Classification of Mixed Plant Samples by Next-Generation Sequencing

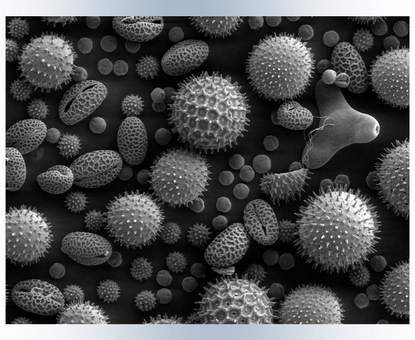
Markus Ankenbrand[1,2], Gudrun Grimmer[1], Stephan Härtel[1], Ingolf Steffan-Dewenter[1], Alexander Keller[1]

[1] Department of Animal Ecology and Tropical Biology (Zoology III), University of Würzburg
[2] Department of Bioinformatics, University of Würzburg

Motivation: Identification of species in mixed plant samples plays an important role in ecology and sheds light on problems from various research fields. Examples for such samples are pollens (also in bee collections and honey), algae in water samples, food or detritus. The advent of high-throughput experiments rendered it possible to obtain sequence data for such samples as an alternative to manual, microscopic classification by experts. But tools for the automated classification of such samples originating from multiple plant species have not been established yet.

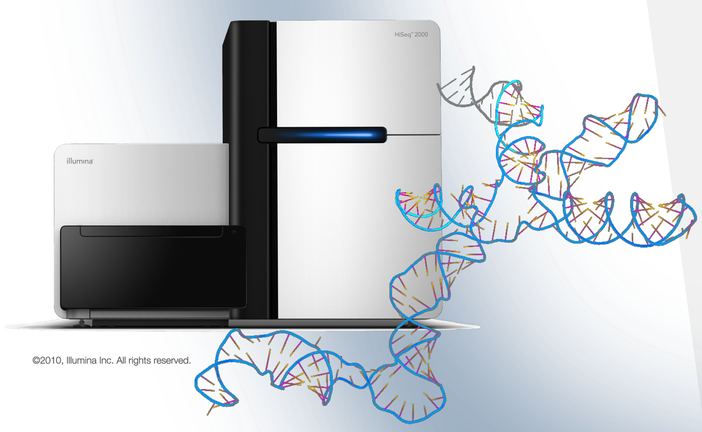
Source

Various biological sources can yield mixed plant samples



Mixed Plant Sample

(e.g. Pollen)



Lab work and sequencing

Next-Generation Sequencing of ITS2 regions

The RDPClassifier [2] was trained using all 73,853 ITS2 Sequences of viridiplantae from the ITS2 database [3]

Classification

ITS2 database trained RDPClassifier

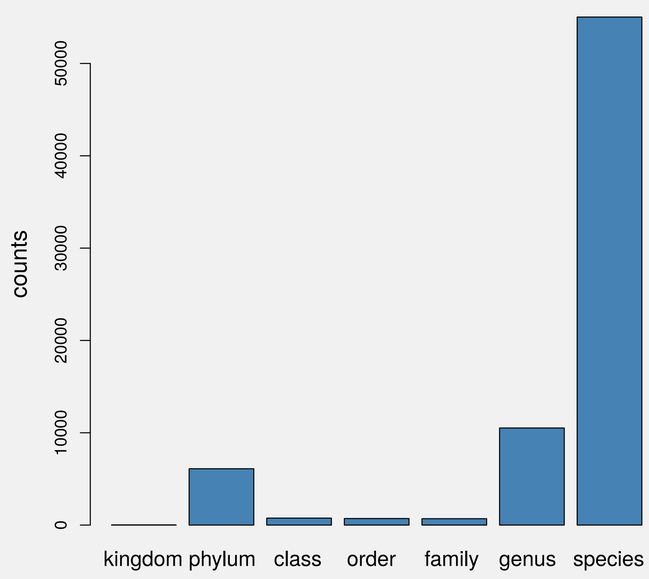
Analysis and Interpretation

(e.g. Krona) [4]

Classification Pipeline

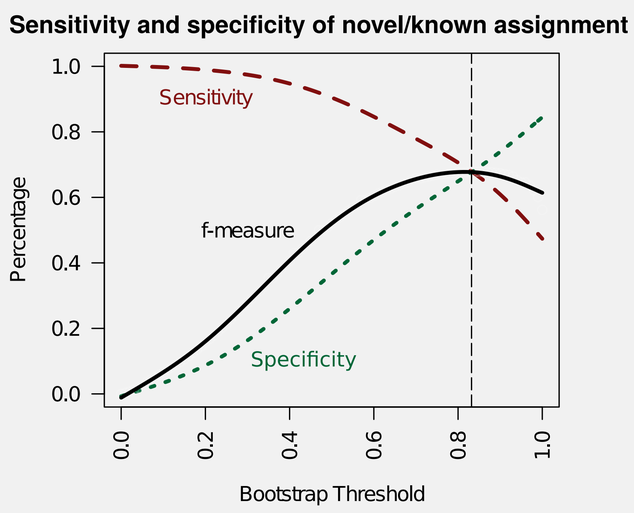
We developed a bioinformatical workflow to analyze sequences of mixed plant samples through the highly variable species-specific internal transcribed spacer 2 (ITS2) region of the nuclear ribosomal DNA. ITS2 sequences are classified with the naïve bayesian RDPclassifier specifically trained with reference sequences from the ITS2 database. To evaluate the performance, we (1) self-validated the database sequences, (2) identified specificity and sensitivity and (3) compared results from classical identification based on light microscopy with our sequencing results for 16 bee collected pollen samples.

Depth of correct assignment (full database)



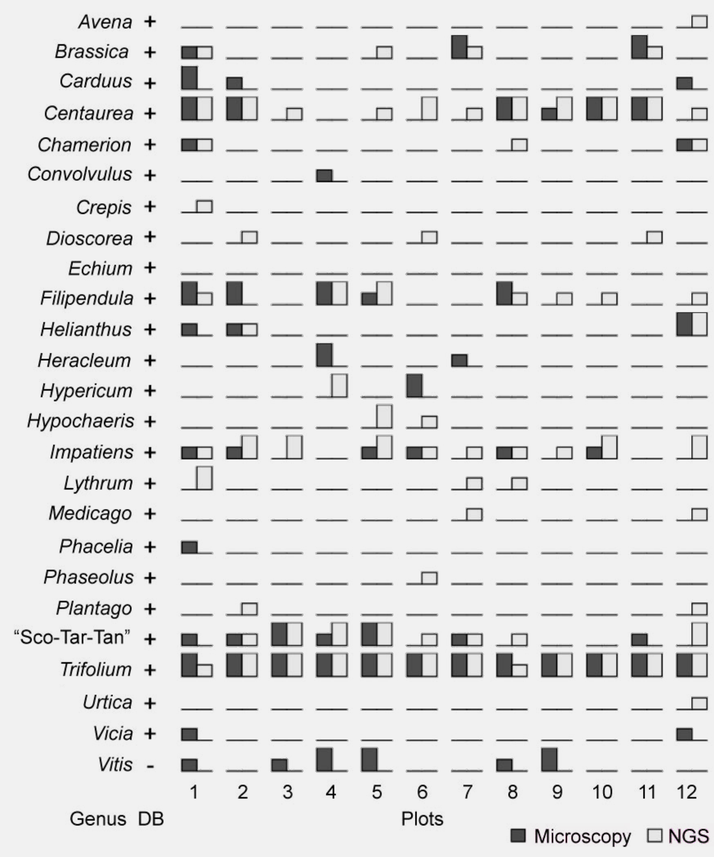
(1) 65,546 of the 73,853 database sequences were identifiable at species or genus level. 6,104 sequences could only be identified to phylum level.

Sensitivity and specificity of novel/known assignment

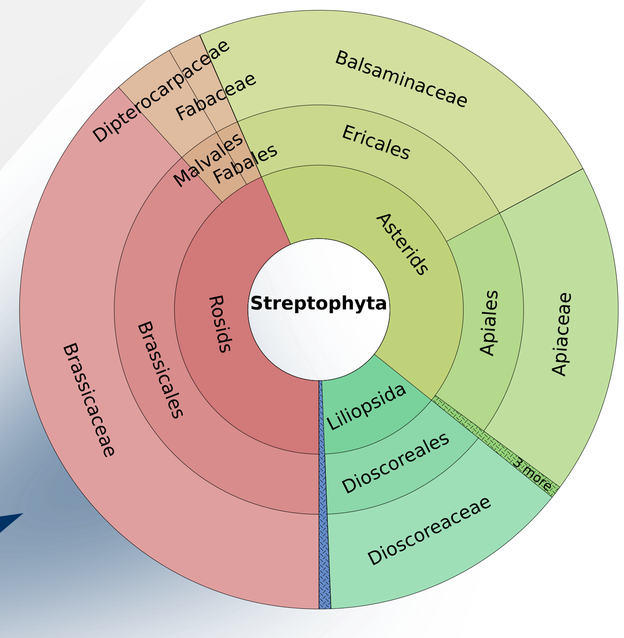


(2) The classifier is able to distinguish between novel and known sequences with 70% sensitivity and specificity.

Comparison of manual and automatic classification



(3) Classification through optical microscopy and NGS of 12 pollen samples yield largely consistent results.



The classification algorithm was computationally validated and its application to biological samples resulted in higher taxon richness (deeper assignments and more identified taxa) compared to light microscopy. The pipeline presents a useful and efficient workflow to identify plants at the genus and species level without requiring specialized expert knowledge and with high throughput.[1]

[1] Keller, A., Danner, N., Grimmer, G., Ankenbrand, M., von der Ohe, K., von der Ohe, W., Rost, S., Härtel, S. and Steffan-Dewenter, I. (2014). Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. Plant Biology, in press.
[2] Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology, 73(16), 5261–7.
[3] Schultz, J., Müller, T., Achtziger, M., Seibel, P. N., Dandekar, T., & Wolf, M. (2006). The internal transcribed spacer 2 database--a web server for (not only) low level phylogenetic analyses. Nucleic Acids Research, 34(Web Server issue), W704–7.
[4] Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. BMC Bioinformatics, 12(1), 385.

Graduate School Life Sciences

markus.ankenbrand@uni-wuerzburg.de