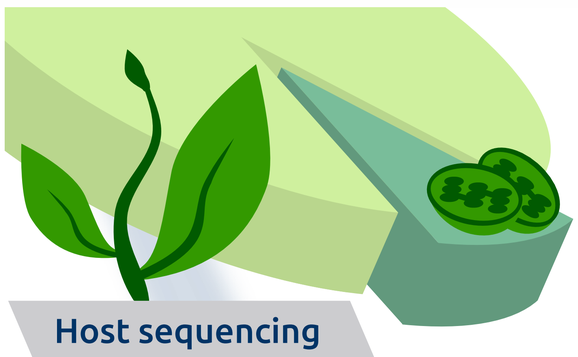
Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

ChloroExtractor: A fully automated plastid assembly pipeline reveals dozens of novel plastid genomes

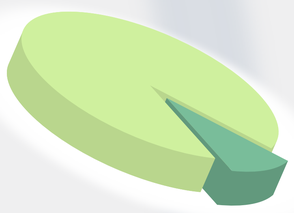
Thomas Hackl[1,2], Markus Ankenbrand[1], Niklas Terhoeven[1], Clemens Weiß[1], Frank Förster[1]

[1] Department of Bioinformatics, AG Genomics, University of Wuerzburg
[2] Department of Molecular Plant Physiology and Biophysics, University of Wuerzburg

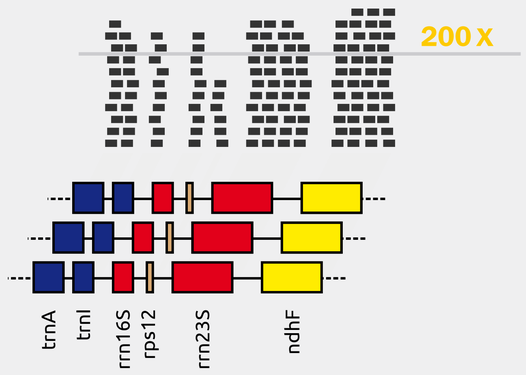
In times of large scale high troughput sequencing, novel plastid genomes mostly emerge from host genome sequencing projects. Here we present **ChloroExtractor**, a fully automated pipeline designed to generate high quality plastid assemblies from heterogeneous short read data. By applying our software to 100 publicly available plant sequencing libraries, we salvaged **27 novel and complete plastid genomes** across a wide range of plant taxa. Thus, we consider ChloroExtractor a valuable tool in the process of further unraveling plastid biology and evolution.



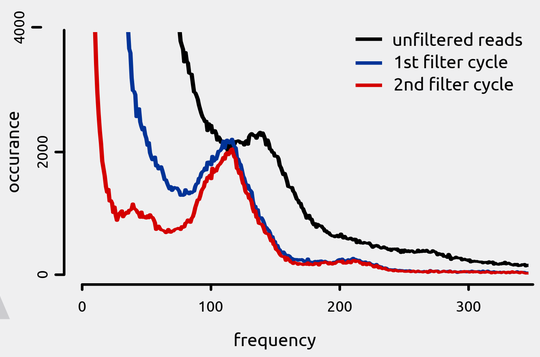
Host sequencing short read data

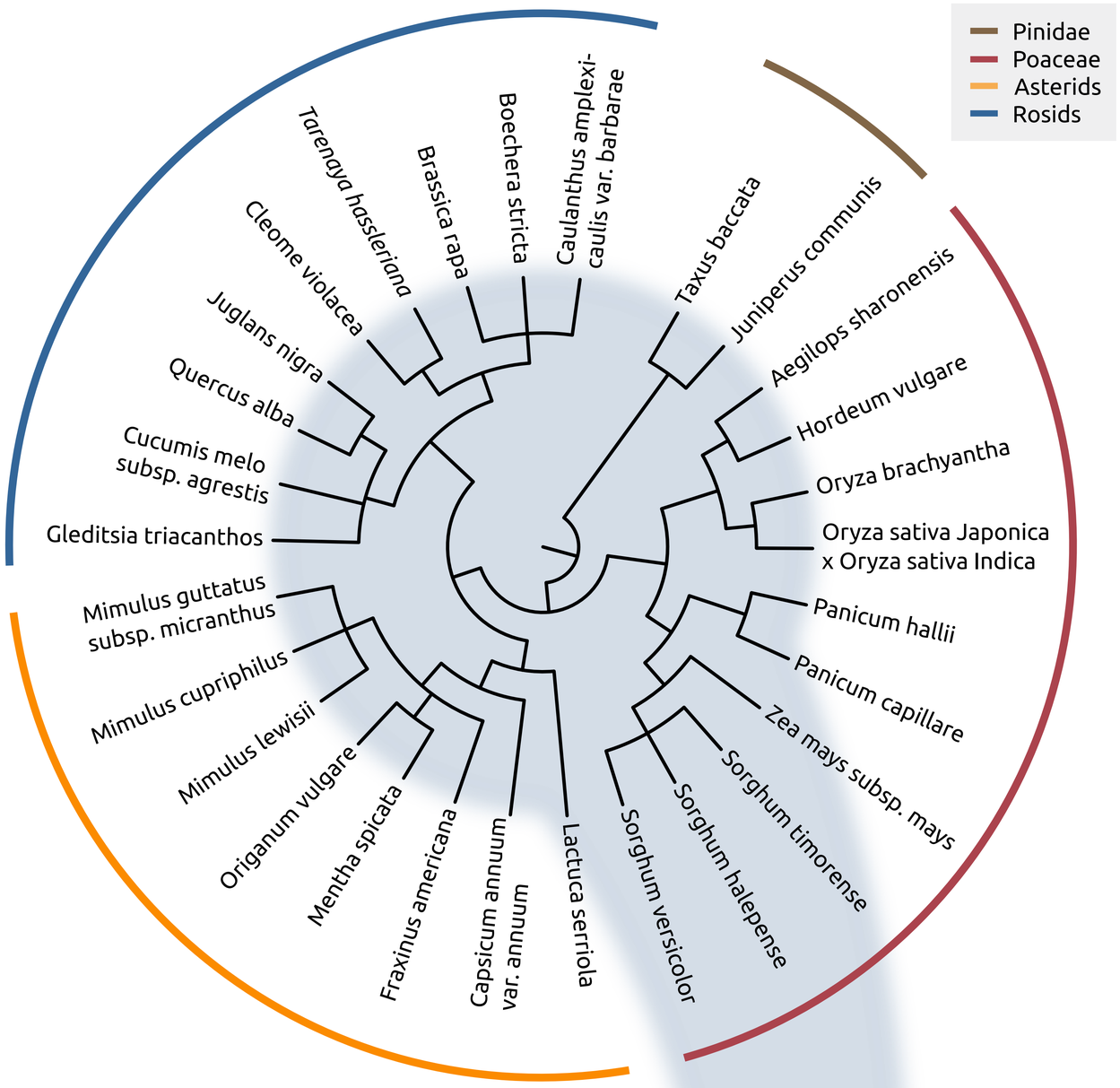


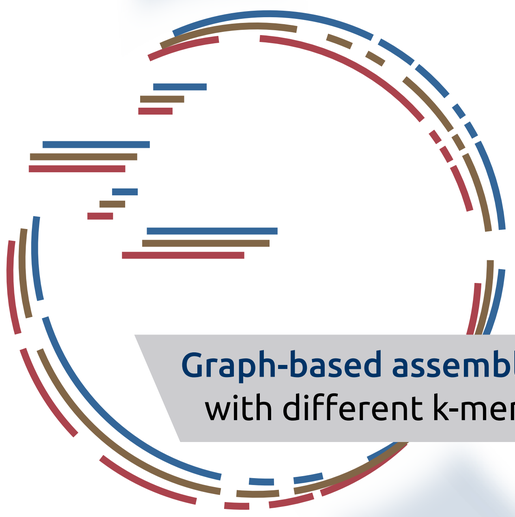
Scaling of data set based on **CDS cluster** coverage



Enrichment of plastid reads by **k-mer coverage**



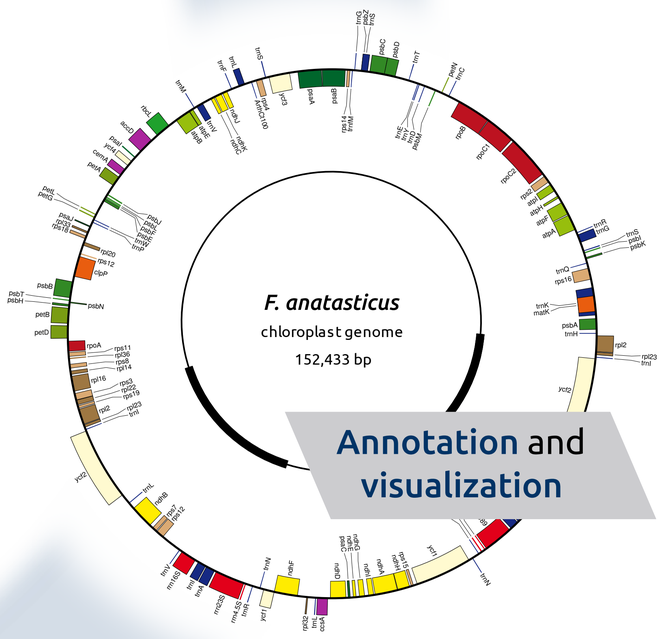




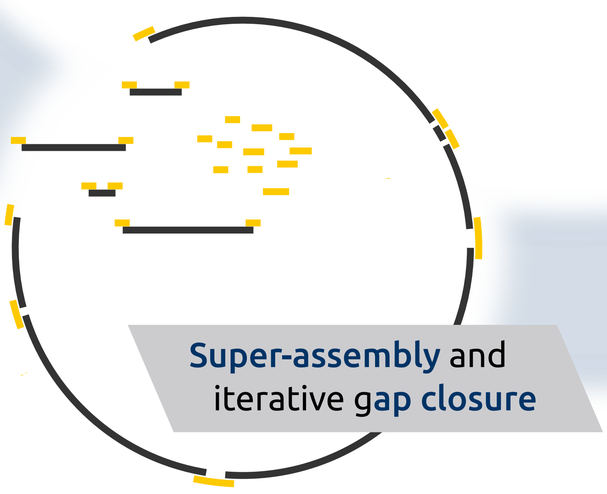
Graph-based assembly with different k-mer sizes

The relative content of plastid DNA is estimated from reads mapped with Bowtie2[1] onto clusters of plastid-gene coding sequences. The initial data set is scaled to an approximate 200-fold plastid coverage. Using k-mer frequencies counted with Jellyfish[2], the set is purged of host sequences by iterative removal of reads with low frequency k-mers.

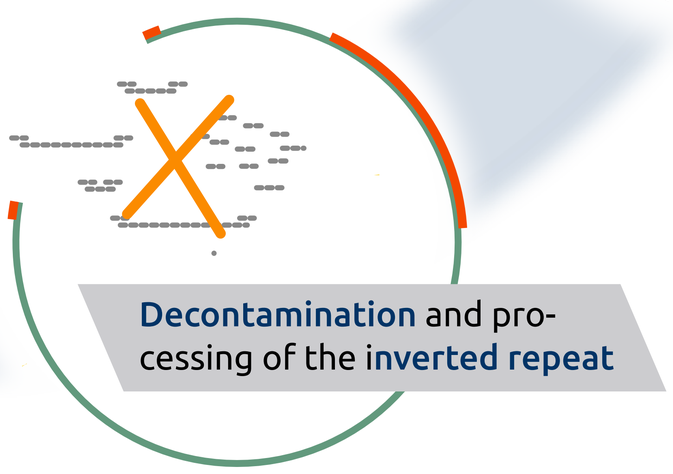
Primary assemblies are generated using Velvet[3] with k-mer sizes between 33 and 93. A contiguous super-assembly is calculated with Phrap[4]. Contig extensions and gap patches are obtained through a newly developed algorithm from paired end mapping information and local micro-assemblies. Contaminations are filtered by homology and the collapsed inverted repeat sequence is recovered. The final genome assembly is annotated with CpGAVAS[5] and visualized with OGDraw[6].



Annotation and **visualization**



Super-assembly and iterative g**ap closure**



Decontamination and pro-cessing of the i**nverted repeat**

thomas.hackl@uni-wuerzburg.de

[1] Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods.
[2] Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics (Oxford, England), 27, 764–770.
[3] Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research, 18, 821–829.
[4] Green, P. (2009). Phrap, version 1.090518. Retrieved from http://phrap.org
[5] Liu, C., Shi, L., Zhu, Y., Chen, H., Zhang, J., Lin, X., & Guan, X. (2012). CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics, 13, 715.
[6] Lohse, M., Drechsel, O., & Bock, R. (2007). OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Current Genetics, 52(5-6), 267–74.

European Research Council
Established by the European Commission