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).

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 GBOL 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. It scans 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).

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 (matK, 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.

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 Berlin Botanical Garden and Museum Berlin and the Stuttgart State Museum of Natural History for contributing data to the GBOL5 web application.

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

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.

Figure 6 Genetic distance Histograms illustrating the intra-(green) and interspecific (grey) distances of species in the order Caryophyllales for the marker gene regions matK (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), paraphylic (white) and polyphyletic (blue) for the marker gene regions matK (left) and rpl16 (right) as calculated with the MonophylyBar tool (Müller et al. 2016-).}

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