Identification of Fungal DNA Barcode Targets and PCR Primers Based on Pfam Protein Families and Taxonomic Hierarchy

Christopher T. Lewis[*,1], Satpal Bilkhu[1,2], Vincent Robert[3], Ursula Eberhardt[3], Szaniszlo Szoke[3], Keith A. Seifert[1] and C. André Lévesque[1]

[1]*Eastern Cereals and Oilseed Research Centre, Agriculture and Agri-food Canada, 960 Carling Ave, Ottawa, ON K2E 5H3, Canada*

[2]*School of Computer Science, Carleton University, 1125 Colonel By Drive, Ottawa, ON K1S 5B6, Canada*

[3]*CBS-KNAW Fungal Biodiversity Centre, P.O. Box 85167, 3508 AD Utrecht, the Netherlands*

Abstract: DNA barcoding is the application of DNA sequences of standardized genetic markers for the identification of eukaryotic organisms. We attempted to identify alternative candidate barcode gene targets for the fungal biota from available fungal genomes using a taxonomy-aware processing pipeline. Putative-protein coding sequences were matched to Pfam protein families and aligned to reference Pfam accessions. Conserved sequence blocks were identified in the resulting alignments and degenerate primers were designed. The processing pipeline is described and the resulting candidate gene targets are discussed. The pipeline allows analysis of subsets at various hierarchical, taxonomic levels (selectable by GenBank taxonomy ID or scientific name) of the available reference data, allowing discrete taxonomic groups to be combined into a single subset, or for subordinate taxa to be excluded from the analysis of higher-level taxa. Putative degenerate primer pairs were designed as high as the superkingdom rank for the set of organisms included in the analysis. The identified targets have essential housekeeping functions, like the well known phylogenetic or barcode markers, and most have a better resolution potential to differentiate species among fully sequenced genomes than the most presently used markers. Some of the commonly used species-level phylogenetic markers for fungi, especially *tef1-α* and *rpb2,* were not recovered in our analysis because of their existence in multiple copies in single organisms, and because Pfam families do not always correlate with complete proteins.

Keywords: Fungi, barcoding, internal transcribed spacer (ITS), translation elongation factor 1A (tef1a), ribosomal polymerase B2 (rpb2), cytochrome oxidase 1 (cox1, COI).

1. INTRODUCTION

Barcoding is the use of DNA sequences to facilitate the identification of species of all domains of eukaryotic life [1]. The marker of choice should be amplified reliably with universal primers, sequenced easily because of appropriate length and GC content, and should provide sufficient interspecific variation when compared to intraspecific variation. The barcode standard is administered by the Consortium for the Barcode of Life (CBOL). It requires that reference sequences from a formally accepted barcode marker be linked to on-line sequence traces, be derived from permanent voucher specimens or cultures identified by a taxonomic expert, and be associated with on-line metadata. Cytochrome oxidase 1 (*cox1* or COI) is the default barcode widely used for most groups of animals. Two barcode regions have been accepted, *rbcL* and *matK*, for vascular plants [2] because *cox1* is not variable enough in plants. These three genes comprise the majority of sequences in the Barcode of Life Database (BOLD, www.barcodinglife.org) [3] created and maintained by the Canadian Centre for DNA Barcoding at the University of Guelph, and GenBank entries with the keyword "barcode". For Fungi, the major kingdom of eukaryotic microorganisms with more than 1.5 million species [4], a barcode region has yet to be proposed and defining one or more barcodes for Fungi is critical for their representation in the International Barcode of Life project (iBOL, www.ibol.org), which was launched in September 2010.

Cox1 is of limited use as a barcode for true Fungi. Although there is adequate interspecies variability in some groups [5, 6], there is inadequate variation in others [7]. There is also sporadic occurrence of introns that can result in dramatic length variation in amplicons, which interfere with the success of PCR and direct sequencing, and paralogues exist in some group than confound barcode utility [7, 8]. Most mycologists expect that the internal transcribed spacer (ITS) of the nuclear ribosomal DNA will be sanctioned as the first barcode for the kingdom Fungi [8, 9]. This marker exists in multiple copies in most fungal cells and is retrievable by relatively robust primers with an established record of reliability [10]. Because of this sensitivity and universality of the primers, a large reference database exists in GenBank. Unfortunately, the individual copies within a cell are not always identical, increasing the technical difficulty for sequencing, and much of the available data does not meet formal barcode standards and/or is from unarchived strains/specimens of questionable identity [11].

*Address correspondence to this author at the Eastern Cereals and Oilseed Research Centre, Agriculture and Agri-food Canada, 960 Carling Ave, Ottawa, ON K2E 5H3, Canada; Tel: 613-759-1232; Fax: 613-759-1701; E-mail: christopher.lewis@agr.gc.ca

1874-1363/11

Furthermore, the ITS is shorter than ideal for a barcode for some subgroups of Fungi; with a very conserved 5.8S domain between the two ITS subunits, the region often lacks species level resolution [9], especially in several genera of the Ascomycota (referred to here as Ascomycetes), which include about 60% of presently recognized fungal species. For the Ascomycetes, at least, there will be a need for a second barcode with sufficient variability to allow robust species identification in a broader range of species. The Assembling the Fungal Barcode of Life (AFTOL) project, completed in 2008 in the U.S.A., provided a multi-gene phylogeny of the kingdom Fungi based on up to six genes [12]. Some of these genes, especially ribosomal polymerase B2 (*rpb2*) and translation elongation factor 1-α (*tef1*-α, or EF1 or TEF), provide good resolution at the species level for many fungal groups, but designing broadly useful PCR and sequencing primers has so far not been achieved, although interesting candidates were identified for *tef1*-α in Robert *et al.* [13].

The initial barcode markers, in particular *cox1*, were selected based on practical criteria (e.g. multiple copies in the genome, high AT ratios for enhanced stability in voucher specimens, availability of universal primers) before significant numbers of genomic sequences were available to optimize selection. Barcodes need not reflect deep phylogeny accurately when aligned and analyzed, but must exhibit significantly more variability between species than within species. The availability of more than 100 fungal genomes [14-16] and more than 100 additional in-progress genomes [15] provides an opportunity for *in silico* screening of a much broader range of genes as potential barcode marker for the Fungi in general, and for the Ascomycetes in particular. As part of our efforts on barcoding the mycota of the indoor environment, many of which are Ascomycetes, we are searching for alternative barcode markers that will enhance the precision of environmental metagenomics of the indoor environment, and which may also have a broader application for Fungi in other environments.

In an accompanying paper, Robert *et al.* [13] outline one bioinformatics approach to locating potential barcode markers from fungal genomes. In this paper, we describe an alternative approach that exploits annotated protein families from the Pfam protein families database [17] to guide the formation of fungal gene clusters in order to identify conserved, single copy (putatively orthologous) fungal genes that might contain regions suitable for use as a barcode. While a barcode region need not be phylogenetically informative, it would be of added value if a robust barcode region were found which did reflect phylogeny. To this end we've restricted our analysis to single copy nuclear genes to reduce the likelihood of amplifying paralogous genes.

Described herein are the results of an *in silico* analysis applying a taxonomy-based approach to identifying alternative barcode genes at various hierarchical levels within the kingdom Fungi. The approach involves matching translated nucleotide sequences to Pfam protein families, aligning the translated sequences with the protein family reference alignment, and designing degenerate primers within conserved regions of the alignment. Efforts were made to target putatively orthologous sequences by analyzing only single copy gene families, which decreases the likelihood of paralogous gene sequences confounding analysis results.

The pipeline for this analysis was developed using several existing bioinformatics tools and was generally inspired by the CARMA [18] algorithm for the taxonomic identification of metagenomic sequences. The pipeline allows subsets (selectable by GenBank taxonomy ID or scientific name) of the available reference data to be analyzed, for particular taxonomic groups to be combined into a single subset, or for sub-taxa to be excluded from the analysis of higher-level taxa. Putative degenerate primer pairs were designed as high as the superkingdom rank for the set of organisms included in the analysis, which included several representatives of the Fungi/Metazoa Group and other fungal-like organisms.

2. MATERIALS AND METHODS

Source Data

For our analyses, transcript sequences for a set of 48 Fungi or fungal-like organisms were downloaded from the Broad Institute on or before Feb 2, 2010. Automated processing failed for 19 genomes, which were excluded from the analysis and reserved for future *in silico* validation of predicted results. The taxonomy of the remaining 29 processed organisms was heavily weighted towards the target Ascomycetes (24), which are least well resolved using the existing markers. Also represented were the Basidiomycetes (2), Chytridiomycetes (1), Oomycota (1) and unclassified (1) genomes (Table **1**).

A supplementary analysis of additional genome sequences downloaded on Nov 9, 2010 included 72 Fungi or fungal-like organisms available from the Broad Institute (Fig. **1**). Of those, 14 failed to automatically process and were reserved for subsequent *in silico* validation of barcode candidates. Represented among the remaining 58 were the Ascomycetes (46), Basidiomycetes (4), Chytridiomycetes (2), Choanoflagellida (1), Blastocladiomycetes (1), Ichthyosporea (1), Zygomycetes (all Mucorales) (1), and Oomycota (1) (Table **1**).

Table **1** presents the lineage of the full set of organisms at the Kingdom, Phylum, Class and Order levels as ranked in the NCBI taxonomy database; all organisms are classified within the Eukaryota superkingdom. A file was created with the version and download URL for each of the data sets used in this analysis (Supplementary Table **1**), and a script was used to download the target data sets.

Reference Data

Fungal sequences were clustered based on Pfam protein families. Version 23.0 of the Pfam database [19] was used for both iterations of the analysis. This analysis was based on the Pfam-A data set, which consists of semi-automated alignments for 10,340 protein family accessions prepared using curated Hidden Markov Models (HMMs). Taxonomic processing was enabled by the NCBI taxonomy. The version of the NCBI taxonomy database available on Feb 1, 2010 was used for the initial analysis and the version available on Nov 22, 2010 was used for the update.

Table 1. Fungal and Fungal-Like Organisms Included in Analysis

Kingdom	Phylum	Class	Order	NCBI Taxon ID	Species/Strain	Included[1]
Other	-	-	-	461836	Thecamonas trahens ATCC 50062	F , N/A
Other	Oomycota	-	Peronosporales	4787	Phytophthora infestans T30-4	F , I
Fungi/Metazoa Group	-	-	Choanoflagellida	431895	Monosiga brevicollis MX1	X , N/A
Fungi/Metazoa Group	-	-	Choanoflagellida	222925	Salpingoeca rosetta ATCC 50818	F , N/A
Fungi/Metazoa Group	-	Ichthyosporea	-	595528	Capsaspora owczarzaki ATCC 30864	F , N/A
Fungi	-	-	Mucorales	246409	Rhizopus oryzae RA 99-880	F , I
Fungi	Ascomycota	Dothideomycetes	Pleosporales	13684	Stagonospora nodorum SN15	F , X
Fungi	Ascomycota	Dothideomycetes	Pleosporales	45151	Pyrenophora tritici-repentis Pt-1C-BFP	F , I
Fungi	Ascomycota	Eurotiomycetes	Eurotiales	5057	Aspergillus clavatus NRRL 1	X , X
Fungi	Ascomycota	Eurotiomycetes	Eurotiales	5059	Aspergillus flavus AAIH01000000	F , I
Fungi	Ascomycota	Eurotiomycetes	Eurotiales	5085	Aspergillus fumigatus Af293	X , X
Fungi	Ascomycota	Eurotiomycetes	Eurotiales	227321	Aspergillus nidulans FGSC A4	F , I
Fungi	Ascomycota	Eurotiomycetes	Eurotiales	5061	Aspergillus niger CBS 513.88	F , I
Fungi	Ascomycota	Eurotiomycetes	Eurotiales	5062	Aspergillus oryzae ATCC 42149	X , X
Fungi	Ascomycota	Eurotiomycetes	Eurotiales	33178	Aspergillus terreus NIH2624	F , X
Fungi	Ascomycota	Eurotiomycetes	Eurotiales	36630	Neosartorya fischeri NRRL 181	X , X
Fungi	Ascomycota	Eurotiomycetes	Onygenales	544712	Histoplasma capsulatus H143	F , N/A
Fungi	Ascomycota	Eurotiomycetes	Onygenales	544711	Histoplasma capsulatus H88	F , N/A
Fungi	Ascomycota	Eurotiomycetes	Onygenales	663331	Arthroderma benhamiae CBS 112371	X , N/A
Fungi	Ascomycota	Eurotiomycetes	Onygenales	535722	Microsporum gypseum CBS 118893	F , I
Fungi	Ascomycota	Eurotiomycetes	Onygenales	554155	Microsporum canis CBS 113480	F , I
Fungi	Ascomycota	Eurotiomycetes	Onygenales	653446	Blastomyces dermatitidis ATCC 18188	F , N/A
Fungi	Ascomycota	Eurotiomycetes	Onygenales	559297	Blastomyces dermatitidis ER-3	F , X
Fungi	Ascomycota	Eurotiomycetes	Onygenales	559298	Blastomyces dermatitidis SLH14081	F , X
Fungi	Ascomycota	Eurotiomycetes	Onygenales	396776	Coccidioides immitis H538.4	X , N/A
Fungi	Ascomycota	Eurotiomycetes	Onygenales	404692	Coccidioides immitis RMSCC 2394	X , N/A
Fungi	Ascomycota	Eurotiomycetes	Onygenales	454286	Coccidioides immitis RMSCC 3703	X , N/A
Fungi	Ascomycota	Eurotiomycetes	Onygenales	246410	Coccidioides immitis RS	F , I
Fungi	Ascomycota	Eurotiomycetes	Onygenales	454284	Coccidioides posadasii RMSCC 3488	F , N/A
Fungi	Ascomycota	Eurotiomycetes	Onygenales	443226	Coccidioides posadasii Silveira	F , X
Fungi	Ascomycota	Eurotiomycetes	Onygenales	447093	Histoplasma capsulatum G186AR	F , N/A
Fungi	Ascomycota	Eurotiomycetes	Onygenales	339724	Histoplasma capsulatum NAm1	F , I
Fungi	Ascomycota	Eurotiomycetes	Onygenales	502779	Paracoccidioides brasiliensis Pb01	F , N/A
Fungi	Ascomycota	Eurotiomycetes	Onygenales	482561	Paracoccidioides brasiliensis Pb03	F , N/A
Fungi	Ascomycota	Eurotiomycetes	Onygenales	502780	Paracoccidioides brasiliensis Pb18	F , I
Fungi	Ascomycota	Eurotiomycetes	Onygenales	559882	Trichophyton equinum CBS 127.97	F , I
Fungi	Ascomycota	Eurotiomycetes	Onygenales	559305	Trichophyton rubrum CBS 118892	F , N/A

(Table 1) contd…..

Kingdom	Phylum	Class	Order	NCBI Taxon ID	Species/Strain	Included[1]
Fungi	Ascomycota	Eurotiomycetes	Onygenales	34387	Trichophyton tonsurans CBS 112818	F , N/A
Fungi	Ascomycota	Eurotiomycetes	Onygenales	663202	Trichophyton verrucosum HKI 0517	X , N/A
Fungi	Ascomycota	Eurotiomycetes	Onygenales	33188	Uncinocarpus reesii 1704	F , I
Fungi	Ascomycota	Leotiomycetes	Helotiales	40559	Botrytis cinerea B05-10	F , N/A
Fungi	Ascomycota	Leotiomycetes	Helotiales	5180	Sclerotinia sclerotiorum ATCC 18683	F , I
Fungi	Ascomycota	Saccharomycetes	Saccharomycetales	237561	Candida albicans sc5314	X , X
Fungi	Ascomycota	Saccharomycetes	Saccharomycetales	294748	Candida albicans W0-1	F , X
Fungi	Ascomycota	Saccharomycetes	Saccharomycetales	5480	Candida parapsilosis CDC 317	F , I
Fungi	Ascomycota	Saccharomycetes	Saccharomycetales	5482	Candida tropicalis MYA-3404	F , I
Fungi	Ascomycota	Saccharomycetes	Saccharomycetales	36911	Candida lusitaniae ATCC 42720	F , N/A
Fungi	Ascomycota	Saccharomycetes	Saccharomycetales	4959	Debaryomyces hansenii CBS 767	F , X
Fungi	Ascomycota	Saccharomycetes	Saccharomycetales	36914	Lodderomyces elongisporus NRRL YB-4239	F , I
Fungi	Ascomycota	Saccharomycetes	Saccharomycetales	4929	Candida guilliermondii ATCC 6260	F , X
Fungi	Ascomycota	Saccharomycetes	Saccharomycetales	285006	Saccharomyces cerevisiae RM11-1a	F , I
Fungi	Ascomycota	Schizosaccharomycetes	Schizosaccharomycetales	866546	Schizosaccharomyces cryophilus OY26	F , N/A
Fungi	Ascomycota	Schizosaccharomycetes	Schizosaccharomycetales	4897	Schizosaccharomyces japonicus yFS275	F , I
Fungi	Ascomycota	Schizosaccharomycetes	Schizosaccharomycetales	483514	Schizosaccharomyces octosporus yFS286	F , X
Fungi	Ascomycota	Schizosaccharomycetes	Schizosaccharomycetales	4896	Schizosaccharomyces pombe 972H	F , I
Fungi	Ascomycota	Sordariomycetes	Hypocreales	5507	Fusarium oxysporum 4287	F , I
Fungi	Ascomycota	Sordariomycetes	Hypocreales	117187	Fusarium verticillioides 7600	F , X
Fungi	Ascomycota	Sordariomycetes	Hypocreales	5518	Fusarium graminearum PH-1	F , I
Fungi	Ascomycota	Sordariomycetes	Magnaporthales	242507	Magnaporthe oryzae 70-15	X , X
Fungi	Ascomycota	Sordariomycetes	Phyllachorales	526221	Verticillium albo-atrum VaMs.102	F , I
Fungi	Ascomycota	Sordariomycetes	Phyllachorales	498257	Verticillium dahliae VdLs.17	F , I
Fungi	Ascomycota	Sordariomycetes	Sordariales	38033	Chaetomium globosum CBS 148.51	F , I
Fungi	Ascomycota	Sordariomycetes	Sordariales	367110	Neurospora crassa OR74A	F , I
Fungi	Basidiomycota	-	Pucciniales	5297	Puccinia graminis f._sp_.tritici CRL 75-36-700-3	F , I
Fungi	Basidiomycota	Agaricomycetes	Agaricales	240176	Coprinopsis cinerea okayama7#130	F , X
Fungi	Basidiomycota	Agaricomycetes	Agaricales	486041	Laccaria bicolor S238N-H82	X , N/A
Fungi	Basidiomycota	Tremellomycetes	Tremellales	37769	Cryptococcus gattii R265	F , I
Fungi	Basidiomycota	Tremellomycetes	Tremellales	235443	Cryptococcus neoformans var. grubii H99	F , X
Fungi	Basidiomycota	Ustilaginomycetes	Ustilaginales	5270	Ustilago maydis 521	X , X
Fungi	Blastocladiomycota	Blastocladiomycetes	Blastocladiales	578462	Allomyces macrogynus ATCC 38327	F , N/A
Fungi	Chytridiomycota	Chytridiomycetes	Chytridiales	109871	Batrachochytrium dendrobatidis JEL423	F , I
Fungi	Chytridiomycota	Chytridiomycetes	Spizellomycetales	109760	Spizellomyces punctatus DAOM BR117	F , N/A

[1]Data set(s) including the organism: either full (F) or initial (I); Organisms which were present but failed processing (X) and unavailable (N/A).

Data Processing

The approach applied in this analysis was inspired by the CARMA algorithm for taxonomic identification of meta-genomic sequences. The objective in CARMA is to match short environmental gene fragments to Pfam protein families and to use the matches to construct a taxonomic profile for the sample. The objective in this analysis is to assign translated input nucleotide sequences to Pfam accessions,

which are used to identify single copy gene regions in the source organisms and design degenerate primers from the alignment of these putatively orthologous sequences. The common elements of the two applications are pre-processing the Pfam data set, translating and assigning DNA sequences to Pfam groups, and adding the sequences to the reference Pfam alignments. To save effort, the CARMA pipeline was adapted for use in this project, although the objectives and end result are quite different. Applications of and modifications to the CARMA pipeline are noted below.

The data is processed in two stages (Fig. **2**). The first stage translates the input sequences and assigns them to Pfam accessions and the second stage prepares alignments and attempts primer design for the selected set of taxa. Adding additional organisms for processing or updating

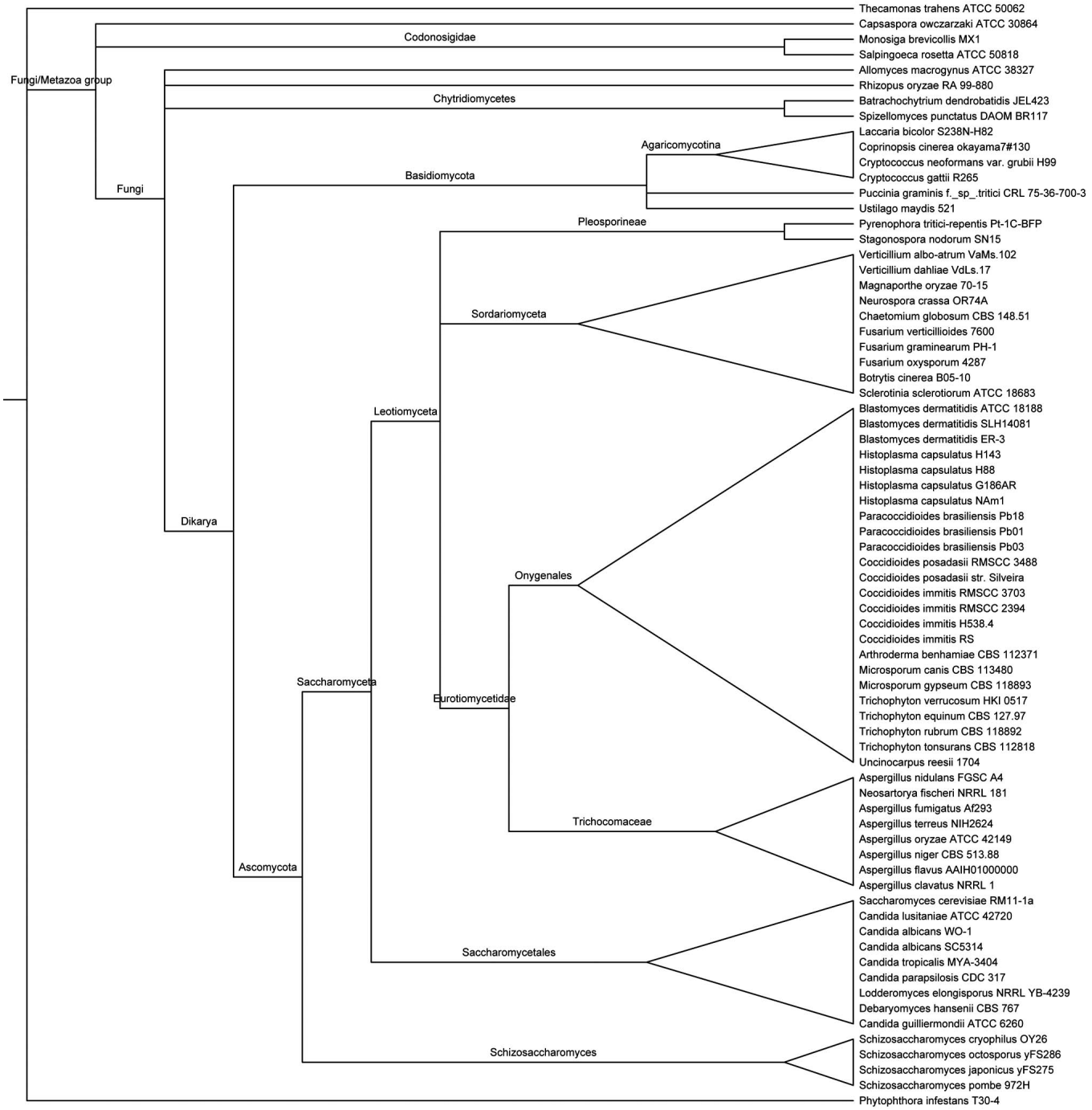

Fig. (1). Taxonomy of organisms included in the analysis. The NCBI taxonomy of the full set of organisms downloaded for the updated analysis is presented in this tree. Subsets of the selected organisms were selected for analysis using the above tree as a guide. Our main group of interest, the Ascomycetes, is well represented within the full data set, with representation from 6 major Orders, although the Onygenales are best represented.

existing organisms requires that the first stage of processing be repeated for the new data. Assaying a taxonomic subset of the data requires that the final stage be repeated for the selected set of organisms. The pipeline requires a recent version of the Pfam-A data set and an NCBI taxonomy export. The input data sets should be available in a single folder for processing, with each organism in its own directory named according to the formula of the organism's "genus_species_strain". Updating to a more recent version of the Pfam data set requires that the first stage of processing be repeated.

The first stage of the pipeline involves processing the source nucleotide sequences to identify the Pfam protein family domains within each sequence and to create clusters of conserved domains. Sequences are translated into protein sequences using the orientation and frame from a translated BLAST [20] search against the Pfam protein family sequences. Each translated sequence is then processed to identify the region corresponding to the matched protein family and the protein sequence is written to a file. This functionality is as originally implemented in the CARMA pipeline.

The second stage of the pipeline involves screening the Pfam matches for a user defined subset of organisms to identify the families that contain a single match for each of the selected organism. Such families are considered to represent conserved, single copy gene regions in the selected organisms. Next, these putatively orthologous sequences are added to the corresponding Pfam family multiple sequence alignment. Adding the sequences to an existing curated alignment should result in a better quality alignment than *de novo* alignment of the sequences. At this point, the original sequences are removed from the reference alignments so that only the targeted organisms remain and the resulting alignment is screened for conserved blocks for primer design. The goal is to identify conserved regions on either side of a variable region and design primers that amplify the flanked variable region. To this end, only alignments with two or more conserved blocks are selected for primer design and pairs of forward and reverse primers are sought in separate blocks. The final set of resulting primers can then screened for protein families containing primer pairs with additional desirable characteristics such as minimum length between primer pairs.

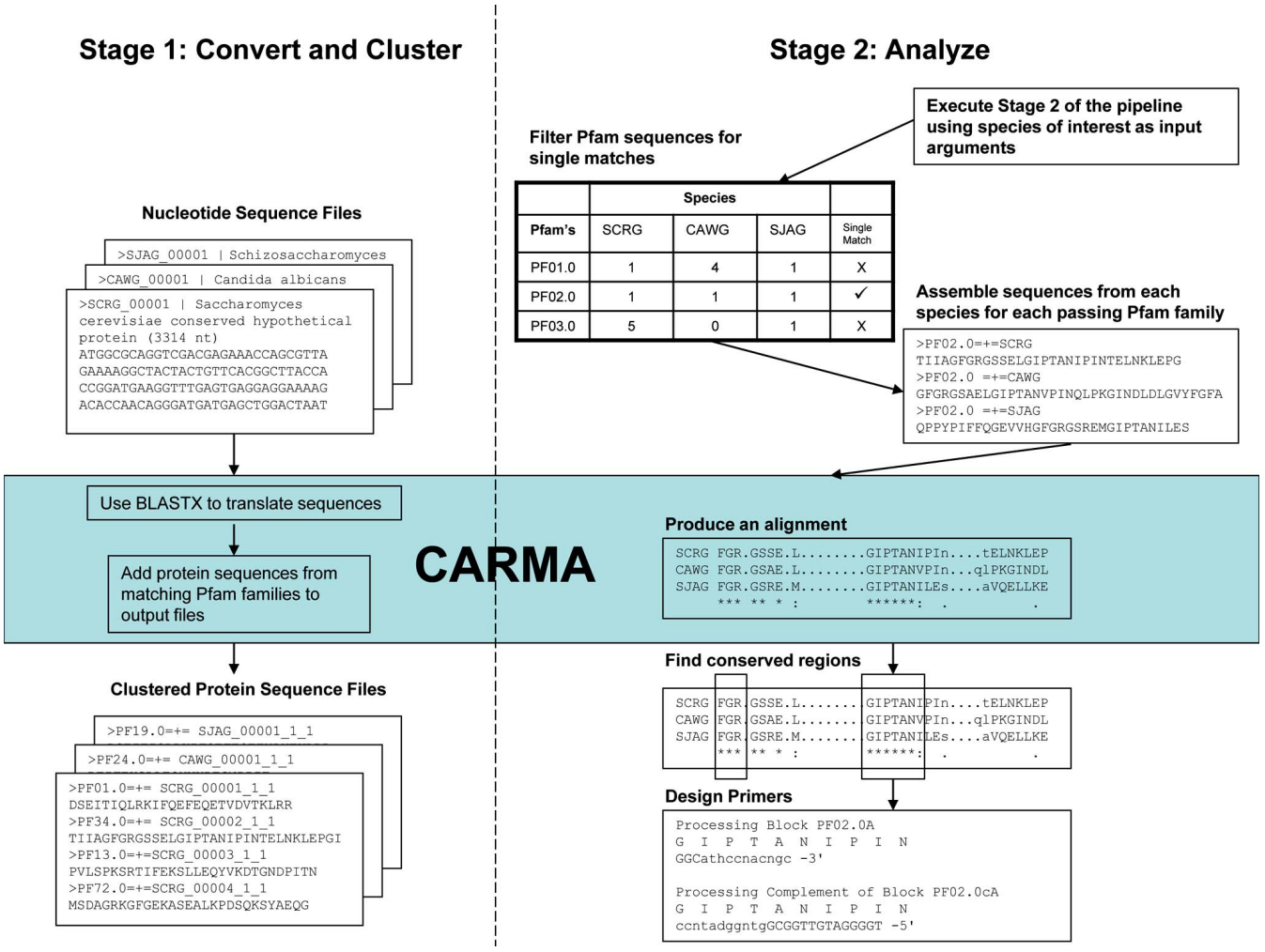

Fig. (2). Diagrammatic representation of the analysis pipeline. Stage 1 involves converting the nucleotide sequences to protein sequences and assigning them to Pfam accessions. Stage 2 involves executing the pipeline with species of interest, filtering the Pfam families for single matches occurring in each species, assembling sequences from each species for each passing Pfam family, adding them to the reference Pfam alignments, identifying conserved blocks of sequences and attempting degenerate primer design.

Producing Trees and Distance Matrices

For the newly identified candidate target regions, the alignment was trimmed just inside the outermost primer pairs. For the known barcode regions (*rpb2*, *tef1-α*), the alignment was trimmed at the edge of the outer most conserved blocks of 3 or more nucleotides.

Software Versions and Parameters

Pipeline development was based on version 1.2 of CARMA. Identification of the Pfam regions within the source sequences (stage 1) and their addition to the reference alignment (stage 2) was accomplished by CARMA using the BLASTX, hmmpfam and hmmalign utilities. The BLASTX utility is a part of the BLAST package. A BLASTX search involves finding matches for the 6-frame translations of the input nucleotide sequence in a database of protein sequences. BLAST version 2.2.23 was used for this analysis. The hmmpfam and hmmalign utilities are distributed as part of the hmmer package [21, 22]. The hmmpfam utility scans sequences for significantly similar matches to a set of reference Hidden Markov Models (HMMs) [21]. The hmmalign utility aligns a set of sequences to a reference profile HMM and outputs the resulting alignment. Hmmer version 2.3.2 was used in this analysis. The design of degenerate PCR primers (stage 2) was accomplished using MABLOCK [23] and CODEHOP [24], which are distributed as part of the BLIMPS software package [25]. The MABLOCK utility identifies gap free, highly conserved regions (blocks) within a protein multiple sequence alignment. The CODEHOP utility designs degenerate primers from input protein blocks. BLIMPS version 3.9 was used in this analysis.

The output generated from this analysis was generated using the following optional/non-default parameters. Stage 1: carma -s -o -v, which runs BLASTX with -w 15 -e 1 -F F -b 20 -v 20 and hmmpfam with -A 0. Stage 2: carma -p, which runs hmmalign with -q -withali (carma has been modified to exit after creating the alignment); mablock with parameter min width=7; CODEHOP with all default parameters.

Manual editing of multiple sequence alignments to prepare phylogenetic trees for this project was accomplished using version 4.2.1-1 of the SeaView application [26]. The alignments were produced using Muscle [27] with the default settings. Trees included in this publication were prepared from NEXUS format inputs using version 1.3.1 of the FigTree [28] application. Distance matrices were calculated using the Jukes-Cantor distance method in PAUP version 4.0b10 [29].

2. RESULTS AND DISCUSSION

Candidate Barcode Regions

The pipeline developed in this study allows users to target subsets of the available organisms by GenBank names or taxonomic identifiers. This functionality was applied to a series of increasingly restrictive taxonomic groups within a set of 72 Fungi or fungal-like organisms (Table **1**, Fig. **2**) and the *in-silico* analysis identified several Pfam accessions for which degenerate primer design succeeded even at the highest (Eukaryota/Fungi) taxonomic levels (Table **2**). Each column of the Pfam accessions section contains first, the number of Pfam accessions within the initial subset of 29 organisms and second, separated by a slash, the number of Pfam accessions within the full set of 58 organisms that are present at least once in all organisms (universal), present only once within each organism (single copy), contain multiple conserved blocks, and had PCR primers designed for them. For example, within the 29 organisms at the Eukaryota level in the initial subset, 604 universal Pfam accessions were identified. Of those, 51 exist as single copies in each organism, 22 contain multiple conserved blocks, and primers were designed for two. Within the Ascomycetes, the group of primary concern for barcode development, the analysis produced degenerate primer pairs for 17 candidate regions in the initial subset and for seven in the full data set. As might be expected, with decreasing taxonomic breadth the number of conserved and single copy gene families available for analysis increases, which results in a corresponding increase in the number of protein families containing conserved blocks within which primers can be designed (Table **2**).

Eukaryota-level primer pairs were designed for two Pfam accessions (Table **2**), PF00218 and PF08572, in the initial subset. However, no Eukaryota-level primer pairs were designed from the larger updated data set because no single copy universal Pfam accessions were identified (Table **2**). The summary report for the Eukaryota-level run with the full data set (not shown) showed that PF00218 was excluded because the accession is missing in *Thecamonas trahens* (a newly included organism from the basal lineages), and because the Pfam accession is multi-copy (n=2) in *Coprinopsis cinerea* (a Basidiomycete which was present for the initial analysis, but which failed to process). Excluding *T. trahens* and *C. cinerea* from the analysis using the taxonomic capabilities built into the pipeline allowed primers to be designed for accession PF00218. This suggests that PF00218 should have potential not only within the key Ascomycete group, but also within many other true Fungi and related organisms. The summary report also showed that PF08572 was excluded because it is a multi-copy (n=2) locus in *Phaeosphaeria nodorum* (an Ascomycete organism, which was excluded in the initial analysis) and because the Pfam accession is multi-copy (n=2) in *Salpingoeca rosetta* (a newly included organism in from the Basal lineages). Removing *P. nodorum* and *S. rosetta* allowed the analysis to proceed for PF08572, but primer design failed despite the presence of conserved blocks in the protein alignment, and continued to fail even after many of the most divergent organisms were removed. This suggests that PF08572 will probably not have the same broad potential within the Fungi and fungal-relatives as PF00218.

The full set of Ascomycete and higher-level primer sequences are present in Supplementary Table **2**. The forward and reverse primer sequences, including their estimated melting temperatures and target Pfam accession are provided for each taxonomic subset. The degenerate core is lower case while the non-degenerate clamp is upper case. Only degenerate primer design was attempted in this analysis, however, design could be attempted for non-degenerate primers. We presume that this would be most effective when attempting primer design for lower-level targets, which were not the focus of this analysis.

Table 2. Single-Copy Genes with Conserved Amino Acid Blocks Across Taxa for Primer Design in the Initial/Full Data Set

Taxa	Organisms	Pfam Accessions			
		Universal[1]	Single Copy[2]	Conserved[3]	Primers[4]
Eukaryota	29/58	604/408	51/0	22/0	2/0
+ Fungi	28/54	681/535	65/2	25/0	2/0
+ Dikarya (Ascomycota & Basidiomycota)	26/50	774/630	142/72	63/18	7/1
* Dikarya & Mucorales	27/51	734/606	79/42	30/10	2/0
* Dikarya & Chytridiomycota	27/52	706/571	110/47	48/11	5/1
* Dikarya & Oomycota	27/51	662/550	94/48	42/13	4/0
+ Ascomycota	24/46	873/736	210/121	105/50	17/7
* Ascomycota & Mucorales	25/47	803/684	119/73	56/32	6/4
* Ascomycota & Chytridiomycota	25/48	758/628	156/75	76/30	11/4
* Ascomycota & Oomycota	25/47	719/615	128/75	57/29	6/2
+ Saccharomyceta	21/42	940/793	244/140	131/66	24/12
+ Leotiomyceta	17/34	1192/1016	416/270	234/136	60/29
+ Dothideomycetes	1/2	2307/2164	1542/1276	1012/872	676/571
+ Eurotiomycetes	14/23	1359/1266	565/475	344/270	162/118
+ Leotiomycetes	2/2	1829/2073	1098/1235	761/830	471/539
+ Sordariomyceta	7/9	1585/1500	727/630	466/397	228/182
+ Sordariomycetes	6/7	1640/1605	775/734	526/493	285/264
+ Saccharomycetes	8/8	1373/1373	609/609	461/461	223/223
+ Schizosaccharomycetes	2/4	1692/1648	1125/1060	862/811	597/507
+ Basidiomycota	2/4	1560/1391	823/639	509/392	264/169
+ Chytridiomycota	1/2	1802/1570	1265/891	792/607	499/356

[1]The number of accession present at least once in each organism within the given taxonomic set.
[2]The number of accessions which appear only once per organism.
[3]The number of accessions with two or more conserved blocks for primer design.
[4]Number of accessions for which forward and reverse primer design succeeded.

Single Pair or Multiple Pairs of Primers for the Barcodes

At the outset of our analysis we expected that it would be possible to design a small set of PCR Primer pairs to amplify all the members of the Fungi even if it was not possible to identify a single pair of PCR primers for this purpose. PCR primer cocktails have been employed to facilitate amplification of *cox1* barcodes from unknown fishes [30], however, as far as we are aware the use of multiplex PCR reactions has not been explored in the fungal barcode context. Thus, despite identifying candidate Pfam accessions amenable to amplification using a single pair of primers, we still sought out regions within which a set of primers might be applied. This was accomplished by identifying the gene regions for which PCR primer design succeeded within each of the lower-level clades of an overarching high-level taxa (Table **3**). The taxonomic capabilities built into the pipeline made it easy to target different taxonomic levels and explore this possibility. We refer to such accessions as multiplex PCR accessions because the set of PCR primers has the potential to enable a multiplex PCR reaction at the higher-level.

These multiplex PCR accessions expand the set of candidate barcode regions (Table **3**). For each of the listed high-level groups, the specified sub-taxa were processed by the pipeline and the common accessions for which PCR primer design succeeded are reported for each of the initial and full data sets. For example, primer pairs were designed for 7 accessions within the initial Dikarya-level set and for 1 accession within the full set of organisms (Table **2**), while two additional multiplex PCR candidate accessions were identified in the initial Dikarya-level set and 1 was identified in the full data set (Table **3**). This provides a total of 9 and 2 Dikarya-level candidates respectively within the initial and final data sets when the single-pair and multiplex primer results from Tables **2** and **3** are combined. At lower taxonomic levels, the increase in potential targets is even greater, corresponding to the availability of additional candidate loci, for example 9 and 10 additional candidates were identified at the Ascomycete-level in the initial and full dataset respectively (Table **3**).

Table 3. Additional Pfam Accessions in the Initial/Full Data Set where Primer Design Failed in the Higher-Level Taxa that Might be Amplified by a Combination of Primer Pairs from Each Sub-Taxa

High-Level Taxa	Sub-Taxa	Common	Multiplex
Ascomycota	Dothideomycetes, Eurotiomycetes, Leotiomycetes, Saccharomycetes, Schizosaccharomycetes, Sordariomycetes	26/17	9/10
Dikarya	Ascomycota, Basidiomycota	9/2	2/1
Dikarya + Mucorales	Ascomycota, Basidiomycota, Mucorales	3/0	1/0
Dikarya + Chytridiomycota	Ascomycota, Basidiomycota, Chytridiomycota	7/2	2/1
Dikarya + Oomycota	Ascomycota, Basidiomycota, Oomycota	6/1	2/1
Fungi	Ascomycota, Basidiomycota, Chytridiomycota, Mucorales	3/0	1/0
Eukaryota	Ascomycota, Basidiomycota, Chytridiomycota, Mucorales, Oomycota	3/0	1/0

Taxonomic Breadth of Candidate Barcode Loci and Primers

To test the potential breadth of the Ascomycete- and Fungi-level barcode candidates these groups were processed in combination with the more distant Zygomycete and Chytridiomycetes (basal Fungal lineages) and the completely different Oomycota (a relative of brown algae and diatoms). The addition of the Zygomycetes to the Dikarya resulted in the loss of five accessions in the initial data set and the single Dikarya-level candidate accession in the full data set (PF00162). Similar results were observed for the Oomycota, however, primer design still succeeded for PF00162 when the Chytridiomycetes were included (Table **4**). The conservation of PF00162 in the more distant Chytridiomycetes suggests that this accession should have good applicability within the Dikarya and may also have applicability in lineages considered more basal. Repeating this process for the Ascomycetes within the full data set, four Pfam accessions conserved within the Ascomycetes and Zygomycete/Chytridiomycete subsets, and two within the Ascomycete and Oomycota subset were found (Table **2**). Like PF00162, the corresponding accessions are expected to have good applicability within and some applicability outside the Ascomycete because they are conserved within their more distant relatives. A consequence of including more distant relatives is that fewer primer candidates are identified for accessions that are maintained within the broader set of organisms. For example, there are 15 forward and 13 reverse primer candidates for PF00162 within the Ascomycetes compared with 12 forward candidates and 9 reverse primer candidates when the Chytridiomycetes are included. This general trend holds for all the candidates within the expanded Ascomycetes and Dikarya subsets (data not shown).

Function of the Candidate Barcode Regions

The protein families for which degenerate primer pairs were identified at the Ascomycetes and higher taxonomic levels are described in Table **4**. The level reported for each accession defines the broadest taxonomic set within which primers were designed for the target Pfam accession and whether or not the accession requires a single primer or a set of primers. The Pfam and Interpro identifiers, Interpro type and Name, and Gene Ontology Biological Process are provided for each accession, and the number of forward and reverse primers for each target is specified. This is the number of forward and reverse primers designed for this broadest taxonomic group, and as described above, the values will generally increase if a narrower taxonomic group is considered. The composition of the sets examined to identify multiplex candidates is available in Table **3**. The Eukaryota- and Fungi-level targets required the removal of organisms from the full dataset as described earlier. To the best of our knowledge, these regions have not been previously explored as candidate regions for barcoding or phylogenetics.

Are there biological reasons that the candidate genes are conserved at different phylogenetic levels? To explore this, we examined annotations associated with the Pfam accessions in Table **4**. In general, one might expect housekeeping genes, those genes which provide a necessary function within the cell, to have a higher chance of being universal than genes involved in secondary metabolism or other processes that might characterize narrow taxonomic groups. Within our nine candidate regions, three are associated with DNA/RNA processing, two are associated with ATP synthesis and hydrolysis, one is associated with purine biosynthesis, another is involved in protein repair, and the remaining two have a variety of functions depending on the protein containing the domain (Table **4**). Based on these functional annotations, all nine candidates seem to clearly fall within the "housekeeping" category.

Searching for Already Known Potential Barcode Regions

It is notable that the standard accepted or proposed barcoding markers (*coxI* and ITS, or the AFTOL genes *rpb2*, and *tef1-α*) were not identified, nor were primers automatically designed, as part of this analysis. The process targets nuclear protein coding genes, which precludes the possibility of identifying primers for ITS (ribosomal) or *coxI* (mitochondrial). It is not surprising that an automated process was unable to design primers for *rpb2* and *tef1-α* because robust universal primers have yet to be identified by other means for these targets. However, not only were the loci not identified with candidates barcode primers, they did not appear anywhere in the output of the initial analysis. One possible explanation for these results reflects the fact that Pfam protein families do not necessarily correspond to full protein sequences; they often refer to shorter domains present in various classes of proteins. Thus, it is possible that

Table 4. Pfam Protein Families for which Candidate Forward (F) and Reverse (R) PCR Primers were Designed at the Ascomycete-Level or Higher and Multiplex Candidates at the Dikarya-Level or Higher

Level	Single/Multiplex[3]	Pfam/Interpro ID	Interpro Type	Interpro Name	GO Biological Process	Primer Candidates (F/R)
Eukaryota [1]	single	PF00218/IPR013798	Domain	Indole-3-glycerol phosphate synthase (IGPS)	Varies	6/5
Eukaryota [1]	single	PF08572/IPR013881	Domain	Pre-mRNA-splicing factor 3 (PRP3)	nuclear mRNA splicing, via spliceosome [2]	2/2
Eukaryota - Mucorales	multiplex	PF00731/IPR000031	Domain	Phosphoribosylaminoimidazole carboxylase, core	'de novo' IMP biosynthetic process	50/52
Fungi [1]	multiplex	PF08145/IPR012953	Domain	BOP1, N-terminal	rRNA processing	70/71
Dikarya + Chytridiomycota	single	PF00162/IPR001576	Family	Phosphoglycerate kinase (PGK)	Gluconeogenesis, Glycolysis	12/9
Ascomycota + Chytridiomycota	single	PF01625/IPR002569	Family	Peptide methionine sulphoxide reductase (MsrA)	Cellular response to oxidative stress	4/3
Ascomycota + Mucorales + Oomycota	single	PF02919/IPR008336	Domain	Eukaryotic DNA topoisomerase I	DNA topological change	5/4
Ascomycota + Mucorales	single	PF08235/IPR013209	Domain	LNS2, Lipin/Ned1/Smp2	Varies	7/7
Ascomycota + Mucorales	single	PF01992/IPR002843	Family	Vacuolar ATPase (vATP)	ATP hydrolysis	6/6

[1]Primers were derived from a subset of the final data set.
[2]GO Biological Process annotations from the Saccharomyces Genome Database (SGD) rather than the Interpro database.
[3]See Table **3** for the composition of the sets used to identify multiplex target accession.

the *rpb2* and *tef1-α* sequences matched a Pfam protein family corresponding to a protein domain that does not obviously indicate the source protein. Fortunately, Pfam is integrated with the Interpro database, which differentiates between "domain" and "family" type accessions. Within the Interpro database a "family" accession covers all domains in the matching proteins, covers more than 80% of the proteins length and has no adjacent domains, whereas a "domain" accession consists of a conserved sub-unit. Through the interpro database it is possible to link sibling Pfam accessions that correspond to Interpro "domain" type accessions that are contained within multi-domain proteins or Interpro "family" accessions.

For example, Interpro accession "IPR004539 Translation elongation factor EF1A, eukaryotic/archaeal" (Fig. **3**) is a "family" type accession, which has no corresponding Pfam family, but seems to best correspond to *tef1-α*. The Interpro accession IPR004539 contains three other accessions: IPR000795 Protein synthesis factor, GTP-binding; IPR004160 Translation elongation factor EFTu/EF1A, C-terminal; IPR004161 Translation elongation factor EFTu/EF1A, domain 2 (Fig. **3**), which correspond to Pfam accessions PF00009, PF03143, and PF03144, respectively. However, no matches were reported to these accessions within the initial analysis at any of the processed taxonomic levels. The Interpro record indicates that the accession is known to be present in 819 Fungal protein sequences in the Pfam database, and we know from the sister analysis with Robert *et al* that *tef1-α* appears to contain many conserved regions within the Fungi, so what happened to the *tef1-α* sequences in this analysis? Using a manual BLAST search with sequences from GenBank, *tef1-α* sequences were identified in the source data set and a handful of these were used to search the Pfam reference database used by CARMA. These input sequences matched sequences contained in Pfam protein family PF00009 (Elongation factor Tu GTP binding domain), which corresponds to Interpro entry IPR000795 (Protein synthesis factor, GTP-binding), a sub-domain of IPR004539 (Translation elongation factor EF1A, eukaryotic/archaeal). This confirmed that *tef1-α* sequences are present in the source data and that meaningful matches should be expected as part of the pipeline. This process was repeated for *rpb2*, which also had good quality matches to PF00562; which corresponds to IPR007120, a domain inside IPR015712 (DNA-directed RNA polymerase, subunit 2), which means that matches to *rpb2* should also have been located. Further investigation revealed that *tef1-α* and *rpb2* matches were excluded because of a filter in CARMA that excluded very common Pfam accessions (discussed further below).

Removing this filter and re-running the analysis for *Schizosaccharomyces pombe* (selected for its small size), confirmed that the *tef1-α* annotated transcript sequences matched to PF00009 and *rpb2* annotated sequences matched PF00562. These two regions appear to be multi-copy in *Schizosaccharomyces pombe*, which rules them out as universal barcode candidates for all groups containing the *Schizosaccharomyces*, although they could be usable for groups which exclude *Schizosaccharomyces*. After reprocessing the excluded Pfam accessions for all organisms, matches were found for the *tef1-α* Pfam protein families: PF00009 (all organisms), PF03143 (11 organisms) and

PF03144 (9 organisms); and matches were found for several of the *rpb2* Pfam protein families PF00562 (all organisms), PF04563 (57 organisms), PF04561 (1 organism) and PF04565 (1 organism). The universal Pfam protein families PF00009 (*tef1-α*) and PF00562 (*rpb2*) are multi-copy in all of the organisms (data not shown), and thus were excluded as candidates by this analysis.

Use of Conserved Domains Versus Full Proteins

The Pfam families may correspond to domains rather than full protein sequences; indeed 6 of the 9 candidate Pfam accessions identified in this analysis correspond to Interpro "domains" rather than "families" (Table **3**). This has two implications for this analysis, which are highlighted in the search for *rpb2* noted above. First, a domain can be multi-copy in the organism though the proteins that contain it may be single-copy. This is because proteins with unique functions may share common structural or functional elements. Although none of the individual domains matched by *rpb2* are single copy within the organisms, the full *rpb2* protein is thought to be single copy [31], which is generally supported by the annotations on the fungal transcript sequences within the full data (the lone exception being *Coccidioides immitis*, which has two annotated *rpb2* sequences). Second, designing to a single domain within a multi-domain protein restricts the available region within which to attempt primer design. While all of the fungal *rpb2* sequences matched PF00562, matches were also observed other *rpb2* protein domains. Considering a longer sequence containing multiple consecutive domains and any intervening sequence(s) adds potential primer design sites. In future, using the Interpro "family" type annotations, it should be possible to identify consecutive domains in the translated protein sequences and search for primer pairs in the longer concatenated regions if additional, longer or more variable candidate regions are required.

These two considerations suggest that targeting conserved domains in this analysis results in a conservative estimate of the number of potential barcode candidates. Further, while the use of a single domain rather than the full protein limits the number of candidate regions, it also removes domain shuffling as a potential source of PCR primer failure when the primers are applied to a broader sampling of organisms. This may make the primer candidates generated by this analysis more robust than those produced by an analysis focused on full gene/protein sequences. Lastly, the observation that the *rpb2* domains appear to be multi-copy while the full sequence is single copy suggests that care must be taken when designing PCR primers from full gene/protein sequences because the region they target within the gene may not be single copy.

Reprocessing the Pfam Accessions Excluded by the CARMA Filter

The Pfam data set used in this analysis contains 10,340 accessions, of which 613 are present in more than 1,500 known unique sequences within the Pfam database. By default CARMA filters out these accessions. Recall that the purpose of CARMA is to provide a taxonomic profile of metagenomics samples, and for this purpose it is reasonable to exclude common accessions that would provide little or no taxonomic information. However, for the purposes of identifying barcode loci, such common accessions might actually be good candidate regions for universal barcodes. Indeed, all three *tef1-α* domains and all but one (PF04560) of the 7 *rpb2* domains (PF04563, PF04561, PF04565, PF04566, PF04567, PF00562, PF04560) fall within this set of Pfam protein families. However, while these Pfam accessions are extremely common, they are often also multi-copy within the organisms in the Pfam data set. Rather than reprocessing the entire data set, all transcript sequences that had already matched to a Pfam accession were removed from the source data and the process was repeated against a BLAST database containing only sequences from the excluded accessions. The values in Supplementary Table **3** clearly demonstrate that while these common Pfam accessions are frequently universally present in the available fungal genomes, they are rarely single-copy. This suggests that there was little harm in excluding these common loci from our analyses or from subsequent analysis targeting high-level taxonomic ranks.

Resolution of the Candidate Barcode Regions and Phylogenetic Potential

In order to assess the taxonomic power of the candidate targets, the outermost degenerate primers were located on the nucleotide alignment of the input sequences for each of the

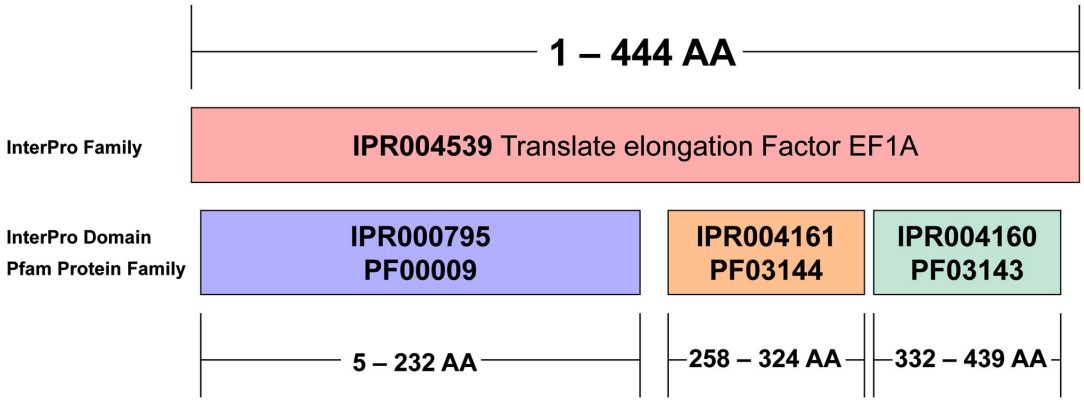

Fig. (3). Structure of IPR004539 Translation elongation factor EF1A, eukaryotic/archaeal. This accession is an Interpro "Family" type accession, which indicates that it covers all domains in the matching proteins and spans more than 80% of the protein length. Within IPF004539, there are three other domains, IPR000795, IPR004160, and IPR004161. From the representative example proteins in the Interpro record it appears that proteins in the IPR004539 family tend to maintain a linear ordering of the sub-domains.

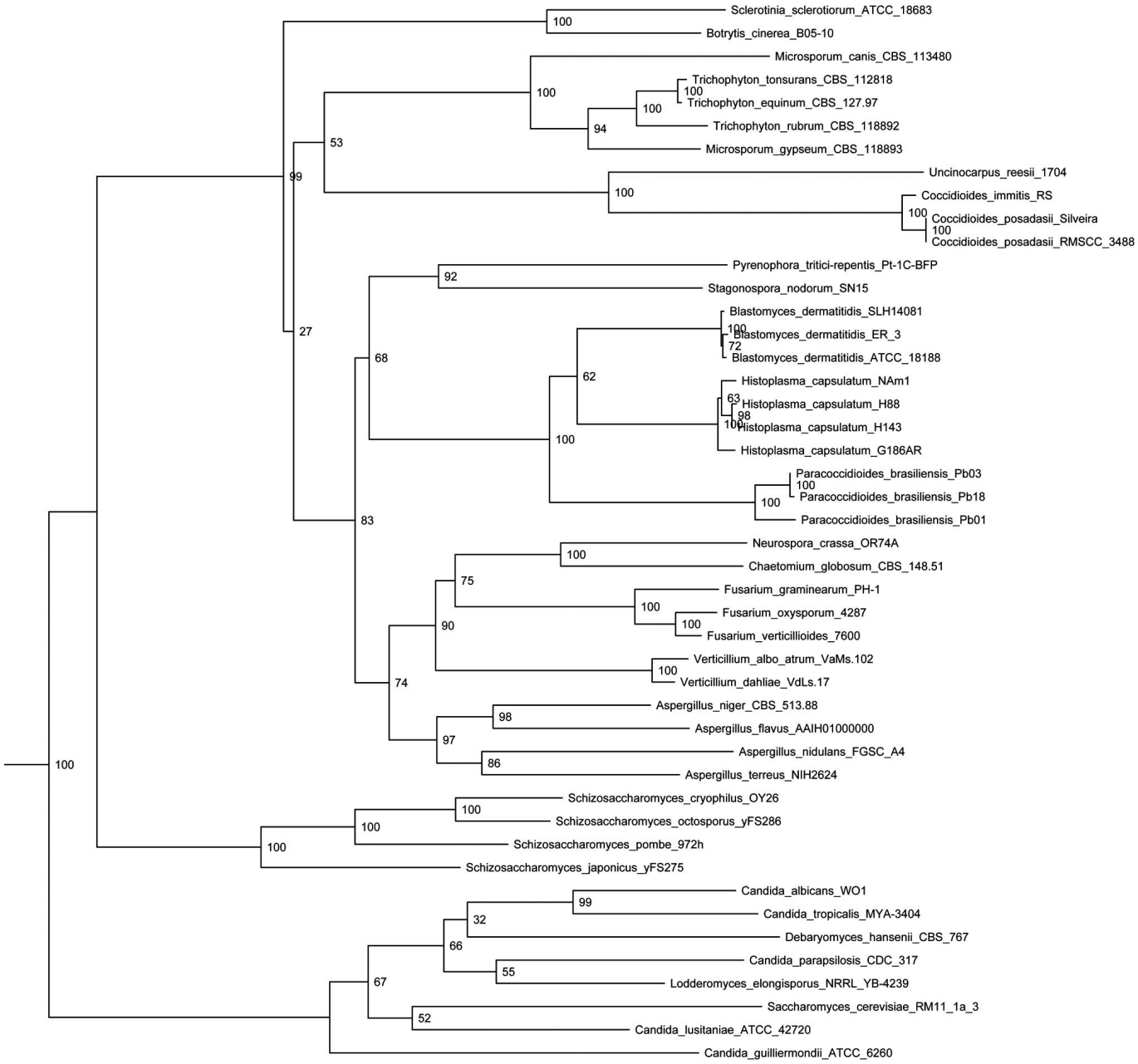

Fig. (4). Neighbour Joining Tree for accession PF00162.11. Accession PF00162.11 provides the ability to differentiate between the Ascomycete species used in this analysis and in some cases can differentiate among strains. Nodes are labelled with bootstrap values (n=1000).

Ascomycete and higher-level candidate regions. This required that the source nucleotide sequences be translated and aligned so that the primer locations could be located using the information in the MABLOCK output file. The alignments were then converted back to nucleotide alignments and the amplified region was excised. The resulting trimmed alignment was used as input to create a bootstrapped (n=1000) neighbour joining trees such as that for PF00162 (Phosphoglycerate kinase) which is presented in Fig. (**4**). This tree demonstrates that sequence differences exist within the region capable of differentiating between all the species and several of the strains within in the Ascomycete subset.

In order to more precisely compare the barcoding potential of the nine regions within the Ascomycetes with the known barcode regions (*coxI*, *rpb2*, and *tef1-α*), the average between species pairwise distances for each of the Genus-level targets containing multiple species and/or strains was prepared (Table **5**). Distances for the standard barcode loci were calculated based on the alignment of the full translated coding sequences, and then trimmed as detailed in the Materials and Methods section. For multi-

Table 5. Within Genus Jukes-Cantor Distance of Candidate Ascomycetes Targets and Known Barcode Targets Compared to the Average JC Distance of Candidates. Distances in Green are Greater than the Average Distance between the Candidate Loci while Distances in Red are below the Average Distance between the Candidate Loci

Genus	Candidates										
	Average	PF01625	PF00218	PF08235	PF02919	PF00731	PF01992	PF00162	*rpb2*	*tef1-α*	*coxI* [3]
Candida	0.365	0.515	0.429	0.390	0.362	0.278	0.334	0.243	0.218	0.088	
Schizosaccharomyces	0.305	0.362	0.350	0.320	0.316	0.318	0.317	0.152	0.227	0.112	
Aspergillus	0.242	0.276	0.280	0.257	0.250	0.228	0.222	0.179	0.208	0.112	
Microsporum	0.169	0.198	0.137	0.132	0.179	0.209	0.192	0.137	0.131	0.092	0.073
Fusarium	0.123	0.112	0.137	0.158	0.148	0.160	0.100	0.049	0.121	0.037	
Verticilium [4]	0.029	0.034	0.030	0.038	0.024	0.020	0.034	0.022	0.027		0.008
Trichophyton	0.027	0.017	0.034	0.032	0.012	0.028	0.038	0.032	0.024	0.009	
Coccidiodes[2]	0.007	0.006	0.000	0.015	0.006	0.006	0.002	0.014	0.004	0.002	
Paracoccidiodes[1]	0.015	0.028	0.011	0.008	0.015	0.014	0.012	0.020	0.011	0.017	
Histoplasma [1,4]	0.012	0.012	0.007	0.009	0.010	0.019	0.017	0.010	0.014	0.011	
Blastomyces[1]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.001	
Average Characters [5]	639	392	307	353	462	312	328	639	2015	787	243

[1]Average distance of all pairwise distance between strains (only a single species in test dataset).
[2]Average distance between *C. immitis* and each of the two available *C. posadasii* strains.
[3]Distance not calculated for all groups owing to difficulties identifying mitochondrial sequences for reference strains.
[4]The RPB2 sequence for *H. capsulatus H143* and Tef1a sequence for *V. albo-atrum VaMs.102* were excluded because they were extremely divergent from the related sequences.
[5]The average number of the characters used to compute the distance matrix for each barcode target.

copy regions, the copy that best fit the alignment was retained. In the event that two or more copies fit the alignment equally well, copies were removed arbitrarily until only a single copy from each organism remained.

Based on the results for the candidate regions (Table **5**), it appears that PF01625 and PF08235 might be the best candidate regions because they have the fewest groups with a below average Jukes-Cantor distance among the candidate loci. However, if we consider only the groups containing multiple species, we see that PF08235 becomes the best candidate with above average distances for all groups but *Microsporum* and PF00218 becomes the second best candidate with above average distances for all groups but *Microsporum* and *Coccidiodes*. We further observe that PF01625, PF00218 and PF08235 are the three most variable targets (in that order) based on the average JC distances for each group. Table **4** shows that these regions also have applicability outside the target Ascomycetes, with PF00218 being universally present in all but two of the organisms included in this study. We see that the known barcode or phylogenetic loci all have a lower than average JC distance within all the groups that contain more than a single species. Table **5** shows that our top three candidate targets should perform better than the known barcode or phylogenetic loci in all circumstances, excepting PF01625 in the *Trichophyton*, where *rpb2* has a greater distance. We also see that of the established loci, *rpb2* has greater between species distances for all groups in this study compared to *tef1-α*.

CONCLUSIONS AND FUTURE WORK

The pipeline is not restricted to Fungi and can work for any data set that is classified by the NCBI taxonomy. The ability to add and extract targets by scientific name or taxonomic identifier allows the user to tailor the analysis to address their target(s) of interest. For example, within a known/suspected polyphyletic group, one might target the full polyphyletic group for analysis and exclude one or more sub-clades to see what impact their exclusion has on the analysis. Alternatively, one might select two clades and analyze them in isolation from the other clades that fall under the same common ancestor. The analysis is fully automatic, which means that it can easily be repeated as the NCBI taxonomy is updated or additional source organisms become available.

While this analysis identified several promising candidate barcode regions, it is not without limitations. As noted below, the majority of the limitations actually provide further avenues for exploration if additional candidate gene regions are required. The limitations were known from the onset of this research, which led to the two very different but complementary bioinformatics approaches that were used to find an optimal barcode (see Robert *et al.* in this special issue).

Introns. The presence of introns is not taken into consideration when selecting primer pairs. This is a consequence of designing PCR primers at the protein level. Introns within primer sites would result in PCR failures. Significant length variation in some members of the target might cause the PCR reaction to fail in those organisms while succeeding in others, giving the impression that the PCR primers are not conserved in all members. During validation we will manually check for introns within the

amplified regions, however, in future it should be possible to add this test into the processing pipeline.

Reliance on annotation. Using the annotated transcript sequences means that the quality of the output is dependent on the quality of the annotation. Alternatives to this include using the annotated genes or fragmenting the genomic sequences and using fragments as input (recall that CARMA was originally intended for the taxonomic classification of metagenomic sequences). This approach may be explored in future if we find that additional candidate gene regions are required. Conversely, using the transcript sequences rather than the predicted/annotated protein sequences required the extra expense of a translated BLAST search to guide their translation into amino acid sequences before assigning the sequences to Pfam protein families. If the quality of the predicted amino acid sequences is deemed to be acceptable, the computation time required by the process could be decreased. Comparison of the results achieved using each of the potential inputs to the pipeline would be an interesting exercise, and it is an approach could also be explored if additional candidate gene regions are required.

Degeneracy of primers & codon usage. Finally, there is no consideration of the nucleotide sequences when designing the primer pairs. In some cases, it might be possible to include fewer degenerate bases in the PCR primers, which are designed using the protein sequences. This could be especially important when deliberately designing primers with limited taxonomic breadth. In such cases excess degeneracy could cause the primer to anneal with templates outside the target group and lose specificity. In future, one could imagine modifying the pipeline such that it searches for primer pairs that are non-degenerate or that are specific to target sets.

In general, we see the expected semi-quantitative results, that higher taxonomic ranks have fewer available candidates while lower taxonomic levels have a wealth of potential targets. We see evidence that it may be possible to design a small set of primers to amplify by multiplex PCR a region where a single degenerate primer pair can not be designed to amplify organisms across the desired taxonomic range. The pipeline has potential as a general purpose tool for taxonomy-based degenerate primer design. That an automated analysis was able to design PCR primers for such a diverse group suggests that a universal barcode region and corresponding primer pairs may exist for Fungi. However, laboratory validation of the candidate primer pairs, which is currently in progress, will be the final and most important step to demonstrate the usefulness of this approach.

ACKNOWLEDGEMENTS

Funding for this project was provided by the Alfred P. Sloan Foundation project "Barcoding the Indoor Mycota" and by the European Community's Seventh Framework Program (FP7, 2007-2013), Research Infrastructures action, under the grant agreement No. FP7-228310 (EMbaRC project). We thank Frank Dehne for his contributions as an honours project supervisor for S. Bilkhu.

CONFLICT OF INTEREST

The authors are unaware of any conflict of interest relating to this work.

SUPPLEMENTARY MATERIAL

This article is also accompanied with supplementary material and it can be viewed at publisher's website.

REFERENCES

[1] P. Hebert, A. Cywinska, S. Ball, and J. deWaard, "Biological identifications through DNA barcodes," *Proc. R. Soc. Lond. B Biol. Sci.,* vol. 270, pp. 313-321, 2003.

[2] P. Hollingsworth, L. Forrest, J. Spouge, M. Hajibabaei, S. Ratnasingham, M. van der Bank, M. Chase, R. Cowan, D. Erickson, A. Fazekas, S. Graham, K. James, K.-J. Kim, J. Kress, H. Schneider, J. van AlphenStahl, S. Barrett, C. van den Berg, D. Bogarin, K. Burgess, K. Cameron, M. Carine, J. Chacón, A. Clark, J. Clarkson, F. Conrad, D. Devey, C. Ford, T. Hedderson, M. Hollingsworth, B. Husband, L. Kelly, P. Kesanakurti, J. Kim, Y.-D. Kim, R. Lahaye, H.-L. Lee, D. Long, S. Madriñán, O. Maurin, I. Meusnier, S. Newmaster, C.-W. Park, D. Percy, G. Petersen, J. Richardson, G. Salazar, V. Savolainen, O. Seberg, M. Wilkinson, D.-K. Yi, and D. Little, "A DNA barcode for land plants," *Proc. Natl. Acad Sci. USA,* vol. 106, pp. 12794-12797, 2009.

[3] S. Ratnasingham, and P. Hebert, "Bold: the barcode of life data system (http://www.barcodinglife.org)," *Mol. Ecol. Notes,* vol. 7, pp. 355-364, 2007.

[4] D. L. Hawksworth, "The magnitude of fungal diversity: the 1.5 million species estimate revisited," *Mycol. Res.,* vol. 105, pp. 1422-1432, 2001.

[5] K. A. Seifert, R. A. Samson, J. R. deWaard, J. Houbraken, C. A. Lévesque, J.-M. Moncalvo, G. Louis-Seize, and P. D. N. Hebert, "Prospects for fungus identification using CO1 DNA barcodes, with Penicillium as a test case," *Proc. Natl. Acad Sci. USA,* vol. 104, pp. 3901-3906, 2007.

[6] H. D. T. Nguyen, and K. A. Seifert, "Description and DNA barcoding of three new species of Leohumicola from South Africa and the United States," *Persoonia – Mol. Phylogeny Evol. Fungi,* vol. 21, pp. 57-69, 2008.

[7] S. R. Gilmore, T. Graefenhan, G. Louis-Seize, and K. A. Seifert, "Multiple copies of cytochrome oxidase 1 in species of the fungal genus *Fusarium*," *Mol. Ecol. Resour.,* vol. 9, pp. 90-98, 2009.

[8] K. A. Seifert, "Integrating DNA barcoding into the mycological sciences," *Persoonia – Mol. Phylogeny Evol. Fungi,* vol. 21, pp. 162-166, 2008.

[9] K. A. Seifert, "Progress towards DNA barcoding of fungi," *Mol. Ecol. Resour.,* vol. 9, pp. 83-89, 2009.

[10] T. White, T. Bruns, S. Lee, and J. Taylor, "Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics," *PCR Protocols: A Guide to Methods and Applications*, M. Innis, D. Gelfand, J. Shinsky, and T. White, Eds., Academic Press, 1990, pp. 315-322.

[11] R. H. Nilsson, M. Ryberg, E. Kristiansson, K. Abarenkov, K. H. Larsson, and U. Kõljalg, "Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective," *PLoS One,* vol. 1, e.59, 2006.

[12] J. W. Spatafora, "Assembling the fungal tree of life (AFTOL)," *Mycol. Res.,* vol. 109, pp. 755-756, 2005.

[13] V. Robert, S. Szöke, U. Eberhardt, G. Cardinali, W. Meyer, K. A. Seifert, C. A. Lévesque, and C. T. Lewis, "The quest for a general and reliable fungal DNA barcode," *Open Appl. Inform J.,* vol. 5, pp. 55-71, 2011.

[14] J. E. Stajich, "Fungal Genome Links," Dec 24, 2010; http://fungalgenomes.org/wiki/Fungal_Genome_Links.

[15] NCBI. "Eukaryotic Genome Sequencing Projects," Dec 24, 2010; http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi.

[16] J. E. Stajich, M. L. Berbee, M. Blackwell, D. S. Hibbett, T. Y. James, J. W. Spatafora, and J. W. Taylor, "The fungi," *Curr. Biol.,* vol. 19, pp. R840-845, 2009.

[17] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, and A. Bateman, "The Pfam protein families database," *Nucleic Acids Res.,* vol. 38, pp. D211-222, 2010.

[18] L. Krause, N. N. Diaz, A. Goesmann, S. Kelley, T. W. Nattkemper, F. Rohwer, R. A. Edwards, and J. Stoye, "Phylogenetic classification of short environmental DNA fragments," *Nucleic Acids Res.,* vol. 36, pp. 2230-2239, 2008.

[19] Pfam. "Pfam Release 23.0," Feb 4, 2011; ftp://ftp.sanger.ac.uk/pub/databases/Pfam/releases/Pfam23.0/.

[20] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.,* vol. 215, pp. 403-410, 1990.

[21] S. R. Eddy, "Profile hidden Markov models," *Bioinformatics (Oxford, England),* vol. 14, pp. 755-63, 1998.

[22] S. R. Eddy. "HMMER: biosequence analysis using profile hidden Markov models," 2010; http://hmmer.janelia.org/

[23] S. Henikoff, J. G. Henikoff, W. J. Alford, and S. Pietrokovski, "Automated construction and graphical presentation of protein blocks from unaligned sequences," *Gene,* vol. 163, pp. GC17-GC26, 1995.

[24] T. M. Rose, E. R. Schultz, J. G. Henikoff, S. Pietrokovski, C. M. McCallum, and S. Henikoff, "Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences," *Nucleic Acids Res.,* vol. 26, pp. 1628-1635, 1998.

[25] S. Pietrokovski, J. G. Henikoff, and S. Henikoff, "The Blocks database--a system for protein classification," *Nucleic Acids Res.,* vol. 24, pp. 197-200, 1996.

[26] M. Gouy, S. Guindon, and O. Gascuel, "SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building," *Mol. Biol. Evol.,* vol. 27, pp. 221-224, 2010.

[27] R. C. Edgar, "MUSCLE: a multiple sequence alignment method with reduced time and space complexity," *BMC Bioinform.,* vol. 5, p. 113, 2004.

[28] A. Rambaut. "FigTree Tree Figure Drawing Tool," 2011; http://tree.bio.ed.ac.uk/software/figtree/

[29] D. L. Swofford, *PAUP*: Phlogenetic analysis using parsimoney (* and other methods). Version 4.0b10*: Sinauer, Sunderland, Massachusetts, 2002.

[30] N. V. Ivanova, T. S. Zemlak, R. H. Hanner, and P. D. N. Hebert, "Universal primer cocktails for fish DNA barcoding," *Mol. Ecol. Notes,* vol. 7, pp. 544-548, 2007.

[31] Y. J. Liu, S. Whelen, and B. D. Hall, "Phylogenetic relationships among ascomycetes: evidence from an RNA polymerse II subunit," *Mol. Biol. Evol.,* vol. 16, pp. 1799-1808, 1999.

Received: January 8, 2011 Revised: February 20, 2011 Accepted: April 22, 2011