# ITS2 database V: Twice as much[*]

Markus J. Ankenbrand,[1] Alexander Keller,[1]
Matthias Wolf,[2] Jörg Schultz[2] and Frank Förster[*,2]

[1]Department of Animal Ecology and Tropical Biology, Julius Maximilian University, Würzburg, Germany

[2]Department of Bioinformatics, Julius Maximilian University, Würzburg, Germany

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.

## 1 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/pseudo-cryptic species via compensatory base

[*] This is a pre-copyedited, author-produced PDF of an article accepted for publication in Molecular Biology and Evolution following peer review. The version of record: Markus J. Ankenbrand, Alexander Keller, Matthias Wolf, Jörg Schultz, and Frank Förster, *ITS2 database V: Twice as much* Molecular Biology and Evolution, first published online August 6, 2015, doi:10.1093/molbev/msv174 is available online at: `http://mbe.oxfordjournals.org/content/early/2015/08/05/molbev.msv174`

changes (CBCs) (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 re-annotation, 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). 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.

## 2 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 **??**).

The taxonomic lineage for each sequence was updated to the current NCBI Taxonomy and also showed some major changes. The NCBI TaxIDs for 7,464

Table 1: Number of sequences and percent change (n.d. means not defined) for main groups of the revised classification of eukaryotes according to Adl *et al.* (2012), data comparison based on 2011 and 2015 (Last accessed 2015-06-14). Group names mapped onto current NCBI taxonomy database (Supplementary Table S3). The taxon "others" comprises all eukaryotic sequences which could not be mapped into the group names defined by Adl *et al.* (2012).

| Taxon | 2011 | 2015 | change |
|---|---:|---:|---:|
| Alveolata | 5,733 | 10,431 | +81.9 % |
| Ancyromonadida | 19 | 28 | +47.4 % |
| Apusomonadida | 3 | 4 | +33.3 % |
| Breviatea | 1 | 1 | 0.0 % |
| Centrohelida | 0 | 1 | n.d. |
| Cercozoa | 206 | 310 | +50.5 % |
| Chloroplastida | 122,497 | 208,822 | +70.5 % |
| Choanomonada | 8 | 8 | 0.0 % |
| Collodictyonidae | 0 | 0 | n.d. |
| Cryptophyceae | 82 | 234 | +185.4 % |
| Dictyostelia | 207 | 365 | +76.3 % |
| Discoba | 823 | 1,284 | +56.0 % |
| Foraminifera | 265 | 265 | 0.0 % |
| Fungi | 206,777 | 405,445 | +96.1 % |
| Glaucophyta | 0 | 20 | n.d. |
| Haptophyta | 38 | 51 | +34.2 % |
| Ichthyosporea | 469 | 1,217 | +159.5 % |
| Kathablepharidae | 5 | 6 | +20.0 % |
| Malawimonadidae | 0 | 0 | n.d. |
| Metamonada | 299 | 502 | +67.9 % |
| Metazoa | 27,859 | 55,645 | +99.7 % |
| Nucleariida | 2 | 2 | 0.0 % |
| Polycystinea | 4 | 43 | +975.0 % |
| Rhodophycea | 764 | 1,278 | +67.3 % |
| Rigifilida | 1 | 1 | 0.0 % |
| Stramenopila | 12,005 | 20,728 | +72.7 % |
| Telonema | 2 | 2 | 0.0 % |
| Tubulinea | 2 | 8 | +300.0 % |
| others | 695 | 4,338 | +524.2 % |

sequences were changed since the last update. 3,743 entries present in 2011 are altered in the current update (Supplementary Table S2).

## 3 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 four 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 2015-05-29).

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 USA, as provided by the Biodiversity Information Serving Our Nation website (`http://bison.usgs.ornl.gov`). Now 72 % of the listed species are covered in the ITS2 workbench which shows an increase of more than 20 % compared to 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.

## 4 Supplementary Material

Supplementary material comprising material and methods section, tables and figures are available at Molecular Biology and Evolution online (`http://www.mbe.oxfordjournals.org/`).

## 5 Acknowledgments

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

## References

Adl, S. M., Simpson, A. G., Lane, C. E., Luke, J., Bass, D., Bowser, S. S., Brown, M., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., leGall, L., Lynn, D. H., McManus, H., Mitchell, E. A. D., Mozley-Stanridge, S. E., Parfrey, L. W., Pawlowski, J., Rueckert, S., Shadwick, L., Schoch, C., Smirnov, A., and Spiegel, F. W. 2012. The revised classification of eukaryotes. *J Eukaryot Microbiol*, 59(5): 429–514.

Baldwin, B. G. 1992. Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: An example from the compositae. *Mol Phylogenet Evol*, 1(1): 3–16.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y., and Leon, C. 2010. Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species. *PLoS ONE*, 5(1).

Coleman, A. W. 2003. ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends Genet*, 19(7): 370–375.

Coleman, A. W. 2009. Is there a molecular key to the level of "biological species" in eukaryotes? A DNA guide. *Mol Phylogenet Evol*, 50(1): 197–203.

Federhen, S. 2012. The NCBI Taxonomy database. *Nucleic Acids Res*, 40(Database issue): D136–D143.

Keller, A., Schleicher, T., Schultz, J., Müller, T., Dandekar, T., and Wolf, M. 2009. 5.8s-28s rRNA interaction and HMM-based ITS2 annotation. *Gene*, 430(12): 50–57.

Keller, A., Förster, F., Müller, T., Dandekar, T., Schultz, J., and Wolf, M. 2010. Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees. *Biology Direct*, 5(1): 4.

Keller, A., Danner, N., Grimmer, G., Ankenbrand, M., von der Ohe, K., von der Ohe, W., Rost, S., Härtel, S., and Steffan-Dewenter, I. 2015. Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. *Plant Biol*, 17(2): 558–566.

Koetschan, C., Förster, F., Keller, A., Schleicher, T., Ruderisch, B., Schwarz, R., Müller, T., Wolf, M., and Schultz, J. 2010. The ITS2 Database III:

sequences and structures for phylogeny. *Nucleic Acids Res*, 38(Database issue): D275–D279.

Koetschan, C., Hackl, T., Müller, T., Wolf, M., Förster, F., and Schultz, J. 2012. ITS2 Database IV: Interactive taxon sampling for internal transcribed spacer 2 based phylogenies. *Mol Phylogenet Evol*, 63(3): 585–588.

Markham, N. R. and Zuker, M. 2008. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol*, 453: 3–31.

Merget, B., Koetschan, C., Hackl, T., Förster, F., Dandekar, T., Müller, T., Schultz, J., and Wolf, M. 2012. The ITS2 Database. *J Vis Exp*, (61).

Müller, T., Philippi, N., Dandekar, T., Schultz, J., and Wolf, M. 2007. Distinguishing species. *RNA*, 13(9): 1469–1472.

NCBI Resource Coordinators 2015. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 43(Database issue): D6–D17.

Pang, X., Shi, L., Song, J., Chen, X., and Chen, S. 2012. Use of the potential DNA barcode ITS2 to identify herbal materials. *J Nat Med*, 67(3): 571–575.

Ratnasingham, S. and Hebert, P. D. N. 2007. bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol Ecol Notes*, 7(3): 355–364.

Ruhl, M. W., Wolf, M., and Jenkins, T. M. 2010. Compensatory base changes illuminate morphologically difficult taxonomy. *Mol Phylogenet Evol*, 54(2): 664–669.

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., and Fungal Barcoding Consortium 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Natl Acad Sci U S A*, 109(16): 6241–6246.

Schultz, J. and Wolf, M. 2009. ITS2 sequencestructure analysis in phylogenetics: A how-to manual for molecular systematics. *Mol Phylogenet Evol*, 52(2): 520–523.

Schultz, J., Maisel, S., Gerlach, D., Müller, T., and Wolf, M. 2005. A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. *RNA*, 11(4): 361–364.

Schultz, J., Müller, T., Achtziger, M., Seibel, P. N., Dandekar, T., and Wolf, M. 2006. The internal transcribed spacer 2 databasea web server for (not only) low level phylogenetic analyses. *Nucleic Acids Res*, 34(suppl 2): W704–W707.

Selig, C., Wolf, M., Müller, T., Dandekar, T., and Schultz, J. 2008. The ITS2 Database II: homology modelling RNA structure for molecular systematics. *Nucleic Acids Res*, 36(Database issue): D377–D380.

Sickel, W., Ankenbrand, M. J., Grimmer, G., Holzschuh, A., Härtel, S., Lanzen, J., Steffan-Dewenter, I., and Keller, A. 2015. Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. *BMC Ecol.*, 15(1): 20.

Wolf, M., Achtziger, M., Schultz, J., Dandekar, T., and Müller, T. 2005. Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures. *RNA*, 11(11): 1616–1623.

Wolf, M., Koetschan, C., and Müller, T. 2014. ITS2, 18s, 16s or any other RNA  simply aligning sequences and their individual secondary structures simultaneously by an automatic approach. *Gene*, 546(2): 145–149.

Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y., Pang, X., Xu, H., Zhu, Y., Xiao, P., and Chen, S. 2010. Use of ITS2 Region as the Universal DNA Barcode for Plants and Animals. *PLoS ONE*, 5(10): e13102.