De novo Assembly and Comparative Genomics on Eukaryotic Species Mixtures

Bastian Greshake*[1], Andreas Blaumeiser[1], Simonida Zehr[1], Francesco Dal Grande[2], Anjuli Meiser[3], Imke Schmitt[2,3], Ingo Ebersberger[1]

[1]Department for Applied Bioinformatics, Institute for Cell Biology and Neuroscience, Goethe University, Frankfurt am Main, Germany
[2]Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany
[3]Institute of Ecology, Evolution and Diversity, Goethe University Frankfurt, Frankfurt, Germany

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 in such communities 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, the insert size distribution, read number and read length, 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 different coverage ratios affect the assembly outcomes. We benchmarked 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.

Our results show – as expected – that the performance of all assemblers depends on the absolute coverage of the individual genomes. Interestingly, however, we see that the qualities of the individual assemblies is also affected by the relative ratios with which the two organisms are present in the sequence data. In particular, standard de Bruijn Graph methods appear particularly sensitive to uneven coverage ratios. We furthermore show that the common procedure of assembly parameter choice using the maximization of the assembly N50 value as objective is not advisable for eukaryotic species mixtures. In extreme cases this strategy can lead to the preclusion of an entire genome from the assembly.

On this basis we have begun investigating the metagenome of *Lasallia pustulata* using a combined approach of Illumina short read and PacBio long read sequencing. From this data we have assembled the genome of the mycobiont, a major fraction of the algal genome – which appears underrepresented in the metagenome -, as well as partial genomes from the microbiome that is associated with this lichen. In particular the high quality draft genome (N50: 1.5 Mbp) of the fungus *L. pustulata* represents a relevant addition to the clade of the Lecanoromycetes, for which so far very little genomic data is available. Integrating this data with genome sequences of closely related non-lichenized fungi now facilitates a high resolution analysis of how lichenization affects genome evolution. In particular, we discuss lineage-specific gene loss and changes in the mode of selection in lichenized vs. non-lichenized fungi.