckiae*, sister to *C. kikuchii*. The ITS sequence of *Colletotrichum fructicola* Nara gc5 (Supplementary Fig. S17) blasted correctly with high similarity to *C. fructicola* sequences, except for a *C. gloeosporioides* sequence (Supplementary Table S10). The multilocus approach correctly identified the queried sample as *C. fructicola* placing it in a clade exclusively composed of *C. fructicola* samples (Supplementary Fig. S18), except for the same (possibly misidentified) *C. gloeosporioides* sample present in the ITS-based barcoding. *Fusarium fujikuroi* IMI5828 ITS blasted with a perfect similarity score closer to many other *F.* species, such as *F. proliferatum*, *F. mangiferae*, *F. globosum*, *F. fujikuroi*, and *F. annulatum*, therefore it was not possible to barcode it correctly only using ITS (Supplementary Table S20). However, the

multilocus phylogeny correctly identified the query sample as *F. fujikuroi* (Supplementary Fig. S19).

DISCUSSION

This study focuses on developing a universal set of single-copy orthologous genes to enhance phylogenomic analyses of Pezizomycotina. The set of 195 genes developed in this study delivered reliable phylogenetic inferences, comparable in quality to inferences that used thousands of loci. The UnFATE pipeline and its integrated database provide a new bioinformatic tool, that will help accelerate the production and analysis of phylogenomic data sets from various sources, including target enrichment, whole-genome sequencing, and assemblies. In addition to phylogenetic reconstructions, a multilocus barcode approach of UnFATE provides a new tool, which can be used to assign species names in problematic taxa when single-locus barcoding has insufficient species-level resolving power. Furthermore, the universal gene set eliminates the need to sequence thousands of markers, which can be challenging when processing hundreds to thousands of samples. Sequencing genomes with low coverage and skimming for the markers of interest is also an option, but it produces unnecessary data and a higher computational burden. While UnFATE can perform genome skimming, it can be avoided for projects aiming primarily to reconstruct phylogenetic relationships. A more efficient approach is targeted capture, coupled with multiplexing using advanced indexing methods (e.g., Glenn et al. 2019), which can produce hundreds of markers from potentially thousands samples at high coverage even on a small bench-top Illumina sequencer.

The clustering method and the filtering steps applied to get the final 195 gene set, starting from the widely used BUSCO genes, allowed a reduced representation approach, identifying loci whose phylogenetic performance was evaluated by comparisons with high-quality phylogenies on a wide range of phylogenetic depth. The comparisons between UnFATE results and reference published phylogenies highlighted the effectiveness of the gene set at different evolutionary scales. The topological distances from the reference trees were similar for all the tested data sets (0.10–0.14 nRF).

Despite an overall excellent similarity of UnFATE phylogenies to previously published phylogenies, a

small number of incongruences were discovered. The Aspergillaceae phylogeny from Steenwyk et al. (2019) is mostly focused on the genera *Aspergillus* and *Penicillium*, however, 2 samples from the *Monascus* and *Xeromyces* genera in the UnFATE trees are either basal to the whole *Penicillium-Aspergillus* clade (Steenwyk et al. 2019) or basal only to *Aspergillus* (UnFATE DNA ML tree). In both positions, these taxa maintain their sister relationship. However, the results change in UnFATE phylogenies based on amino-acid alignments where these taxa are placed in accordance with the published phylogeny. However, also in the original phylogenomic inference of Steenwyk et al. (2019) these alternative placements of the 2 genera were proposed.

The Parmeliaceae phylogenies also differed in the placement of a few samples. *Relicina intertexta* is the most striking one, compared to the previously published tree by Pizarro et al. (2018) and Grewe et al. (2020). However, the position in the UnFATE concatenated phylogeny lacks support, and the taxon is placed congruently in the UnFATE ASTRAL phylogeny. The UnFATE placement of the Usneoid clade as a sister to all remaining Parmeliaceae is consistent with the tree in Grewe et al. (2020). However, this clade is sister to the Anzioid clade in Pizarro et al. (2018). *Arctoparmelia centrifuga* and *Pseudevernia furfuracea* are basal to each other but switched in the UnFATE tree compared to the tree in Pizarro et al. (2018). Both taxa are in a weakly supported sister relationship in Grewe et al. (2020).

Dothideomycetes phylogenies were consistent across different taxon sampling, data types, and methods applied. The incongruences found highlight the difficulty of reconstructing some phylogenetic relationships in Dothideomycetes. The trees show similar problems to the phylogenies in Ametrano et al. (2019) and Haridas et al. (2020), which used comparable taxon samplings. For example, *Lineolata rhizophorae*, which is basal to Aulographales and Venturiales in Ametrano et al. (2019) and instead basal to the rest of Pleosporomicetidae and sister to *Patellaria atrata* in Haridas et al. (2020), is sister to *Coniosporium* (another problematic taxon) in the UnFATE phylogeny. While some of these positions are only weakly supported in Ametrano et al. (2019) and Haridas et al. (2020), the positions of these taxa have strong support in the UnFATE DNA ML phylogeny.

The comparison of Pezizomycotina phylogenies also highlighted some differences between the UnFATE phylogeny and the published tree. For example, the placement of *Cryomyces*, which is consistent in Ametrano et al. (2019), Li et al. (2021), and Shen et al. (2020), differs in the UnFATE phylogeny when DNA data is used. Interestingly, *Cryomyces* in the UnFATE DNA concatenated ML tree is reported as sister to *Coniosporium* on long terminal branches. This placement could be caused by long-branch attraction in the presence of incorrectly modeled multiple substitutions. Amino-acid alignments do not have to deal with abundant and possibly saturated synonymous substitutions, which can occur in deep phylogenies. Therefore, this taxon is placed

consistently with the reference phylogenies when amino-acid alignments are used. Other similar examples are Venturiales, *Byssoschlamys*, and Erysiphaceae, whose placement is congruent in all of the previously mentioned phylogenies but is consistently placed by UnFATE only when amino-acid data is used.

Differences were also detected among UnFATE phylogenies, driven by data type (DNA or aa) and reconstruction method. For example, it was highlighted by previous studies (Zhang et al. 2018, 2020; Wang et al. 2021; Widhalm et al. 2021) that coalescence-based summary species trees can have a high level of node conflict even when advanced summary reconstruction methods are used. Even though we discarded gene trees that were topological outliers to decrease conflict, other methods could be applied to filter out problematic loci with low signal-to-noise ratio, and further improve species tree reconstruction and conflict level in specific lineages. Metrics such as Partitioned Coalescence Support (Gatesy et al. 2017, 2019), or methods to collapse poorly supported node in single-locus phylogenies before summarizing them into a species tree, could help decrease the problematic nodes. This, or other similar approaches, would also possibly improve the concordance between the results obtained by concatenation and coalescence-based species trees.

The final gene set is constrained by the need to limit the size of the bait set. It is therefore composed of many conserved protein-coding genes, possibly more suitable for deeper rather than shallower phylogenies. However, we achieved a satisfactory trade-off between bait set size, gene number, and phylogenetic performance of the gene set. At larger time scales, some taxa were misplaced possibly due to sequence divergence and saturation. Therefore, the UnFATE phylogenies should be interpreted with care, using the most suitable method/data type case by case.

When the universal bait set was applied to an *in vitro* experiment to capture and sequence the 195 genes from various quality herbarium specimens, both the power of the method and some critical points, that must be taken into consideration, emerged. The bait set was able to capture at least 1 sequence from nearly every gene in every sample and it showed a fairly homogeneous behavior in capturing a wide phylogenetic range of fungi, nonetheless few samples resulted in an high amount of missing data, possibly because competing with more efficiently captured sequences (even within the sample itself). Since the bait set was designed to be as universal as possible, and the samples' DNA did not come from pure cultures, often multiple contaminant copies were obtained. In more than half of these, coverage depth helped to identify the copy belonging to the target organism. When this was not possible, the markers were precautionary discarded in those samples, increasing the amount of missing data.

Some samples, often from smaller lineages (Arthoniomycetes, Lichinomycetes, Coniocybomycetes) with fewer genomic resources available, showed a low

fraction of on-target reads, though other samples from the same lineages performed consistently better. While this is expected for such a universal probe set, additional enrichment tests using high-quality DNA from pure cultures—eliminating intra-sample competition and DNA quality issues—could help better elucidate any relevant phylogenetic bias. However, a conscious approach to sample grouping and customized hybridization conditions can mitigate this issue. The test using herbarium samples highlighted the crucial role of DNA quality and purity, particularly regarding contaminant fungi, which can be captured as efficiently as—or even more efficiently than—the target organism’s DNA, in obtaining consistent results.

If abundant paralogous genes are suspected in the data, the Hybpiper `paralog_retriever` can be run on any UnFATE run to assess the gene copy number, since UnFATE is based on the same folder structure of Hybpiper2 (Johnson et al. 2016). The user can then decide if the amount of missing data generated by the strict filtering in UnFATE is worth being re-assessed and loci partially reintroduced into the workflow by a paralogy resolution pipeline such as, ParaGone (Jackson et al. 2023). ParaGone is designed to be compatible with Hybpiper2, and therefore also works with the UnFATE output, however, it is not integrated in the pipeline. At a stage, where UnFATE highlighted that multiple contaminant copies can affect analyses more than the missing data, we adopted a precautionary deletion approach. This is crucial when working with a universal bait set and specimens from unknown fungal species, as the only help in detecting the right copy among many is provided by the coverage. Possible real paralogs (genes in exactly 2 copies of similar, high coverage) should be rare among the orthologs we selected, as shown by the very low amount of missing data generated from assembled genomic resources. Ploidy and duplication events can potentially complicate the picture, although these fungi are usually haploid in their vegetative phase. Despite these limitations, the phylogenies produced by the bait set and the UnFATE pipeline were consistent with the currently accepted phylogenetic relationship of the main classes in Pezizomycotina.

In silico UnFATE runs based on genomic resources proved the concatenation as a still useful approach (Gatesy and Springer 2014; Gatesy et al. 2017), especially in the context where single loci are producing highly conflicting trees. However, as this is often the case, an advanced coalescence-based species tree method was also implemented, providing valuable insight into the data. This was especially true for the proof of concept phylogeny produced by our bait set. The phylogeny inferred by this method using nucleotide data proved to be the closest to the currently accepted relationships among the 10 major classes we selected to try our baits and phylogenetic pipeline. One major node proved to be particularly difficult to resolve, but this holds for a recent phylogenomic analysis that proposed 3 alternative topologies for the placement of Lichinomycetes

and their relationship to Lecanoromycetes and Eurotiomycetes (Díaz-Escandón et al. 2022).

Our testing effort managed to find a suitable number of genes, trying to minimize missing data during the capture process (e.g., reference sequences for bait design identified by clustering) and data analysis (e.g., mapping vs “assembly first”), while maintaining phylogenetic accuracy, discarding putative paralogous genes and contaminant copies.

Most of the protein-coding genes we selected for the universal gene set lacked the sequence variability to be applied on their own as effective barcodes. However, when used collectively in a multilocus approach, they outperformed the single-locus barcode approach based on the universal barcode for Fungi (ITS). In general, although there will be cases, that will not be solved even using this 195 genes multilocus approach, such as species described using other species concepts (e.g., morphological) without genetic evidence, the same species described by different names (anamorph/teleomorph), or (putative) species only distinguished by few Single Nucleotide Polymorphisms (SNPs) across the genome (Payne et al. 2006; Han et al. 2024), this multilocus approach improves the resolutive power of the barcoding approach, as shown by the case studies we reported. However, it is worth noting that the accuracy of the barcoding results is influenced by the quality and completeness of the reference database, which is the case for any barcoding approach. Being based on publicly available genomic resources, our database is currently way less comprehensive than the ITS database from NCBI GenBank or any curated ITS database (such as UNITE, Nilsson et al. 2019). However, the difference between species with genomic resources and with classical barcodes available will decrease in the long run. With this increasing database, UnFATE and its universal gene set can effectively be applied as a multilocus barcoding approach, to reveal the identity of unknown samples or it can be used to get both a comprehensive phylogeny and the closest related samples from the UnFATE database in the same phylogenetic tree.

CONCLUSIONS

Our baits set, the command line wrapper script and the support database provide a framework for a collaborative approach to build a robust phylogeny of Pezizomycotina at different phylogenetic levels. Phylogenomics by reduced representation could be applied routinely for systematics and integrative taxonomy and present an alternative to the classically adopted markers, or whole-genome sequencing.

This first version of UnFATE provides a fast, general workflow that can be applied by most scientists without specific bioinformatics expertise. UnFATE produces phylogenetic trees, launching a single command. The pipeline allows further customization and additional analyses by the user, as every step is described, and intermediate results are retained and organized in the

output folder. The universality of the bait set allows the building of deep phylogenies but also proved to be robust for shallower phylogenies. The high resolving power of the universal gene set was also highlighted by its capability to resolve sister and closely related species, which ITS failed to resolve. Hence, we also developed the multilocus barcoding approach, which is integrated into UnFATE. As the efficacy of this approach relies on the database completeness, we aim to keep the database updated, collecting both data sets from target enrichment experiments and mining the UnFATE genes from newly sequenced genomes periodically.

In summary, our program and bait set provide a novel approach to improve filamentous ascomycetes phylogeny and systematics based on molecular data.

SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository: <https://dx.doi.org/10.5061/dryad.tht76hf1x>

ACKNOWLEDGMENTS

Thanks to Matt Johnson for the inspiration to create this tool and the bait set for filamentous ascomycetes fungi, as he did for plants. His clearly explained methods and tools made our work much easier. Thanks to Kathleen Delany for contributing to the start of this project during her internship at the Grainger Bioinformatics Center at the Field Museum. Thanks to *Alejandrina Barcena-Peña* for preparing the herbarium samples. Thanks to Yukun Sun, *Alejandrina Barcena-Peña* and Holly Robertson for testing the software. Thanks to Yukun Sun for technical help and the fruitful discussions during my stay at the Field Museum.

AUTHOR CONTRIBUTIONS

FG and TL conceived the ideas and designed the methodology; CGA and FG developed the pipeline; CGA and JJ wrote the code; CGA and JJ collected and analyzed the data; CGA, FG, and TL led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest.

DATA AVAILABILITY

UnFATE tool and database available at CGA GitHub page: <https://github.com/claudioametrano/UnFATE>.

Scripts for multilocus barcoding test available at JJ GitHub page: https://github.com/jacobnjensen1/UnFATE_paper_scripts. Data generated in this study have been deposited in the NCBI Sequence Read Archive (SRA) under the BioProject ID PRJNA1216762. Data files (bait sequences, alignments, and trees) can be found in the Dryad data repository: <https://doi.org/10.5061/dryad.tht76hf1x>. All genomic data used to benchmark the pipeline are available at <https://www.ncbi.nlm.nih.gov/assembly>, accessions of the assemblies we used are reported in [Supplementary Information](#). Target enrichment data from [Grewe et al. \(2020\)](#), [Widhelm et al. \(2019\)](#) and [Widhelm et al. \(2021\)](#) are available at <https://www.ncbi.nlm.nih.gov/sra>.

REFERENCES

- Allio R., Schomaker-Bastos A., Romiguier J., Prosdociimi F., Nabholz B., Delsuc F. 2020. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol. Ecol. Resour.* 20:892–905.
- Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.
- Ametrano C.G., Grewe F., Crous P.W., Goodwin S.B., Liang C., Selbmann L., Lumbsch H.T., Leavitt S.D., Muggia L. 2019. Genome-scale data resolve ancestral rock-inhabiting lifestyle in Dothideomycetes (Ascomycota). *IMA Fungus* 10:1–12.
- Ametrano C.G., Lumbsch H.T., Di Stefano I., Sangvichien E., Muggia L., Grewe F. 2022. Should we hail the Red King? Evolutionary consequences of a mutualistic lifestyle in genomes of lichenized ascomycetes. *Ecol. Evol.* 12:e8471.
- Andermann T., Cano A., Zizka A., Bacon C., Antonelli A. 2018. SECAPR—a bioinformatics pipeline for the rapid and user-friendly processing of targeted enriched Illumina sequences, from raw reads to alignments. *PeerJ* 6:e5175.
- Baker W.J., Dodsworth S., Forest F., Graham S.W., Johnson M.G., McDonnell A.J., Pokorny L., Tate J., Wicke S., Wickett N.J. 2021. Exploring Angiosperms353: An open, community toolkit for collaborative phylogenomic research on flowering plants. *Am. J. Bot.* 08:1059–1065.
- Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Pribelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N., Tesler G., Alekseyev M.A., Pevzner P.A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19:455–477.
- Beimforde C., Feldberg K., Nylander S., Rikkinen J., Tuovila H., Dörfelt H., Gube M., Jackson D.J., Reitner J., Seyfullah L.J., Schmidt A. 2014. Estimating the Phanerozoic history of the Ascomycota lineages: combining fossil and molecular data. *Mol. Phylogenet. Evol.* 78:386–398.
- Bengtsson-Palme J., Ryberg M., Hartmann M., Branco S., Wang Z., Godhe A., De Wit P., Sánchez-García M., Ebersberger I., de Sousa F., Amend A., Jumpponen A., Unterseher M., Kristiansson E., Abarenkov K., Bertrand Y.J.K., Sanli K., Eriksson K.M., Vik U., Veldre V., Nilsson R.H. 2013. Improved software detection and extraction of ITS1 and ITS 2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods Ecol. Evol.* 4:914–919.
- Blaxter M.L. 2004. The promise of a DNA taxonomy. *Philos. Trans. R. Soc. London, Ser. B* 359(1444):669–679.
- Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.
- Breinolt J.W., Carey S.B., Tiley G.P., Davis E.C., Endara L., McDaniel S.F., Neves L.G., Sessa E.B., von Konrat M., Chantanaorrapint S., Fawcett S., Ickert-Bond S.M., Labiak P.H., Larrain J., Lehnert M., Lewis L.R., Nagalingum N.S., Patel N., Rensing S.A., Testo W.,

- Vasco A., Villarreal J.C., Williams E.W., Burleigh J.G. 2021. A target enrichment probe set for resolving the flagellate land plant tree of life. *Appl. Plant Sci.* 9:e11406.
- Brewer G.E., Clarkson J.J., Maurin O., Zuntini A.R., Barber V., Bellot S., Biggs N., Cowan R.S., Davies N.M., Dodsworth S., Cowan R.S., Davies N.M.J., Dodsworth S., Edwards S.L., Eiserhardt W.L., Epitawalage N., Sue F., Grall A., Kersey P.J., Pokorný L., Leitch I.J., Forest F., Baker W.J. 2019. Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. *Front. Plant Sci.* 10:1102.
- Campana M.G. 2018. BaitsTools: software for hybridization capture bait design. *Mol. Ecol. Resour.* 18:356–361.
- Capella-Gutiérrez S., Silla-Martínez J.M., Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17:540–552.
- Chafin T.K., Douglas M.R., Douglas M.E. 2018. MrBait: universal identification and design of targeted-enrichment capture probes. *Bioinformatics* 34:4293–4296.
- Chang P.K. 2019. Genome-wide nucleotide variation distinguishes *Aspergillus flavus* from *Aspergillus oryzae* and helps to reveal origins of atoxigenic *A. flavus* biocontrol strains. *J. Appl. Microbiol.* 127:1511–1520.
- Davey J.W., Blaxter M.L. 2010. RADSeq: next-generation population genetics. *Brief. Funct. Genom.* 9:416–423.
- Di Franco A., Poujol R., Baurain D., Philippe H. 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol. Biol.* 19:1–17.
- Díaz-Escandón D., Tagirdzhanova G., Vanderpool D., Allen C.C., Aptroot A., Češka O., Hawksworth D.H., Huereca A., Knudsen K., Kocourková J., Lücking R., Resl P., Spribille T. 2022. Genome-level analyses resolve an ancient lineage of symbiotic ascomycetes. *Curr. Biol.* 32(23):5209–5218.
- Eddy S.R. 2009. A new generation of homology search tools based on probabilistic inference. In: (Ed.), *Genome Informatics 2009: Genome Informatics Series Vol. 23*, World Scientific.
- Eisen J.A. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8(3):163–167.
- Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* 61:717–726.
- Finn R.D., Clements J., Eddy S.R. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39:W29–W37.
- Forrest L.L., Hart M.L., Hughes M., Wilson H.P., Chung K.F., Tseng Y.H., Kidner C.A. 2019. The limits of Hyb-Seq for herbarium specimens: impact of preservation techniques. *Front. Ecol. Evol.* 7:439.
- Gatesy J., Meredith R.W., Janečka J.E., Simmons M.P., Murphy W.J., Springer M.S. 2017. Resolution of a concatenation/coalescence kerfuffle: partitioned coalescence support and a robust family-level tree for Mammalia. *Cladistics* 33:295–332.
- Gatesy J., Sloan D.B., Warren J.M., Baker R.H., Simmons M.P., Springer M.S. 2019. Partitioned coalescence support reveals biases in species-tree methods and detects gene trees that determine phylogenomic conflicts. *Mol. Phylogenet. Evol.* 139:106539.
- Gatesy J., Springer M.S. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalence conundrum. *Mol. Phylogenet. Evol.* 80:231–266.
- Glenn T.C., Nilsen R.A., Kieran T.J., Sanders J.G., Bayona-Vásquez N.J., Finger J.W., Pierson T.W., Bentley K.E., Hoffberg S.L., Louha S., García-De León F.J., del Río Portilla M.A., Reed K.D., Anderson J.L., Meece J.K., Aggrey S.E., Rekaya R., Alabady M., Belanger M., Winker K., Faircloth B.C. 2019. Adapterama I: universal stubs and primers for 384 unique dual-indexed or 147,456 combinatorially-indexed Illumina libraries (iTru & iNext). *PeerJ* 7:e7755.
- Good J.M. 2012. Reduced representation methods for subgenomic enrichment and next-generation sequencing. *Mol. Methods Evol. Genet.* 85–103.
- Grewe F., Ametrano C., Widhalm T.J., Leavitt S., Distefano I., Polyam W., Pizarro D., Wedin M., Crespo A., Divakar P.K., Lumbsch H.T. 2020. Using target enrichment sequencing to study the higher-level phylogeny of the largest lichen-forming fungi family: Parmeliaceae (Ascomycota). *IMA Fungus* 11:1–11.
- Grewe F., Huang J.P., Leavitt S.D., Lumbsch H.T. 2017. Reference-based RADseq resolves robust relationships among closely related species of lichen-forming fungi using metagenomic DNA. *Sci. Rep.* 7:1–11.
- Grewe F., Lagostina E., Wu H., Printzen C., Lumbsch H.T. 2018. Population genomic analyses of RAD sequences resolves the phylogenetic relationship of the lichen-forming fungal species *Usnea antarctica* and *Usnea aurantiacoatra*. *MycKeys* 43:91–113.
- Hale H., Gardner E.M., Viruel J., Pokorný L., Johnson M.G. 2020. Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants. *Appl. Plant Sci.* 8(4):e11337.
- Han D.M., Baek J.H., Choi D.G., Jeon M.S., Eyun S.I., Jeon C.O. 2024. Comparative pangenome analysis of *Aspergillus flavus* and *Aspergillus oryzae* reveals their phylogenetic, genomic, and metabolic homogeneity. *Food Microbiol.* 119:104435.
- Haridas S., Albert R., Binder M., Bloem J., LaButti K., Salamov A., Andreopoulos B., Baker S., Barry K., Bills G., Bluhm B.H., Cannon C., Castanera R., Culley D.E., Daum C., Ezra D., González J.B., Henrissat B., Kuo A., Liang C., Lipzen A., Lutzoni F., Magnuson J., Mondo S.J., Nolan M., Ohm R.A., Pangilinan J., Park H.-J., Ramírez L., Alfaro M., Sun H., Tritt A., Yoshinaga Y., Zwiers L.-H., Turgeon B.G., Goodwin S.B., Spatafora J.W., Crous P.W., Grigoriev I.V. 2020. 101 Dothideomycetes genomes: a test case for predicting lifestyles and emergence of pathogens. *Stud. Mycol.* 96:141–153.
- Hawksworth D.L., Lücking R. 2017. Fungal diversity revisited: 2.2 to 3.8 million species. *Microbiol. Spectr.* 5:5–4.
- Hill R., Leitch I.J., Gaya E. 2021. Targeting Ascomycota genomes: what and how big? *Fungal Biol. Rev.* 36:52–59.
- Huerta-Cepas J., Serra F., Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* 33:1635–1638.
- Jackson C., McLay T., Schmidt-Lebuhn A.N. 2023. hybpiper-nf and paragone-nf: containerization and additional options for target capture assembly and paralog resolution. *Appl. Plant Sci.* 11(4):e11532.
- Jain M., Olsen H.E., Paten B., Akeson M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17:1–11.
- Johnson M.G., Gardner E.M., Liu Y., Medina R., Goffinet B., Shaw A.J., Zerega N.J., Wickett N.J. 2016. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Plant Sci.* 4:1600016.
- Johnson M.G., Pokorný L., Dodsworth S., Botigue L.R., Cowan R.S., Devault A., Eiserhardt W.L., Epitawalage N., Forest F., Kim J.T., Leebens-Mack J.H., Leitch I.J., Maurin O., Soltis D.E., Soltis P.S., Wong G.K., Baker W.L., Wickett N.J. 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Syst. Biol.* 68:594–606.
- Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780.
- Kirk P., Cannon P., Minter D., Stalpers J. 2008. *Ainsworth and Bisby's Dictionary of the Fungi*. 10th ed. Wallingford, UK: CAB International. p. 771.
- Kozlov A.M., Darriba D., Flouri T., Morel B., Stamatakis A. 2019. RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35:4453–4455.
- Kück P., Longo G.C. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Front. Zool.* 11:1–8.
- Li Y., Steenwyk J.L., Chang Y., Wang Y., James T.Y., Stajich J.E., Spatafora J.W., Groenewald M., Dunn C.W., Hittinger C.T., Sheng XX, Rokas A. 2021. A genome-scale phylogeny of the kingdom Fungi. *Curr. Biol.* 31:1653–1665.

- Liu J., Jiang J., Song S., Tornabene L., Chabarría R., Naylor G.J., Li C. 2017. Multilocus DNA barcoding—species identification with multilocus data. *Sci. Rep.* 7:1–12.
- Liu Y., Johnson M.G., Cox C.J., Medina R., Devos N., Vanderpoorten A., Hedenäs L., Bell N.E., Shevock J.R., Agüero B., Quandt D., Wickett N.J., Shaw A.J., Goffinet B. 2019. Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear genomes. *Nat. Commun.* 10:1–11.
- Lücking R., Aime M.C., Robbertse B., Miller A.N., Ariyawansa H.A., Aoki T., Cardinali G., Crous P.W., Druzhinina I.S., Geiser D.M., Hawksworth D.L., Hyde K.D., Irinyi L., Jeewon R., Johnston P.R., Kirk P.M., Malosso E., May T.W., Meyer W., Öpik M., Robert V., Stadler M., Thines M., Vu D., Yurkov A.M., Zhang N., Schoch C.L. 2020. Unambiguous identification of fungi: where do we stand and how accurate and precise is fungal DNA barcoding? *IMA Fungus* 11:1–32.
- Mallo D., Posada D. 2016. Multilocus inference of species trees and DNA barcoding. *Philos. Trans. R. Soc. London, Ser. B* 371:20150335.
- Mamanova L., Coffey A.J., Scott C.E., Kozarewa I., Turner E.H., Kumar A., Howard E., Shendure J., Turner D.J. 2010. Target-enrichment strategies for next-generation sequencing. *Nat. Methods* 7:111–118.
- Maurin O., Anest A., Bellot S., Biffin E., Brewer G., Charles-Dominique T., Cowan R.S., Dodsworth S., Epitawalage N., Gallego B., Giarretta A., Goldenberg R., Gonçalves D.J.P., Graham S., Hoch P., Mazine F., Low Y.W., McGinnie C., Michelangeli F.A., Morris S., Penneys D.S., Pérez Escobar O.A., Pillon Y., Pokorny L., Shimizu G., Staggeimeier V.G., Thornhill A.H., Tomlinson K.W., Turner I.M., Vasconcelos T., Wilson P.G., Zuntini A.R., Baker W.J., Forest F., Lucas E. 2021. A nuclear phylogenomic study of the angiosperm order Myrtales, exploring the potential and limitations of the universal Angiosperms353 probe set. *Am. J. Bot.* 108:1087–1111.
- Mayer C., Sann M., Donath A., Meixner M., Podsiadlowski L., Peters R.S., Petersen M., Meusemann K., Liere K., Wägele J.W., Misof B., Bleidorn C., Ohl M., Niehuis O. 2016. BaitFisher: a software package for multispecies target DNA enrichment probe design. *Mol. Biol. Evol.* 33:1875–1886.
- McDonnell A.J., Baker W.J., Dodsworth S., Forest F., Graham S.W., Johnson M.G., Pokorny L., Tate J., Wicke S., Wickett N.J. 2021. Exploring Angiosperms353: developing and applying a universal toolkit for flowering plant phylogenomics. *Appl. Plant Sci.* 9:10–1002.
- Mertes F., ElSharawy A., Sauer S., van Helvoort J.M., Van Der Zaag P., Franke A., Nilsson M., Lehigh H., Brookes A.J. 2011. Targeted enrichment of genomic DNA regions for next-generation sequencing. *Brief. Funct. Genom.* 10:374–386.
- Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., Von Haeseler A., Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37:1530–1534.
- Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30:i541–i548.
- Nelsen M.P., Lücking R., Boyce C.K., Lumbsch H.T., Ree R.H. 2020. No support for the emergence of lichens prior to the evolution of vascular plants. *Geobiology* 18:3–13.
- Nilsson R.H., Larsson K.H., Taylor A.F.S., Bengtsson-Palme J., Jeppesen T.S., Schigel D., Kennedy P., Picard K., Glockner F.O., Tedersoo L., Saar I., Koljalg U., Abarenkov K. 2019. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res.* 47(D1):D259–D264.
- Nurk S., Meleshko D., Korobeynikov A., Pevzner P.A. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27:824–834.
- Payne G.A., Nierman W.C., Wortman J.R., Pritchard B.L., Brown D., Dean R.A., Bhatnagar D., Cleveland T.E., Machida M., Yu J. 2006. Whole genome comparison of *Aspergillus flavus* and *A. oryzae*. *Med. Mycol.* 44(sup1):9–11.
- Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau A. 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12:2825–2830.
- Philippe H., Delsuc F., Brinkmann H., Lartillot N. 2005. Phylogenomics. *Annu. Rev. Ecol. Syst.* 36(1):541–562.
- Pino-Bodas R.M.P., Martín A.R., Burgaz, Lumbsch H.T. 2013. Species delimitation in *Cladonia* (Ascomycota): a challenge for the DNA barcoding philosophy. *Mol. Ecol. Res.* 13:1058–1068.
- Pizarro D., Divakar P.K., Grewe F., Leavitt S.D., Huang J.P., Dal Grande F., Schmitt I., Wedin M., Crespo A., Lumbsch H.T. 2018. Phylogenomic analysis of 2556 single-copy protein-coding genes resolves most evolutionary relationships for the major clades in the most diverse group of lichen-forming fungi. *Fungal Divers.* 92:31–41.
- Prieto M., Wedin M. 2013. Dating the diversification of the major lineages of Ascomycota (Fungi). *PLoS One* 8:e65576.
- Ranwez V., Douzery E.J., Cambon C., Chantret N., Delsuc F. 2018. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.* 35:2582–2584.
- Resl P., Bujold A.R., Tagirdzhanova G., Meidl P., Freire Rallo S., Kono M., Fernández-Brime S., Guðmundsson H., Andrén O.S., Muggia L., Mayrhofer H., McCutcheon J.P., Wedin M., Werth S., Willis L.M., Spribille T. 2022. Large differences in carbohydrate degradation and transport potential among lichen fungal symbionts. *Nat. Commun.* 13(1):2634.
- Rhoads A., Au K.F. 2015. PacBio sequencing and its applications. *Genom. Proteomics Bioinform.* 13:278–289.
- Roe A.D., Rice A.V., Bromilow S.E., Cooke J.E., Sperling F.A. 2010. Multilocus species identification and fungal DNA barcoding: insights from blue stain fungal symbionts of the mountain pine beetle. *Mol. Ecol. Resour.* 10:946–959.
- Ronaghi M. 2001. Pyrosequencing sheds light on DNA sequencing. *Genome Res.* 11(1):3–11.
- Saari T.W., Schroeder A.L., Ankley G.T., Villeneuve D.L. 2017. First-generation annotations for the fathead minnow (*Pimephales promelas*) genome. *Environ. Toxicol. Chem.* 36:3436–3442.
- Salas-Lizana R., Oono R. 2018. Double-digest RAD seq loci using standard Illumina indexes improve deep and shallow phylogenetic resolution of *Lophodermium*, a widespread fungal endophyte of pine needles. *Ecol. Evol.* 8:6638–6651.
- Schoch C.L., Sung G.H., López-Giráldez F., Townsend J.P., Miadlikowska J., Hofstetter V., Robbertse B., Matheny P.B., Kauff F., Wang Z., Spatafora J.W., Andrie R.M., Trippe K., Ciuffetti L.M., Wynns A., Fraker E., Hodkinson B.P., Bonito G., Groenewald J.Z., Arzanlou M., de Hoog G.S., Crous P.W., Hewitt D., Pfister D.H., Peterson K., Gryzenhout M., Wingfield M.J., Aptroot A., Suh S.-O., Blackwell M., Hillis D.M., Griffith G.W., Castlebury L.A., Rossman A.Y., Lumbsch H.T., Lücking R., Büdel B., Rauhut A., Diederich P., Ertz D., Geiser D.M., Hosaka K., Inderbitzin P., Köhlmeier J., Volkman-Kohlmeier B., Mostert L., O'Donnell K., Sipman H., Rogers J.D., Shoemaker R.A., Sugiyama J., Summerbell R.C., Untereiner W., Johnston P.R., Stenroos S., Zuccaro A., Dyer P.S., Crittenden P.D., Cole M.S., Hansen K., Trappe J.M., Yahr R., Lutzoni F., Spatafora J.W. 2009. The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Syst. Biol.* 58:224–239.
- Schott R.K., Panesar B., Card D.C., Preston M., Castoe T.A., Chang B.S. 2017. Targeted capture of complete coding regions across divergent species. *Genom. Biol. Evol.* 9:398–414.
- Shen X.X., Steenwyk J.L., LaBella A.L., Opulente D.A., Zhou X., Kominek J., Li Y., Groenewald M., Hittinger C.T., Rokas A. 2020. Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum Ascomycota. *Sci. Adv.* 6:eabd0079.
- Simão F.A., Waterhouse R.M., Ioannidis P., Kriventseva E.V., Zdobnov E.M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212.
- Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic data sets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* 15:1–15.

- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.
- Stanke M., Keller O., Gunduz I., Hayes A., Waack S., Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34:W435–W439.
- Steenwyk J.L., Shen X.X., Lind A.L., Goldman G.H., Rokas A. 2019. A robust phylogenomic time tree for biotechnologically and medically important fungi in the genera *Aspergillus* and *Penicillium*. *MBio* 10:e00925–e00919.
- Thandapani P., O'Connor T.R., Bailey T.L., Richard S. 2013. Defining the RGG/RG motif. *Mol. Cell* 50:613–623.
- Tice A.K., Žihala D., Pánek T., Jones R.E., Salomaki E.D., Nenarokov S., Burki F., Eliáš M., Eme L., Roger A.J., Rokas A., Shen X.-X., Strasser J.F.H., Kolísko M., Brown M.W. 2021. PhyloFisher: a phylogenomic package for resolving eukaryotic relationships. *PLoS Biol.* 19:e3001365.
- Venter J.C., Adams M.D., Sutton G.G., Kerlavage A.R., Smith H.O., Hunkapiller M. 1998. Shotgun sequencing of the human genome. *Science* 280(5369):1540–1542.
- Wang H.X., Morales-Briones D.F., Moore M.J., Wen J., Wang H.F. 2021. A phylogenomic perspective on gene tree conflict and character evolution in Caprifoliaceae using target enrichment data, with Zabelioideae recognized as a new subfamily. *J. Syst. Evol.* 59(5):897–914.
- Waterhouse R.M., Tegenfeldt F., Li J., Zdobnov E.M., Kriventseva E.V. 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.* 41(D1):D358–D365.
- Weitemier K., Straub S.C., Cronn R.C., Fishbein M., Schmickl R., McDonnell A., Liston A. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Appl. Plant Sci.* 2:1400042.
- Widhalm T.J., Grewe F., Goffinet B., Wedin M., Goward T., Coca L.F., Distefano I., Košuthová A., Lumbsch H.T. 2021. Phylogenomic reconstruction addressing the Peltigeralean backbone (Lecanoromycetes, Ascomycota). *Fungal Divers.* 110:59–73.
- Widhalm T.J., Grewe F., Huang J.P., Mercado-Díaz J.A., Goffinet B., Lücking R., Moncada B., Mason-Gamer R., Lumbsch H.T. 2019. Multiple historical processes obscure phylogenetic relationships in a taxonomically difficult group (Lobariaceae, Ascomycota). *Sci. Rep.* 9(1):8968.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24:1586–1591.
- Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinf.* 19:15–30.
- Zhang C., Scornavacca C., Molloy E.K., Mirarab S. 2020. ASTRAL-Pro: quartet-based species-tree inference despite paralogy. *Mol. Biol. Evol.* 37:3292–3307.