A peer-reviewed version of this preprint was published in PeerJ on 28 March 2017.

Santos L, Alves A, Alves R. 2017. Evaluating multi-locus phylogenies for species boundaries determination in the genus *Diaporthe*. PeerJ 5:e3120 https://doi.org/10.7717/peerj.3120

Evaluating multi-locus phylogenies for species boundaries determination in the genus Diaporthe

Liliana Santos [1], **Artur Alves** [Corresp., 1], **Rui Alves** [Corresp. 2]

[1] Departamento de Biologia, CESAM, Universidade de Aveiro, Aveiro, Portugal

[2] Departament de Ciències Mèdiques Bàsiques, Universitat de Lleida and IRBLleida, Lleida, Spain

Corresponding Authors: Artur Alves, Rui Alves
Email address: artur.alves@ua.pt, ralves@cmb.udl.cat

Background. Species identification is essential for controlling disease, understanding epidemiology, and to guide the implementation of phytosanitary measures against fungi from the genus *Diaporthe*. Accurate *Diaporthe* species separation requires using multi-loci phylogenies. However, defining the optimal set of loci that can be used for species identification is still an open problem. **Methods**. Here, we addressed that problem by identifying five loci that have been sequenced in 142 *Diaporthe* isolates representing 96 species: *TEF1*, *TUB*, *CAL*, *HIS* and ITS. We then used every possible combination of those loci to build, analyse, and compare phylogenetic trees. **Results.** As expected, species separation is better when all five loci are simultaneously used to build the phylogeny of the isolates. However, removing the ITS locus has little effect on reconstructed phylogenies, identifying the *TEF1-TUB-CAL-HIS* four loci tree as almost equivalent to the five loci tree. We further identify the best 3-loci, 2-loci, and 1-locus trees that should be used for species separation in the genus. **Discussion.** Our results question the current use of the ITS locus for DNA barcoding in the genus *Diaporthe* and suggest that *TEF1* might be a better choice if one locus barcoding needs to be done.

Title: Evaluating multi-locus phylogenies for species boundaries determination in the genus *Diaporthe*

Short title: Multi-locus phylogenies for *Diaporthe*

Liliana Santos[1], Artur Alves[1*] and Rui Alves [2*]

Affiliations:

[1] Departamento de Biologia, CESAM, Universidade de Aveiro, Campus de Santiago, 3810-193 Aveiro, Portugal

[2] Dept Ciències Mèdiques Bàsiques, Universitat de Lleida & IRBLleida, Edific de Recerca Biomèdica I, Av. Rovira Roure 80, 25198 Lleida, Spain

* Corresponding authors

Artur Alves, artur.alves@ua.pt

Rui Alves, ralves@cmb.udl.cat

Abstract

Background. Species identification is essential for controlling disease, understanding epidemiology, and to guide the implementation of phytosanitary measures against fungi from the genus *Diaporthe*. Accurate *Diaporthe* species separation requires using multi-loci phylogenies. However, defining the optimal set of loci that can be used for species identification is still an open problem.

Methods. Here, we addressed that problem by identifying five loci that have been sequenced in 142 *Diaporthe* isolates representing 96 species: *TEF1*, *TUB*, *CAL*, *HIS* and ITS. We then used every possible combination of those loci to build, analyse, and compare phylogenetic trees.

Results. As expected, species separation is better when all five loci are simultaneously used to build the phylogeny of the isolates. However, removing the ITS locus has little effect on reconstructed phylogenies, identifying the *TEF1-TUB-CAL-HIS* four loci tree as almost equivalent to the five loci tree. We further identify the best 3-loci, 2-loci, and 1-locus trees that should be used for species separation in the genus.

Discussion. Our results question the current use of the ITS locus for DNA barcoding in the genus *Diaporthe* and suggest that *TEF1* might be a better choice if one locus barcoding needs to be done.

1. Introduction

Species in the ascomycete genus *Diaporthe* have been identified all over the world. Typically, *Diaporthe* species are saprobes, endophytes, or plant pathogens (Webber & Gibbs, 1984; Boddy & Griffith, 1989; Udayanga et al., 2011). Some plant pathogenic *Diaporthe* species are associated with cankers, diebacks, rots, spots and wilts on a wide range of plants, some of which are of economic importance as is the case of citrus, cucurbits, soybeans, eggplant, berries and grapevines (Backman, Weaver & Morgan-Jones, 1985; Merrin, Nair & Tarran, 1995; Farr, Castlebury & Rossman, 2002; Farr et al., 2002; Shishido et al., 2006). Less frequently, *Diaporthe* species can also cause lupinosis and other health problems in humans and other mammals (Van Warmelo & Marasas, 1972; Sutton et al., 1999; Battilani et al., 2011; Garcia-Reyne et al., 2011).

Distinction between *Diaporthe* species has historically been based on an approach that combined morphological information, cultural characteristics, and host affiliation (Udayanga et al., 2011). This approach made it difficult to reliably discriminate between the various members of the genus, because many of these fungi are asexual with low host specificity (Rehner & Uecker, 1994; Murali, Suryanarayanan & Geeta, 2006). As a consequence, an unnecessary increase in the number of proposed *Diaporthe* species occurred. This number currently stands at 977 and 1099 for *Diaporthe* and 980 and 1047 for *Phomopsis* in Index Fungorum and Mycobank, respectively (both accessed 14 November 2016). The extinction of the dual nomenclature system for fungi raised the question about which generic name to use, *Diaporthe* or that of its asexual morph *Phomopsis*. Given that both names are well known among plant pathologists, and have been equally used, Rossman et al. (2015) proposed that the genus name *Diaporthe* should be retained over *Phomopsis* because it was introduced first and therefore has priority.

The problem of incorrect species attribution has practical consequences for the study of this genus, because accurate species identification is essential for understanding the epidemiology,

for controlling plant diseases, and to guide the implementation of international phytosanitary measures (Santos and Phillips, 2009; Udayanga et al., 2011). Therefore, there was an urgent need to reformulate species identification in the genus *Diaporthe* (Santos and Phillips, 2009).

Advances in the areas of gene sequencing and molecular evolution over the last 50 years have led to the notion that ribosomal genes can be used to distinguish between species and study their molecular evolution (Woese & Fox, 1977). The choice of these genes comes from the fact that their function is conserved over all living organisms, which has been assumed to imply that their evolutionary rate should be roughly constant over the tree of life.

The molecular evolution studies mentioned have been used to develop general fungal classifications (Shenoy, Jeewon & Hyde, 2007) and have also been used for species reclassification in the genus *Diaporthe* (Santos and Phillips 2009; Santos et al., 2011; Thompson et al., 2011; Baumgartner et al., 2013; Gomes et al, 2013; Huang et al., 2013; Tan et al., 2013; Gao et al., 2014; Udayanga et al., 2014a; Udayanga et al., 2014b). In fact, recently the ITS region of the ribosomal genes has been accepted as the official fungal barcode (Schoch et al., 2012), and its sequence is frequently used for molecular phylogeny analysis of *Diaporthe* species.

However, assuming that ribosomal gene sequences evolve at a uniform rate, independent of species is sometimes incorrect (Anderson & Stasovski, 1992; O'Donnell, 1992; Carbone & Kohn, 1993). In addition, due to the strong constraints imposed by ribosome function on the mutations in the sequence of ribosomal genes, close microbial species may have identical rDNA sequences, while having clearly different genomes. For example, a comparison between *Cladosporium, Penicillium* and *Fusarium* species at the NCBI Genome and GenBank databases (Schoch et al., 2012) will confirm this statement. Such considerations suggested that phylogenetic trees based on sets of genes are potentially more powerful in solving species boundaries than phylogenetic trees based on any single genes, as the former trees contain information about the simultaneous evolution of various biological processes (Olmstead & Sweere, 1994; Rokas et al., 2003).

The possibility of using full genomes to create phylogenetic trees becomes more feasible as the number of fully sequenced genomes increases. For example, the full genomic complement of

genes/proteins involved in metabolism have been used to reconstruct phylogenies that provide information regarding the evolution of metabolism in various species (Heymans & Singh, 2003; Ma & Zeng, 2004; Forst et al., 2006; Oh et al., 2006). This type of genome wide phylogeny reconstruction is impossible for organisms that have not had their genomes fully sequenced and annotated. This is the case for the genus *Diaporthe,* for which the first genome sequencing project started in 2013 (GOLD project Gp0038530) and until now only *Diaporthe* species have their genome sequenced (*Phomopsis longicolla*, *Diaporthe aspalathi*, *Diaporthe ampelina* and *Diaporthe helianthi*) (Li et al., 2015; Baroncelli et al., 2016; Li et al., 2016; Savitha et al., 2016).

Although full genome sequences are still forthcoming for *Diaporthe* species, current species identification and phylogeny reconstruction in the genus are already largely dependent on molecular sequences (Santos, Correia & Phillips, 2010). The sequences more frequently used for these studies are: large subunit (LSU) of the ribosomal DNA, intergenic spacers (IGS) of the ribosomal DNA, internal transcribed spacer (ITS) of the ribosomal DNA, translation elongation factor 1-α (*TEF1*) gene, ß-tubulin (*TUB*) gene, histone (*HIS*) gene, calmodulin (*CAL*) gene, actin (*ACT*) gene, DNA-lyase (*APN2*) gene, 60s ribosomal protein L37 (FG1093) gene and mating type genes (MAT-1-1-1 and MAT-1-2-1) (Farr, Castlebury & Rossman, 2002; Farr et al., 2002; Castlebury et al., 2003; Pecchia, Mercatelli & Vannacci, 2004; Schilder et al., 2005; Van Rensburg et al., 2006; Kanematsu, Adachi & Ito, 2007; Santos, Correia & Phillips, 2010; Santos et al., 2011; Thompson et al., 2011; Grasso et al., 2012; Sun et al., 2012; Udayanga et al., 2012; Baumgartner et al., 2013; Bienapfl & Balci, 2013; Gomes et al., 2013; Huang et al., 2013; Sun et al., 2013; Tan et al., 2013;Vidić et al., 2013; Gao et al., 2014; Udayanga et al., 2014a; Udayanga et al., 2014b; Wang et al., 2014).

However, multi-locus phylogenies for the genus *Diaporthe* have only been developed in the last few years (Schielder et al., 2005; Van Rensburg et al., 2006; Udayanga et al., 2012; Baumgartner et al., 2013; Gomes et al., 2013; Huang et al., 2013; Tan et al., 2013; Gao et al., 2014; Udayanga et al., 2014a; Udayanga et al., 2014b; Wang et al., 2014). In fact, creating phylogenies that include several loci is still possible only for a limited set of species from the genus *Diaporthe*, because not all genes have been sequenced for all tentative species. This is due to many reasons, among which the lack of resources that prevents unlimited sequencing of samples. Nevertheless,

a multi-locus approach should always be used for accurate resolution of species in the genus *Diaporthe*.

In recent studies the maximum number of loci used was to create multi loci phylogenies seven (*TEF1*, *TUB*, *HIS*, *CAL*, *ACT*, *APN2* and FG1093), simultaneously sequenced across approximately 80 isolates from 9 *Diaporthe* species (Udayanga et al., 2014a). These loci were used to establish the specific limits of *D. eres*. This work provides a good example of how to establish the boundaries for one species within the genus *Diaporthe*. However, if this is to be extended to the other species of the genus, it is important to determine which loci are the most informative to be sequenced and used in a much wider range of *Diaporthe* species.

With this in mind we asked which combination of frequently sequenced loci better discriminate species boundaries in *Diaporthe*. To answer this question, we considered the ITS, *TEF1*, *TUB*, *HIS* and *CAL* loci, which had been sequenced for 96 different *Diaporthe* species. This paper ranks these loci according to their contribution for improving/decreasing the resolution of *Diaporthe* species determination, as they are added/removed from multi-locus phylogenies analysis.

2. Materials & Methods

2.1. Data collection

In-house PERL scripts were used to search the GenBank and download all sequences from *Diaporthe* and *Phomopsis* species for the 11 loci mentioned in the introduction. We then determined that sequences for ITS, *CAL*, *TUB*, *HIS*, and *TEF1* loci were known in 142 *Diaporthe* and *Phomopsis* isolates, corresponding to 96 different species. Adding any other loci would reduce the number of species. Thus, we have chosen to study these five loci in those 96 species, as a way of maximizing the statistical power of our analysis. Species and gene identifications, as well as, the accession numbers are given in SM Table 1. The current study used 142 *Diaporthe* isolates that were selected by choosing two isolates per species (whenever they were available), at least one of them being an ex-type isolate. With these constrains in mind,

we chose the two isolates for which the sequences were more dissimilar within the same species, in order to maximize intraspecific sequence diversity.

Also considering this intraspecific heterogeneity, we used a larger number of sequence sample for *Diaporthe* species complexes (Udayanga et al., 2014a). These are species with a higher than average diversity between individuals. In our case they include *D. sojae, D. foeniculacea*, and *D. eres*. For example, the *D. eres* complex includes strains CBS 113470, CBS 116953, CBS 200.39, and CBS 338.89, some of which were originally classified as *D. nobilis* and later reclassified into the *D. eres* complex (Gomes et al. 2013; Udayanga et al. 2014a). In addition, we used more than one ex-type isolate for the species complexes, because these species are highly heterogeneous. All sequence data used in this study have been validated and published previously (Castleburry et al., 2002; Van Niekerk et al., 2005; Santos et al., 2011, Gomes et al., 2013 and Udayanga et al., 2014a).

As species concept we used the criteria of Genealogical Concordance Phylogenetic Species Recognition (GCPSR) to resolve species boundaries based on individual and combined analyses of the 5 genes.

2.2. Sequence alignment and phylogenetic analyses

Five multiple alignments, one per locus, were created using the software ClustalX2.1 (Larkin et al., 2007), and the following parameters: pairwise alignment parameters (gap opening = 10, gap extension = 0.1) and multiple alignment parameters (gap opening = 10, gap extension = 0.2, transition weight = 0.5, delay divergent sequences = 25 %), and optimized manually with BioEdit (Hall, 1999). The alignments for the individual locus were then concatenated into all possible combinations of 2, 3, 4, and five loci. This generated 31 alternatives multiple alignments, counting the 5 multiple alignments for the individual genes and the alignment for the five concatenated gene sequences. MEGA6 (Tamura et al., 2013) was used to create and analyse phylogenetic trees for each of the 31 alignments, independently using two alternative methods (Maximum Parsimony [MP] and Maximum Likelihood [ML]; Li, 1997). MEGA6 was also used to determine the best evolution models to be used for building the ML tree from each multiple alignment, as described previously (Tamura et al., 2013). These models are listed in Table 1. Each tree was bootstrapped 1000 times, and branches that split in less than 90% of the 1000 trees

were condensed. MP trees were obtained using the Tree-Bisection-Reconnection (TBR) algorithm (Nei & Kumar, 2000) with search level 1, in which the initial trees were obtained by the random addition of sequences (10 replicates). The initial trees for the heuristic ML search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, allowing for some sites to be evolutionarily invariable ([+I], 0.0000% sites). As in Gomes et al. (2013), we choose *Diaporthella corylina* (CBS 121124) as outgroup.

2.3. Comparing trees

2.3.1. Tree scores

MEGA6 was used to create and analyse all MP and ML phylogenetic trees. As a first approximation, we compare the likelihood values between ML trees and the MP scores between MP trees (Table 2 and 3) for identifying the best and worst trees of each type.

The length of an MP tree estimates phylogenetic tree resolution. This value is also dependent on the length of the sequences that are used to build the tree. This means that comparing tree lengths for trees built using a varying number of loci should also consider normalizing the length of the tree by the corresponding size of the aligned sequence (Table 2). This normalization allows us to estimate which loci provide more added value when it comes to species resolution.

ML tree building methods seek the tree that is more likely (the highest likelihood), based on a probabilistic model of sequence evolution. The best ML tree has the lowest - log likelihood scores and worst ML tree has the highest – log likelihood value. This likelihood is also dependent on the length of the alignment. In order to be able to compare all the trees among them we also normalized the values of – log Likelihood in the same way of the MP length (Table 3). This means that comparing tree log likelihoods for trees built using a varying number of loci should also consider normalizing the log likelihood of the tree by the corresponding size of the aligned sequence (Table 3).

2.3.2. Tree distances

All trees we build have the same species. Thus, we are able to measure the difference between every possible pair of trees, based on the analysis of the symmetric distance between equal leafs

in two trees (Robinson & Foulds, 1981). This distance was calculated for all pairs of MP trees using the Treedist methods of the PHYLIP suite of programs (Felsenstein, 1989). The same calculations were made for all pairs of ML trees. For these calculations we used condensed trees with a 90% bootstrap cut-off value. This allows us to measure how adding/removing a locus to/from the multiple alignments causes the resulting phylogenetic tree to change.

2.3.3. Testing Phylogenetic informativeness and identification of species boundaries.

We used PhyDesign (López-Giráldez & Townsed, 2011) to establish the informativeness of the various combinations of loci alignments, as described in (Udayanga et al., 2014a). We also manually analyzed all trees to identify all cases where isolates of the same species did not cluster together. This allowed us to determine the loci that provided the best species resolution.

3. Results

We analyse 142 isolates from 96 *Diaporthe* species for which the ITS, *CAL*, *TUB*, *HIS*, and *TEF1* loci had been sequenced (SM Table 1). The alignments for each locus were then concatenated in all possible 31 combinations of 1, 2, 3, 4, and 5 genes. Alignment characteristics for this study are reported in Table 4. Each combination was used to build a ML and ME phylogenetic trees. Each tree was bootstrapped 1000 times and every tree used is a condensed tree with a 90% cut-off. Alignments and trees were deposited in TreeBase (Study Accession: S20343).

3.1. Best and worst resolving phylogenetic trees

The "quality" (resolution) of the individual phylogenetic trees was determined as described in methods.

Figures 1 and 2 present the condensed MP and ML trees build from the concatenated multiple alignments of the 5 loci, respectively. Phylograms showing all complete trees are given as supplementary figures (SM Figure 1 and 2, respectively). These trees are the best resolving trees built for each method, as indicated by the scores shown in Table 2 for MP trees and in Table 3 for ML trees.

The increase in tree length (Table 2) and log-likelihood scores (Table 3) of the trees with the increase in number of loci indicates that resolution of the trees is directly correlated with the number of loci used to build them. This is also true for the tree scores and log-likelihood scores normalized by alignment length. Thus, the worst trees are built using the multiple alignments for only one locus. Within the one-locus trees, the best MP (Figure 3 and SM – Figure 3) and ML (Figure 4 and SM – Figure 4) condensed trees are shown in Figure 3 and 4. *TEF1* trees have the highest values for length and-log likelihood.

3.2. Choosing the most informative loci for sequencing

The previous results indicate that, whenever possible, all five loci should be sequenced, in order to better differentiate between *Diaporthe* species. However, this might not always be possible. In situations where only a subset of one, two, three, or four out of the five loci can be sequenced, which sequences might be more informative? This can be roughly answered in two steps.

The first step is done by measuring how adding/removing a locus to/from the multiple alignments causes the resulting phylogenetic tree to change. These changes can be measured by calculating the symmetric distance between the two trees and by analysing if species resolution changes when the relevant locus is added or removed. The smaller the changes are, the less informative the locus is. The symmetric distance matrices between every pair of MP (SM Table 2) or ML (SM Table 3) trees were calculated as described in methods. Table 5 summarize these results and show how many changes are observed on average when a specific locus is removed from a multi-locus tree. On average, the ITS locus is the least informative one, closely followed by the HIS locus. The third locus whose removal causes the least changes in the trees is CAL. This is true for both, the MP and the ML trees.

The second step is done by evaluating the changes in the resolution of the trees when a locus is removed from the multiple alignments. A more detailed analysis of Tables 3-5 reveals that removing the ITS locus from any MP or ML multi-loci tree causes the smallest decrease in MP tree length and in ML tree likelihood. Hence, if only four loci can be sequenced these should be *TEF1-TUB-CAL-HIS*. The second locus with the least effect in tree resolution is *TUB*, closely followed by *HIS*. Given that, as measured in step one of the process, average differences between trees when *HIS* is removed are much smaller than differences between trees when *TUB* is

removed, if only three loci can be sequenced these should be *TEF1-TUB-CAL*. If only two loci can be sequenced, we suggest *TEF1-TUB*, as removing *CAL* has the least average effect on trees. Finally, if only one locus can be sequenced tree resolution suggests that this locus should be *TEF1*. *TEF1* trees are the best single locus MP and ML trees (Figure 3 and 4).

3.3. Phylogenetic informativeness and identification of species boundaries

Figure 5 shows that the *TEF1* sequence is the most informative for species separation, both globally and per alignment site. In addition, we also see that the ITS sequence is the least informative to resolve *Diaporthe* species (Figure 5). The five loci can be ranked from most to least informative for *Diaporthe* species separation as follows: *TEF1>HIS>CAL>TUB>ITS*.

The dataset we used for this analysis is as close as we currently can get to a standard set of well separated *Diaporthe* species, taking into account that the five loci we analyse needed to be sequenced for all individuals in the set. Taking this into account, an inspection of the trees is required to understand, on top of all the statistical analyses, if species are well separated or not.

We see that, in general, the addition of a new locus to the alignment decreases the number of isolates from the same species that do not cluster together (separation errors). Therefore, the tree of 5 loci has less separation errors than 4-loci trees, which in turn have less separation errors than the 3-loci trees, and so on. As expected from our previous analysis, the *TEF1* tree provides the best single locus ML tree, *TEF1-TUB* tree provides the best 2-loci ML tree, *TEF1-TUB-CAL* the best 3-locus ML tree. The results from the MP trees are qualitatively similar although, in general, these trees have more separation errors that the ML ones.

4. Discussion

Identifying species boundaries in organisms is a difficult task, as theoretical and practical definitions of species are not always consistent with each other (Doolittle & Zhaxybayeva, 2009; Giraud et al., 2008). While Woese & Fox (1977) suggested using ribosomal sequences to define species borders, such sequences are not always the best choice. For example, searching GenBank will reveal that some *Cladosporium, Penicillium* and *Fusarium* species cannot be differentiated using ITS (Schoch et al., 2012).

More recent work suggests that trees based on multi-loci sequence analysis (MLSA) provide more accurate estimations of phylogeny than single gene trees, if appropriate loci are used (Gadagkar, Rosenberg & Kumar, 2005; Mirarab, Bayzid & Warnow, 2014). Briefly, MLSA concatenates sequence alignments from multiple genes and uses the concatenated sequences to determine phylogenetic relationships. This method appears to more optimally resolve the phylogenetic position of species in the same or in closely related genera (Hanage, Fraser & Spratt, 2006). An increase in the number of loci used to build MLSA phylogenetic trees positively correlates to sensitivity and accuracy in species separation (Rokas et al., 2003; Udayanga et al., 2011). In contrast, increasing the number of species in the alignment leads to a decrease in the ability to separate them accurately, unless a higher number of appropriate loci are used to maintain the quality of that separation (Bininda-Emonds et al. 2001; Kim, 1998; Poe & Swofford, 1999; Rokas et al., 2003; Udayanga et al., 2011). The choice of appropriate loci to be used in such trees can be optimized in genera with a large number of sequenced genomes, because in such cases it is possible to make full genome studies to identify the best set of loci to separate species. Nevertheless, the amount of information that must be analysed for doing so could become prohibitive (Thangaduras & Sangeetha, 2013).

The choice of appropriate loci that optimizes species separation is harder when fully sequenced genomes are not available, as is the case for the genus *Diaporthe*. Nevertheless, MLSA phylogenetic studies of *Diaporthe* species have been done using loci that have been chosen in a more or less *ad hoc* manner, by taking into account how conserved they were in different fungal genus (Baumgartner et al., 2013; Gao et al., 2014; Gomes et al., 2013; Huang et al., 2013; Schielder et al., 2005; Tan et al., 2013; Udayanga et al., 2012; Udayanga et al., 2014a; Udayanga et al., 2014b; Van Rensburg et al., 2006; Wang et al., 2014). In general, these studies show that MLSA phylogenetic trees provide higher resolution for *Diaporthe* species than single locus phylogenetic trees (Huang et al., 2013; Udayanga et al. 2012; Van Rensberg et al., 2006).

The current study addresses the problem of which loci are best for accurate species separation in the genus *Diaporthe* in a systematic manner. Walker et al. (2012) performed a similar study. While we use five non-coding loci to study species separation in *Diaporthe*, those authors employed two single copy protein-coding genes (FG1093 and MS204) to study species separation in Sordariomycetes. While Walker et al. (2012) analysed various aspects of codon

conservation and substitution rates, these analyses are meaningless for our sequence dataset. The use of non-coding sequences is favoured in *Diaporthe* species separation because coding sequences are typically too conserved to allow for appropriate separation within the genus.

The major contributions of this paper are two-fold. First, our work confirms that the quality of species separation in phylogenetic trees increases with the number of loci used to build phylogenetic trees. Second and more importantly it identifies the best combination of loci that one should use for building those phylogenetic trees, if only one, two, three, or four loci can be sequenced. To achieve this, we took the most commonly sequenced loci for 142 *Diaporthe* isolates and studied which loci optimize species differentiation in the genus. We chose only loci that are commonly sequenced for members of the genus. Then, we selected a sequence dataset that was experimentally validated by others (Castleburry et al., 2002; Van Niekerk et al., 2005; Santos et al., 2011 and Gomes et al., 2013) before being deposited in GenBank. Whenever possible we favoured sequences from ex-type isolates and produced via low throughput, high fidelity, sequencing methods. In addition, our sequence selection maximized intraspecific sequence variation, which in turn maximizes the possibility that intra-specific hyperdiversity could be higher than interspecific diversity. Thus, species separation through phylogenetic trees in our sample is made more difficult by our sequence selection, making our analysis more robust. In this paper we only show and analyse condensed MP and ML trees, using a cut-off of 90%, which means that our trees are very robust to gene order, as a significant amount of bootstrapping was used to calculate them. In fact, to test that, we performed a side experiment where we changed the order of the locus sequences in the alignments and recalculated the trees (SM - Figure 5).

We found that species differentiation is optimized by creating phylogenetic trees built from the multiple sequence alignment of five loci: *TEF1*-*TUB*-*HIS*-*CAL*-ITS. However, little information is lost when ITS locus is removed and only the other four loci are used to simultaneously build the phylogeny. In addition, we also provide researchers with a ranking of best loci to sequence if only 1, 2, 3, or 4 of the loci can be sequenced.

It may be surprising that the ribosomal ITS locus is the least informative of the five loci when it comes to separating *Diaporthe* species. However, Santos, Correia & Phillips (2010) found that

the ITS region in *Diaporthe* is evolving at much faster rates than *TEF1* or even *MAT* genes. Hence, what seems to be happening is that ITS sequences present a wider variation than is advisable for creating precise species boundaries. Therefore a slowly evolving gene region should be utilized in order to establish precise species limits (Udayanga et al., 2012).

DNA barcoding (Kress et al., 2014) refers to the use of standard short gene sequences to identify species. The use of DNA barcoding implies that an effort should be made to standardize the use of the loci for phylogenetic studies. ITS is the official DNA barcode region in fungi (Schoch et al., 2012). This work supports previous studies whose results suggest that using ITS as a standard for species separation in fungi should be discontinued (Gomes et al., 2013; Thangaduras & Sangeetha 2013). Our results strongly recommend that TEF1 should be used instead, at least in the genus *Diaporthe*. This is consistent with and further develops previously published results, which proposed either *TEF1*, *HIS*, or *APN2* as alternative locus for barcoding in the genus (Santos, Correia & Phillips, 2010; Udayanga et al., 2014b). However, Gomes et al. (2013), using Bayesian analysis, consider *HIS* and *TUB* as best resolving genes. Nevertheless, considering that Gomes et al. (2013) use shorter sequences than those used here, one is tempted to cautiously analyse and reinterpret their conclusions.

Despite the *TEF* tree appears to be a better species separator than the 5 loci tree, the true is that, the alignment used to build the 5 loci tree is roughly five times larger than that for the *TEF* tree. This means that, with a larger number of positions, there is bound to be more variability in the bootstrapping of the 5 loci tree than in the bootstrapping of the *TEF* tree. Hence, the observation that the *TEF* give better resolution than 5 loci results from a statistical artefact. This fact occurs when focusing on the *D. eres* complex clade. For example, in the case of the *D. eres* complex, all the species are grouped in the same clade in both cases (*D. alleghaniensis*, *D. alnea*, *D. celastrina*, *D. bicincta*, *D. eres*, *D. neilliae* and *D. vaccinii*). However, in the 5-loci trees the resolution of this species complex is better. This is especially important as phylogenetic analyses of the *D. eres* complex often revealed ambiguous clades with short branch and moderate statistic supports due to their high variability. Udayanaga et al. (2014a) studied this problematic by using different genes, whose sequences are not available for the other *Diaporthe* species we consider. Therefore, we could not incorporate their data in our study. We also note that one possible explanation for the observation that some species of the *D. eres* complex do not "group" in the

same clade could be due to the fact that they are not really *D. eres*. However, to test that, we would need to actually obtain samples of the complex, re-sequence and analyse them in order to clarify the species boundaries in this group.

The problem of species boundary identification is very relevant in the genus *Diaporthe*, where a general taxonomic revision based on molecular analysis is probably overdue. Such a revision could then be used to improve the annotation of sequences in public databases, such as GenBank. For example, many of the sequences we use in our analysis are still assigned to species that have already been reclassified. This also emphasizes that a standard procedure with minimal information required for submitting new *Diaporthe* species needs to be put in place in order to avoid unnecessary creation of new species (Udayanga et al., 2014b). Furthermore, as also suggested by Gomes et al. (2013) we feel that this revision should be made using molecular data. Any new *Diaporthe* species report should be accompanied by molecular data that supports the identification of the individual as a new species. In addition, we feel that a proper taxonomic revision of the genus should also consider morphological descriptions and epitypification of species as previously suggested (Gomes et al., 2013; Udayanga et al., 2014b).

5. Conclusions

Our results indicate that:

- In order of effectiveness the best sets of loci for resolving *Diaporthe* species are *TEF1-TUB-CAL-HIS*-ITS, *TEF1-TUB-CAL-HIS*, *TEF1-TUB-CA*L, *TEF1-TUB* and *TEF1*.
- The *TEF1* locus is a better candidate for single locus DNA barcoding in the genus *Diaporthe* than the ITS locus.
- Multi-loci DNA barcoding will provide a more accurate species separation in the genus than single locus barcoding. Furthermore, a four loci barcoding including *TEF1-TUB-HIS-CAL* will be almost as effective as a five loci barcoding including ITS-*TEF1-TUB-HIS-CAL*.

6. Acknowledgements

We thank Anabel Usié and R. Benfeitas for assistance with the creation of the Perl scripts.

7. References

Anderson JB, Stasovski E. 1992. Molecular phylogeny of northern hemisphere species of *Armillaria*. *Mycologia* 84:505-516.

Backman PA, Weaver DB, Morgan-Jones G. 1985. Soybean stem canker: an emerging disease problem. *Plant Disease* 69:641-647.

Baroncelli R, Scala F, Vergara M, Thon MR, Ruocco M. 2016. Draft whole-genome sequence of the *Diaporthe helianthi* 7/96 strain, causal agent of sunflower stem canker. *Genomics Data* 10: 151–152. DOI: http://dx.doi.org/10.1016/j.gdata.2016.11.005

Battilani P, Gualla A, Dall'Asta C, Pellacani C, Galaverna G, Giorni P, Caglieri A, Tagliaferri S, Pietri A, Dossena A, Spadaro D, Marchelli R, Gullino ML, Costa LG. 2011. Phomopsins: an overview of phytopathological and chemical aspects, toxicity, analysis and occurrence. *World Mycotoxin Journal* 4(4):345-359. DOI:10.3920/WMJ2011.1302.

Baumgartner K, Fujiyoshi PT, Travadon, R, Castlebury LA, Wilcox WF, Rolshausen PE. 2013. Characterization of Species of *Diaporthe* from Wood Cankers of Grape in Eastern North American Vineyards. *Plant Disease* 97(7):912-920.

Bienapfl JC, Balci Y. 2013. Phomopsis Blight: A New Disease of Pieris japonica Caused by *Phomopsis amygdali* in the United States. *Plant Disease* 97(11):1403-1407.

Bininda-Emonds OR, Brady SG, Kim J, Sanderson MJ. 2001. Scaling of accuracy in extremely large phylogenetic trees. *Pacific Symposium on Biocomputing* 6:547-558.

Boddy L, Griffith S. 1989. Role of endophytes and latent invasion in the development of decay communities in sapwood of angiospermous trees. *SYDOWIA* 41:41-73.

Carbone I, Kohn LM. 1993. Ribosomal DNA sequence divergence within internal transcribed spacer 1 of the *Sclerotiniaceae*. *Mycologia* 85(3):415-427. DOI:10.2307/3760703.

Castlebury LA, Farr DF, Rossman AY, Jaklitsch W. 2003. *Diaporthe angelicae* comb. nov., a modern description and placement of *Diaporthopsis* in *Diaporthe*. *Mycoscience* 44(3):203-208. DOI:10.1007/s10267-003-0107-2.

Doolittle WF, Zhaxybayeva O. 2009. On the origin of prokaryotic species. Genome Research, 19(5): 744-756. DOI: 10.1101/gr.086645.108.

448 Farr DF, Castlebury LA, Rossman AY. 2002. Morphological and molecular characterization of
449 *Phomopsis vaccinii* and additional isolates of *Phomopsis* from blueberry and cranberry in the
450 eastern United States. *Mycologia* 94(3):494-504.

451 Farr DF, Castlebury LA, Rossman AY, Putnam ML. 2002. A new species of *Phomopsis* causing
452 twig dieback of *Vaccinium vitis-idaea* (lingonberry). *Mycological Research* 106(6):745-752.

453 Felsenstein J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5:164-
454 166.

455 Forst CV, Flamm C, Hofacker IL, Stadler PF. 2006. Algebraic comparison of metabolic
456 networks, phylogenetic inference, and metabolic innovation. *BMC Bioinformatics* 7:67.
457 DOI:10.1186/1471-2105-7-67.

458 Gadagkar SR, Rosenberg MS, Kumar S. 2005. Inferring species phylogenies from multiple
459 genes: concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology*
460 *Part B: Molecular and Developmental Evolution* 304B (1):64-74. DOI: 10.1002/jez.b.21026.

461 Garcia-Reyne A, López-Medrano F, Morales JM, García EC, Martín I, Eraña I, Meije Y, Lalueza
462 A, Alastruey-Izquierdo A, Rodríguez-Tudela JL, Aquado JM. 2011. Cutaneous infection by
463 *Phomopsis longicolla* in a renal transplant recipient from Guinea: first report of human infection
464 by this fungus. *Transplant Infectious Disease* 13:204-207. DOI:10.1111/j.1399-3062.2010.0057.

465 Gao Y, Sun W, Su Y, Cai L. 2014. Three new species of *Phomopsis* in Gutianshan Nature
466 Reserve in China. *Mycological Progress* 13:111-121. DOI:10.1007/s11557-013-0898-2.

467 Giraud T, Refrégier G, Le Gac M, de Vienne DM, Hood ME. 2008. Speciation in fungi. *Fungal*
468 *Genetics and Biology* 45:791-802. DOI: 10.1016/j.fgb.2008.02.001.

469 Gomes RR, Glienke C, Videira S I.R, Lombard L, Groenewald JZ, Crous PW. 2013. *Diaporthe*:
470 a genus of endophytic, saprobic and plant pathogenic fungi. *Persoonia* 31:1-41.
471 DOI:10.3767/003158513X666844.

472 Grasso FM, Marini M, Vitale A, Firrao G, Granata G. 2012. Canker and dieback on *Platanus x*
473 *acerifolia* caused by *Diaporthe scabra*. *Forest Pathology* 42:510-513. DOI:10.1111/j.1439-
474 0329.2012.00785.x.

475 Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis
476 program for Windows 95/98/NT. *Nucleic Acids Symposium* 41:95-98.

477 Hanage WP, Fraser C, Spratt BG. 2006. Sequences, sequence clusters and bacterial species.
478 *Philosophical Transactions of The Royal Society B – Biological Sciences* 361:1917-1927.

Hasegawa M, Kishino H, Yano TA. 1985. Dating of the human ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22:160-174.

Heymans M, Singh AK. 2003. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics, 19 (suppl1)*: i138-i146. DOI: 10.1093/bioinformatics/btg1018

Huang F, Hou X, Dewdney MM, Fu Y, Chen G, Hyde KD, Li H. 2013. *Diaporthe* species occurring on citrus in China. *Fungal Diversity* 61:237-250. DOI:10.1007/s13225-013-0245-6.

Kanematsu, S., Adachi, Y. and Ito, T. 2007. Mating-type loci of heterothallic *Diaporthe* spp.: homologous genes are present in opposite mating-types. *Current Genetics*, 52: 11–22. Doi:10.1007/s00294-007-0132-3.

Kim J. 1998. Large-scale phylogenies and measuring the performance of phylogenetic estimators. *Systematic Biology* 47:43-60.

Kress WJ, García-Robledo C, Uriarte M, Erickson DL. 2014. DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology & Evoltion* 30(1):25-35. DOI: 10.1016/j.tree.2014.10.008.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007 Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947-2948. DOI: 10.1093/bioinformatics/btm404

Li S, Darwish O, Alkharouf N, Matthews B, Ji P, Domier LL, Zhang N, Bluhm BH. 2015. Draft genome sequence of *Phomopsis longicolla* isolate MSPL 10-6. *Genomics Data* 3: 55–56. DOI: http://dx.doi.org/10.1016/j.gdata.2014.11.007

Li S, Song Q, Martins AM, Cregan P. 2016. Draft genome sequence of *Diaporthe aspalathi* isolate MS-SSC91, a fungus causing stem canker in soybean. *Genomics Data* 7: 262–263. DOI: http://dx.doi.org/10.1016/j.gdata.2016.02.002

Li W. 1997. *Molecular Evolution*. Sunderland, Massachusetts, USA: Sinauer and Associates.

López-Giráldez F, Townsend JP. 2011. PhyDesign: an online application for profiling phylogenetic informativeness. *BMC Evolutionary Biology* 11:152. DOI: 10.1186/1471-2148-11-152

Ma H, Zeng A. 2004. Phylogenetic comparison of metabolic capacities of organisms at genome level. *Molecular Phylogenetics and Evolution* 31(1):204-213.

Merrin SJ, Nair NG, Tarran J. 1995. Variation in *Phomopsis* recorded on grapevine in Australia and its taxonomic and biological implications. *Australasian Plant Pathology* 24(1):44-56.

Mirarab S, Bayzid MS, Warnow T. 2014. Evaluating Summary Methods for Multilocus Species Tree Estimation in the Presence of Incomplete Lineage Sorting. *Systematic Biology* syu063. DOI: 10.1093/sysbio/syu063.

Murali TS, Suryanarayanan TS, Geeta R. 2006. Endophytic *Phomopsis* species: host range and implications for diversity estimates. *Canadian Journal of Microbiology* 52(7):673-680. DOI:10.1139/w06-020.

Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. New York, USA: Oxford University Press.

O'Donnell K. 1992. Ribosomal DNA internal transcribed spacers are highly divergent in the phytopathogenic ascomycete *Fusarium sambucinum* (*Gibberella pulicaris*). *Current Genetics* 22(3):213-220. DOI: 10.1007/BF00351728.

Oh SJ, Joung J, Chang J, Zang B. 2006. Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks. *BMC Bioinformatics* 7:284. DOI:10.1186/1471-2105-7-284

Olmstead RG, Sweere JA. 1994. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the *Solanaceae*. *Systematic Biology* 43(4):467-481. DOI: 10.1093/sysbio/43.4.467.

Pecchia S, Mercatelli E, Vannacci G. 2004. Intraspecific diversity within *Diaporthe helianthi*: evidence from rDNA intergenic spacer (IGS) sequence analysis. *Mycopathologia* 157:317-326.

Poe S, Swofford DL. 1999. Taxon sampling revisited. *Nature* 398:299–300. DOI:10.1038/18592

Rehner SA, Uecker FA. 1994. Nuclear ribosomal internal transcribed spacer phylogeny and host diversity in the coelomycetes *Phomopsis*. *Canadian Journal of Botany* 72(11):1666-1674. DOI: 10.1139/b94-204.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53:131-147.

Rokas A, Williams BL, King N, Carroll, SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798-804. DOI: 10.1038/nature02053.

Rossman AY, Adams GC, Cannon PF, Castlebury LA, Crous PW, Gryzenhout M, Jaklitsch WM, Mejia LC, Stoykov D, Udayanga D, Voglmayr H, Walker DM. 2015. Recommendations of generic names in *Diaporthales* competing for protection or use. *IMA Fungus* 6(1): 145-154. DOI: 10.5598/imafungus.2015.06.01.09

Santos JM, Vrandečić K, Ćosić T, Duvnjak T, Phillips AJL. 2011. Resolving the *Diaporthe* species occurring on soybean in Croatia. *Persoonia* 27:9-19.

Santos JM, Correia VG, Phillips A JL. 2010. Primers for mating-type diagnosis in *Diaporthe* and *Phomopsis*: their use in teleomorph induction in vitro and biological species definition. *Fungal Biology* 114:255-270. DOI: 10.1016/j.funbio.2010.01.007

Santos, J. and Phillips, A.J.L. 2009. Resolving the complex of *Diaporthe* (*Phomopsis*) species occurring on *Foeniculum vulgare* in Portugal. *Fungal Diversity*, 34:111–125.

Savitha J, Bhargavi SD, Praveen VK. 2016. Complete Genome Sequence of the Endophytic Fungus *Diaporthe* (*Phomopsis*) *ampelina*. *Genome Announcements* 4 (3): e00477-16.

Schilder AMC, Erincik O, Castlebury L, Rossman A, Ellis MA. 2005. Characterization of *Phomopsis* spp. Infecting Grapevines in the Great Lakes Region of North America. *Plant Disease* 89(7):755-762. DOI: 10.1094/PD-89-0755.

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA and Chen W. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. *Proceedings of the National Academy of Sciences of the United States of America* 109(16):6241-6246. DOI: 10.1073/pnas.1117018109.

Shenoy BD, Jeewon R, Hyde KD. 2007. Impact of DNA sequence-data on the taxonomy of anamorphic fungi. *Fungal Diversity* 26:1-54.

Shishido M, Yoshida N, Usami T, Shinozaki T, Kobayashi M, Takeuchi T. 2006. Black root rot of cucurbits caused by *Phomopsis sclerotioides* in Japan and phylogenetic grouping of the pathogen. *Journal of General Plant Pathology* 72:220-227. DOI: 10.1007/s10327-006-0273-0.

Sun S, Van K, Kim MY, Min KH, Lee Y, Lee S. 2012. *Diaporthe phaseolorum* var. *caulivora*, a Causal Agent for Both Stem Canker and Seed Decay on Soybean. *The Plant Pathology Journal* 28(1):55-59. DOI:10.5423/PPJ.NT.10.2011.0194.

Sun S, Kim MY, Chaisan T, Lee Y, Van K, Lee S. 2013. *Phomopsis* (*Diaporthe*) Species as the Cause of Soybean Seed Decay in Korea. *Journal of Phytopathology* 161:131-134. DOI:10.1111/jph.12034.

Sutton DA, Timm WD, Morgan-Jones G, Rinaldi MG. 1999. Human phaeohyphomycotic osteomyelitis caused by the coelomycete *Phomopsis* saccardo 1905: criteria for identification, case history, and therapy. *Journal of Clinical Microbiology* 37(3):807-811.

Tan YP, Edwards J, Grice KRE, Shivas RG. 2013. Molecular phylogenetic analysis reveals six new species of *Diaporthe* from Australia. *Fungal Diversity* 61:251-260. DOI:10.1007/s13225-013-0242-9.

Tamura K. 1992 Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. *Molecular Biology and Evolution* 9:678-687.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* 10:512-526.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution*. DOI: 10.1093/molbev/mst197

Thangaduras D, Sangeetha J. 2013. Bioinformatics Tools for the Multilocus Phylogenetic Analysis of Fungi. In: Gupta VK, Tuohy MG, Ayyachamy M, Turner KM, O'Donovan A, eds. *Laboratory Protocols in Fungal Biology*. New York: Springer New York, 579-592. DOI:10.1007/978-1-4614-2356-0

Thompson SM, Tan YP, Young AJ, Neate SM, Aitken EAB, Shivas RG. 2011. Stem cankers on sunflower (*Helianthus annuus*) in Australia reveal a complex of pathogenic *Diaporthe* (*Phomopsis*) species. *Persoonia* 27:80-89.

Udayanga D, Castlebury LA, Rossman LA, Chukeatirote E, Hyde KD. 2014a. Insights into the genus *Diaporthe*: phylogenetic species delimitation in the *D. eres* species complex. *Fungal Diversity* 64:203-229. DOI: 10.1007/s13225-014-0297-2.

Udayanga D, Castlebury LA, Rossman AY, Hyde KD. 2014b. Species limits in *Diaporthe*: molecular re-assessment of *D. citri*, *D. cytosporella*, *D. foeniculina* and *D. rudis*. *Persoonia* 32:83-101.

Udayanga D, Liu X, Crous PW, McKenzie EHC, Chukeatirote E, Hyde KD. 2012. A multi-locus phylogenetic evaluation of *Diaporthe* (*Phomopsis*). *Fungal Diversity* 56:157-171. DOI: 10.1007/s13225-012-0190-9.

Udayanga D, Liu X, McKenzie EHC, Chukeatirote E, Bahkali AHA, Hyde KD. 2011. The genus *Phomopsis*: biology, applications, species concepts and names of common phytopathogens. *Fungal Diversity* 50:189-225. DOI: 10.1007/s13225-011-0126-9.

Van Niekerk JM, Groenewald JZ, Farr DF, Fourie PH, Halleen F, Crous PW. 2005. Reassessment of *Phomopsis* species on grapevines. *Australasian Plant Pathology* 34:27-39.

Van Rensburg JCJ, Lamprecht SC, Groenewald JZ, Castlebury LA, Crous PW. 2006. Characterisation of *Phomopsis* spp. associated with die-back of rooibos (*Aspalathus linearis*) in South Africa. *Studies in Mycology* 55:65-74.

Van Warmelo KT, Marasas WFO. 1972. *Phomopsis leptostromiformis*: the causal fungus of lupinosis, a mycotoxicosis, in sheep. *Mycologia* 64:316–324.

608 Vidić M, Petrović K, Đorđević V, Riccioni L. 2013. Occurrence of *Phomopsis longicolla* β
609 Conidia in Naturally Infected Soybean. *Journal of Phytopathology* 161:470-477. DOI:
610 10.1111/jph.12092

611 Walker DM, Castlebury LA, Rossman AY, White Jr. JF. 2012. New molecular markers for
612 fungal phylogenetics: Two genes for species-level systematics in the Sordariomycetes
613 (Ascomycota). *Molecular Phylogenetics and Evolution* 64:500-512.

614 Wang J, Xu X, Mao L, Lao J, Lin F, Yuan Z, Zhang C. 2014. Endophytic *Diaporthe* from
615 Southeast China are genetically diverse based on multi-locus phylogeny analyses. *World Journal*
616 *of Microbiology & Biotechnology* 30:237-243. DOI: 10.1007/s11274-013-1446-6.

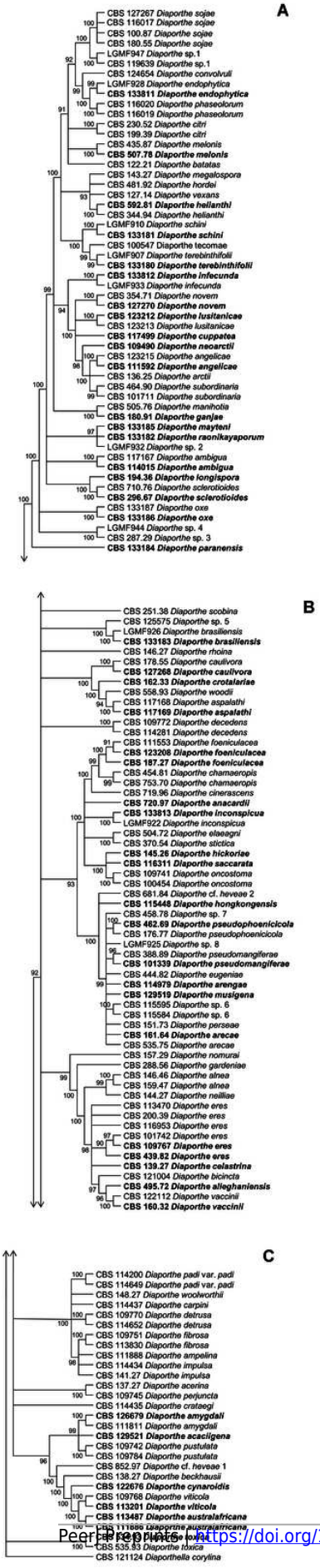
617 Webber JG, Gibbs JN. 1984. Colonization of elm bark by Phomopsis oblonga. *Transactions of*
618 *the British Mycological Society* 82:348-352.

619 Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary
620 kingdoms. *Proceedings of the National Academy of Sciencies of the United States of America*
621 74(11):5088-5090.

Figure 1

Figure 1 - MP condensed tree with a 90% cut-off, build using the five loci *TEF1-TUB-CAL-HIS*-ITS for the 96 *Diaporthe* species.

Ex-type or ex-epitype or isotype isolates are represented in **bold**.



Peer J Preprints | https://doi.org/10.7287/peerj.preprints.2822v1 | CC BY 4.0 Open Access | rec: 21 Feb 2017, publ: 21 Feb 2017

Figure 2

Figure 2 - ML condensed tree with a 90% cut-off, build using the five loci *TEF1-TUB-CAL-HIS*-ITS for the 96 *Diaporthe* species.

The percentage of trees in which the associated taxa clustered together is shown next to the branches. Ex-type, ex-epitype, or isotype isolates are represented in **bold**.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2822v1 | CC BY 4.0 Open Access | rec: 21 Feb 2017, publ: 21 Feb 2017

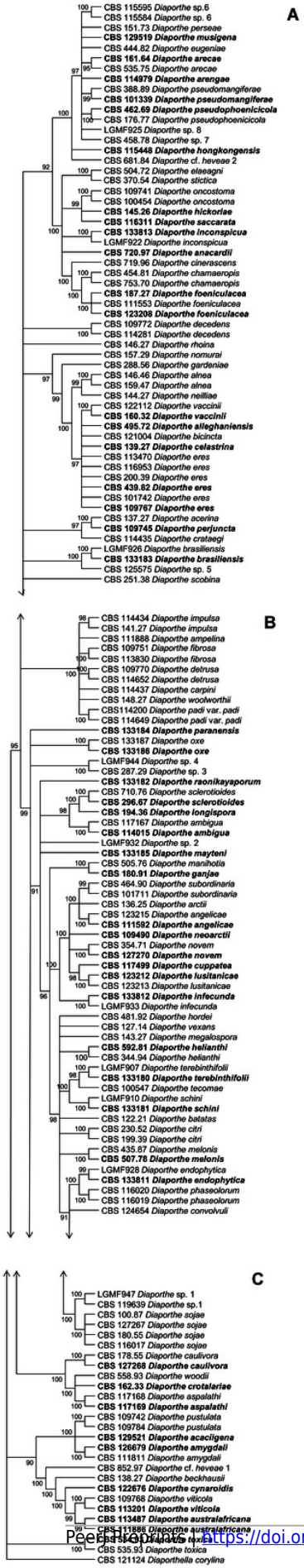


Figure 3

Figure 3 - MP condensed tree with a 90% cut-off build using the *TEF1* locus for the 96 *Diaporthe* species.

This locus generates the best single locus trees for the MP method. Ex-type, ex-epitype, or isotype isolates are represented in **bold**

Figure 4

Figure 4 - ML condensed tree with a 90% cut-off, build using the *TEF1* locus for the 96 *Diaporthe* species.

This locus generates the best single locus trees for the ML method. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Ex-type, ex-epitype, or isotype isolates are represented in **bold.**

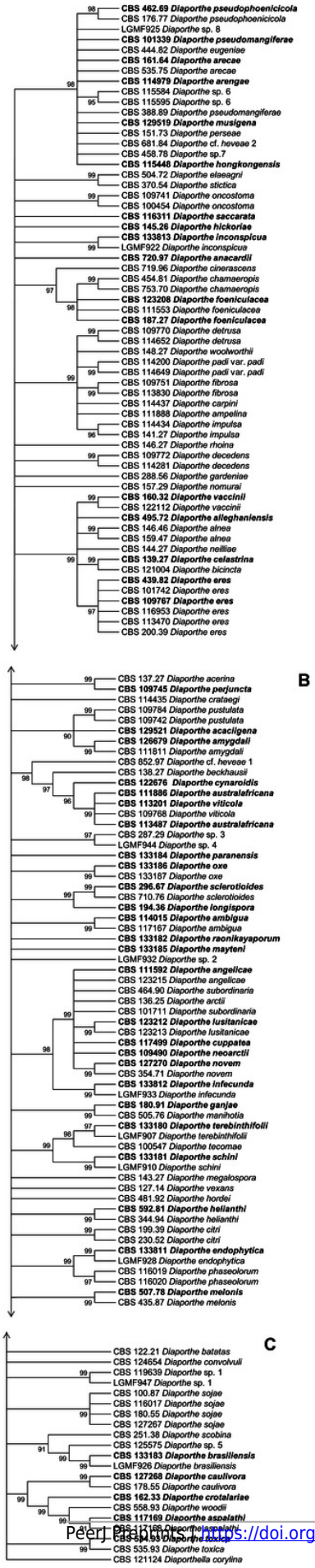


Figure 5

Figure 5 - Profiles of phylogenetic informativeness for the 96 Diaporthe species and 5 loci.

A) Net Phylogenetic informativeness. B) Phylogenetic informativeness per site.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2822v1 | CC BY 4.0 Open Access | rec: 21 Feb 2017, publ: 21 Feb 2017

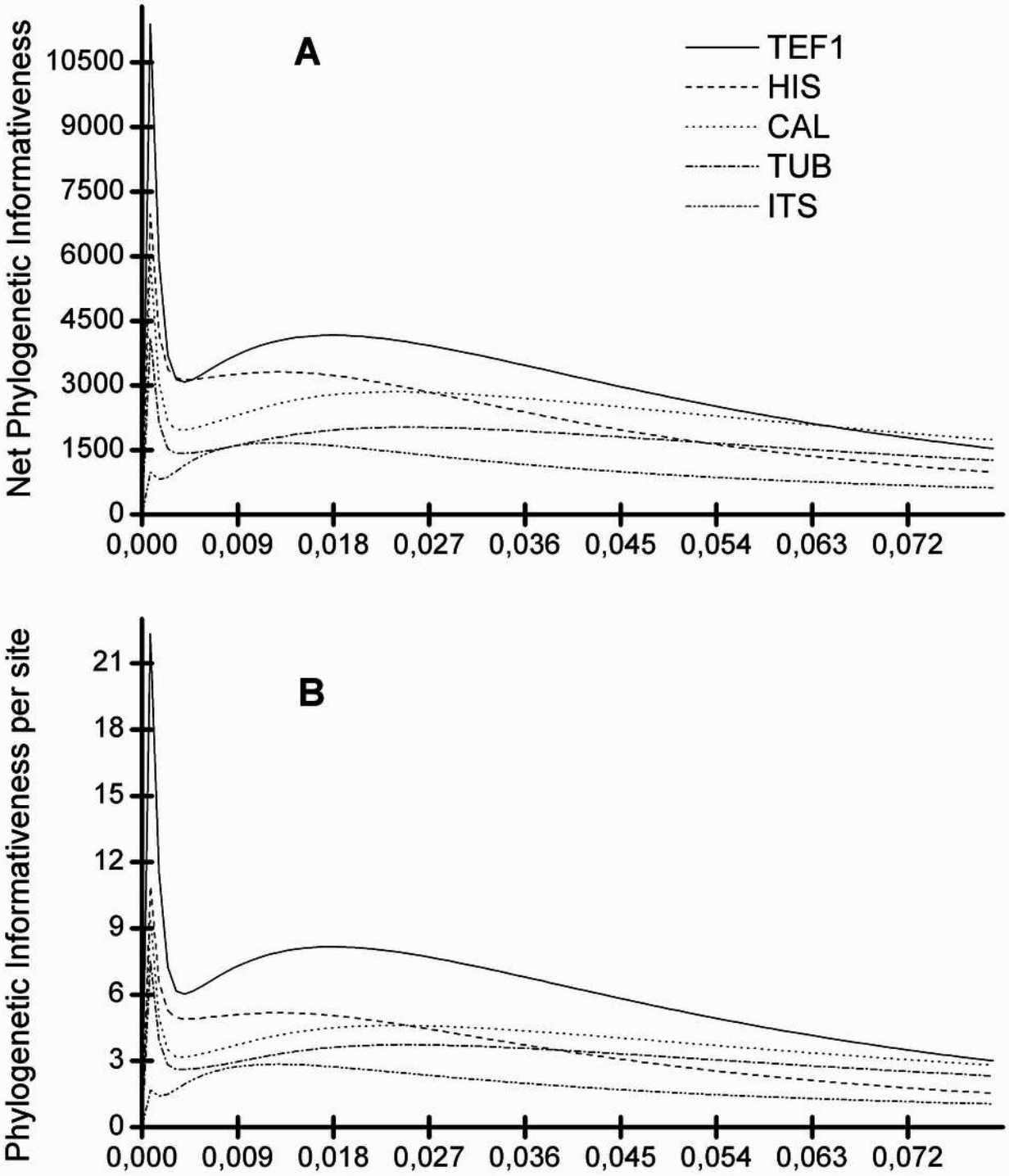


Table 1(on next page)

Models used to construct the ML trees.

Tree	Model	References
ITS	Tamura-Nei	Tamura & Nei, 1993
TEF1	Hasegawa-Kishino-Yano	Hasegawa, Kishino & Yano, 1985
TUB	Hasegawa-Kishino-Yano	Hasegawa, Kishino & Yano, 1985
HIS	General Time Reversible	Nei & Kumar, 2000
CAL	Tamura 3-parameter	Tamura, 1992
ITS - TEF1	Tamura-Nei	Tamura & Nei, 1993
ITS - TUB	Tamura-Nei	Tamura & Nei, 1993
ITS - HIS	Tamura-Nei	Tamura & Nei, 1993
ITS - CAL	Tamura-Nei	Tamura & Nei, 1993
TEF1 - TUB	Hasegawa-Kishino-Yano	Hasegawa, Kishino & Yano, 1985
TEF1 - HIS	Tamura-Nei	Tamura & Nei, 1993
TEF1 - CAL	Hasegawa-Kishino-Yano	Hasegawa, Kishino & Yano, 1985
TUB - HIS	General Time Reversible	Nei & Kumar, 2000
TUB - CAL	Hasegawa-Kishino-Yano	Hasegawa, Kishino & Yano, 1985
HIS - CAL	Hasegawa-Kishino-Yano	Hasegawa, Kishino & Yano, 1985
ITS - TEF1 - TUB	Tamura-Nei	Tamura & Nei, 1993
ITS - TEF1 - HIS	General Time Reversible	Nei & Kumar, 2000
ITS - TEF1 - CAL	Tamura-Nei	Tamura & Nei, 1993
ITS - TUB -HIS	General Time Reversible	Nei & Kumar, 2000
ITS - TUB - CAL	Tamura-Nei	Tamura & Nei, 1993
ITS - HIS - CAL	Tamura-Nei	Tamura & Nei, 1993
TEF1 - TUB - HIS	General Time Reversible	Nei & Kumar, 2000
TEF1 - TUB - CAL	Hasegawa-Kishino-Yano	Hasegawa, Kishino & Yano, 1985
TEF1 - HIS - CAL	Tamura-Nei	Tamura & Nei, 1993
TUB - HIS - CAL	Hasegawa-Kishino-Yano	Hasegawa, Kishino & Yano, 1985
ITS - TEF1 - TUB - HIS	General Time Reversible	Nei & Kumar, 2000
ITS - TEF1 - TUB - CAL	Tamura-Nei	Tamura & Nei, 1993
ITS - TEF1 - HIS - CAL	Tamura-Nei	Tamura & Nei, 1993
ITS - TUB - HIS - CAL	Tamura-Nei	Tamura & Nei, 1993
TEF1 - TUB - HIS - CAL	Hasegawa-Kishino-Yano	Hasegawa, Kishino & Yano, 1985
ITS - TEF1 - TUB - HIS - CAL	General Time Reversible	Nei & Kumar, 2000

Table 2(on next page)

MP trees scores.

Tree	No. trees	Length	Normalized length	Consistency index	Retention index	Composite index	Parsimony-informative sites
1gene							
ITS	1	1 200	1 .970	0 .278906	0 .765634	0 .244365	0 .213540
TEF1	1	2 830	4 .647	0 .280810	0 .773915	0 .229713	0 .217323
TUB	1	1 628	2 .673	0 .349176	0 .785012	0 .289798	0 .274107
HIS	1	1 880	3 .087	0 .285557	0 .729297	0 .224608	0 .208256
CAL	1	2 234	3 .668	0 .355136	0 .816321	0 .304750	0 .289905
2 genes							
ITS-TEF1	1	4 218	6 .926	0 .267368	0 .756266	0 .219278	0 .202201
ITS-TUB	1	2 977	4 .888	0 .303147	0 .758804	0 .250811	0 .230029
ITS-HIS	1	3 268	5 .366	0 .266073	0 .721901	0 .212506	0 .192078
ITS-CAL	1	3 657	6 .005	0 .308194	0 .780338	0 .259686	0 .240496
TEF1-TUB	2	4 535	7 .447	0 .300317	0 .772148	0 .245351	0 .231889
TEF1-HIS	1	4 828	7 .928	0 .275606	0 .749486	0 .220282	0 .206562
TEF1-CAL	1	5 206	8 .548	0 .304724	0 .784949	0 .252402	0 .239193
TUB-HIS	1	3 606	5 .921	0 .306263	0 .746843	0 .244391	0 .228730
TUB-CAL	1	3 975	6 .527	0 .342310	0 .795145	0 .287052	0 .272186
HIS-CAL	1	4 267	7 .007	0 .311460	0 .770357	0 .255101	0 .239935
3 genes							
ITS-TEF1-TUB	1	5 942	9 .757	0 .285318	0 .758687	0 .232892	0 .216467
ITS-TEF1-HIS	3	6 233	10 .235	0 .266876	0 .740786	0 .214166	0 .197698
ITS-TEF1-CAL	1	6 609	10 .852	0 .290524	0 .771371	0 .240083	0 .224102
ITS-TUB-HIS	1	4 989	8 .192	0 .288178	0 .737853	0 .231161	0 .212633
ITS-TUB-CAL	1	5 378	8 .831	0 .315121	0 .775889	0 .262285	0 .244499
ITS-HIS-CAL	2	5 661	9 .296	0 .293677	0 .757132	0 .240207	0 .222352
TEF1-TUB-CAL	1	6 910	11 .346	0 .311702	0 .781378	0 .257255	0 .243557
TEF1-TUB-HIS	1	6 537	10 .734	0 .290338	0 .754311	0 .233090	0 .219005
TEF1-HIS-CAL	1	7 209	11 .837	0 .294419	0 .766557	0 .239569	0 .225689
TUB-HIS-CAL	1	5 962	9 .790	0 .318135	0 .770532	0 .260290	0 .245133
4 genes							
ITS-TEF1-TUB-HIS	1	7 934	13 .028	0 .281222	0 .747108	0 .226279	0 .210103
ITS-TEF1-TUB-CAL	1	8 326	13 .672	0 .298775	0 .770405	0 .245945	0 .230178
ITS-TEF1-HIS-CAL	1	8 622	14 .158	0 .284827	0 .757809	0 .231684	0 .215844
ITS-TUB-HIS-CAL	1	7 364	12 .092	0 .302877	0 .759944	0 .247364	0 .230170
TEF1-TUB-HIS-CAL	2	8 911	14 .632	0 .301867	0 .767101	0 .245686	0 .231563
5 genes							
ITS-TEF1-TUB-HIS-CAL	1	10 327	16 .957	0 .292768	0 .759604	0 .238098	0 .222388

Table 3(on next page)

Data from the likelihood values using ML trees.

Tree	- log Likelihood	Normalizad - log Likelihood
1 gene		
ITS	- 6 778.9324	- 11.1313
TEF1	- 12 771.9747	- 20.9720
TUB	- 8 921.3230	- 14.6491
HIS	- 9 330.7481	- 15.3214
CAL	- 11.756.6407	- 19.3048
2 genes		
ITS-TEF1	- 20 494.0008	- 33.6519
ITS-TUB	- 16 381.1047	- 26.8984
ITS-HIS	- 16 835.4062	- 27.6443
ITS-CAL	- 19 449.5000	- 31.9368
TEF1-TUB	- 22 209.9657	- 36.4696
TEF1-HIS	- 22 707.1478	- 37.2860
TEF1-CAL	- 25 263.7157	- 41.4839
TUB-HIS	- 18 720.0479	- 30.7390
TUB-CAL	- 21 286.5020	- 34.9532
HIS-CAL	- 21 896.7086	- 35.9552
3 genes		
ITS-TEF1-TUB	- 29 959.8491	- 49.1952
ITS-TEF1-HIS	- 30 409.1656	- 49.9329
ITS-TEF1-CAL	- 33 105.3032	- 54.3601
ITS-TUB-HIS	- 26 256.8160	- 43.1146
ITS-TUB-CAL	- 29 008.0228	- 47.6322
ITS-HIS-CAL	- 29 425.9498	- 48.3185
TEF1-TUB-CAL	- 34 699.3754	- 56.9776
TEF1-TUB-HIS	- 32 201.5900	- 52.8762
TEF1-HIS-CAL	- 35 160.1260	- 57.7342
TUB-HIS-CAL	- 31 194.9713	- 51.2233
4 genes		
ITS-TEF1-TUB-HIS	- 39 950.3574	- 65.5999
ITS-TEF1-TUB-CAL	- 42 574.6960	- 69.9092
ITS-TEF1-HIS-CAL	- 42 940.9069	- 70.5105
ITS-TUB-HIS-CAL	- 38 862.9726	- 63.8144
TEF1-TUB-HIS-CAL	- 44 608.0234	- 73.2480
5 genes		
ITS-TEF1-TUB-HIS-CAL	- 52 626.8115	- 86.4151

1

2

3

Table 4(on next page)

Alignments characteristics.

Locus	No. Characters	No. Conserved sites (in %)	No. variable sites (in %)	No. Parsim-info sites (in %)
1 gene				
ITS	609	350 (57)	235 (39)	177 (29)
TEF1	535	128 (24)	382 (71)	328 (61)
TUB	603	220 (36)	323 (54)	279 (46)
HIS	688	329 (48)	311 (45)	259 (38)
CAL	667	194 (29)	425 (64)	370 (55)
2 genes				
ITS-TEF1	1149	478 (42)	617 (54)	505 (44)
ITS-TUB	1217	570 (47)	558 (46)	456 (37)
ITS-HIS	1302	679 (52)	546 (42)	436 (33)
ITS-CAL	1281	544 (42)	660 (52)	547 (43)
TEF1-TUB	1143	348 (30)	705 (62)	607 (53)
TEF1-HIS	1228	457 (37)	693 (56)	587 (48)
TEF1-CAL	1207	322 (27)	807 (67)	698 (58)
TUB-HIS	1296	549 (42)	634 (49)	538 (42)
TUB-CAL	1275	414 (32)	748 (59)	649 (51)
HIS-CAL	1360	523 (38)	736 (54)	629 (46)
3 genes				
ITS-TEF1-TUB	1757	698 (40)	940 (54)	784 (45)
ITS-TEF1-HIS	1842	807 (44)	928 (50)	764 (41)
ITS-TEF1-CAL	1821	672 (37)	1042 (57)	875 (48)
ITS-TUB-HIS	1910	899 (47)	869 (45)	715 (37)
ITS-TUB-CAL	1889	764 (40)	983 (52)	826 (44)
ITS-HIS-CAL	1974	873 (44)	971 (49)	806 (41)
TEF1-TUB-CAL	1815	542 (30)	1130 (62)	977 (54)
TEF1-TUB-HIS	1836	677 (379)	1016 (55)	866 (47)
TEF1-HIS-CAL	1900	651 (34)	1118 (59)	957 (50)
TUB-HIS-CAL	1968	743 (38)	1059 (54)	908 (46)
4 genes				
ITS-TEF1-TUB-HIS	2450	1027 (42)	1251 (51)	1043 (43)
ITS-TEF1-TUB-CAL	2429	892 (37)	1365 (56)	1154 (48)
ITS-TEF1-HIS-CAL	2514	1001 (40)	1353 (54)	1134 (45)
ITS-TUB-HIS-CAL	2582	1093 (42)	1294 (50)	1085 (42)
TEF1-TUB-HIS-CAL	2508	871 (35)	1441 (57)	1236 (49)
5 genes				
ITS-TEF1-TUB-HIS-CAL	3102	1221 (39)	1676 (54)	1413 (46)

1

2

3

Table 5(on next page)

Average changes in tree resolution when a locus is added or removed.

Each row indicates the locus that is added to the trees. Each column indicates the difference between trees build using n or n -1 loci. For example, row ITS, columns 4→3, indicate the average differences between every pair of 3 and 4 loci trees that include the ITS locus, using either a MP or a ML approach. The higher the number, the more different the two trees in the pair are, on average. “Average” columns indicate the average changes for all columns when a specific locus is considered. Darker cells indicate smaller average changes (and thus smaller information losses) when a locus is added from phylogenetic trees

	MP					ML					
	5 → 4	**4 → 3**	**3 → 2**	**2 → 1**	**Overall**		**5 → 4**	**4 → 3**	**3 → 2**	**2 → 1**	**Overall**
ITS	1.416	1.407	31.9	36.875	17.8995	ITS	12	23.75	24.3888889	31.75	22.9722222
TEF1	2.963	2.95075	44.7333333	44.375	23.7555208	TEF1	22	25.25	30.6111111	42.25	30.0277778
TUB	1.705	1.70575	41.4	43.75	22.1401875	TUB	16	24.375	28.6666667	41.375	27.6041667
HIS	2.001	1.998	41.2	40.5	21.42475	HIS	12	19	24.8888889	37	23.2222222
CAL	2.393	29	42.6	42.125	29.0295	CAL	10	24.5	25.3333333	39.125	24.7395833

1