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†

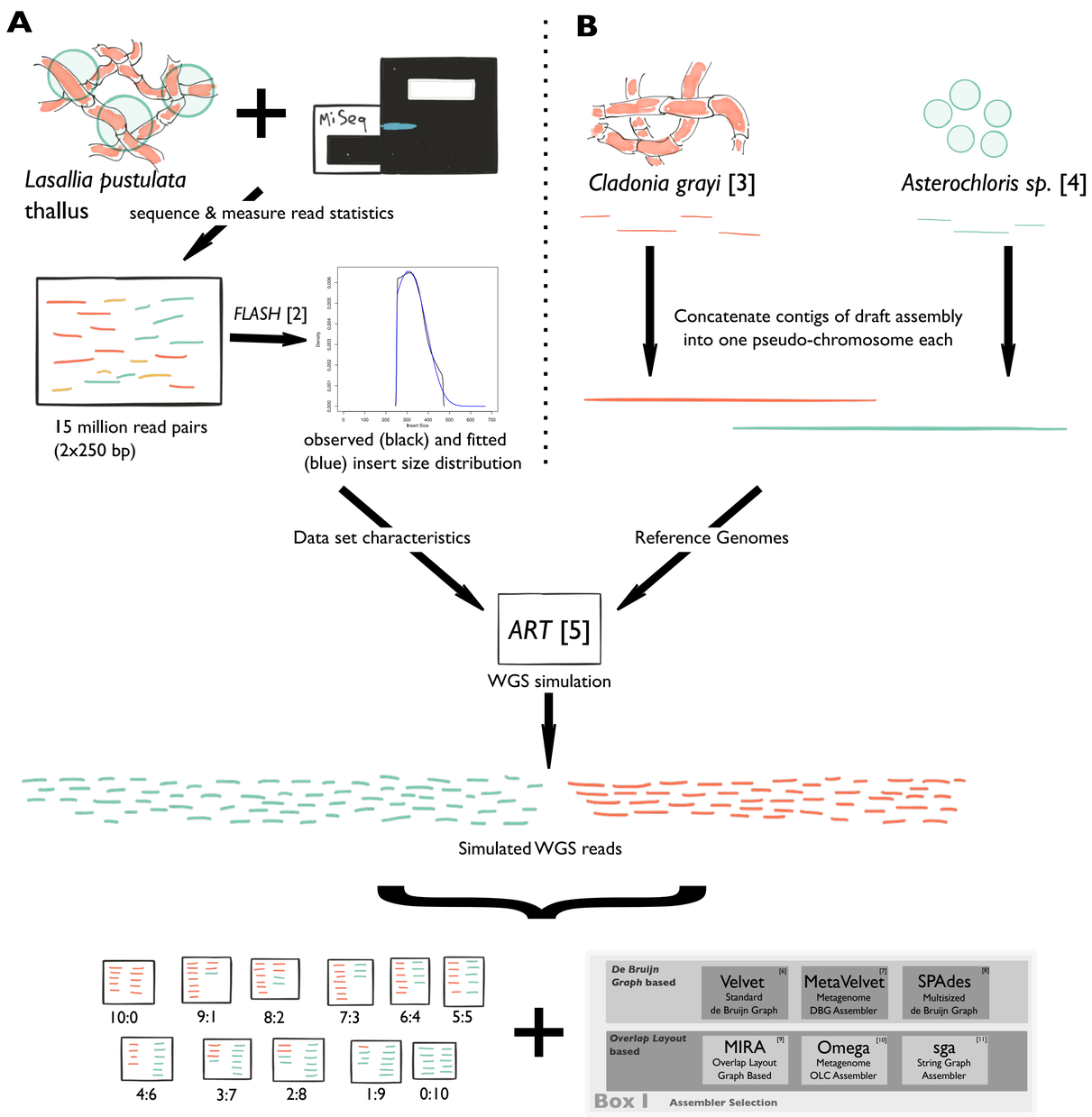
† Department for Applied Bioinformatics, Institute for Cell Biology and Neuroscience, Goethe University, Frankfurt am Main, Germany
* Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany
§ Institute of Ecology, Evolution and Diversity, Goethe 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. The close interdependence in such communities, however, confounds genomic studies. In many cases separate sequencing of the participating organisms is not feasible, leaving metagenomics approaches as the method of choice. Here we address how and to what extent eukaryotic genomes can be reconstructed from such data.

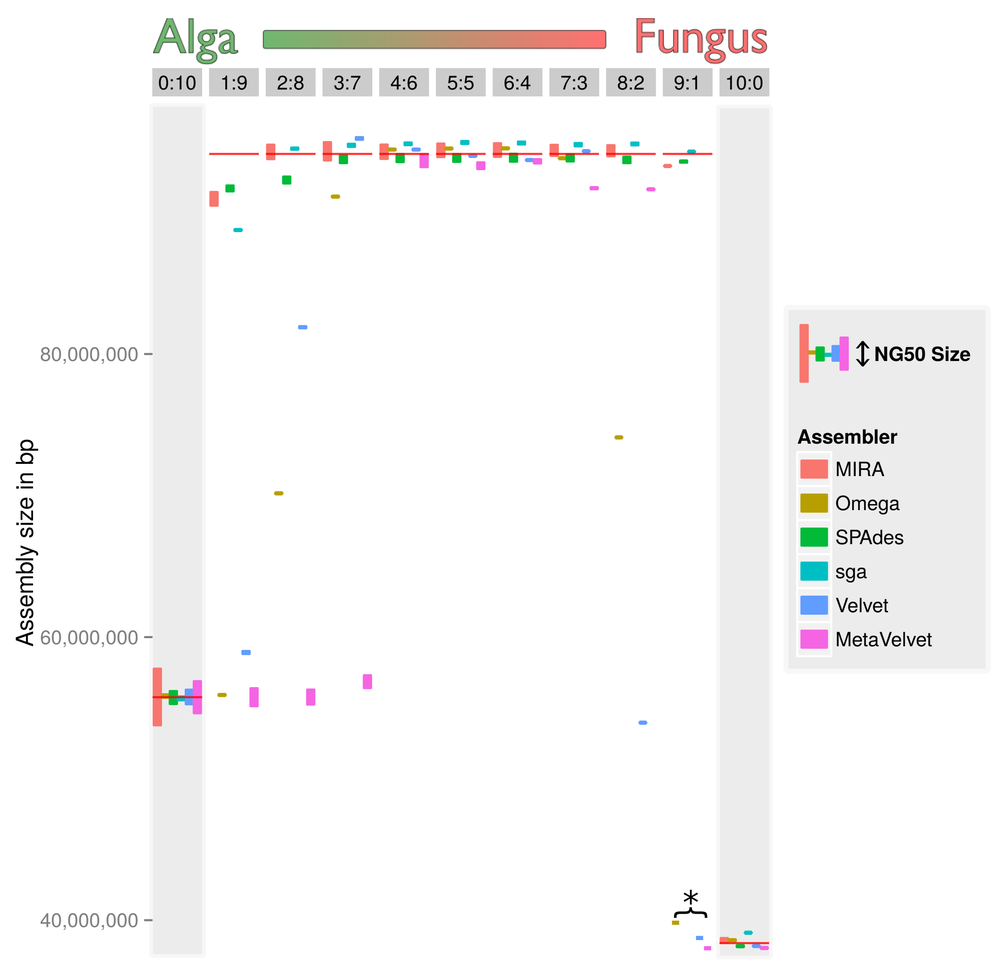
We use in silico-generated data sets to sound out the performance of different assembly paradigms on Whole Genome Shotgun (WGS) data from eukaryotic species mixtures. On this basis we have begun reconstructing the metagenome of the lichen *Lasallia pustulata*. Using a hybrid sequencing approach, that combines Illumina short read and PacBio long read data, we have assembled the genome of the mycobiont and a major fraction of the algal photobiont. We integrate this data with genome sequences of closely related non-lichenized fungi as a first step towards analyzing how lichenization affects genome evolution.

1. Assembler Evaluation with Simulated Twin Sets [1]



Merge reads simulated from either reference genomes to form *L. pustulata* twin sets with varying coverage ratios for the two genomes. Assemble each set with 6 different assemblers.

2. Assembly Results of the Twin Sets



Assembly results for the 11 twin sets. Bars are centered at total assembly length, red lines indicate reference lengths. Height of bars shows the NG50 size. Assemblies marked with an asterisk cover less than 50% of the reference length. A default height was used in those instances.

3. Sequencing the *L. pustulata* metagenome

Pilot Study

- Box II summarizes the assembly results of the metagenome skimming data (1A) with *MIRA*.
- A comparison to the twin set (2) analysis hints at unexpected issues with the reconstruction of the algal genome.
- A qPCR analysis of the lichen thallus reveals a highly biased fungal-to-algal genome ratio of 15:1.

Box II Illumina Assembly Thallus			
Whole Assembly	Number of Contigs: 64,180	Total Length: 119 Mbp	N50: 3.3 Kbp
	Number of Contigs	Total Length	N50
Fungal	6977	37 Mbp	19 kbp
Algal	8872	14 Mbp	2 kbp
Bacteria	19,371	34 Mbp	3 kbp
		Coverage	
Fungal		80x	
Algal		10x	
Bacteria		14x	

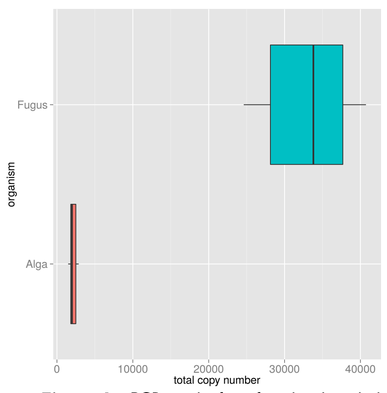
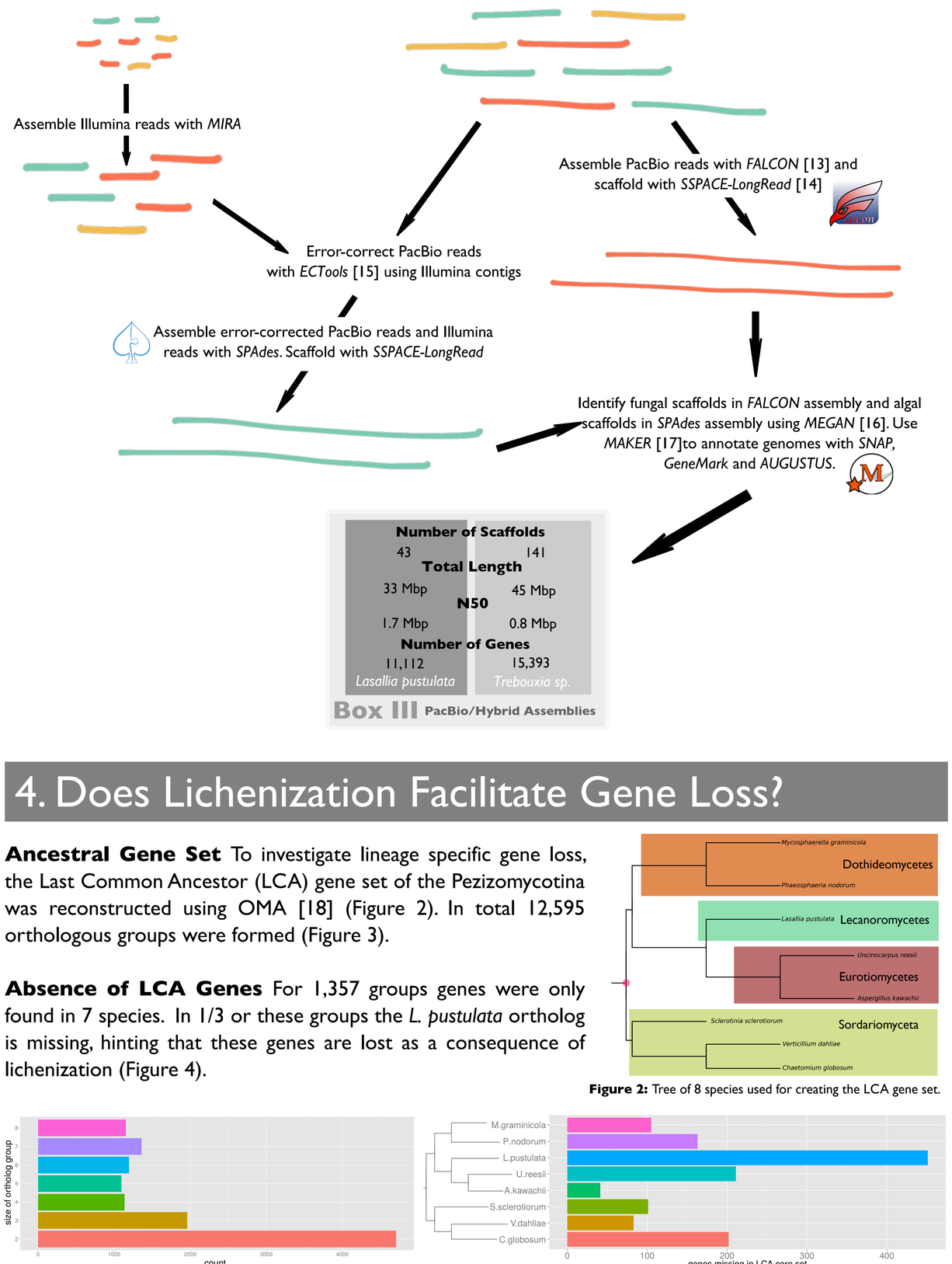


Figure 1: qPCR results for a fungal and an algal single copy. The fungal:algal ratio is around 15:1.

Hybrid Assembly: Short-Read meets Long-Read Using PacBio 2,705,256 polymerase reads with a read N50 of 15kb were sequenced. A 250 bp mate pair library (5kb inserts) of 15 million reads was sequenced using Illumina MiSeq.

To cope with the coverage differences we pursued two different assembly strategies, targeting the fungal and the algal genome respectively. For high coverage data PacBio-only assemblies are state of the art, low-coverage data require hybrid assemblies using Illumina and PacBio data [12].



4. Does Lichenization Facilitate Gene Loss?

Ancestral Gene Set To investigate lineage specific gene loss, the Last Common Ancestor (LCA) gene set of the Pezizomycotina was reconstructed using OMA [18] (Figure 2). In total 12,595 orthologous groups were formed (Figure 3).

Absence of LCA Genes For 1,357 groups genes were only found in 7 species. In 1/3 or these groups the *L. pustulata* ortholog is missing, hinting that these genes are lost as a consequence of lichenization (Figure 4).

Figure 2: Tree of 8 species used for creating the LCA gene set.

Figure 3: Results of the orthology prediction with OMA. For 1153 groups all 8 species were found. For 1357 groups only 7 species were found.

Figure 4: Distribution of genes missing in the LCA set. Lasallia pustulata is missing in twice as many orthologous groups as any other species.

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

doi:10.6084/m9.figshare.1466892

References

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