**Doing *de novo* Assembly on Eukaryotic Species Mixtures – An *in silico* Evaluation**

**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, this symbiosis is so close that the participating organisms are no longer able to survive on their own. Such close interdependences make mutualistic symbionts valuable objects to study the genomic basis of co-evolution. At the same time this close interdependence confounds genomic studies, as the separate sequencing of the participating organisms is not feasible for many closely interacting symbionts. While there has been extensive work on prokaryotic metagenomics, it is still unclear how a eukaryotic species mixture influences the assembly outcomes

We investigate how different assembly paradigms perform on such eukaryotic species mixtures, using *in silico*-generated data sets. These insights from those data sets are then applied to 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 data 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 data sets, to evaluate how the coverage distribution influences the assembly outcomes. We used 6 assemblers (*MIRA*, *Omega*, *sga*, *Velvet*, *MetaVelvet* & *SPAdes*), covering different 2 different *Overlap-Layout-Consensus* methods, 2 *de Bruijn Graph* based methods and 2 dedicated metagenome assemblers, for the evaluation.

Our results show that there are marked differences in the performance of the different assemblers. All assemblers depend on the absolute coverage, with a minimum coverage needed to yield good results. Furthermore the results are also dependent on the relative ratio with which the two organisms are present in the data set. We also see that the standard procedure of maximizing the N50 value is not advisable for eukaryotic species mixtures as it yields suboptimal assemblies.

Applying the different strategies inferred from the twin sets to our real data of *Lasallia pustulata*, we find that the relative assembler performance by and large stays the same, validating our simulation approach. Nevertheless the assembly quality in terms of assembly length and contiguity stays behind the expectations generated from the twin sets. This is possibly a result of the more complex species composition in the real data. Ameliorating this by additionally using mate pair and single molecule sequencing, we are able to reconstruct and characterize the genome of *L. pustulata*.