ITS2 Database V: Twice as Much

Markus J. Ankenbrand,¹ Alexander Keller,¹ Matthias Wolf,² Jörg Schultz,² and Frank Förster*²
¹Department of Animal Ecology and Tropical Biology, Julius Maximilian University, Würzburg, Germany
²Department of Bioinformatics, Julius Maximilian University, Würzburg, Germany
*Corresponding author: E-mail: frank.foerster@biozentrum.uni-wuerzburg.de.
Associate editor: Koichiro Tamura

Abstract

The internal transcribed spacer 2 (ITS2) is a well-established marker for phylogenetic analyses in eukaryotes. A reliable resource for reference sequences and their secondary structures is the ITS2 database (http://its2.bioapps.biozentrum.uni-wuerzburg.de/). However, the database was last updated in 2011. Here, we present a major update of the underlying data almost doubling the number of entities. This increases the number of taxa represented within all major eukaryotic clades. Moreover, additional data has been added to underrepresented groups and some new groups have been added. The broader coverage across the tree of life improves phylogenetic analyses and the capability of ITS2 as a DNA barcode.

Key words: barcoding, database, internal transcribed spacer, phylogeny, sequence-structure, toolbox.

Introduction

The internal transcribed spacer 2 (ITS2) of the ribosomal cistron is a well-established marker in eukaryotic molecular systematics (Schultz and Wolf 2009). With a relatively variable sequence it is well suited for low-level analyses, yet limited for distantly related taxa (Baldwin 1992). However, ITS2 exhibits a common core of secondary structure (Schultz et al. 2005) making it a valuable marker also on higher taxonomic levels (Coleman 2003). Furthermore, inclusion of the secondary structure improves the accuracy and robustness of phylogenetic tree reconstructions (Keller et al. 2010) and allows for distinguishing cryptic/pseudocryptic species via compensatory base changes (Müller et al. 2007; Coleman 2009; Ruhl et al. 2010). Recently, it has also been applied in DNA (meta-)barcoding (Chen et al. 2010; Yao et al. 2010; Pang et al. 2012; Keller et al. 2015).

In 2006, we developed the ITS2 database to provide a central resource for ITS2 sequences and their individual secondary structures (Schultz et al. 2006). In the following years, the ITS2 database was further expanded from a data repository to a rather full featured interactive workbench (Selig et al. 2008; Koetschan et al. 2010, 2012; Wolf et al. 2014). Data of the ITS2 workbench consist of sequences extracted from NCBI (NCBI Resource Coordinators 2015) that are automatically trimmed using Hidden Markov Models (Keller et al. 2009). The workbench determines complete individual secondary structures for ITS2 sequences based on energy minimization (Markham and Zuker 2008) or iterative homology modelling (Wolf et al. 2005). Additionally, partial structures are predicted for entries with as few as two helices (Koetschan et al. 2010). Finally, ITS2 sequences without a predicted structure are included as sequence-only entities (Koetschan et al. 2010). During the automatic structure validation all entries have to match the four helix core. Thus, other ITS2 structures are not represented in our database. Basic analyses like reannotation, secondary structure prediction, sequence-structure alignment, and tree calculation can be directly performed in the web-based database (Merget et al. 2012). The last update of the underlying data was performed in 2011. Meanwhile, the NCBI database experienced a drastic increase in sequence content (supplementary table S1, Supplementary Material online). Moreover, the NCBI Taxonomy (Federhen 2012) is continuously revised to reflect the current knowledge of the evolutionary history of represented taxa. We thus performed a major update on the ITS2 workbench to benefit from this increased amount of data and make it available to the scientific ITS2 communities.

In the following, we report the most prominent improvements in terms of stored data, taxonomic coverage, and changes in major lineages.

Results

The new version of the database now contains 711,172 sequences, which nearly doubles the 379,329 of the previous release. In detail, the number of entries matching the eukaryotic core structure increased by 84%, and those with a partial structure increased by 217%. In contrast, the number of sequences without structure decreased by 11%. Similarly, the number of different species and genera represented in the database increased by 59% and 23% respectively. Overall, the proportional increase in number of new sequences was distributed across all major groups of eukaryotes (table 1).

The taxonomic lineage for each sequence was updated to the current NCBI Taxonomy and also showed some major changes. The NCBI TaxDs for 7,464 sequences were changed since the last update. 3,743 entries present in 2011 are altered in the current update (supplementary table S2, Supplementary Material online).

Discussion

When calculating reliable phylogenetic trees or when performing DNA barcoding analyses, it is essential to have a
trustworthy reference database with good coverage over all major taxonomic groups of interest. With this update of the ITS2 workbench, we were able to increase the number of taxa represented within all major eukaryote clades by a large amount of newly included species and genera. Besides the actual underlying sequence data, this update also aimed to revise the taxonomic status from the last 4 years according to current knowledge, as reflected on the NCBI Taxonomy database.

The ITS region has not only been used for phylogenetic reconstruction, but also as a DNA barcode to identify fungal species (Schoch et al. 2012) and plant species (Chen et al. 2010; Yao et al. 2010; Keller et al. 2015). Basic DNA barcoding is already applicable through the integrated BLAST search on the workbench or by downloading the reference data to train barcoding classifiers (Sickel et al. 2015). Besides the ITS2 workbench, only the original NCBI databases and the BOLD system (Ratnasingham and Hebert 2007) allow identification of ITS2 barcodes. For the latter, it is stated that it is an unvalidated database with very few entries, limited to fungal species (http://www.boldsystems.org/index.php/IDS_OpenIdEngine, last viewed May 29, 2015).

The ITS2 workbench includes all of the necessary features to be used as a reference database and is thus a valuable resource beyond the use of phylogenetics. This is reflected in the good coverage of currently known plant species that have been mapped in the United States, as provided by the Biodiversity Information Serving Our Nation website (http://bison.usgs.gov/). Now, 72% of the listed species are covered in the ITS2 workbench which shows an increase of more than 20% compared with the previous version.

To summarize, the update of the ITS2 workbench facilitates and broadens the usage of ITS2 as a phylogenetic marker and, additionally, as a DNA barcode.

### Supplementary Material

Supplementary material is available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

### Acknowledgments

The authors thank Christian Koetschan for curation and maintenance of the database. Additionally, the authors thank Shirley J. McCaffrey, Patricia M. Barassi, and Paul J. Bartels for proofreading our manuscript. MJA was supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg.

### References


