Sarah Wiechers[1], Kai F. Müller[1]
1) Evolution and Biodiversity of Plants Group, Institute for Evolution and Biodiversity, WWU Münster, Hüfferstr. 1, 48149 Münster, Germany

The GBOL5 Web App – Plant barcode management and analysis

DNA barcoding

DNA barcoding allows to identify species using **predefined target genes**. It proves especially useful when samples are **degraded**, **fragmented** or consist of **hard to identify** parts, e. g. larval stages of insects or seeds and roots of plants (Reviewed in Hollingsworth et al. 2016).

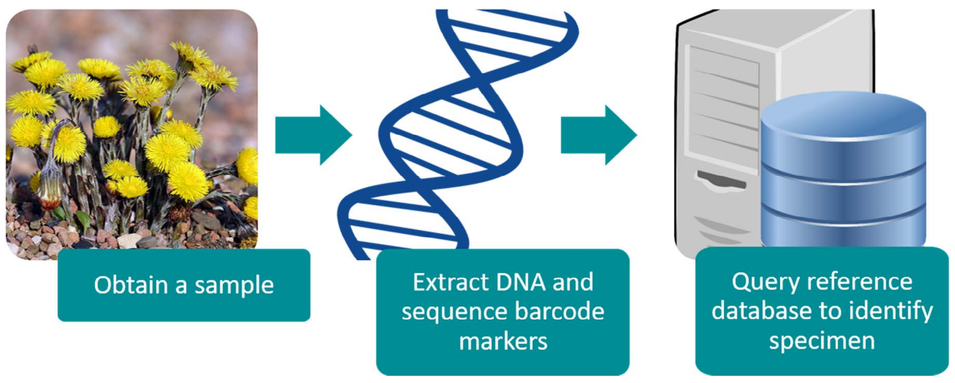

Figure 1 After obtaining a sample and sequencing one or multiple marker genes a reference database can be queried with the resulting sequence or sequences to assign one or (in case of mixed samples and metabarcoding studies) multiple species to the sample.

GBOL

Multiple national and international groups and consortia work on building a **reference database** of these marker gene sequences. The **German Barcode of Life (GBOL)** project aims to create such a database of all plants, animals and fungi in Germany and develop applications of DNA barcoding. The responsibility of the **GBOL5** sub-project is barcoding the approximately 4,800 **land plant** species native to Germany.

Barcode analyses

Integrated into the structure of the web application is an automated, **taxonomy driven quality assessment** of generated marker sequences to scan for misidentification of specimen and errors during sequencing before this data is used for analyses or uploaded to other reference databases. This is done using the **SATIVA** pipeline (Kozlov et al. 2016).

Featured analyses include the assessment of a taxon set for the existence of a barcoding gap (the lack of overlap between variations in infraspecific distances on the one hand and interspecific distances on the other), species-level **monophyly** for the gene regions used as markers, and the **identification success** of these markers.

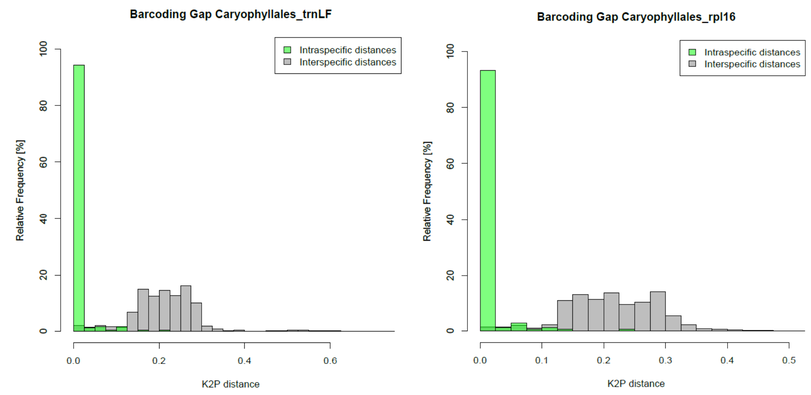

Figure 6 Genetic distance histograms illustrating the intra- (green) and interspecific (grey) distances of species in the order Caryophyllales for the marker gene regions trnLF (left) and rpl16 (right) as calculated with the R package *spider* (Brown et al. 2012).

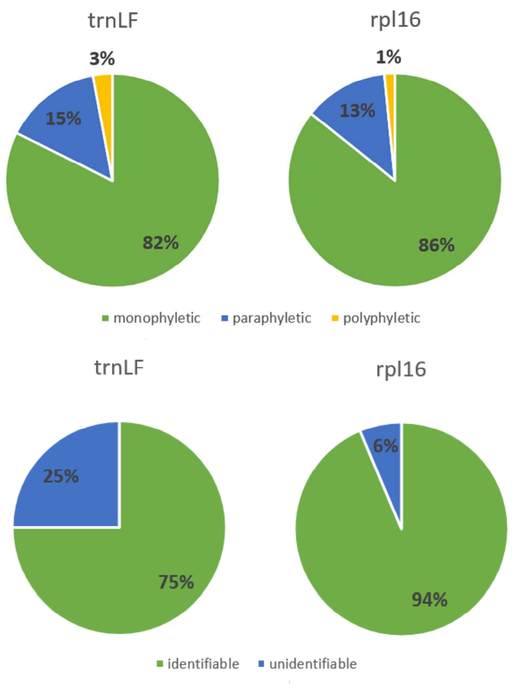

Figure 7 Monophyly assessment. Diagrams show the fraction of species in the order Caryophyllales that are mono- (green), para- (blue) or polyphyletic (yellow) for the marker gene regions trnLF (left) and rpl16 (right) as calculated with the *Monophylizer* tool (Mutanen et al. 2016).

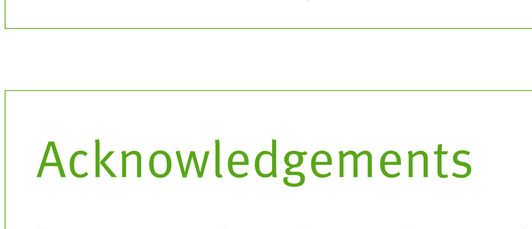

Figure 8 Identification success. Diagrams show the fraction of species that are (uniquely) identifiable by sequences of the marker gene regions trnLF (left) and rpl16 (right). A species was considered identifiable, if all sequences representing it were not identical to any sequences representing other analyzed species (Korotkova et al. 2011).

The GBOL5 Web App

Coordinating the efforts of the participating institutes from different parts of Germany created the need for a **shared information management system**. **Data** relevant to the project, e. g. target species as well as uploaded data, e.g. sequences for the four barcode markers featured in the project (trnK-matK, rpl16, trnLF and ITS), are available from all devices through an **online interface**. The app also provides extensive functionality for **laboratory and taxonomic management** as well as features such as **automatic primer read trimming and assembly**. A selection of the available tools and features is presented in detail below.

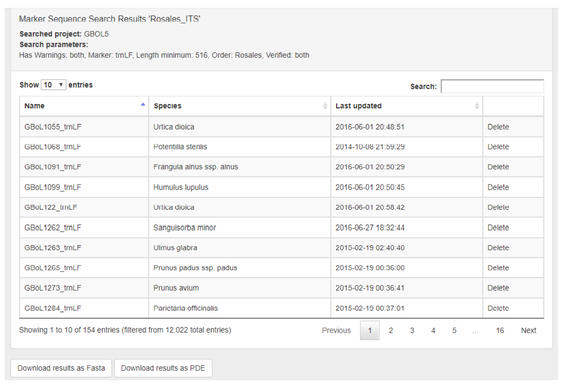

Figure 2 Advanced search. This feature allows users to filter contigs, marker sequences and specimen using various parameters and download the results.

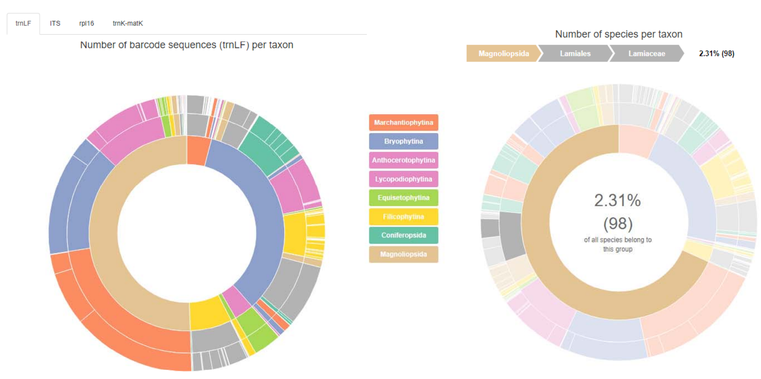

Figure 3 Interactive progress visualization. The left diagram shows the number of available barcode sequences for a selected taxon. In the right diagram the number of target species per taxon is shown.

Figure 4 Primer read view. Various parameters can be changed here to influence automatic read trimming. Manual trimming and editing of called bases is possible as well from this view.

Figure 5 Contig view. Assembled reads can be viewed, edited and verified by users here. Parameters for the automatic assembly can be changed in this view as well.

Acknowledgements

The funding of the German Barcode of Life project (currently in its second phase) by the Federal Ministry of Education and Research is highly appreciated. The authors are very thankful to the GBOL5 project partners from the Nees Institute (Bonn), the Botanical Garden and Museum Berlin and the Stuttgart State Museum of Natural History for contributing data to the GBOL5 web application.

Citations: Peter M. Hollingsworth, De-Zhu Li, Michelle van der Bank, Alex D. Twyford (2016). Telling plant species apart with DNA: from barcodes to genomes. *Phil. Trans. R. Soc. B*, 371(1702); Alexey M. Kozlov, Jiajie Zhang, Pelin Yilmaz, Frank Oliver Glöckner and Alexandros Stamatakis (2016). Phylogeny-aware Identification and Correction of Taxonomically Mislabeled Sequences. *Nucleic Acids Research;* Korotkova, N., Borsch, T., Quandt, D., Taylor, N. P., Müller, K. F., & Barthlott, W. (2011). What does it take to resolve relationships and to identify species with molecular markers? An example from the epiphytic Rhipsalideae (Cactaceae). *American Journal of Botany*, 98(9), 1549–1572; Mutanen, M. et al. (2016). Species-Level Poly- and Paraphyly in DNA barcode Gene Trees: Strong Operational Bias in European Lepidoptera. *Systematic Biology*; Brown et al. (2012). Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources*; DNA artwork retrieved March 15, 2018 from https://thecliparts.com/the-best-free-clipart-24647/; Database server artwork retrieved March 15, 2018 from https://openclipart.org/image/2400px/svg_to_png/163711/database-server.png; Photo of Coltsfoot retrieved March 15, 2018 from https://de.m.wikipedia.org/wiki/Datei:Coltsfoot.jpg