Chapter 10

A Bioinformatics Pipeline for Sequence-Based Analyses of Fungal Biodiversity

D. Lee Taylor and Shawn Houston

Abstract

The internal transcribed spacer (ITS) is the locus of choice with which to characterize fungal diversity in environmental samples. However, methods to analyze ITS datasets have lagged behind the capacity to generate large amounts of sequence information. Here, we describe our bioinformatics pipeline to process large fungal ITS sequence datasets, from raw chromatograms to a spreadsheet of operational taxonomic unit (OTU) abundances across samples. Steps include assembling of reads originating from one clone, identifying primer "barcodes" or "tags," trimming vectors and primers, marking low-quality base calls and removing low-quality sequences, orienting sequences, extracting the ITS region from longer amplicons, and grouping sequences into OTUs. We expect that the principles and tools presented here are relevant to datasets arising from ever-evolving new technologies.

Key words: Fungi, Next-generation sequencing, Biodiversity, ITS, Ribosomal, Automated, Pipeline

1. Introduction

Studies of microbial diversity in complex communities present in natural environments have exploded in the last 20 years, in large part due to the advent of the polymerase chain reaction. Analyses of prokaryotic diversity have led the way, starting with the pioneering studies of Woese et al. (1–3) while parallel studies of microbial eukaryotes, including fungi, have blossomed more recently (4–6). Methods for prokaryotic microbes have matured to the point that streamlined, Web-accessible software pipelines are available (7, 8), although these tools continue to advance rapidly (9). Unfortunately, many methods developed for prokaryotes are not appropriate for fungi, due largely to the fact that the ribosomal small subunit (SSU), the linchpin of prokaryotic diversity

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722,
DOI 10.1007/978-1-61779-040-9_10, © Springer Science+Business Media, LLC 2011

analyses, is insufficiently variable to precisely discriminate fungal taxa. Instead, the highly variable internal transcribed spacer (ITS) region of the nuclear ribosomal operon is more appropriate for fungal diversity analyses (10, 11). The ITS region is well suited for species discrimination, but is too variable to be aligned across fungal genera. As a result, many methods to analyze prokaryotic diversity that depend upon global alignments of SSU sequences cannot be applied to fungi. Although efforts to develop standardized protocols and tools to analyze fungal diversity are underway, at present, most research groups either develop their own sets of tools or are confounded by the lack of tools.

Sequence-based studies of microbial biodiversity face two diametrically opposed technical challenges. The first challenge is posed by the overestimation of diversity that arises from artifactual DNA sequences. Sequence artifacts arise through a number of mechanisms, the most common being mistaken base calls derived from low-quality sequence reads (12) and sequences derived from two different organisms as a result of chimera formation during PCR (13, 14). Taq error and pseudogenes can also contribute to false diversity (15). Thus, strict quality controls and careful analyses are critical for accurate assessments of diversity. The tremendous diversity of microbes poses the second challenge: massive sequencing efforts are required to enumerate a significant fraction of total diversity (16). Depth of sequencing is perhaps most pressing when analyzing prokaryotic diversity, but also applies to fungi and other microbial eukaryotes (4, 6, 17). Adequately large datasets are beyond the reach of manual quality control and analysis. In short, the opposing imperatives of large sequence datasets and strict quality control evoke the need for carefully conceived and implemented bioinformatics pipelines.

Here, we describe steps developed by our lab to process large fungal ITS sequence datasets. We emphasize quality control and the underlying biological principles. The steps are not linked in an end-to-end pipeline, so user intervention is required at multiple steps. However, our approach relies on readily available programs and Web sites, most of which are freeware; unique scripts have been made publicly available through a Web portal. This text assumes no background in bioinformatics; our pipeline obviates command line programs and scripting entirely. Our target audience is those new to sequence-based analysis of fungal diversity rather than the experienced practitioner.

Examples are given from our experiences with high-throughput Sanger sequencing of clone libraries generated from PCR of soil DNA extracts. Although Sanger sequencing is a somewhat dated technology given the current explosion of next-generation sequencing (18), we hope that the principles and tools discussed here will be useful regardless of sequencing methodology. Our bench methods have been presented elsewhere (19–22).

In brief, we amplify a fungal nuclear ribosomal region spanning ITS1, the 5.8S, ITS2, and roughly 700 bp of the large subunit (LSU) using the primers ITS1-FL and tagged versions of TW13 (20, 23). By obtaining LSU data, we are able to infer the broad relationships of novel fungi that lack closely related, known fungi in public databases, which is difficult using the ITS alone.

2. Methods

2.1. Chromatogram Processing

Quality scores ("Q scores") for each base call are a critical element of any automated sequence cleanup and analysis pipeline. In the past, sequence reads from the same template, clone, or amplicon were assembled, and the chromatograms inspected manually to insure sequence quality, to correct base calls, and to remove regions of low confidence. This approach is clearly untenable with large datasets, and automated approaches that utilize aspects of chromatogram peak height and shape to determine base confidence values were elaborated in the early 1990s to support the growing needs of genome sequencing efforts (24, 25). Researchers carrying out phylogenetic and biodiversity analyses have adopted these automated quality control approaches slowly, but today's massive datasets are forcing the transition. Due to the statistical nature of the derived base call confidence values, there is less need for multiple coverage of a given sequence region (e.g., the traditional bidirectional and completely overlapping reads often used for phylogenetic studies). For example, if the probability that a base call is incorrect is below 1/100th of 1% (phred score of 40), there is little need to cover the same base from an additional read. On the other hand, our lab routinely generates two reads per clone in order to obtain adequate coverage of the 1,200–1,500 bp region. In this case, an assembly step is still required in order to join these overlapping reads.

We use the program CodonCode Aligner (http://www.codoncode.com/aligner/) to carry out initial sequence assembly and processing. Aligner provides a Mac OSX graphical user interface (GUI) for the popular freeware package phred/phrap/consed that is written in C (http://www.phrap.org/phredphrapconsed.html). Phred is a basecaller that includes the key feature of determining a confidence score ("phred score") for each base call (24, 25). A phred score of ten implies that there is a one in ten chance that the base call is incorrect. A phred score of 20 implies a one in one hundred chance of an incorrect base call, and so on. Analogous scores can now be generated with pyrosequencing platforms, despite the rather different patterns of basecall errors in comparison to Sanger sequencing (26). Phrap carries out assembly of reads into "contigs" utilizing phred scores to derive the

best supported consensus sequence from overlapping reads (27). Usually, assembly refers to the joining of numerous reads from random shotgun sequencing of genomes into consensus sequences spanning contiguous regions of the genome. However, in our procedure, we simply wish to assemble the reads derived from a single clone. Usually, these are two paired-end reads obtained using the vector primers flanking the inserted PCR product. With small numbers of clones, the reads to be assembled together can be selected manually in Aligner. However, this is not practical for large datasets. Fortunately, phrap and Aligner offer a variety of options for automatically selecting sets of reads to be joined based on shared elements within the read names. Note that pyrosequencing circumvents the cloning step and creates reads in only a single direction from a particular template (paired-end read technologies are being developed, but are not currently available for the sequencing of 400–800 bp PCR products), meaning that assembly is not required in the analysis of pyrosequence datasets. This is also a good point at which to eliminate short reads, most of which likely represent either (1) poor quality sequences, or (2) sequences of undesired clone inserts, particularly primer-dimers. An appropriate cutoff for sequence length after assembly depends on the locus. We typically use 200 bp as the cutoff after trimming the ends.

1. It is often necessary to rename your sequence reads before importing the chromatograms to a base caller/assembler in order to allow automated joining of reads arising from the same clone (see Note 1).
2. CodonCode aligner is able to instruct phrap to assemble reads that have a shared component of the name that is set off by a specified delimiter (often dot "." or underscore "_"). This component that unites your target reads must not be found in any other reads within the Codoncode Project; consult the Codoncode Help for more detail.
3. Once your reads are named properly, import all of the reads you wish to assemble into Aligner using the File → Import → Add Folder command.
4. Save the project with an appropriate file name and remember to save after every step along the way, especially before and after sequence assembly.
5. Trim the dirty ends of the reads using Sample → Clip Ends. The default clipping settings are usually adequate, as we will further trim ends and deal with low-quality base calls in subsequent steps.
6. Remove sequences with low numbers of high-quality base pairs. Sort sequences by clicking on the "quality" column heading, then visually inspect chromatograms for sequences with low-quality scores. Choose your threshold (<300 quality

bases is often used), select all the sequences you wish to remove, then navigate to Edit → Move to Trash.

7. Remove vector sequences using Sample → Trim Vector. Note that you must specify which cloning vector you are using in the Vector Trimming settings under CodonCode Aligner → Preferences. As with end trimming, there are later opportunities to remove vector regions that were missed by Aligner.
8. Save your project, and then assemble sets of reads derived from a single clone using the Contig → Assemble with Options command. Click the radio button next to "assemble in groups." Note that you need to click the button "Define Groups" to instruct Aligner how to recognize reads that should be joined. Aligner parses components of the read names between delimiters from left to right. Click the "Preview" button to confirm that your definitions accurately specify the sets of reads that should be assembled.
9. It is likely that some reads will not assemble into contigs due to poor sequence quality, insufficient read overlap, or other problems. We generally keep these reads because the phred scores allow us to screen out poor quality sequence and keep high-quality sequence, even for regions lacking bidirectional coverage.
10. Save your project again, and then export both your sequences and phred scores. Select all the contigs created at the assembly step, then use File → Export → Consensus Sequences as Single File command, making sure to check the box for "export quality scores" (see Note 2).
11. You need to separately export any unassembled reads by selecting them and using File → Export → Sequences.
12. Assuming that you had some unassembled reads, you should now have four text files: (1) a .contig file of the contig sequences, (2) a matching .qual file of the contig quality scores, (3) a file of the unassembled sequences, and (4) a file of the unassembled sequence quality scores.
13. For simplicity in downstream steps, combine the contig and unassembled .fasta files into one file. Also combine the corresponding .qual files.
14. From this point forward, a series of text files is used instead of the chromatogram data. It is important to keep these files in a simple "txt" format – do not open them in Word and save them in a default Word format, as the various bioinformatics programs are not able to open and read Word documents. On a MacIntosh platform, we use TextEdit and TextWrangler to work with these text files. WordPad can be used on the PC platform.

2.2. Masking

The next goal is to permanently mark low-quality base calls so that questionable bases are not evaluated as reliable sequence data. At this point, the phred base call quality scores are essential. We use an "in-house" Perl script to change consensus bases with phred scores below a certain threshold, usually 20, to Ns or to lowercase letters.

1. Direct your browser to our Fungal Metagenomics Web portal at http://www.borealfungi.uaf.edu/. Under Search ITS, click on the "Mask" link.
2. Submit your sequence file and corresponding file of phred scores using the upload links on the Web page. Set your desired phred score threshold and decide whether to convert low-quality bases to Ns or to lowercase letters (see Note 3).
3. Download the resulting "masked" files from the portal. The process is repeated for each pair of text files containing sequences and corresponding phred scores (see Note 3).

2.3. Finding Tags

For automated, high-throughput sequencing of clone libraries, we pool many samples into single-clone libraries. To attribute clone sequences to source samples after sequencing, we add 10-bp "tags" to the PCR primers that uniquely identify each sample in a pooled clone library (20). To identify these tags after sequencing, we utilize an approach that allows some sequence error within the tagged primer region, yet accurately assigns sequences to the correct source sample.

1. Direct your browser to our Fungal Metagenomics Web portal at http://www.borealfungi.uaf.edu/. Under "Search ITS," click on the "Tag Finder" link.
2. Submit each of the masked files together with a text file listing all of the primer sequence tags to the appropriate upload boxes. Additional instructions are provided on the Web site.
3. Open the output files in a text editor, and rename reads to include the information from the tags. This can be accomplished easily in Textedit by searching for ">" with the Edit → Find command, then placing "*>sample_name_or_tag_*" in the "Replace with" box and clicking "Replace All." The greater than sign occurs at the beginning of each sequence name when the sequences are in fasta format, and provides a convenient key for adding or changing components of sequence names. Similar commands are available in TextWrangler and Wordpad.

2.4. Sequence Orientation

The Invitrogen TOPO TA pcr4 cloning vector is not directional, meaning that the PCR product can be cloned in either of the two possible orientations. To obtain high-quality sequence reads through the primers, which contain the critical tag region, we

sequence with vector primers. A given vector primer may read into either the ITS or the LSU end of a particular amplicon, meaning that we must analyze the reverse complement of any contigs that are in the wrong orientation with respect to the ribosomal operon. We accomplish this by comparing our sequences to a series of short motifs corresponding to conserved regions of the SSU, 5.8S and LSU. If a strong match is found to a particular motif, the input sequence is returned unaltered. If a match to the reverse complement of the motif is found, the input sequence is returned as the reverse complement. If no strong match is found, the input sequence is returned in a separate file. The motifs are used consecutively until nearly all sequences are in the proper orientation. In our experience, the few sequences for which no strong matches to our motifs are found are usually nonfungal, and should be discarded from the dataset.

1. Submit fasta files of sequences to the Orient page of our Fungal Metagenomics Web site, http://www.borealfungi.uaf.edu/.
2. Download the resulting "oriented.contigs" and "orientation_unknown" files.
3. Analyze any remaining unoriented sequences for the presence of nonfungal sequences and check their orientation by performing a standard nucleotide BLAST search on the NCBI Web site (http://blast.ncbi.nlm.nih.gov/Blast.cgi, click nucleotide blast, then check "others" under Database).

2.5. Splitting ITS from LSU

Our amplicons span the entire ITS plus about 700 bp of the LSU. However, species-level discrimination of fungi using percent identity thresholds is currently based on the highly variable ITS regions, rather than the LSU. Hence, we must split ITS from LSU prior to carrying out clustering of our clones into operational taxonomic units (OTUs).

1. Align your cleaned sequences using a fast multiple sequence alignment program, such as Clustal (28) or Muscle (29); there are numerous Web servers for these programs, e.g., http://www.ebi.ac.uk/Tools/clustalw2/index.html.
2. For Clustalw, we use the following nondefault settings: -gapext = 1; transweight = 0.2; -pwgapext = 1; apply –kimura.
3. Convert the alignment from clustal format to fasta format using the public Web tool Format Converter at http://hcv.lanl.gov/content/sequence/FORMAT_CONVERSION/form.html (see Note 4).
4. Open the alignment in an editor. For Mac OS, we recommend SeAl (http://tree.bio.ed.ac.uk/software/seal/) and for Windows, we recommend BioEdit (30) http://www.mbio.ncsu.edu/BioEdit/bioedit.html).

5. In SeAl, use Alignment → Find to locate the conserved motif of the ITS4 primer site (reverse complement): GCATATCAATAAGCGGAGG. Under the menu for Treat Gaps, select "as if removed from sequence." If using Bioedit, the analogous menus are Edit → Search → Find.
6. When properly aligned, large blocks of downstream sequence can be selected and deleted (see Note 5).
7. The 5′ end of the sequences should be well aligned in the forward primer ITS1-F region: CTTGGTCATTTAGAGG AAGTAA (see Note 6).
8. This is also a convenient opportunity to remove the forward primer sequence, in our case ITS1-F. Since primers do not necessarily agree with the template sequence of the organism, base calls from primer regions should not be included in sequence submissions to GenBank. This step also ensures that any remaining vector sequence is removed.
9. Select blocks of sequence from the 5′ end to remove everything up to the 3′ end of ITS1-F (or the primer used in your studies).

2.6. Additional Sequence Cleanup

At this point, there are likely to be a number of relatively low-quality sequences remaining in the dataset. This is a good point at which to identify and remove these sequences.

1. If you used lowercase, rather than N, to replace low-quality bases at the masking step (2.2), change the lowercase bases to Ns. Open the file with Textedit, open the Find dialog and unselect the Ignore Case box. Search consecutively for "a," "c," "t," and "g," replacing with "N" in each case using the Replace All command. Beware of any a, c, t, or gs in your read names, as these are also replaced.
2. With Sanger sequencing, most of the poorer base calls are located at the beginning and ends of the sequence reads. It is helpful to trim the sequences from the beginning and end to remove these low-quality regions. We use the program TrimSeq, which is available as a Web tool at many locations, including http://imed.med.ucm.es/cgi-bin/emboss.pl?_action=input&_app=trimseq. Set the window size to 40 and the percent ambiguity to 5% (or lower), and select "yes" for "trim at the start?" and "trim at the end?".
3. Now that the ends are trimmed, sequences that still have a number of Ns should be deleted from the dataset. We use BioEdit to accomplish this. Open the fasta file of sequences in BioEdit, then use Edit → Search → Find/Replace → Replace With command to replace Ns with gaps (dash "-"). Then use Sequences → Filter Out Sequences Containing → Greater than X% Gaps to delete sequences with greater than 2% gaps.

A more stringent cutoff is acceptable, less stringent cutoffs are not recommended, since a 3% difference would cause two sequences to be placed in different OTUs (see Subheading 2.7). Finally, the gaps must be converted back to Ns using the Edit → Search → Find/Replace command in reverse.

2.7. Grouping ITS Sequences into OTUs

In a large sequence dataset, there are many identical and nearly identical sequences that likely represent the same fungal species. With large datasets, the most practical approach is to group or cluster sequences based upon a percent identity threshold (i.e., the percentage of bases that are identical throughout the entire overlapping region in a pairwise sequence alignment) with a sequence assembly program. For smaller datasets, this can be accomplished directly in sequence editors such as Sequencher, CodonCode Aligner, or Bioedit. However, for large datasets and to achieve greater control over the grouping behavior, we use the program TGICL (31) which carries out a first-pass grouping using BLAST and a second, finer grouping via the genome assembler Cap3 (32). A wide range of percent identity values for the ITS region have been used in various fungal studies, ranging from 90 to 99%. A balance must be struck between lumping discrete species when using a low percent threshold and splitting a single species due to base call errors, polymerase error, and intraspecific variation when applying a stringent threshold. A consensus of 95–98% seems to be emerging in the recent literature (6, 17, 33).

The next task is to reformat the output from Cap3 in such a way that it is useful for subsequent analyses of diversity. Cap3 creates a number of files for each run. The '.singletons' and '.contigs' files contain DNA sequences that are rarely used, since the contig sequences are a consensus of all reads that were grouped in that OTU, and thus represent an artificial sequence, not one that necessarily occurs in nature. The key file is the Cap3.out file, which includes both a list of reads that were grouped into specific contigs and the multiple alignments comprising those contigs. It is worth visually checking several of the multiple alignments to ensure satisfaction with the way Cap3 grouped the input sequences. The list of reads by contig is the most important data for downstream ecological analyses, but must be 'parsed' in order to clean up the file format. We use a freeware program called TextWrangler to accomplish this initial parsing.

1. Submit fasta file of ITS sequences to the OTU_Grouping page of our Fungal Metagenomics Web site, http://www.borealfungi.uaf.edu/.
2. Open the resulting Cap.out files in TextWrangler.
3. Copy the top part of the file containing the list of contigs (not the multiple sequence alignments) into a new text file.

4. In the new text file, replace all instances of "**** Contig" by clicking the "grep" check box and inserting the following text in the Find box under the Search menu: *+\s*Contig*\s, then typing \t*SampleName* in the Replace With box (where SampleName is something of your choosing to indicate the source of the sequence), then selecting "replace all" (see Note 7).
5. Remove unneeded spaces by unclicking the "grep" box and simply placing a space in the Find box and deleting all contents from the Replace With box, then selecting "replace all."
6. Remove unneeded asterisks the same way spaces were removed, but using * in the Find box.
7. Remove unneeded read name after "is in" using by clicking "grep" and inserting the following text in the Find box: *sis*\ *sin*\s[0-9A-Za-z_+-?]+\r, then inserting \r in the Replace With box and clicking "replace all."
8. Remove the unneeded "+" and "–" symbols by checking the grep box and inserting the text [+-] in the Find box, emptying any contents from the Replace With box, and selecting "replace all."
9. Save this 'parsed' text file.

Now that you have simplified the Cap3 output, you are ready to import the data into Excel (or a spreadsheet program of your choice).

1. In Excel, select File > Open, and select your parsed Cap3 text file.
2. When the dialog box comes up asking you how to interpret the data, select "Delimited" then click "Next."
3. In the choice of delimiters given in the dialog box, select Tab, and then click OK.
4. At this point, Excel should present you with a spreadsheet in which your read names (clones, sequences) are in the first column, and the OTU groupings or contigs are in the second column. The only remaining task is to fill out the contig names for every row, since they are now only at the top of each sequence group.
5. To fill out the contig information, select a contig name and pull down the column to select up to the the next contig name, then select "fill down" under the Edit menu; repeat for all contigs.
6. We routinely incorporate information about the source site or sample into each read name. Then, in Excel, this information can easily be propagated to a new column using the "text to columns" command under the Data menu. This allows one to sort sequences by sample or OTU.

7. Lastly, the resulting long list of sequences, OTU grouping and site or sample can be summarized by selecting the relevant columns then navigating to Data → Pivot Table Report. Click Next twice, then click Layout, then drag OTU to the "rows" area, site or sample to the "columns" area and OTU to the central "data" area in the Format dialog box to create a useful summary (see Note 8).

3. Discussion

The steps outlined above describe our analysis pipeline from sequence chromatograms to a spreadsheet of OTU abundances across samples with high confidence in the quality of the sequences and groupings. There are many possible directions for downstream ecological analyses, such as diversity analyses with programs like EstimateS (34), community ordination with programs, including PC Ord (35) or Vegan (36), or phylogenetic community analysis with Unifrac (37) or Phylocom (38).

Several important steps are not covered here. One is the identification of the fungal taxa represented by the OTUs. The traditional approach to this is a simple BLAST search of GenBank. However, the large numbers of unidentified and misidentified sequences in the international databases limit the utility of these searches. Pipelines and curated databases that help overcome these issues have been described elsewhere (39, 40). We also provide tools for fungal identification at our Fungal Metagenomics Web site.

Another issue not addressed in this chapter is the detection of chimeric sequences within the dataset. This issue is not trivial, as estimated proportions of chimeric sequences in fungal ITS clone libraries have been as high as 30% (41). In our datasets, we commonly find roughly 3% chimeras. As each chimera is typically a "singleton" with regard to OTU grouping in Cap3, these chimeras obviously lead to a considerable inflation of estimated diversity. Unfortunately, methods developed for prokaryotic 16S sequences are not appropriate for fungal ITS datasets because the prokaryotic methods depend on broad multiple sequence alignments (14, 42). We have developed a multistep BLAST approach to identify chimeras that is not amenable to automation. Due to the complexity of the method, it is beyond the scope of this chapter. We hope that automated detection methods appropriate for fungal ITS datasets will become available in the near future.

With regard to areas for future development, it is clear that there are shortcomings of the "one size fits all" percent sequence identity approach (21, 22). Different clades of fungi have ITS sequences that evolve at different rates. Thus, any arbitrary identity threshold may lump discrete species in some clades while

splitting a single species in other clades. Eventually, we expect that phylogenetic approaches in which well-supported clades are used to distinguish OTUs will supplant current percent identity approaches. However, at present, automated pipelines for fungal phylogenetic analysis of very large datasets are not publicly available.

4. Notes

1. Renaming of your reads can be accomplished using BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) on a PC or the "Replace text in names" script in the Finder Scripts folder for the application Applescript on a Mac. If you have control over the read naming convention, you should not need to rename reads – just use an appropriate convention at the beginning.
2. While the assignment of base call phred scores to bases within a single read follows a strict algorithm developed by Phil Green, there are several ways that these scores can be combined to approximate a phred score for a consensus base call derived from assembled reads. These options can be set in Aligner Preferences. We use the subtract scores for conflicting bases option.
3. If you have incorporated "tags" or "barcodes" in your primers, as described in the next section, it will be helpful to convert low-quality base calls to lowercase rather than N so that the tags can be identified in a higher proportion of your sequences.
4. If this Web tool is unavailable, most versions of the freeware Readseq should be able to perform the desired conversion.
5. It may be convenient at this point to paste this LSU components into a new alignment so that they can be used for subsequent phylogenetic analyses.
6. There are a number of Ascomycetes that have a large and highly variable intron just downstream of the ITS1-F primer (43); if such taxa are present in your dataset, they may cause the ITS1-F region to be located erratically in the alignment. As long as the ITS1-F primer region can be located and removed, these introns should not cause major problems. Because they can be present or absent among isolates of the same putative species, it is reasonable to remove them prior to OTU grouping.
7. If you would like to better understand and perhaps modify these commands, look up Grep in the online TextWrangler Help menu.

8. These steps are for Excel 2000–2004; the menus and commands are organized differently in Excel 2009, but pivot table reports are still available.

Acknowledgments

We thank James Long for writing several of the original pipeline scripts and Dan Cardin for writing the tag-finder script. Niall Lennon and Chad Nusbaum of the Broad Institute, MA, spearheaded high-throughput Sanger sequencing of our fungal clone libraries. Lab members Michael Booth, Robert Burgess, Ian Herriott, Jack McFarland, and Ina Timling have assisted with testing and improving our pipeline and also provided valuable comments on earlier drafts of the chapter. This work was supported in part by the National Science Foundation under grant numbers EF-0333308 and ARC-0632332. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation. This publication was also made possible by grant number 2P20RR016466 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH).

References

1. Fox, G. E., Stackebrandt, E., Hespell, R. B., Gibson, J., Maniloff, J., Dyer, T. A., Wolfe, R. S., Balch, W. E., Tanner, R. S., Magrum, L. J., Zablen, L. B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B. J., Stahl, D. A., Luehrsen, K. R., Chen, K. N., and Woese, C. R. (1980) The phylogeny of prokaryotes, *Science* ***209***, 457–463.
2. Pace, N. R., Stahl, D. A., Lane, D. J., and Olsen G. J. (1985) Analyzing natural microbial populations by rRNA sequences, *ASM American Society for Microbiology News* ***51***, 4–12.
3. Giovannoni, S. J., Britschgi, T. B., Moyer, C. L., and Field, K. G. (1990) Genetic diversity in Sargasso Sea bacterioplankton, *Nature* ***345***, 60–63.
4. Vandenkoornhuyse, P., Baldauf, S. L., Leyval, C., Straczek, J., and Young, J. P. W. (2002) Evolution – extensive fungal diversity in plant roots, *Science* ***295***, 2051–2051.
5. Schadt, C. W., Martin, A. P., Lipson, D. A., and Schmidt, S. K. (2003) Seasonal dynamics of previously unknown fungal lineages in tundra soils, *Science* ***301***, 1359–1361.
6. O´Brien, H. E., Parrent, J. L., Jackson, J. A., Moncalvo, J. M., and Vilgalys, R. (2005) Fungal community analysis by large-scale sequencing of environmental samples, *Appl Environ Microb* ***71***, 5544–5550.
7. Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker, C. T., Saxman, P. R., Farris, R. J., Garrity, G. M., Olsen, G. J., Schmidt, T. M., and Tiedje, J. M. (2001) The RDP-II (Ribosomal Database Project), *Nucleic Acids Res* ***29***, 173–174.
8. DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB, *Appl Environ Microb* ***72***, 5069–5072.
9. Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., and Weber, C. F. (2009) Introducing mothur: open-source, platform-independent, community-supported

software for describing and comparing microbial communities, *Appl Environ Microb* **75**, 7537–7541.

10. 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.
11. Seifert, K. A. (2009) Progress towards DNA barcoding of fungi, *Mol Ecol Resour* **9**, 83–89.
12. Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates, *Environ Microbiol* **12**, 118–23.
13. Meyerhans, A., Vartanian, J. P., and Wainhobson, S. (1990) DNA recombination during Pcr, *Nucleic Acids Res* **18**, 1687–1691.
14. Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., and Weightman, A. J. (2006) New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras, *Appl Environ Microb* **72**, 5734–5741.
15. Valentini, A., Miquel, C., Nawaz, M. A., Bellemain, E., Coissac, E., Pompanon, F., Gielly, L., Cruaud, C., Nascetti, G., Wincker, P., Swenson, J. E., and Taberlet, P. (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach, *Mol Ecol Resour* **9**, 51–60.
16. Sogin, M. L., Morrison, H. G., Huber, J. A., Mark Welch, D., Huse, S. M., Neal, P. R., Arrieta, J. M., and Herndl, G. J. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere", *P Natl Acad Sci USA* **103**, 12115–12120.
17. Buee, M., Reich, M., Murat, C., Morin, E., Nilsson, R. H., Uroz, S., and Martin, F. (2009) 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity, *New Phytol* **184**, 449–456.
18. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z. T., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P. G., Begley, R. F., and Rothberg, J. M. (2005) Genome sequencing in microfabricated high-density picolitre reactors, *Nature* **437**, 376–380.
19. Taylor, D. L., Herriott, I. C., Long, J., and O'Neill, K. (2007) TOPO TA is A-OK: a test of phylogenetic bias in fungal environmental clone library construction, *Environ Microbiol* **9**, 1329–1334.
20. Taylor, D. L., Booth, M. G., Mcfarland, J. W., Herriott, I. C., Lennon, N. J., Nusbaum, C., and Marr, T. G. (2008) Increasing ecological inference from high throughput sequencing of fungi in the environment through a tagging approach, *Mol Ecol Resour* **8**, 742–752.
21. Geml, J., Laursen, G. A., and Taylor, D. L. (2008) Molecular diversity assessment of arctic and boreal Agaricus taxa, *Mycologia* **100**, 577–589.
22. Geml, J., Laursen, G. A., Timling, I., Mcfarland, J. M., Booth, M. G., Lennon, N., Nusbaum, C., and Taylor, D. L. (2009) Molecular phylogenetic biodiversity assessment of arctic and boreal ectomycorrhizal Lactarius Pers. (Russulales; Basidiomycota) in Alaska, based on soil and sporocarp DNA, *Mol Ecol* **18**, 2213–2227.
23. White, T. J., Bruns, T., Lee. S., Taylor, J. (1990) Amplification and direct sequencing of fungal ribosomal RNA Genes for phylogenetics, *PCR protocols: a guide to methods and applications* **42**, 315–322.
24. Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment, *Genome Res* **8**, 175–185.
25. Ewing, B., and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities, *Genome Res* **8**, 186–194.
26. Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W. L., Russ, C., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems, *Genome Res* **18**, 763–770.
27. Gordon, D., Abajian, C., and Green, P. (1998) Consed: A graphical tool for sequence finishing, *Genome Res* **8**, 195–202.
28. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) Clustal-W – improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res* **22**, 4673–4680.
29. Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res* **32**, 1792–1797.

30. Hall, T. A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT, *In*: Nucleic acids symposium series. p. 95–98.
31. Pertea, G., Huang, X. Q., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., and Quackenbush, J. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets, *Bioinformatics **19***, 651–652.
32. Huang, X. Q., and Madan, A. (1999) Cap3: A DNA sequence assembly program, *Genome Res **9***, 868–877.
33. Higgins, K. L., Arnold, A. E., Miadlikowska, J., Sarvate, S. D., and Lutzoni, F. (2007) Phylogenetic relationships, host affinity, and geographic structure of boreal and arctic endophytes from three major plant lineages, *Mol Phylogenet Evol **42***, 543–555.
34. Colwell, R. K., and Coddington, J. A. (1994) Estimating terrestrial biodiversity through extrapolation, *Philos T Roy Soc B **345***, 101–118.
35. McCune, B., Mefford, M. J. (1999) PC-ord. Multivariate analysis of ecological data, version 4(0).
36. Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M. H. (2007) vegan: Community Ecology Package. R package version 1.8-8. Online at: http://r-forge.r-project.org/projects/vegan.
37. Lozupone, C., and Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities, *Appl Environ Microb **71***, 8228–8235.
38. Webb, C. O., Ackerly, D. D., and Kembel, S. W. (2008) Phylocom: software for the analysis of phylogenetic community structure and trait evolution, *Bioinformatics **24***, 2098–2100.
39. Koljalg, U., Larsson, K. H., Abarenkov, K., Nilsson, R. H., Alexander, I. J., Eberhardt, U., Erland, S., Hoiland, K., Kjoller, R., Larsson, E., Pennanen, T., Sen, R., Taylor, A. F. S., Tedersoo, L., Vralstad, T., and Ursing, B. M. (2005) UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi, *New Phytol **166***, 1063–1068.
40. Nilsson, R., Bok, G., Ryberg, M., Kristiansson, E., Hallenberg, N. (2009) A software pipeline for processing and identification of fungal ITS sequences. *Source Code Biol Med* **4**, 1.
41. Jumpponen, A. (2003) Soil fungal community assembly in a primary successional glacier forefront ecosystem as inferred from rDNA sequence analyses, *New Phytol **158***, 569–578.
42. Huber, T., Faulkner, G., and Hugenholtz, P. (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments, *Bioinformatics **20***, 2317–2319.
43. Perotto, S., Nepote-Fus, P., Saletta, L., Bandi, C., and Young, J. P. W. (2000) A diverse population of introns in the nuclear ribosomal genes off ericoid mycorrhizal fungi includes elements with sequence similarity to endonuclease-coding genes, *Mol Biol Evol **17***, 44–59.