OXFORD

Phylogenetics

T-BAS: Tree-Based Alignment Selector toolkit for phylogenetic-based placement, alignment downloads and metadata visualization: an example with the Pezizomycotina tree of life

Ignazio Carbone[1,*], James B. White[1], Jolanta Miadlikowska[2],
A. Elizabeth Arnold[3,4], Mark A. Miller[5], Frank Kauff[6], Jana M. U'Ren[3],
Georgiana May[7] and François Lutzoni[2]

[1]Center for Integrated Fungal Research, Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC 27695 USA, [2]Department of Biology, Duke University, Durham, NC 27708, USA, [3]School of Plant Sciences, The University of Arizona, Tucson, AZ 85721, USA, [4]Department of Ecology and Evolutionary Biology, The University of Arizona, Tucson, AZ 85721, USA, [5]San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA, [6]Department of Medicine, University of Giessen, Giessen, Germany and [7]Department of Ecology, Evolution and Behavior, University of Minnesota, Saint Paul, MN 55108, USA

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Abstract

Motivation: High-quality phylogenetic placement of sequence data has the potential to greatly accelerate studies of the diversity, systematics, ecology and functional biology of diverse groups. We developed the Tree-Based Alignment Selector (T-BAS) toolkit to allow evolutionary placement and visualization of diverse DNA sequences representing unknown taxa within a robust phylogenetic context, and to permit the downloading of highly curated, single- and multi-locus alignments for specific clades.

Results: In its initial form, T-BAS v1.0 uses a core phylogeny of 979 taxa (including 23 outgroup taxa, as well as 61 orders, 175 families and 496 genera) representing all 13 classes of largest subphylum of Fungi—Pezizomycotina (Ascomycota)—based on sequence alignments for six loci (nr5.8S, nrLSU, nrSSU, mtSSU, *RPB1, RPB2*). T-BAS v1.0 has three main uses: (i) Users may download alignments and voucher tables for members of the Pezizomycotina directly from the reference tree, facilitating systematics studies of focal clades. (ii) Users may upload sequence files with reads representing unknown taxa and place these on the phylogeny using either BLAST or phylogeny-based approaches, and then use the displayed tree to select reference taxa to include when downloading alignments. The placement of unknowns can be performed for large numbers of Sanger sequences obtained from fungal cultures and for alignable, short reads of environmental amplicons. (iii) User-customizable metadata can be visualized on the tree.

Availability and Implementation: T-BAS Version 1.0 is available online at http://tbas.hpc.ncsu.edu. Registration is required to access the CIPRES Science Gateway and NSF XSEDE's large computational resources.

Contact: icarbon@ncsu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The capacity to generate sequence data has far surpassed the capability of biologists to describe new species, especially in hyperdiverse lineages. Such challenges are particularly acute in the Fungi: although rich in a tradition of morphological systematics and among the first to adopt molecular tools for phylogenetic studies (White et al., 1990), mycologists have described to date only ca. 100,000 (Kirk *et al.*, 2008) of the estimated 1.5–5 million fungal species (Blackwell, 2011; Hawksworth, 1991, 2001; O'Brien *et al.*, 2005). In mycology, as in other disciplines in the life sciences, the number of taxonomists with expertise needed to describe new species has not increased with the expanding understanding of fungal microbiomes generated by new research initiatives and sequencing technologies, and has led to a severe information bottleneck even as enthusiasm for understanding the scope of fungal biodiversity has grown (Crous *et al.*, 2015; Hawksworth, 2001).

Increasingly, fungal biologists are generating datasets consisting of sequence data from axenic cultures or environmental samples (e.g. Barberán *et al.*, 2015; Kemler *et al.*, 2013; Schmidt *et al.*, 2013; Tedersoo *et al.*, 2014; U'Ren *et al.*, 2014; Zimmerman and Vitousek, 2012). Such efforts often recover many 'unknowns' and suggest new species. However, integration of these datasets into traditional taxonomic frameworks has lagged due to the absence of powerful tools for reliable phylogenetic classification. In parallel, systematists may lack the bioinformatic tools necessary to integrate and keep track of the deluge of new sequence datasets, which range from single-locus (e.g. nuclear ribosomal loci) to multi-locus datasets, genomes, transcriptomes and environmental metagenomic datasets.

Accordingly, there is need for bioinformatic tools that centralize diverse streams of sequence data and metadata, and that allow integration of this information within a robust phylogenetic framework. To this end we have developed the Tree-based Alignment Selector (T-BAS) toolkit, a publically available, online toolkit designed to link sequence data with established phylogenetic information. The toolkit provides users with phylogenetic placement of unknown sequences (Berger *et al.*, 2011; Berger and Stamatakis, 2011; Matsen *et al.*, 2010) at the clade level without the uncertainty inherent to using BLAST (Altschul *et al.*, 1990) approaches alone.

The core concept of T-BAS is that a user can submit sequence data for integration within reference alignments and a robust reference phylogeny, and download carefully curated, single- or multi-gene alignments for close relatives. Such alignments can then be used for phylogenetic analyses with increased precision and for species identification, either immediately or via pointing out important loci and taxon sampling needed to diagnose novelty and facilitate new species descriptions. Thus T-BAS is structured to allow taxonomists to uncover novelty and describe new taxa in clades in which they have expertise, and to obtain the necessary voucher- and metadata needed for species descriptions. In turn, by providing a single online portal that provides voucher tables, metadata visualization, designation of operational taxonomic units (OTU), curated alignments, and a curated, updating tree, T-BAS can facilitate the work and collaborations of systematists and ecologists, enhancing new species descriptions, perspectives on the evolution of ecological modes, clade-level placement for unknown biodiversity, data archiving, and insights on the robustness of particular loci for phylogenetic reconstructions of specific clades. In these contexts the T-BAS toolkit is designed to be flexible, providing users with customizable visualization tools and customizable, downloadable metadata, and voucher tables. By linking ecological metadata with phylogenetic placements, T-BAS generates metadata-enriched phylogenies that can be shared online to enhance phylogenetic analyses of known groups with novel taxa and their traits.

The current multi-locus tree of reference species in T-BAS (v1.0) includes 867 named species of the most species-rich subphylum of Fungi, the Pezizomycotina (Ascomycota). It represents all 13 classes of the Pezizomycotina, including taxa in 12 subclasses, 61 orders, 175 families and 496 genera. It is based on hand-curated alignments for six genes (nuclear ribosomal [nr] 5.8S, large subunit [nrLSU], small subunit [nrSSU], mitochondrial ribosomal small subunit [mtSSU], RNA polymerase II largest subunit gene [*RPB*1], and RNA polymerase II second largest subunit gene [*RPB*2]) (Fig. 1, Supplementary Fig. S1, Supplementary Table S1, also visible online with T-BAS at http://tbas.hpc.ncsu.edu/). Thus in its first release, T-BAS is designed specifically for alignment selection and download, phylogenetic placement, and metadata visualization in the context of the Pezizomycotina, but the architecture is readily scalable to other clades.

Overall, T-BAS provides tools designed to open the information bottleneck born from the time lag between the discovery of new biodiversity suggested by sequence data and the description of new species, by integrating unknown sequences into curated alignments and phylogenies, which can be used to download comprehensive alignments for these unknown taxa and their closest relatives. The long-term goal of the T-BAS toolkit is to offer a central phylogenetic information hub for systematists to track and obtain the necessary data and specimens or strains to describe new species, and to connect ecologists to the rapidly advancing frontier of eco-phylogenetics.

2 Methods

2.1 Functions of T-BAS v1.0

T-BAS v1.0 has two primary functions, both of which are available to users via the public web portal. The first function ('Option 1') is the web-based, interactive visualization of the multi-locus phylogeny for Pezizomycotina (Fig. 2). This option enables the user to select clades and individual species of interest by clicking directly on the tree. After selecting clades and/or species, the user may download curated alignments of sequence data for the selected taxa. Alignments may include up to six loci, reflecting the data on which the multi-locus tree is based. A dashboard window displays the number of sequences available for each locus, and the number of loci available for each species can be displayed on the tree. These interactive features allow the user to control the amount of missing data in the resulting supermatrix (Gaya *et al.*, 2012; Miadlikowska *et al.*, 2006, 2014). Unaligned nrITS1 and nrITS2 sequences also can be downloaded for each terminal species or clade and queried by BLAST, providing an opportunity for users to expand taxon sampling to include those in T-BAS. The broad phylogenetic context enabled in T-BAS is useful for making well-informed selections of outgroup taxa and conducting hypothesis testing in a phylogenetic framework.

Because alignments in the T-BAS database span the entire Pezizomycotina, many regions of the nrLSU are ambiguously aligned (Lutzoni *et al.*, 2000) and are delimited as such. Reference alignments were hand-curated and secondary structure information for rRNA sequences was used to improve global alignment. Separate or concatenated alignments can be downloaded as PHYLIP, FASTA, XML, CSV and NEXUS files. Downloaded as a NEXUS file, alignments might be visualized in Mesquite, for example (Maddison and
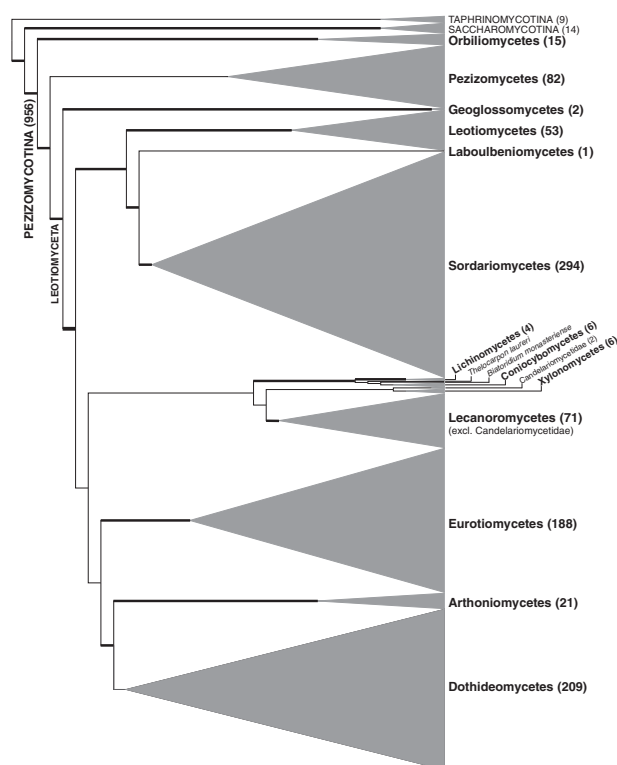
Fig. 1. Schematic summary of a six-locus (nr5.8S, nrLSU, nrSSU, mtSSU, RPB1, RPB2) phylogenetic tree of 979 taxa representing all 13 classes of the Pezizomycotina (12 subclasses, 61 orders, 175 families and 496 genera). See Supplementary Figure S1 for the full version of this tree and methods used for its inference. The Pezizomycotina tree is displayed online by T-BAS and can be used to phylogenetically place unknowns and/or download alignments for the six loci currently included in T-BAS. Numbers in parentheses represent the total number of species examined for each fungal class or subphylum. Thick branches represent highly supported internodes (see Supplementary Fig. S1)

Maddison, 2011; http://mesquiteproject.org), and the alignments of specific regions can be optimized. The narrower the clade of interest (i.e. groups of taxa with low levels of divergence) within the Pezizomycotina, the more likely regions of the alignment will be unambiguously alignable even though these same regions may be ambiguously aligned globally. Thus, through Option 1 users can obtain an updated view of curated sequence data for particular groups, providing a tool for primer design and a core reference alignment for subsequent studies at user-selected levels of taxonomic diversity.

T-BAS offers a second main function ('Option 2'), which contains two analysis pipelines and addresses issues of unknown taxa. Users first upload a FASTA formatted file with unknown query sequences and, if desired, a comma-separated file with metadata. Then, the user may BLAST the nrITS, nrLSU or the entire nrITS-LSU (or any other loci used to construct the Pezizomycotina tree for T-BAS), and visualize the BLAST result in the form of their placements on the T-BAS tree. Alternatively, the user can use a phylogeny-based placement method such as EPA (Berger et al., 2011; Berger and Stamatakis, 2011) with visualization of the results on the T-BAS tree. Both the BLAST and phylogeny-based pipelines provide the user with an output file containing class-level placement and closest reference taxa. The resulting tree has all of the alignment-download functions described above, as well as the ability to visualize metadata in a phylogenetic context.

The rationale for implementing the phylogeny-based pipeline reflects the challenges associated with estimating clade-level placement for unknown Fungi. Although systematists often have a hypothesis for the phylogenetic relationships of taxa under consideration, results from many studies often generate a large number of unknowns. These include (1) specimens or strains for which the taxonomic affiliation is uncertain even though the nrITS and/or nrLSU has been sequenced and a BLAST analysis is performed (e.g. Chen et al., 2015; Gazis et al., 2012; U'Ren et al., 2016) and (2) the high number of unknown species or OTU that are often encountered with surveys of fungal communities (e.g. Massimo et al., 2015). Although a frequently used and important resource, BLAST comparisons with GenBank can provide erroneous taxonomic identification because of limited taxonomic availability in the database, misidentified species, or limited robustness due to a non-evolutionary inference. More broadly, analyses based on OTU can be more insightful when placed within the hierarchical structure provided by phylogenetic analyses.

In turn, phylogeny-based placement methods (Berger et al., 2011; Berger and Stamatakis, 2011; Matsen et al., 2010) are useful when unknown sequences are similar to reference taxa. However, their performance is unreliable when query taxa are distantly related to reference taxa (Berger and Stamatakis, 2011). To deal with varying levels of DNA sequence divergence, the T-BAS approach encompasses a two-step process: first, a BLAST search (using curated sequences as reference) is performed to assign unknown sequences to clades. Evolutionary placement can then be performed within each clade for finer taxonomic resolution. Although other methods to carry out these analyses are available, BLAST searches typically are not coupled with a phylogeny in other approaches, and evolutionary placement often is performed on a global (i.e. phylogeny) rather than on a local (i.e. clade) level (Stamatakis, 2014).

T-BAS gives the user the option to map BLAST placements on the tree and allows users to download FASTA formatted files containing all BLAST sequences that fall in a particular clade. In parallel the user also has the option to perform phylogeny-based placement of the targeted set of BLAST sequences in T-BAS using the more accurate pairwise-alignment and iterative-refinement processes implemented in MAFFT. By using both methods to estimate potential affiliations, T-BAS will help users select the right array of taxa for their alignments to download for more detailed phylogenetic study on a specific group. A final outcome of future versions of T-BAS will be the opportunity for biologists to request focal strains diagnosed as novel, in part through tapping existing culture collections for systematics, functional analyses, or ecological studies (e.g. endophytic and endolichenic fungi at the University of Arizona's Robert L. Gilbertson Mycological Herbarium, and other established collections worldwide).

By using phylogeny-based placement methods, alignable sequences (e.g. nr5.8S and/or nrLSU sequences) uploaded to T-BAS by the user can be aligned to reference taxa and can be included in the alignment downloaded by the user. Therefore, T-BAS v1.0 can also serve as a tool to align large numbers of nr5.8S and nrLSU sequences to reference alignments and further, to extend this function to perform a phylogenetic placement using sequences for any of the six reference loci.

T-BAS v1.0 integrates the designation of operational taxonomic units (OTU) for unknown sequences into this analysis framework, providing users with the opportunity to set sequence similarity thresholds and speed placement of large datasets by collapsing redundant sequences into OTU groups. This option has been tested successfully for datasets including >38,000 nrITS-LSU sequences representing diverse endophytic and endolichenic fungi (Arnold,
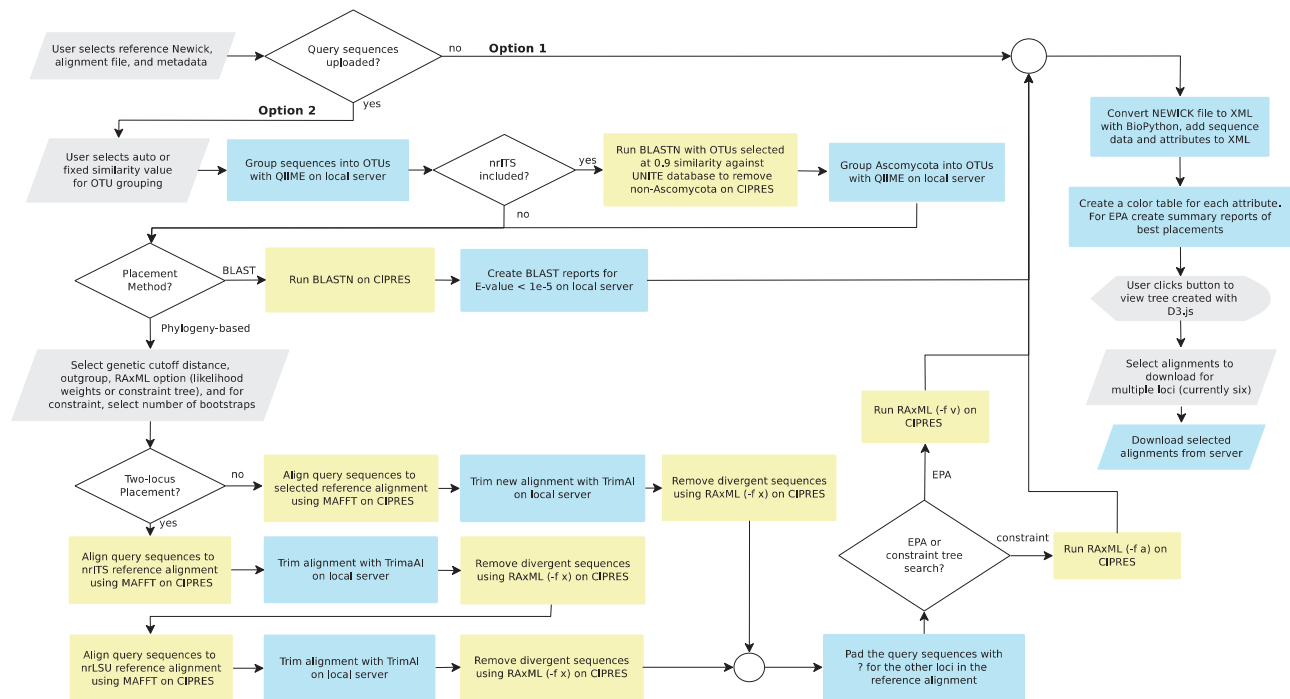
Fig. 2. A schematic flowchart showing the major analysis steps and their partitioning between the T-BAS server and CIPRES REST API. In T-BAS v1.0, two primary functions are available. Option 1 is used to view the reference tree, alignment data, and metadata, as well as download alignments. Option 2 is used for placement of unknown query sequences using BLAST, single locus phylogeny-based placement, or two-locus (nr5.8S-nrLSU) placement, and then can be used to download alignments that include unknown query sequences. Metadata can be displayed on trees visualized using both Option 1 and Option 2. All actions that are performed in the browser (e.g. uploading files and runtime options) are indicated in grey boxes. Individual steps alternate between local server (blue) and CIPRES (yellow), with computationally expensive steps performed on CIPRES resources. T-BAS scripts provide uninterrupted functionality across both platforms. Abbreviations in the flowchart are as follows: EPA = Evolutionary Placement Algorithm, CIPRES = Cyberinfrastructure for Phylogenetic Research, OTU = Operational Taxonomic Unit, UNITE = Unified system for the DNA based fungal species linked to the classification, BLAST = Basic Local Alignment Search Tool, TrimAl = a tool for automated alignment trimming, RAxML = Randomized Axelerated Maximum Likelihood, MAFFT = Multiple Alignment using Fast Fourier Transform, Newick = a tree format using parentheses and commas, D3.js = a JavaScript library for visualizing data with HTML, SVG and CSS, XML = Extensible Markup Language. Specific command line parameters for RAxML runs are indicated in parentheses

unpublished data; results not shown). A wide range of sequence-similarity thresholds in T-BAS creates flexibility for the researcher to explore the effects of different stringencies in OTU designation, often critical for ecological inference (U'Ren *et al.*, 2009). Outputs are informed by phylogeny-based or BLAST-based placement and are compatible with diverse biodiversity informatics tools, including QIIME (Caporaso *et al.*, 2010; see below). Details of OTU designation are described below.

Finally, any metadata associated with uploaded sequences by the user, such as phenotypic and environmental data, can be visualized and downloaded for further ecological analysis in an evolutionary context. T-BAS has several options for displaying metadata for both reference and query sequences, providing flexibility in visualization and the ability to simultaneously view multiple levels of phylogenetic and ecological information.

Although in its initial focus, T-BAS provides a tool that is restricted to the Pezizomycotina, the T-BAS architecture allows multiple T-BAS v1.0 portals to be deployed for different reference trees using a single configuration system. In addition to the multi-locus Pezizomycotina tree, we plan to deploy phylogenies with greater taxon coverage in the Xylariaceae and Lecanoromycetes, as well as a nrITS-LSU phylogeny of the Pezizomycotina, including endolichenic and endophytic fungi isolated in pure culture. Contrary to the multi-locus trees, which include only reference taxa for which sequences are available for multiple loci, the nrITS-LSU tree will include reference taxa but also published, unknown sequences for

which ecological data are available. The availability of these portals can guide mycologists to determine which unknown and potentially new species need additional loci, or genomes, to be sequenced; and to evaluate when enough information has been gathered to move forward with new species delimitations and descriptions.

Importantly, the architecture on which T-BAS is based can be extended beyond the Pezizomycotina. We anticipate that T-BAS v2.0 will allow users interested in other taxonomic groups to upload trees of reference taxa and alignments from focal taxa as needed (see Supplementary Material on background and future development for T-BAS).

2.2 T-BAS web server

Tree visualization and manipulation in T-BAS are seamlessly integrated with data analysis web services accessible through the CIPRES RESTful application (CRA) programmer interface (Miller *et al.*, 2015). The CRA provided by the CIPRES Science Gateway grants public access to computational resources through the National Science Foundation's Extreme Science and Engineering Discovery Environment (XSEDE) project. New T-BAS users are required to register to use CIPRES REST Services, and then sign in to T-BAS using their system-generated four-digit PIN and email address. After signing in, users are directed to the T-BAS main page. T-BAS uses PHP and Python scripts to load pages and process data. Files are uploaded by the user as HTML multipart/form-data and

progress is monitored using the XMLHttpRequest progress event. The schematic in Figure 2 shows the options and parameters that are available to users. Once analyses are complete, links for results reports and the URL for the tree are emailed to users, and results will be stored on the T-BAS server for 10 days. An archive containing all the results files required to recreate the tree can also be downloaded by the user and re-uploaded at a later date. T-BAS is optimized to run on Google Chrome, which offers faster JavaScript performance for highly interactive webpages.

After files are uploaded and preliminary processing is complete, the main PHP script is called from the command line to carry out all the runtime options specified by the user. This allows the scripts to run without using web server resources. The T-BAS pages use AJAX (http://www.adaptivepath.com/ideas/ajax-new-approach-web-appli cations/) to send and receive data to and from the server and cookies to store user preferences and configuration information in the browser. The PHP pages use the PHP shell_exec command to call the Python scripts that process the uploaded data, including calls to the CIPRES RESTful API. While the PHP and Python scripts are running, the calling page polls the server every few seconds for the status, and when completed, it displays metadata and a link to the viewer. T-BAS uses the NGINX web server running FastCGI (http://php-fpm.org/) to serve the PHP pages. A custom python script is used to read a NEWICK file using the Biopython Phylo.NewickIO module (http://biopython.org/DIST/docs/api/Bio.Phylo.NewickIO. Parser-class.html) and save it as XML. Further processing of CSV and PHYLIP/NEXUS files with Python was facilitated using the lxml module (http://lxml.de) to create a single phyloXML file including sequence data from multiple loci and colored attributes for voucher data. The script then creates a JSON file with all the data except the sequences to send to the browser where the tree is displayed using the JavaScript library D3.js (https://d3js.org/), which converts the JSON to SVG. When T-BAS needs to access sequence data, a JavaScript function makes a request to the server for the data. Printing the tree and legend uses SVG-CROWBAR (http://nytimes.github.io/svg-crowbar/) to create SVG files, which are converted to PDF and PNG using INKSCAPE, SVGUTILS (http://neuro science.telenczuk.pl/?p=331) and BATIK-1.7 and FPDF library (http://www.fpdf.org/).

2.3 OTU clustering

Unknown query sequences that are pre-assembled and quality filtered are uploaded as unaligned FASTA files. Sequences are first assigned to OTUs using the pick_otus.py script within QIIME Version 1.9.1 (Caporaso et al., 2010), which utilizes the UCLUST heuristic clustering algorithm (Edgar, 2010). Since UCLUST assigns the first sequence in a new centroid as the representative of a cluster, input sequences are sorted in order of descending length and the longest is selected as the representative OTU sequence. The default OTU clustering step reduces (i.e. de-replicates) the number of unknown queries to less than 4000, a number that is optimal for subsequent alignment and placement. Alternatively, the user has the option of selecting clustering at the 0.65, 0.75, 0.80, 0.90, 0.93, 0.95, 0.97, 0.98, 0.99 or 1.00 sequence similarity level. The number of OTUs is capped at 4000, which—including reference taxa—yields 4979 tips in the tree. This number optimizes visual representation of the tree such that taxon names and bootstrap support values are readable at higher magnifications and allows for real-time tree manipulation and printing. Outputs of this process include OTU files ready for analysis in QIIME as well as lists of query sequences with OTU designations, coupled in turn with taxonomic information based on

BLAST or evolutionary placement analyses (sections 2.4 and 2.5, below).

2.4 BLAST function

In T-BAS, BLAST searches of unknown query sequences against reference databases are performed using the BLASTN web service accessible through the CRA. Reference databases are based on DNA sequences of 979 OTU for each of five loci (nrITS-LSU, nrSSU, mtSSU, RPB1, RPB2). The combined nrITS-LSU database is created by aligning and extracting consensus sequences from taxa with overlapping regions. In addition, the public UNITE fungal nrITS database (Release date: 2015-08-01; http://unite.ut.ee/) (Koljalg et al., 2013) is used to pre-filter non-Ascomycota unknown sequences. This feature is only available if the user uploads nrITS sequences. For the filtering step, unknown nrITS sequences are grouped into OTUs using a default similarity value of 0.90. A BLAST search of OTUs against UNITE is performed to filter out non-Ascomycota OTU. BLAST searches of known nrITS sequences showed that there was minimal improvement in filtering accuracy in setting a higher similarity (S) value for OTU clustering, but accuracy (F-score) decreased for values less than $S = 0.90$ ($F = 0.95$, $S = 0.75$; $F = 0.95$, $S = 0.80$; $F = 0.97$, $S = 0.90$; see section 3.2 for explanation of F-scores). The Ascomycota OTU are collapsed into taxa and another round of clustering at a user-selected similarity level is performed. A BLAST search of representative sequences then is performed using the default BLASTN search parameters. Seven reports are generated: (1) a summary report showing the top three hits, bit-scores and E-values, (2) a detailed report showing the hits for each unknown query sequence, bit-score, E-value, and the start and end positions of the query and reference sequences (3) a file listing the best class-level match, the top three class-level matches and the top three taxon matches for each query sequence and OTU, (4) a report similar to (3) but with all duplicates removed, (5) a summary file showing the counts and percentage of unknown queries in each class, (6) a report of BLAST misses where the E-value was > 1e-5 (or for which BLAST matches were not obtained) and (7) a report of BLAST hits to the UNITE fungal nrITS database. All reports are displayed in T-BAS with links for downloading. A link to the CIPRES job folder (which contains all the output of the job) is also provided. The user has the option of displaying the best hits on the reference tree. The number of unknowns with best matches to reference taxa are also displayed in a dashboard window. In future iterations of T-BAS we anticipate that these BLAST functions can be performed on taxa outside of the Pezizomycotina using user-supplied or other curated sequence databases.

2.5 Evolutionary placement

In T-BAS, phylogeny-based evolutionary placement of unknown query sequences can be performed for either single loci (nrITS, nrLSU, nrSSU, mtSSU, RPB1, RPB2) or for multiple loci (currently nrITS-LSU, the amplicon from primer ITS1F/ITS5/ITS1 [Gardes and Bruns, 1993] to primer LR3 [Vilgalys and Hester, 1990]). These approaches leverage the reference tree provided by T-BAS and generate multiple sequence alignments (MSA) that contain the reference and unknown sequences. Evolutionary placement begins with alignment of unknown OTU sequences to the reference MSA using MAFFT (Katoh and Toh, 2010), which is accessible through the RESTful services at CIPRES.

For these analyses, unaligned unknown sequences in FASTA format are added into an existing reference MSA using the auto method in MAFFT. The auto setting automatically selects between

progressive and iterative refinement alignment methods depending on the size of the dataset. These distance-based methods use the reference MSA as a backbone to build a larger MSA containing unknown query sequences. The advantage of this approach is that it allows the reference MSA to include sequences that are correctly aligned over a portion of their lengths and unalignable in other regions, thereby increasing the length of sequence available for alignment with unknowns. An option is included to correctly align reverse-complement sequences. For optimal speed and accuracy the number of unknown OTU sequences defaults to a maximum of 4000, as described above.

The new alignment of the unknowns to the reference MSA is further subjected to a second run in MAFFT, which results in an improved global alignment. This works well with unknown sequences that are similar to reference sequences, but is problematic with highly divergent or nonhomologous sequences, which should be excluded as described below.

To deal with potential misalignments, positions in the MSA with gaps or unknowns are removed using TrimAl v1.4 at a threshold of 0.30 (Capella-Gutierrez *et al.*, 2009). Because TrimAl only removes gaps, missing sequence data (shown as question marks) in the alignment are first converted to gaps and then converted back by the *col-numbering* option of TrimAl. This outputs the positions of selected columns in the original alignment and allows writing back the question marks from the original file.

Pairwise maximum likelihood (ML) distances between sequences are then calculated for the trimmed reference MSA with unknowns using the RAxML (Stamatakis, 2014) service at CIPRES. To identify misalignments, a base distribution is calculated and a score for each unknown query is compared to it using a user-selected standard deviation interval value. The minimum distances among all pairwise comparisons of reference sequences are used to calculate an overall mean and standard deviation threshold score. These scores specify the standard deviation interval cutoffs selected by the user when running phylogeny-based placement methods in T-BAS. The minimum pairwise distance for each unknown query and the reference sequences is calculated as the unknown query score. If the query score falls above the user-selected threshold score, it is excluded from the alignment. If the unknown query sequences are very divergent from the reference taxa, selecting a larger standard deviation interval will ensure that they are not excluded; however leaving them in may result in incorrect placements using phylogeny-based alignment methods (Berger and Stamatakis, 2011). Highly divergent sequences also yield very long branches in the phylogeny, making global tree visualization difficult when branches are drawn to scale.

Two strategies for evolutionary placement of unknown sequences have been implemented in T-BAS v1.0: (1) a slow, constrained-tree search, using the reference tree as a backbone constraint tree, and (2) a fast, phylogeny-based ML search strategy. In the constrained tree search, the reference tree is treated as a binary backbone tree and the unknowns are added to the best positions based on their likelihood. Branch support values are based on rapid bootstrapping and ML searches under the GAMMA model of rate heterogeneity. The closest taxon match in the reference tree is determined using XPATH queries to traverse up tree from the terminal edges until a node containing reference taxa is reached: the first reference taxon in that node set is chosen as the closest query relative. Bootstrap support values are shown in the tree to allow the user to evaluate the placements.

In the fast ML searches, unknown query sequences are mapped to the reference tree using the evolutionary placement algorithm (EPA) implemented in RAxML (Berger and Stamatakis, 2011). Under this scenario, EPA will compute the likelihood for the insertion and removal of each unknown query sequence in the reference tree. Unknowns are successively assigned to inner and terminal edges of the reference tree and the best placement is the edge where the accumulated likelihood weight has reached a value of 0.95 out of a total of 1.0. Summary reports are created using the most likely edge match. When these scores are <0.95, it is difficult to decide on the exact placement. This is addressed through a likelihood weights report, which can identify subtrees into which sequences fall with an accumulated weight of 0.95. Smaller likelihood values indicate weaker placements that may be unreliable. These alternative placements can be visualized in T-BAS.

From these processes, five reports are generated: (1) a detailed report that shows, for each unknown, the most common class-level assignment, the equivocal class-level and taxa assignments, EPA placement probabilities (likelihood weights) and OTU assignment, (2) a report similar to (1) but with all duplicates removed, (3) a summary file showing the counts and percentage of unknown queries placed into representative classes, (4) a file showing the overall reference mean and standard deviation threshold score, the user selected cutoff score, and the calculated unknown query score and (5) a likelihood report that shows the edges needed to reach an accumulated likelihood weight of 0.95. All other analysis reports are based on the best scoring edge. The tree with all placements can be visualized interactively in T-BAS and downloaded as a PNG, SVG, or PDF (single and multi-page format).

3 Results

3.1 The most comprehensive multi-locus phylogeny for Pezizomycotina

Methods used to generate the 979 taxon tree displayed in T-BAS v1.0 are described in the caption of Supplementary Figure S1. The comprehensive phylogenetic tree more than doubles the number of taxa in previous multi-locus phylogenies of the Pezizomycotina (e.g. Schoch *et al.*, 2009). The Pezizomycotina ingroup is rooted using 23 outgroup taxa representing the two other subphyla of the Ascomycota: Saccharomycotina (14 taxa representing one class) and Taphrinomycotina (eight taxa representing four classes, and one taxon unclassified at the class level). One taxon was unclassified at the subphylum level for a grand total of 979 taxa included in the data matrix (Supplementary Table S1). Of the 979 included taxa, the supermatrix underlying T-BAS has 331 taxa for which sequences of at least five of the six loci are represented; 702 taxa with sequence data for at least four loci; 912 with sequence data for at least three loci, and 966 taxa with sequence data for at least two loci. DNA sequences of the nrITS1 and nrITS2 spacer regions are also available (but unaligned) for all Pezizomycotina taxa for which we have also included nr5.8S sequence data in the supermatrix (i.e. 681 Pezizomycotina OTU).

The metadata in T-BAS associated with the 979 OTU tree include the classification (all main ranks) and accession numbers for each sequence found in GenBank (Supplementary Table S1). The number of loci for which sequences are available in T-BAS is recorded and reported by T-BAS. This information is linked to the phylogenetic topology representing the optimal tree. The phylogenetic uncertainty for each internode based on bootstrap values is reported using a cumulative supermatrix approach (see Gaya *et al.*, 2012; Miadlikowska *et al.*, 2006, 2014) in Supplementary Figure S1.

The multi-locus tree at the core of T-BAS v1.0 is well supported despite including nearly one thousand OTU, and it is in agreement with well-established relationships within the Pezizomycotina (see Hibbett et al., 2007; Schoch et al., 2009). Here we recover a well-supported paraphyletic relationship where Orbiliomycetes diverged first, followed by the divergence of the Pezizomycetes, in agreement with James et al. (2006, but unsupported) and Spatafora et al. (2006, but unsupported) and in disagreement with Schoch et al. (2009, but unsupported) and Gazis et al. (2016; well supported). These relationships have been particularly difficult to resolve (see Hibbett et al., 2007). The genus *Lecophagus* (represented in the tree by a single specimen), which is currently unclassified at the class level, is well supported as a member of the Orbiliomycetes. *Lecophagus* includes predaceous hyphomycetes (McInnes, 2003), a hallmark of Orbiliomycetes (Spatafora et al., 2006). The Leotiomyceta is well supported as monophyletic, as is the divergence of Geoglossomycetes as the first split within the Leotiomyceta (suggested by Schoch et al., 2009). Our results strongly confirm the sister relationship of the Laboulbeniomycetes to the Sordariomycetes and the sister relationship of this group to Leotiomycetes, as well as the well-established sister relationship of the Arthoniomycetes with Dothideomycetes (see Hibbett et al., 2007; Schoch et al., 2009). The relationships of the Candelariomycetidae, Eurotiomycetes, Lecanoromycetes and Xylonomycetes remain uncertain within the Leotiomyceta. For the first time, the Lichinomycetes forms a well-supported monophyletic group with Coniocybomycetes, *Thelocarpon* and *Biatoridium*. This tree forms the basis for the primary functions of T-BAS v1.0.

3.2 Examples

Arnold et al. (2009) published a multi-locus analysis of the Pezizomycotina with the aim of reconstructing the phylogenetic relationships and evolutionary origins of unknown endophytic and endolichenic fungi. This study included 108 strains for which the nrITS-LSU amplicon was sequenced (Supplementary Table S2). We used these sequences, for which phylogenetic placement was estimated by Arnold et al. (2009), to test the efficacy and visualization tools of T-BAS under Option 2, described above (Fig. 2).

The nrITS-LSU sequence data were submitted to the T-BAS web portal as separate nrITS and LSU FASTA files (provided as examples in T-BAS) with metadata provided as a comma-separated (CSV) file (Supplementary Table S3). The check-boxes listed under T-BAS Option 1 were selected to enable T-BAS to access the reference NEWICK file, the reference alignment file, and the reference metadata. Under Option 2, phylogeny-based placement was chosen. For this example, the similarity filter was set to 1.0, with default values for standard deviations and the fast likelihood weights EPA method selected. The outgroup option was used with the default outgroup taxa. The analysis was run in Google Chrome for optimal performance.

Results are shown in Figure 3. The tree and color legends based on pseudo-color assignments were created automatically by the T-BAS interface. The tree is drawn in circular layout (rectangular also possible) without branch lengths drawn to scale (a scale bar is shown if drawn to scale). In T-BAS v1.0, reference and query metadata can be displayed as colorized outer rings in the tree. In Figure 3, the tree leaves or taxa (innermost ring) are colorized according to habitat, host type (middle ring) and origin (outermost ring). The branches are colorized by taxonomy (class-level assignment). Clustering at a similarity level of 1.0 yielded 103 OTUs. Two OTUs had multiple sequences: OTU 42 (6714_endo, 6749_endo,

6741_endo, 151_endo, 412_endo) and OTU 98 (6533_endo, 1449_endo) (Supplementary Table S4).

Likelihood weights for OTU placements ranged from 0.197060 (2706_endo had the weakest placement) to greater than 0.96 (50 OTU) (Supplementary Table S4). Placements based on constrained tree searches (Supplementary Table S5) were consistent with likelihood weights. One advantage of the likelihood weights strategy is that it allows for placement of shorter reads (e.g. nrITS fragments) into subtrees without having to decide on the exact position; bootstrapping is more informative with longer reads where additional sequence variation will improve placements. Also, assigning likelihood weights with EPA allows quantification of uncertainty and highlights where additional taxon sampling and/or sequencing is required for increasing phylogenetic resolution.

We further compared results for the same dataset based on the nrITS-LSU locus vs. its component loci using both T-BAS (BLAST and phylogeny-based) and BLAST analyses conducted in GenBank. Results are shown in Supplementary Table S6 (class level) and Supplementary Table S7 (closest reference sequences). T-BAS analyses of the nrITS, partial nrLSU (through LR3), nr5.8S+ partial nrLSU and nrITS-LSU correctly placed the 108 query sequences in the proper classes relative to results presented in Arnold et al. (2009). However, EPA analyses based only on the nr5.8S were less informative, reflecting the limitations of that locus (Supplementary Table S6). Extending the nr5.8S sequences to include the unalignable portions of the nrITS improved the placement of the 108 query sequences over nr5.8S alone such that only two taxa (1442_endo and 1504_endo) were inconsistent with placements reported in Arnold et al. (2009). The likelihood weights for 1442_endo and 1504_endo are low, 0.153786 and 0.338378, respectively, with many alternative placements with smaller likelihoods indicating that their placements may be unreliable. BLAST analyses in GenBank were less frequently correct for all main data partitions (Supplementary Table S6). Importantly, T-BAS is not designed for species-level identification, but results assessing top matches or closest reference sequences generally revealed clade-level (above species level) congruence between the Arnold et al. (2009) results and those generated by T-BAS analyses of nrITS (T-BAS BLAST and phylogeny-based placements), partial nrLSU (through LR3; T-BAS BLAST and phylogeny-based placements), nr5.8S + partial nrLSU (T-BAS BLAST and phylogeny-based placements) and nrITS-LSU (T-BAS BLAST and phylogeny-based placements).

As a second example, we compared the performance of the T-BAS BLAST and phylogeny-based placement methods (EPA-ML and EPA-BS) for a set of taxonomically annotated nrITS sequences that were randomly selected from GenBank. A total of 1000 sequences were examined for each of three regions: nrITS1, nrITS2 and nrITS1-5.8S-nrITS2 (Supplementary Table S8). For each region, class placements were categorized as true positive (TP), false positive (FP; type I error) and false negative (FN; type II error). The accuracy of taxonomic placement was measured by calculating precision, defined as $TP/(TP + FP)$, which is the proportion of sequences correctly placed out of all sequences that the placement method deemed positive, and recall defined as $TP/(TP + FN)$, which is the proportion of correctly placed sequences out of all predicted to be positive. An overall weighted average of precision and recall was computed using an F-score, where $F = 2*Precision*Recall/(Precision + Recall)$. Overall F-scores closer to 1 for nrITS1-5.8S-nrITS2 sequences ($F = 0.97$ [BLAST]; $F = 0.97$ [EPA-ML]; $F = 0.95$ [EPA-BS]) indicate better accuracy in placement compared to only partial nrITS1 ($F = 0.82$ [BLAST]; $F = 0.70$ [EPA-ML]; $F = 0.70$ [EPA-BS]) or nITS2 ($F = 0.86$ [BLAST]; $F = 0.86$ [EPA-ML]; $F = 0.87$ [EPA-BS]) sequences.
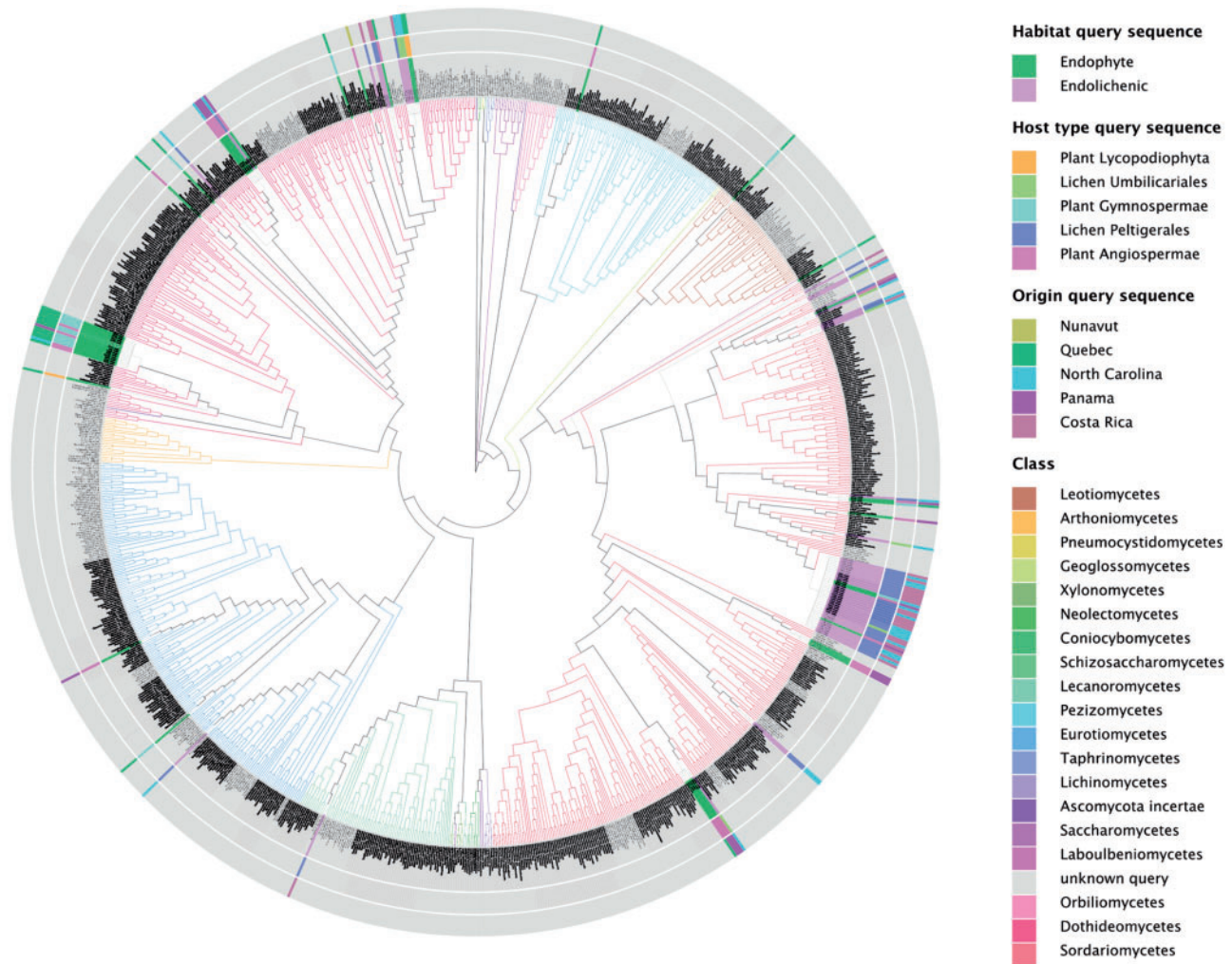
Fig. 3. Phylogenetic placements of unknown endophytic and endolichenic fungi from Arnold *et al*. (2009) as a case study and demonstration of T-BAS v1.0. Branches are colored to indicate classes; rings indicate metadata for input (query) sequences. A total of 108 query sequences were submitted (nrITS-LSU); these were collapsed into OTU based on 100% sequence identity

4 Discussion

The first example shows that commonly sequenced loci for unknown Pezizomycotina (in this case, the nrITS-partial LSU) can be used in T-BAS to recapitulate placement based on multi-locus alignments and analysis. The example data used here, which encompass a ca. 1000–1200 base pair fragment, are easily generated from Sanger sequencing of cultures using primers ITS1/ITS1F/ITS5 and LR3 (the latter as an alternative to ITS4; Arnold *et al.*, 2007; U'Ren *et al.*, 2010, 2012, 2014). This approach yields a single fragment that is more informative than nrITS alone, and is recommended for studies of cultured Pezizomycotina whenever possible. At present, such long reads are not yet readily available from next-generation sequencing approaches (Barberán *et al.*, 2015; U'Ren *et al.*, 2014; Zimmerman and Vitousek, 2012), which typically encompass short fragments such as only the nrITS1 or nrITS2 regions (in part or in full). These sequences remain challenging to place in an evolutionary framework, and sequences that contain only the nr5.8S or no conserved region are not yet easily placed at a subphylum scale. Although nrITS1-5.8S-nrITS2 sequences have higher *F*-scores and better placement accuracy compared to partial nrITS1 or nrITS2 sequences, overall *F*-scores above 0.80 for BLAST placements of partial nrITS1 and nrITS2 are generally considered good performance. We expect improvements in accuracy as we continue to

populate T-BAS with partial nrITS (and LSU) sequences from our culture collection, as well as taxonomically validated top BLAST hits for short read sequences from other sources. In future T-BAS versions we anticipate developing a hierarchical placement system that would allow such sequences to be placed based on BLAST within particular clades, then (using nrITS-enriched trees) to be aligned and placed within terminal groups with which they affiliate. User-supplied or T-BAS reference trees, enriched by Sanger sequences, will provide a powerful framework for understanding the phylogenetic placement and novelty of next-gen sequences, complementing and extending the work of systematists and highlighting hyperdiverse clades in need of revision or further species description, or that are yet to be explored.

References

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Arnold,A.E. *et al.* (2007) Diversity and phylogenetic affinities of foliar fungal endophytes in loblolly pine inferred by culturing and environmental PCR. *Mycologia*, **99**, 185–206.

Arnold,A.E. *et al.* (2009) A phylogenetic estimation of trophic transition networks for ascomycetous fungi: are lichens cradles of symbiotrophic fungal diversification? *Syst. Biol.*, **58**, 283–297.

Barberán,A. *et al.* (2015) Continental-scale distributions of dust-associated bacteria and fungi. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 5756–5761.

Berger,S.A. *et al.* (2011) Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst. Biol.*, **60**, 291–302.

Berger,S.A. and Stamatakis,A. (2011) Aligning short reads to reference alignments and trees. *Bioinformatics*, **27**, 2068–2075.

Blackwell,M. (2011) The fungi: 1, 2, 3 … 5.1 million species? *Am. J. Bot.*, **98**, 426–438.

Capella-Gutierrez,S. *et al.* (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.

Caporaso,J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Chen,K.H. *et al.* (2015) Phylogenetic analyses of eurotiomycetous endophytes reveal their close affinities to Chaetothyriales, Eurotiales, and a new order – Phaeomoniellales. *Mol. Phylogenet. Evol.*, **85**, 117–130.

Crous,P.W. *et al.* (2015) Identifying and naming plant-pathogenic fungi: past, present, and future. *Annu. Rev. Phytopathol.*, **53**, 247–267.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Gardes,M. and Bruns,T.D. (1993) ITS primers with enhanced specificity for basidiomycetes–application to the identification of mycorrhizae and rusts. *Mol. Ecol.*, **2**, 113–118.

Gaya,E. *et al.* (2012) Implementing a cumulative supermatrix approach for a comprehensive phylogenetic study of the Teloschistales (Pezizomycotina, Ascomycota). *Mol. Phylogenet. Evol.*, **63**, 374–387.

Gazis,R. *et al.* (2016) The genome of *Xylona heveae* provides a window into fungal endophytism. *Fungal Biol.*, **120**, 26–42.

Gazis,R. *et al.* (2012) Culture-based study of endophytes associated with rubber trees in Peru reveals a new class of Pezizomycotina: Xylonomycetes. *Mol. Phylogenet. Evol.*, **65**, 294–304.

Hawksworth,D.L. (1991) The fungal dimension of biodiversity: magnitude, significance, and conservation. *Mycol. Res.*, **95**, 641–655.

Hawksworth,D.L. (2001) The magnitude of fungal biodiversity: the 1.5 million species estimate revisited. *Mycol. Res.*, **105**, 1422–1432.

Hibbett,D.S. *et al.* (2007) A higher-level phylogenetic classification of the Fungi. *Mycol. Res.*, **111**, 509–547.

James,T.Y. *et al.* (2006) Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature*, **443**, 818–822.

Katoh,K. and Toh,H. (2010) Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics*, **26**, 1899–1900.

Kemler,M. *et al.* (2013) Ion Torrent PGM as tool for fungal community analysis: a case study of endophytes in *Eucalyptus grandis* reveals high taxonomic diversity. *PLoS One*, **8**, e81718.

Kirk,P.M. *et al.* (2008) *Ainsworth & Bisby's Dictionary of the Fungi*. 10th edn. Cabi Series. CAB International, Wallingford, UK.

Koljalg,U. *et al.* (2013) Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.*, **22**, 5271–5277.

Lutzoni,F. *et al.* (2000) Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. *Syst. Biol.*, **49**, 628–651.

Maddison,W.P. and Maddison,D.R. (2011) Mesquite: A Modular System for Evolutionary Analysis. Version 2.75. http://mesquiteproject.org.

Massimo,N.C. *et al.* (2015) Fungal endophytes in aboveground tissues of desert plants: infrequent in culture, but highly diverse and distinctive symbionts. *Microb. Ecol.*, **70**, 61–76.

Matsen,F.A. *et al.* (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, **11**, 538.

McInnes,S.J. (2003) A predatory fungus (Hyphomycetes: *Lecophagus*) attacking Rotifera and Tardigrada in maritime Antarctic lakes. *Polar Biol.*, **26**, 79–82.

Miadlikowska,J. *et al.* (2006) New insights into classification and evolution of the Lecanoromycetes (Pezizomycotina, Ascomycota) from phylogenetic analyses of three ribosomal RNA- and two protein-coding genes. *Mycologia*, **98**, 1088–1103.

Miadlikowska,J. *et al.* (2014) A multigene phylogenetic synthesis for the class Lecanoromycetes (Ascomycota): 1307 fungi representing 1139 infrageneric taxa, 317 genera and 66 families. *Mol. Phylogenet. Evol.*, **79**, 132–168.

Miller,M.A. *et al.* (2015) A RESTful API for Access to Phylogenetic Tools via the CIPRES Science Gateway. *Evol. Bioinform. Online*, **11**, 43–48.

O'Brien,H.E. *et al.* (2005) Fungal community analysis by large-scale sequencing of environmental samples. *Appl. Environ. Microbiol.*, **71**, 5544–5550.

Schmidt,P.A. *et al.* (2013) Illumina metabarcoding of a soil fungal community. *Soil Biol. Biochem.*, **65**, 128–132.

Schoch,C.L. *et al.* (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.

Spatafora,J.W. *et al.* (2006) A five-gene phylogeny of Pezizomycotina. *Mycologia*, **98**, 1018–1028.

Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Tedersoo,L. *et al.* (2014) Fungal biogeography. Global diversity and geography of soil fungi. *Science*, **346**, 1256688.

U'Ren,J.M. *et al.* (2009) Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples. *Mycol. Res.*, **113**, 432–449.

U'Ren,J.M. *et al.* (2010) Community analysis reveals close affinities between endophytic and endolichenic fungi in mosses and lichens. *Microb. Ecol.*, **60**, 340–353.

U'Ren,J.M. *et al.* (2012) Host and geographic structure of endophytic and endolichenic fungi at a continental scale. *Am. J. Bot.*, **99**, 898–914.

U'Ren,J.M. *et al.* (2016) Contributions of North American endophytes to the phylogeny, ecology, and taxonomy of Xylariaceae (Sordariomycetes, Ascomycota). *Mol. Phylogenet. Evol.*, **98**, 210–232.

U'Ren,J.M. *et al.* (2014) Tissue storage and primer selection influence pyrosequencing-based inferences of diversity and community composition of endolichenic and endophytic fungi. *Mol. Ecol. Resour.*, **14**, 1032–1048.

Vilgalys,R. and Hester,M. (1990) Rapid genetic identification and mapping of enzymatically amplified ribosomal DNA from several *Cryptococcus* species. *J. Bacteriol.*, **172**, 4238–4246.

White,T.J. *et al.* (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. InL Innis,M. *et al.* (eds) *PCR Protocols: A Guide to Methods and Applications*. Academic Press, Orlando, Florida, pp. 315–322.

Zimmerman,N.B. and Vitousek,P.M. (2012) Fungal endophyte communities reflect environmental structuring across a Hawaiian landscape. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 13022–13027.