De novo Assembly and Comparative Genomics on Eukaryotic Species Mixtures

Bastian Greshake†, Andreas Blaumeiser†, Simonida Zehr†,
Francesco Dal Grande*, Anjuli Meiser§, Imke Schmitt*§, Ingo Ebersberger†

† Department for Applied Bioinformatics, Institute for Cell Biology and Neuroscience, Goethe University, Frankfurt am Main, Germany
* Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany
§ Institute of Ecology, Evolution and Diversity, Goethe University, Frankfurt am Main, Germany

Summary

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. 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 and a major fraction of the algal genome. Integrating this data with genome sequences of closely related non-lichenized fungi now facilitates a high resolution analysis of how lichenization affects genome evolution.

1. *in silico* Sequencing & Assembler Evaluation

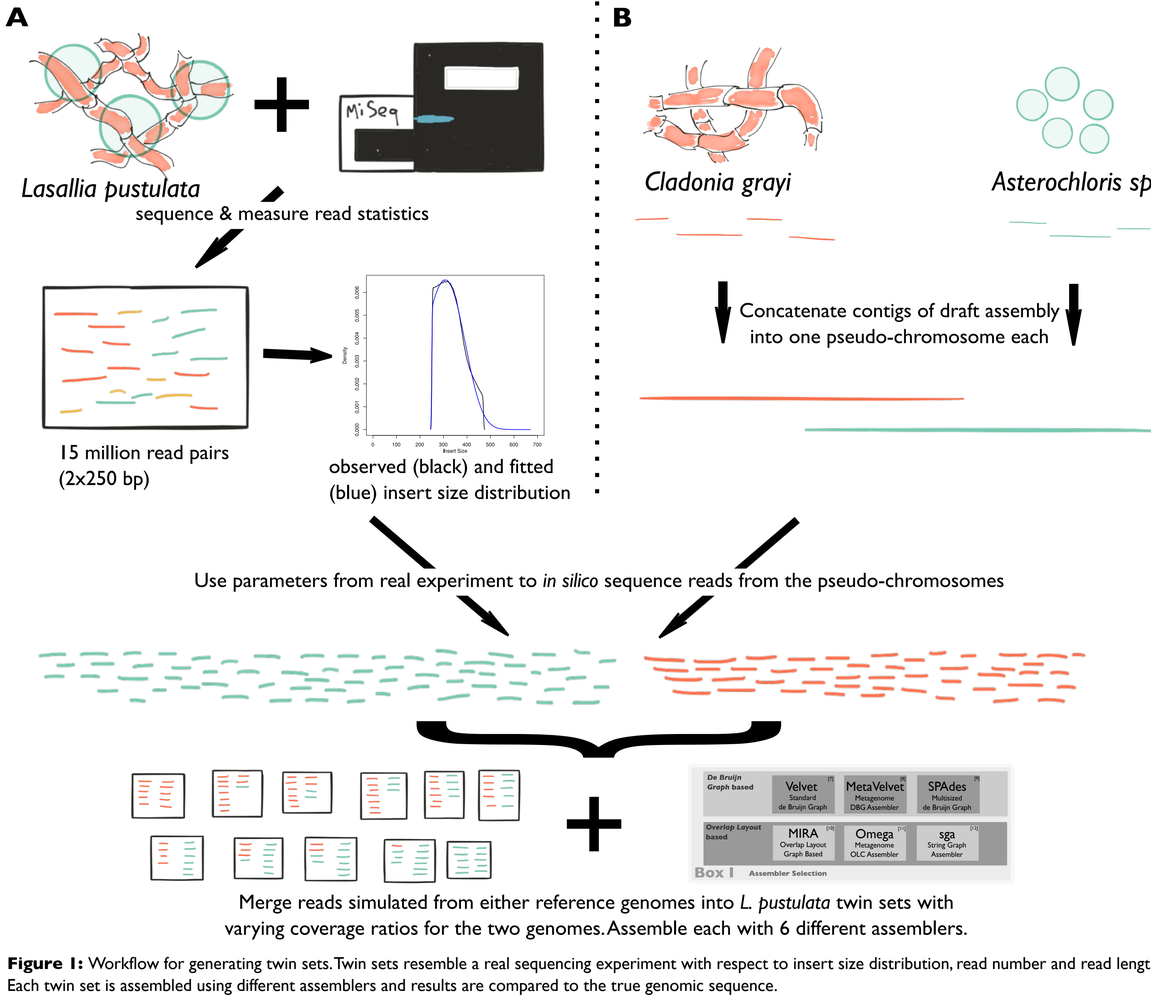


Figure 1: Workflow for generating twin sets. Twin sets resemble a real sequencing experiment with respect to insert size distribution, read number and read length. Each twin set is assembled using different assemblers and results are compared to the true genomic sequence.

Sequencing & Parameter Estimation DNA from *L. pustulata* was sequenced using Illumina MiSeq. The insert size distribution was estimated overlapping read pairs using *FLASH* [1] and fitting a Weibull distribution to the observed insert size distribution (Fig. 1, A). Scaffolds of *Cladonia grayi* [2] and *Asterochloris sp.* [3] were each concatenated to create a contiguous pseudo-chromosome, respectively (Fig. 1, B).

Simulation Using the pseudo-chromosomes as templates, we simulated reads with *ART* [6], parameterized with the values estimated from the *L. pustulata* data. Subsequently, we compiled 11 twin sets by mixing simulated fungal & algal reads at varying ratios (Table 1).

Assembly Twin sets were assembled using different programs, including dedicated metagenome assemblers (Box I). Assemblies were compared for contiguity and correctness of assembly.

2. Assembly Results of the Twin Sets

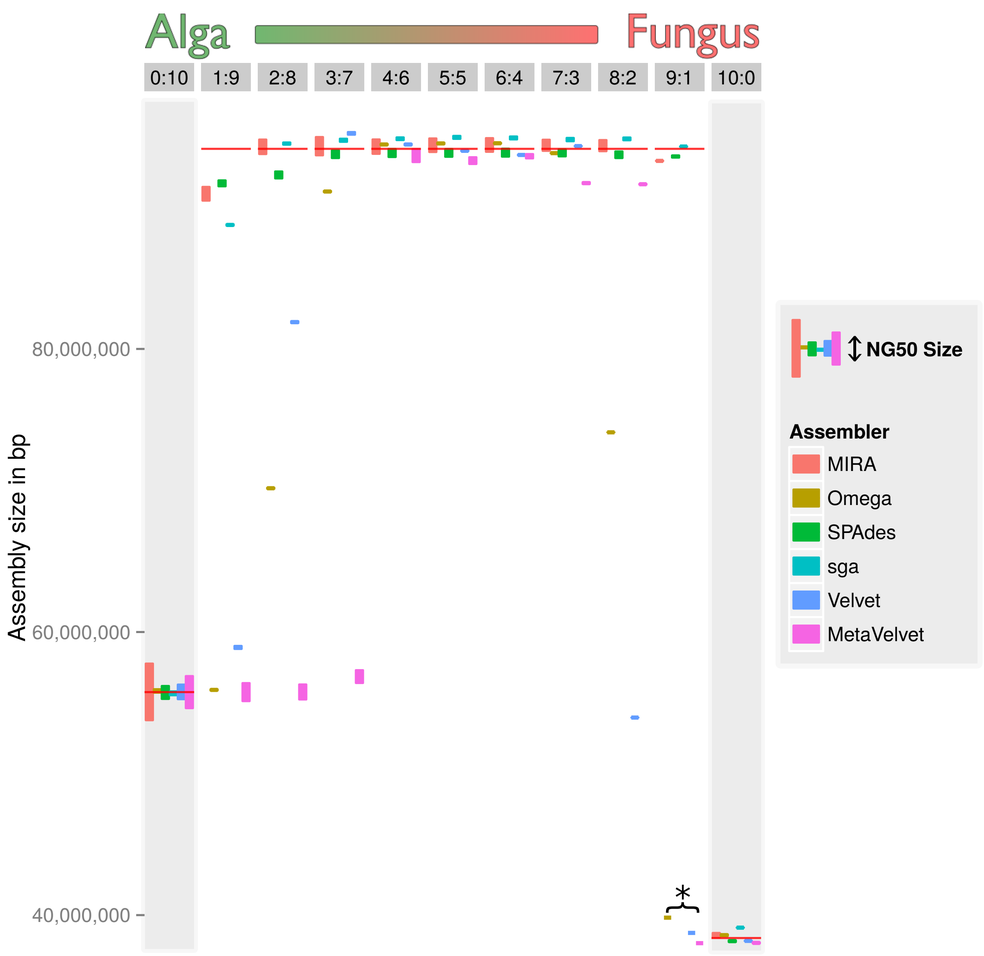
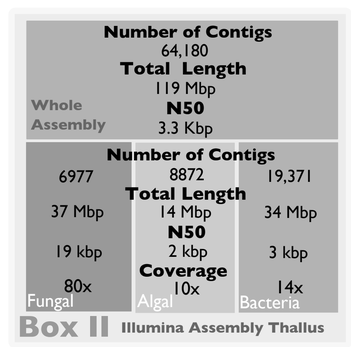


Figure 2: Assembly results for the 11 twin sets and the different assemblers. Bars are centered at total assembly length, red lines are reference lengths. Height of bars shows the NG50 size. For the assemblies with the asterisk the total assembly length was below 50% of the reference length. A default height was used in those instances.

3. Assembly Strategy for *L. pustulata*



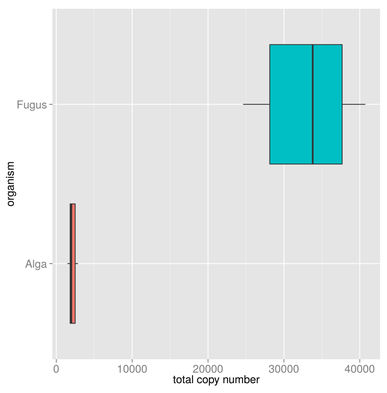


Figure 3: qPCR results for a fungal and an algal single copy. The fungal:algal ratio is around 15:1.

Illumina Sequencing of *L. pustulata* The WGS data was assembled with *MIRA* and contigs were taxonomically assigned using *MEGAN* [13]. The algal assembly is much more fragmented than expected given the in silico study (Box II). It appears that this is a result of a biased library preparation, yielding a highly uneven read coverage for the algal genome, or the large difference in DNA in the lichen thallus (Figure 3).

PacBio Sequencing Additional long-read sequencing was performed to ameliorate the low coverage for the algal genome and to solve repeats in the fungal genome. A total of 18 SMRT cells were sequenced, yielding 2,705,256 polymerase reads with a N50 of 15kb.

PacBio Assemblies For high coverage data sets PacBio-only assemblies are state of the art. For low-coverage data hybrid assemblies using Illumina and PacBio data are needed [14]. Because of this we performed two different assemblies, targeting the fungal and the algal genome respectively (Figure 4).

Gene Predictions For both assemblies scaffolds were taxonomically assigned using *MEGAN* and genes predicted using *Maker*.

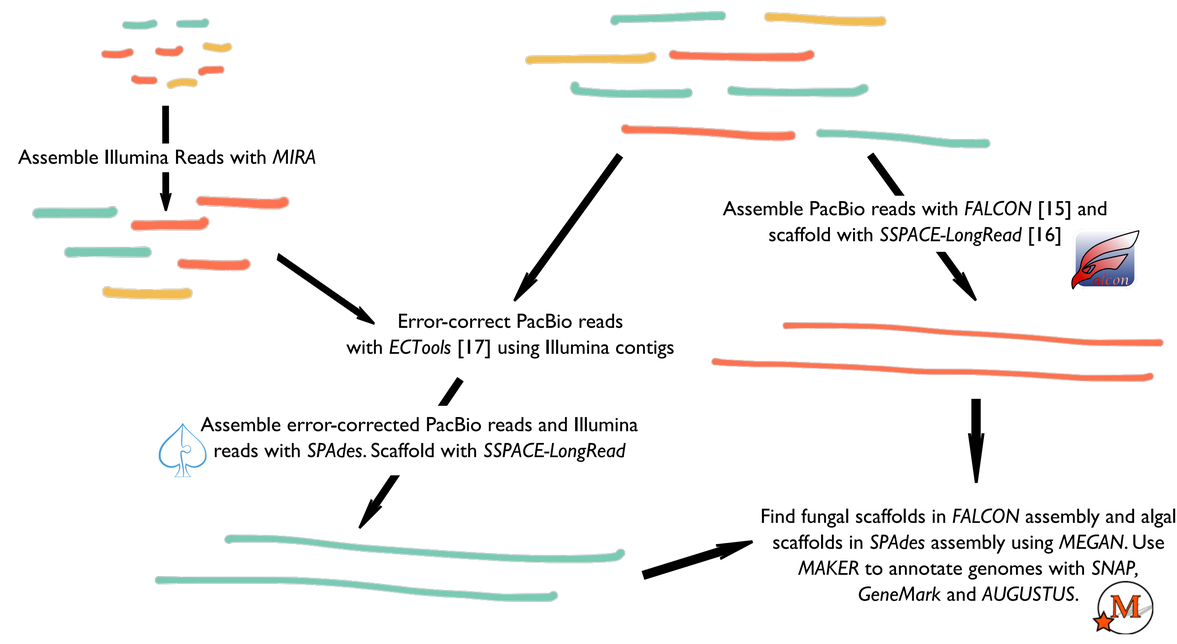


Figure 4: Outlining the assembly workflow for the genomes of the mycobiont and the photobiont. The mycobiont can be assembled directly from the PacBio data. The photobiont is assembled using a hybrid approach using PacBio and Illumina data.

3. Analysing *Lasallia pustulata*

Assemblies Both genomes could be assembled to a high level of completeness and contiguity with our two-pronged approach. Both total size and contiguity are in line with expectations given by related organisms, as is the number of genes (Box III).

Pezizomycotina Gene Set The Last Common Ancestor (LCA) gene set of the Pezizomycotina was reconstructed using OMA [19]. Besides *L. pustulata*, 7 further species - representing 3 clades - were used (Figure 5). Genes found in at least 7 species were assumed to be ancestral for the Pezizomycotina. In total 12,595 orthologous groups were formed (Figure 6).

Absence of LCA Genes All 8 species were found for 1153 orthologous groups. For 1357 further groups genes were only found in 7 species. For these groups it is *L. pustulata* which is absent most often, hinting that these genes are lost over time in lichenization (Figure 7).

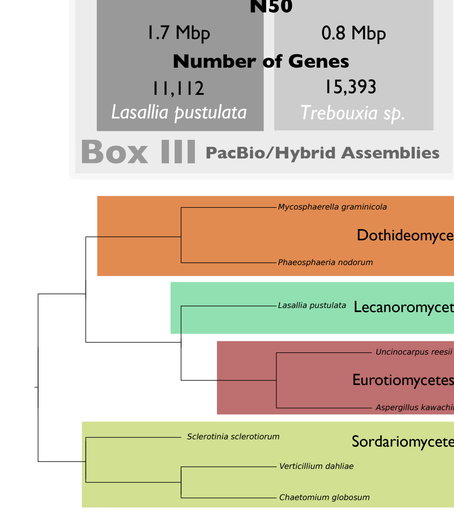


Figure 5: Tree of 8 species used for creating the LCA gene set.

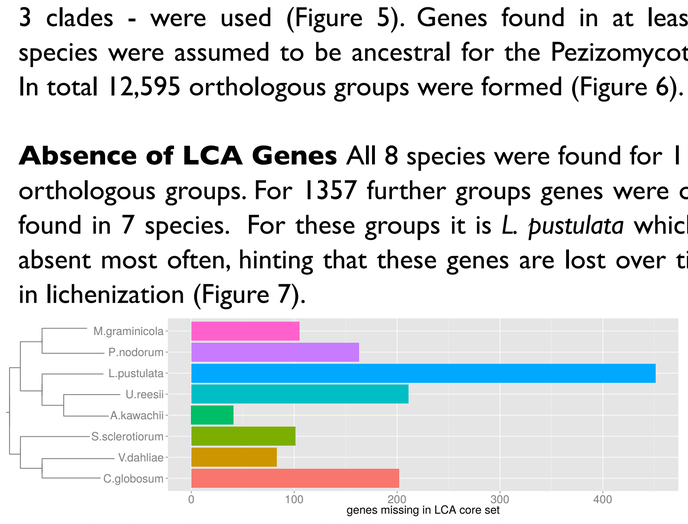


Figure 7: Distribution of genes missing in the LCA set. Lasallia pustulata is missing in twice as many orthologous groups as any other species.

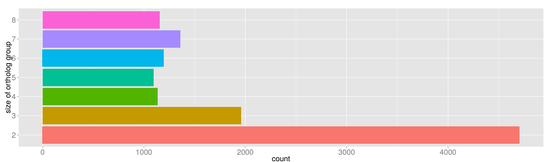


Figure 6: Results of the orthology prediction with OMA. For 1153 groups all 8 species were found. For 1357 groups only 7 species were found.

doi:10.6084/m9.figshare.1466892

Contact

Bastian Greshake
bgreshake@gmail.com
Goethe University, Frankfurt am Main, Germany
Max-von-Laue-Straße 13, 60438 Frankfurt am Main

References

[1] Magoc T and Salzberg S. Bioinformatics (2011) 27 (21):2957-63
[2] http://genome.jgi.doe.gov/Clagr2/
[3] http://genome.jgi.doe.gov/Astpho2/
[4] Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0 2013-2015
[5] Krumsiek J, Arnold R, Rattei T. Bioinformatics (2007) 23 (8): 1026-1028
[6] Huang W, Li L, Myers JR, Marth GT. Bioinformatics (2012) 28 (4):593-594
[7] Zerbino DR and Birney E. Genome Research (2008) 18:821-829.
[8] Namiki T, Hachiya T, Tanaka H, Sakakibara Y. Nucleic Acids Res, (2012) 40(20), e155
[9] Bankevich A, Nurk S, Antipov D et al. Journal of Computational Biology (2012) 19(5):455-477
[10] http://sourceforge.net/projects/mira-assembler/
[11] Haider B, Ahn T, Bushnell B et al. Bioinformatics (2014) btu395
[12] Simpson JT and Durbin R. Bioinformatics (2010) 26 (12): i367-i373
[13] Huson DH, Mitra S, Ruscheweyh H et al. Genome Research (2011) 21: 1552-1560
[14] Mike Schatz, PAG 2014 (http://schatzlab.cshl.edu/presentations/2014-01-14.PAG.Single%20Molecule%20Assembly.pdf)
[15] https://github.com/PacificBiosciences/FALCON
[16] Boetzer M and Pirovano W. BMC Bioinformatics (2014) 15:211
[17] https://github.com/jgurtowski/ectools
[18] Campbell MS, Holt C, Moore B, Yandell M. Curr Protoc Bioinformatics (2014) 48:4.11.1-4.11.39
[19] http://omabrowser.org/standalone/