**Doing *de novo* Assembly and Comparative Genomics on Eukaryotic Species Mixtures**

**Bastian Greshake\*1**, Simonida Zehr1, Francesco Dal Grande2, Anjuli Meiser3, Imke Schmitt2,3, Ingo Ebersberger1

1Department for Applied Bioinformatics, Institute for Cell Biology and Neuroscience, Goethe University, Frankfurt am Main, Germany

2Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany

3Institute of Ecology, Evolution and Diversity, Goethe University Frankfurt, Max- von-Laue Str. 13, D-60438 Frankfurt, Germany

\* bgreshake@googlemail.com

Mutualistic symbiotic relationships are found across organisms of all complexity. In extreme instances, as in some lichens, the interaction appears so close that the participating organisms grow only poorly – or even not at all – when cultivated in isolation. This renders mutualistic symbionts valuable objects to study the genomic basis of adaptation and co-evolution. At the same time, however, the close interdependence confounds genomic studies. The separate sequencing of the participating organisms is not feasible in many cases, leaving metagenomics approaches as the method of choice..While there has been extensive work on prokaryotic metagenomics, it is still unclear to what extent larger and more complex eukaryotic genomes can be reconstructed from metagenomic data. Here we use *in silico*-generated data sets to sound out the performance of different assembly paradigms on such eukaryotic species mixtures,. The insights from the simulation study form then the basis for guiding the sequencing and analysis of a real lichen model.

To emulate real data sets as closely as possible, we sequenced a eukaryotic species mixture, generated from the lichen *Lasallia pustulata*. Based on the observed parameters, such as the insert size distribution and read number, we generated 11 twin sets based on the draft genomes of the lichenized fungus *Cladonia grayi* and its photobiont *Asterochloris sp.* The ratio of sequencing reads stemming from the two organisms was varied between the twin sets, to evaluate how the coverage distribution influences the assembly outcomes. We used a diverse set of six assemblers (*MIRA*, *Omega*, *sga*, *Velvet*, *MetaVelvet* & *SPAdes*), covering 2 different *Overlap-Layout* methods, 2 *de Bruijn Graph* based methods and 2 dedicated metagenome assemblers, for the evaluation.

Our results show that assemblers are not only dependent on the absolute coverage, but results are also dependent on the relative ratio with which the two organisms are present in the data set, with standard de Bruijn Graph methods being especially sensitive to uneven ratios. We furthermore see that the standard procedure of maximizing the N50 value is not advisable for eukaryotic species mixtures as it yields suboptimal assemblies.

On this basis we have begun investigating the metagenome of *Lasallia pustulata*. Combining Illumina short read and PacBio long read sequencing, we succeeded in assembling the genome of the mycobiont. With this we contribute a high quality draft genome to the clade of the Lecanoromycetes, for which so far very little genomic data is available. Comparing the gene content and selection pressures in the mycobiont to a conserved gene set of closely related non-lichenized fungi, we evaluate which core genes are potentially lost or change their mode of evolution due to the formation of the symbiosis. Additionally we are also able to recover large parts of the photobiont’s genome and get insights into the microbial diversity living in and on lichens.