\section*{Methods}

Are similarity- or phylogeny-based methods more appropriate for classifying internal transcribed spacer (ITS) metagenomic amplicons?

Teresita M. Porter and G. Brian Golding
Department of Biology, McMaster University, Hamilton, ON L8S 4L8, Canada

\textbf{Summary}

- The internal transcribed spacer (ITS) of the nuclear ribosomal DNA region is a widely used species marker for plants and fungi. Recent metagenomic studies using next-generation sequencing, however, generate only partial ITS sequences. Here we compare the performance of partial and full-length ITS sequences with several classification methods.
- We compiled a full-length ITS data set and created short fragments to simulate the read lengths commonly recovered from current next-generation sequencing platforms. We compared recovery, erroneous recovery, and coverage for the following methods: best BLAST hit classification, MEGAN classification, and automated phylogenetic assignment using the Statistical Assignment Program (SAP).
- We found that summarizing results with more inclusive taxonomic ranks increased recovery and reduced erroneous recovery. The similarity-based methods BLAST and MEGAN performed consistently across most fragment lengths. Using a phylogeny-based method, SAP runs with queries 400 bp or longer worked best. Overall, BLAST had the highest recovery rates and MEGAN had the lowest erroneous recovery rates.
- A high-throughput ITS classification method should be selected, taking into consideration read length, an acceptable tradeoff between maximizing the total number of classifications and minimizing the number of erroneous classifications, and the computational speed of the assignment method.

\section*{Introduction}

The internal transcribed spacer (ITS) region of nuclear-encoded ribosomal DNA (ITS rDNA) is a popular species-level marker in fungal and plant phylogenetic studies (Bruns et al., 1991; Baldwin et al., 1995; Alvarez & Wendel, 2003). ITS rDNA is widely used to facilitate the identification of plant pathogens and ectomycorrhizal symbionts and has recently been proposed as a potential ‘barcoding’ marker for fungi (Seifert, 2009; Begerow et al., 2010). ITS rDNA is comprised of internal transcribed spacer region 1 (ITS1), the 5.8S rRNA gene, and internal transcribed spacer region 2 (ITS2).

As ITS rDNA is often a target in metagenomic studies, an array of resources is available to facilitate sequence-based identifications. The AFTOL and UNITE databases contain ITS sequences from fungi that have been identified by experts (Spatafora, 2005; McLaughlin et al., 2009; Abarenkov et al., 2010). Many of the sequences from these databases are also available in GenBank, which is by far the most widely used resource for classifying ITS sequences. Online tools have also been developed to facilitate sequence-based identification (Nilsson et al., 2005, 2009a; Ryberg et al., 2009; Abarenkov et al., 2010; Koetschan et al., 2010; Taylor & Houston, 2011).

At least two independent studies have suggested that a greater number of short reads captures more diversity than a smaller number of longer reads (Liu et al., 2007; Jumpponen et al., 2010b). Although greater amounts of
sequence diversity are certainly being recovered using next-generation sequencing (NGS) approaches (Buee et al., 2009; Jumpponen & Jones, 2009; Opik et al., 2009; Caporaso et al., 2010; Jumpponen et al., 2010b), and for the first time fungal collector’s curves from large-scale studies are being saturated (Opik et al., 2009; Amend et al., 2010), identifying the species source of this diversity with any measure of confidence is still an issue.

The gold standard for fungal ITS sequence-based classification is based on the concept of phylogenetic species recognition (Hugenholtz & Pace, 1996; Taylor et al., 2000). This approach is commonly used with conserved small- or large-subunit rDNA across taxonomically diverse fungi (O’Brien et al., 2005; Arnold et al., 2007; Porter et al., 2008b; Jumpponen et al., 2010a). As the inclusion of taxonomically diverse ITS sequences into a single alignment is difficult because of high amounts of sequence variation, normally only the highly conserved 5.8S region can be retained for phylogenetic analysis (Pringle et al., 2003; Nagano et al., 2010). More often, ITS sequences are aligned and analyzed among only closely related species (Geml et al., 2009, 2010). Because of the speed and ease of BLAST best-hit analyses (the BH method), this has become a common approach for ITS sequence-based classifications. With the significantly higher throughput of NGS platforms, automated tools such as MEGAN are now being used to classify metagenomic ITS amplicons (Huson et al., 2007; Buee et al., 2009; Amend et al., 2010). To our knowledge there has been no direct comparison among the BH method, MEGAN and phylogeny-based methods for ITS sequence classification, nor has there been a systematic assessment of how sequence length or taxonomic level of classification may affect classification accuracy.

In this study, we compared the performance of similarity- and phylogeny-based methods for classifying ITS metagenomic sequences. We classified partial and full-length ITS sequences using several approaches: the BH method, MEGAN classification, the Statistical Assignment Program (SAP; an automated phylogeny-based taxon assignment method) with neighbor joining, and SAP with Bayesian inference. The results of this study represent a best-case scenario for classifying metagenomic ITS sequences, as we do not consider here the effect of sequencing errors (Kunin et al., 2010; Tedersoo et al., 2010), nor do we consider complications arising from the analysis of chimeric sequences (Hugenholtz & Huber, 2003; Jumpponen, 2007; Nilsson et al., 2010).

Materials and Methods

Assembling an ITS data set

We compiled a control ITS data set using BioPerl (Stajich et al., 2002) modules with a GenBank query: ‘internal tran-

scribed’ AND ‘AFTOL’ NOT ‘uncultured’ NOT ‘sp.’ NOT ‘aff.’ NOT ‘cf.’ (14 February 2011). We targeted submissions from the AFTOL project so as to gather a set of sequences identified by experts where species names reflect the current state of taxonomic knowledge. A total of 531 sequences were retrieved from GenBank, of which 448 putatively complete sequences were retained for further analysis. Each sequence spanned the ITS1, 5.8S rDNA, and ITS2 regions, and we made sure that the ITS1 and ITS2 regions were at least 100 bp long. Any sequences from the flanking ribosomal genes were discarded to focus our analyses on the barcoding region. The complete scientific name and taxonomic lineage were retrieved for each sequence using BioPerl. Although several subregions of ITS rDNA can be targeted by different primer sets (Bellemain et al., 2010), we focused on the most inclusive region targeted in metagenomic studies using the ITS1F and ITS4 primers (White et al., 1990; Gardes & Bruns, 1993; O’Brien et al., 2005; Amend et al., 2010). From our control data set, we generated eight sets of short ITS fragments that were 50, 100, 200, and 400 bp in length, similar to the read lengths produced by current NGS platforms (Supporting Information Fig. S1 and Table S1). Fragments trimmed from the 5’-end of the ITS extend into the ITS1 region and fragments trimmed from the 3’-end extend into the ITS2 region.

Two search scenarios

We tested for the recovery of correct classifications using two different search scenarios. Under the first scenario, query accessions were permitted to be a valid classification result. This allowed us to gauge the suitability of classifying ITS sequences using a reference database known to contain at least one ITS sequence per species. Under the second scenario, we simulated classifying newly generated ITS sequences by systematically excluding the query accession from the search results using a ‘leave one out’ approach similar to that used by Liu et al. (2008). In this situation correct classifications require more than one sequence per species in the reference database. Incorrect classifications would then be attributable to a lack of sequence variation among closely related species, misidentified database sequences, or insufficient database coverage. This is a more realistic scenario that allowed us to gauge the suitability of classifying ITS sequences using a reference database that is potentially incomplete.

BH classification

Using only the top BLAST hit to classify an unknown sequence is known to be potentially misleading (Altschul et al., 1997; Koski & Golding, 2001). Despite this, BLAST analyses are often used as the basis for metagenomic
annotations or as a starting point for classifications using other methods such as MEGAN (Huson et al., 2007). To simulate a high-throughput method for annotating metagenomic amplicon sequences, we used BLAST 2.2.24+ with the blastn algorithm with default parameter settings to search a local installation of the GenBank ‘nt’ database (14 February 2011). We used custom perl scripts to check if the top BLAST hit contained a full scientific binomial; otherwise, the next best hit with a complete scientific name was retrieved. Results were summarized using three measures: recovery was calculated as the percentage of correctly classified queries; erroneous recovery was calculated as the percentage of incorrectly classified queries; and coverage was calculated as the percentage of queries for which a classification (correct or incorrect) could be made. Coverage at different taxonomic ranks can vary depending on GenBank taxonomy, reference database sequence coverage and annotation, as well as classification method. Recovery, erroneous recovery, and the number of unclassified queries sum to 100% at each taxonomic rank. We did not correct for synonyms or anamorph–teleomorph relationships. Using the ‘leave one out’ approach we excluded the query accession from the ‘nt’ database using the ‘negative gilist’ option for each blastn search. We also checked the recovery consistency of 5’ and 3’ fragments that were trimmed from the same full-length sequence.

MEGAN classification

We used MEGAN version 3.9 to parse the raw BLAST results initially retrieved using the BH method (as described in the previous section). MEGAN uses a lowest common ancestor (LCA) algorithm to parse BLAST output and assign reads to the GenBank taxonomic level that presumably reflects the level of sequence conservation (Huson et al., 2007). The LCA settings we used were minimum support = 1, minimum score = 50 (for 50-bp fragments) or 100 (for all other sequence lengths), top per cent = 1.0, and win score = 0.0. We disabled all taxa in the National Centre for Biotechnology Information (NCBI) taxonomy that MEGAN uses except for Eukaryotes to try to avoid parsing insufficiently identified sequences from environmental samples.

SAP taxonomic assignment

SAP automates the process of conducting BLAST searches, compiling homologs, and conducting phylogenetic analyses to automate phylogeny-based taxonomic assignments of ITS rDNA sequences (Munch et al., 2008). After testing numerous variations of parameters to compile the homolog set, we ultimately used the following settings: hits were retained if the local BLAST alignment similarity of the best hit was at least 70%; homologs were compiled that represent at least one phyllum, one class, one order, one family, three genera, five individuals of the same species, and one sequence from each subspecies if possible. We used both the default recommended Bayesian Barcoder algorithm and the neighbor-joining ConstrainedNJ algorithm. SAP is distinguished from the previous methods by providing a measure of statistical support for each taxonomic assignment. We used the default 0.95 Bayesian posterior probability or 95% neighbor-joining bootstrap proportion to filter results considered good taxonomic assignments. To facilitate comparisons with methods that do not filter results according to any measure of support, we also compiled SAP results based on phylogenetic clustering, without enforcing a minimum level of support. We repeated these analyses for the ‘leave one out’ search scenario by using the ‘forceexcludegilist’ option for each query.

Results

ITS data set

As shown in Table 1, the average length of the full ITS region in the control data set was 591 bp. These sequences were of fungal origin, except for one nonfungal sequence from Anoebidium parasiticum (Ichthyosporea). Generally, the Ascomycota (497 bp) and the Glomeromycota (485 bp) had the shortest average ITS sequence lengths, and the Chytridiomycota (726 bp), Kickxellomycotina (833 bp), Blastocladiomycota (983 bp), Entomophthoromycotina (975 bp), and Zoopagomycotina (914 bp) had the longest annotated ITS sequence lengths. A complete list of GenBank accessions used is shown in Table S2.

Full-length ITS classification

Recovery at the genus and species ranks using full-length ITS sequences is shown in Fig. 1. Each method performed

Table 1 Taxonomic and the internal transcribed spacer (ITS) sequence length breakdown for the test data set

<table>
<thead>
<tr>
<th>Taxonomic group</th>
<th>Number of sequences</th>
<th>Average length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ITS1</td>
<td>5.85</td>
</tr>
<tr>
<td>Basidiomycota</td>
<td>284</td>
<td>214.9</td>
</tr>
<tr>
<td>Ascomycota</td>
<td>111</td>
<td>177.9</td>
</tr>
<tr>
<td>Chytridiomycota</td>
<td>24</td>
<td>300</td>
</tr>
<tr>
<td>Mucoromycotina</td>
<td>7</td>
<td>226.6</td>
</tr>
<tr>
<td>Kickxellomycotina</td>
<td>7</td>
<td>315.3</td>
</tr>
<tr>
<td>Blastocladiomycota</td>
<td>4</td>
<td>461.3</td>
</tr>
<tr>
<td>Glomeromycota</td>
<td>4</td>
<td>115.3</td>
</tr>
<tr>
<td>Entomophthoromyocotina</td>
<td>3</td>
<td>210</td>
</tr>
<tr>
<td>Zoopagomycotina</td>
<td>2</td>
<td>341</td>
</tr>
<tr>
<td>Neocallimastigomyocotina</td>
<td>1</td>
<td>198</td>
</tr>
<tr>
<td>Ichthyosporea*</td>
<td>1</td>
<td>199</td>
</tr>
<tr>
<td>Total</td>
<td>448</td>
<td>213.4</td>
</tr>
</tbody>
</table>

*The only nonfungal eukaryote included in this data set.
best under the first search scenario when query accessions were permitted to be a result. The alternate search scenario using the ‘leave one out’ approach greatly reduced the recovery of correct classifications. For each search scenario, recovery was highest using the BH method, followed by SAP using the ConstrainedNJ or Barcoder algorithms (no cutoff enforced). Enforcing a statistical cutoff with the SAP method further decreased recovery. As the results from the ConstrainedNJ and Barcoder algorithms were similar, only results for the ConstrainedNJ method (no cutoff enforced) are presented in subsequent analyses. Query sequences with problematic classifications were compared across methods (Table S3). Using the search scenario that included query accessions, 2% of queries were incorrectly classified by all three methods: the BH method, MEGAN, and SAP ConstrainedNJ. Under the ‘leave one out’ search scenario, when the only representative of a species was excluded as a possible hit, 40% of queries were incorrectly identified by all three methods (Table S4).

Partial ITS classification

Fig. 2 compares the performance of three classification methods using partial ITS sequences. Average values are plotted and error bars indicate the minimum and maximum value from 5’ and 3’ fragments. For each method, recovery increased when classifications were summarized with more inclusive taxonomic ranks, although this came at the price of taxonomic precision. The highest recovery rates were generated with the BH method. Generally, erroneous recovery decreased when classifications were summarized with more inclusive taxonomic ranks. MEGAN showed the lowest rates of erroneous recovery for fragment lengths of 400 bp or shorter. The BH method had the highest coverage at all taxonomic ranks. Coverage, and subsequent recovery values, from MEGAN and SAP were generally lower than for the BH method because short query sequences cannot always be classified by the LCA algorithm in MEGAN or because a homolog set cannot always be compiled with SAP.

In Table 2, classifications using 5’ and 3’ fragments simulated from the same full-length ITS sequences using the BH method with the ‘leave one out’ approach are compared. At the species rank, as fragment size increased, the number of 5’ and 3’ fragments that were both correctly classified increased. Conversely, the number of 5’ fragments alone (or 3’ fragments alone) that were correctly classified decreased. A similar trend was found at the genus rank.

The times required to classify amplicons using the BH method and SAP were also compared (Methods S1, Fig. S2). SAP with the ConstrainedNJ algorithm took < 20 s longer to classify each sequence compared with the BH method. SAP with the Barcoder algorithm, however, took up to 9 min longer per full-length ITS query compared with the BH method.

Discussion

Until now there has been no benchmark available for the classification accuracy of ITS rDNA sequences across classification methods. Liu et al. (2008) tested the effect of using short 16S rDNA read lengths with a ‘leave one out’ approach. They concluded that correct classifications are sensitive to the region of 16S rDNA targeted and that consistently accurate results were found only at the phylum rank for Bacteria. In our study, we showed that > 70% of ITS classifications are correct at the genus rank using the BH method and at the class rank for MEGAN and SAP, but only with longer ITS sequences for the latter.
There is a tradeoff between the significantly higher number of short reads generated by the Illumina platform compared with fewer but longer reads generated by 454 pyrosequencing. In a recent study, short 200-bp bacterial 16S rDNA amplicons were sequenced from soil (Bartram et al., 2011). The advantage of the Illumina paired-end sequencing approach was not only the depth of sampling, but also the ability to assemble forward and reverse reads, a step not normally performed with 454 ITS amplicons. If the short read approach is used with ITS rDNA, the variable ITS1 or ITS2 regions can be targeted to maximize the information content for classification (Nilsson et al., 2009b), at least for fungal groups with average-sized ITS1/ITS2 regions (Feibelman et al., 1994). This approach should be used with caution, however, as Nilsson et al. (2009b) also showed that, when using the BH method, ITS1 or ITS2 sequences hit to the same species as the original full-length ITS sequence only 49% of the time. We also show here that, although recovery rates are similar among 5' and 3' fragments, the species classification of these fragments may differ, particularly for short query lengths.

Strategies for improving ITS classifications include expanding reference databases and improving the taxonomic annotation of sequences already present, as erroneous annotations can degrade the performance of all downstream analyses (Bidartondo et al., 2008). A major improvement towards increasing the number of correct ITS classifications may come from ongoing work to define and formally name molecular operational taxonomic units (MOTUs) known only from environmental sampling studies (Hibbett et al., 2011). The reasons for this are compelling: only a fraction of the 1.5–5.1 million fungal species estimated to exist worldwide have been named (Hawksworth, 2001; O’Brien et al., 2005); only a fraction of all named fungal species are represented by a GenBank sequence (Brock et al., 2009; Hibbett et al., 2011; Nagy et al., 2011); and the proportion of sequence clusters exclusively comprised of unidentified environmental sequences in 2008 and 2009 exceeded the number of clusters from identified samples in GenBank (Hibbett et al., 2011). Examples of phylogenetically divergent lineages represented exclusively or mainly by environmental sequences are the Soil Clone Group I (SCGI) clade (Schadt et al., 2003; Portet et al., 2008a), the Deep sea fungi (DSF) - Group 1 (Nagano et al., 2010), the 'Rozellida' clade (Lara et al., 2010), and small subunit rDNA clades I–V (Vandenkoonhuyse et al., 2002).

Most of the methods we tested here are suitable for processing NGS metagenomic ITS amplicons. BLAST is already used for high-throughput classifications and MEGAN was specifically developed for metagenomic data sets (Huson et al., 2007). In terms of computation time per query, SAP with the ConstrainedNJ algorithm performs similarly to the BH method, whereas SAP with the Barcoder algorithm is probably best suited for use with smaller data sets (Munch et al., 2008). SAP ConstrainedNJ makes phylogeny-based ITS taxonomic assignments tractable even with large data sets. In terms of query sequence length, the similarity-based BH method and MEGAN showed consistent recovery across most fragment lengths. By contrast, the phylogeny-based SAP method performed best with ITS lengths of at least 400 bp. When the largest number of classified queries was the goal, BLAST performed best with our data. When the lowest number of erroneous classifications was more important, MEGAN performed best with ITS fragments of 400 bp or less. Thus, whether similarity- or phylogeny-based methods perform best depends on ITS fragment length and a tradeoff between maximizing the total number of classifications and minimizing erroneous classifications.

Acknowledgements

This study was funded by an NSERC discovery grant and by a Genome Canada grant to GBG.

References


Table 2  Comparison of classifications using 5' and 3' fragments simulated from the same full-length internal transcribed spacer (ITS) sequences using the BLAST best-hit (BH) method with the ‘leave one out’ approach

<table>
<thead>
<tr>
<th>Classification rank</th>
<th>Fragment size (bp)</th>
<th>5' correct</th>
<th>5' incorrect</th>
<th>3' correct</th>
<th>3' incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>50</td>
<td>41</td>
<td>8</td>
<td>7</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>45</td>
<td>6</td>
<td>6</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>40</td>
<td>5</td>
<td>6</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>51</td>
<td>2</td>
<td>4</td>
<td>43</td>
</tr>
<tr>
<td>Genus</td>
<td>50</td>
<td>65</td>
<td>6</td>
<td>6</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>70</td>
<td>5</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>72</td>
<td>3</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>75</td>
<td>2</td>
<td>2</td>
<td>21</td>
</tr>
</tbody>
</table>


Porter TM, Schadt CW, Rizvi L, Martin AP, Schmidt SK, Scott-Denton L, Vilgalys R, Moncalvo J-M. 2008a. Widespread occurrence and


**Supporting Information**

Additional supporting information may be found in the online version of this article.

**Fig. S1** Schematic diagram of the internal transcribed spacer (ITS) rDNA region and simulated fragments.

**Fig. S2** Comparison of the time required to run BLAST searches and Statistical Assignment Program (SAP) analyses.

**Table S1** Summary of next-generation sequencing platforms, number of reads produced, and read length

**Table S2** List of species and corresponding GenBank accession included in the test data set

**Table S3** Query recovery comparison using BLAST, MEGAN, and the Statistical Assignment Program (SAP) ConstrainedNJ (no cutoff enforced)

**Table S4** Problematic sequences that were not correctly classified by BLAST, MEGAN, or the Statistical Assignment Program (SAP) ConstrainedNJ (no cutoff enforced)

**Methods S1** Comparison of the time required to classify amplicons using the BH method and the Statistical Assignment Program (SAP).