Sequencing and Characterizing the Genome of the Lichen *Lasallia pustulata*

Bastian Greshake†, Francesco Dal Grande*, 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

Motivation

Lichens are composite organisms comprising a fungal mycobiont and one or several species of green algae or cyanobacteria as photobionts. Fossil evidences for lichens date back to the Early Devonian approximately 400 MYA. As an effect of this long standing interaction both mycobiont and photobiont often grow poorly without their partner. In some cases, such as for *Lasallia pustulata*, a solitary cultivation of the mycobiont has been impossible so far. The molecular basis for this reciprocal dependence remains yet to be determined, and understanding the evolutionary implications of lichenization for the interacting partners in general is still in its infancies. This circumstance is partly due to the scarcity of both genome sequences and transcriptome data for lichens. Here we present an initial analysis of the *Lasallia pustulata* genome and transcriptome. *De novo* assembly quality is highly dependent on the input data, the chosen algorithm and the parameter settings. Using a simulation approach, we first explored the performance of different assembly strategies on simple meta-genomes of varying coverage ratios. The best performing strategy was then taken to reconstruct the genome sequences of the lichen *Lasallia pustulata* from a set of 30 million MiSeq reads. The resulting data for the mycobiont was then used for initial gene prediction, phylogenetic placement and functional annotation.

1. Assembly Strategy

Shotgun sequencing of the lichen *L. pustulata* obtained 15 million MiSeq read pairs of 250 bp in length, with a mean insert size of 336 bp (Figure 1). We simulated twin sets resembling the *L. pustulata* data in insert size distribution, read count and length using *ART* [1] and the draft genomes of the lichenized fungus *Cladonia grayi* and its photobiont *Asterochloris sp.* Coverage ratios for the alga and the fungus varied between 10:0 and 0:10 in the 11 twin sets. We then assessed the performance of four different assemblers on this mixed species species data (Figure 2).

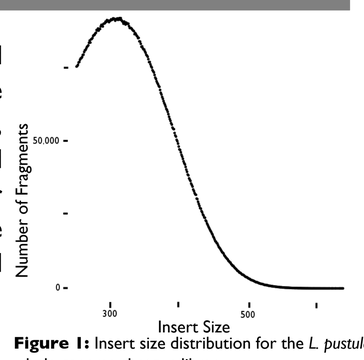


Figure 1: Insert size distribution for the *L. pustulata* whole genome shotgun library.

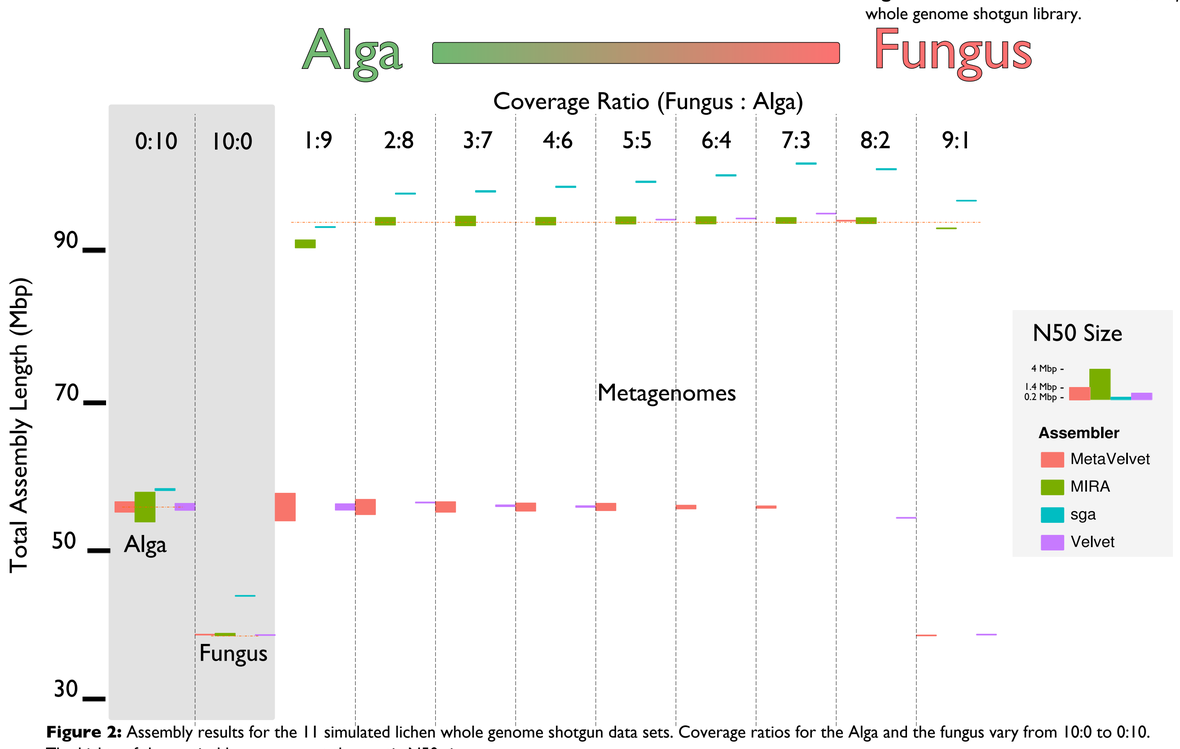


Figure 2: Assembly results for the 11 simulated lichen whole genome shotgun data sets. Coverage ratios for the Alga and the fungus vary from 10:0 to 0:10. The hights of the vertical bars represent the contig N50 sizes.

For data derived from a single species (10:0;0:10) all assemblers perform comparable. Only a single assembler, *MIRA*, is unaffected by the varying coverage ratios and outperforms all other assemblers. Thus the assembly of *L. pustulata* was done with MIRA (Box 1).

Box 1 *Lasallia pustulata* Assembly

Number of Contigs	64,180
Total Assembly Length	119,028,408 bp
Largest Contig	520,743 bp
N50 Size	3,373 bp

2. Taxonomic Assignment & Gene Prediction

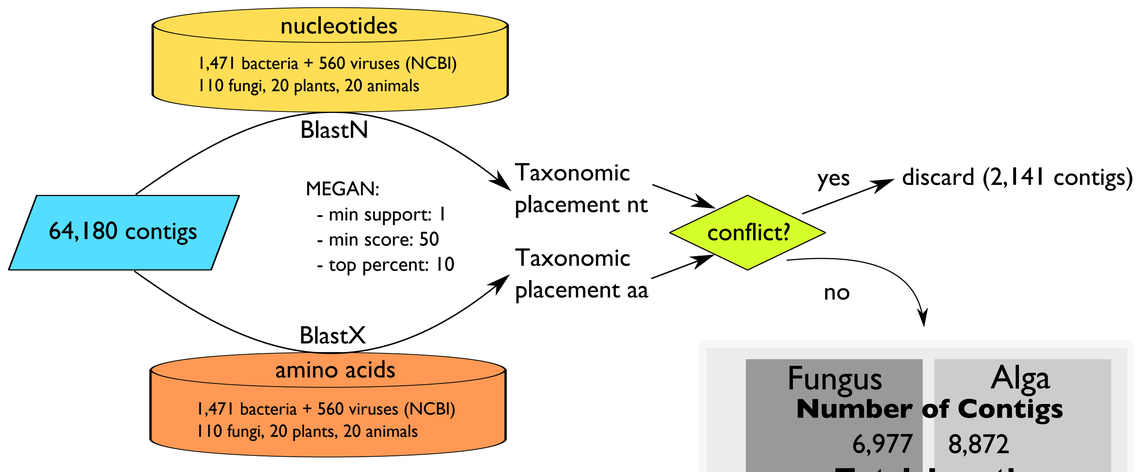


Figure 3: Workflow used for our *MEGAN* analysis.

The taxonomic assignment of the *L. pustulata* meta-genome contigs was done with *MEGAN* [6] (Box II & Figure 3).

Gene prediction was done with *AUGUSTUS* [7] only for the fungal contigs, using a two-step procedure. First, *AUGUSTUS* was trained using *L. pustulata* genes found by *CEGMA* [8] (298 proteins in 231 groups, 93.15% completeness). Second, complementary RNASeq data from *L. pustulata* was mapped to the contigs with Tophat [9] and was used to create intron evidences for a refined gene prediction. *AUGUSTUS* annotated 8,156 genes with an average length of 418 amino acids (Figure 4).

	Fungus	Alga
Number of Contigs	6,977	8,872
Total Length	37,469,368 bp	14,839,567 bp
Largest Contig	161,762 bp	21,823 bp
N50 Size	19,048 bp	2,158 bp

Box II Taxonomic Assignment

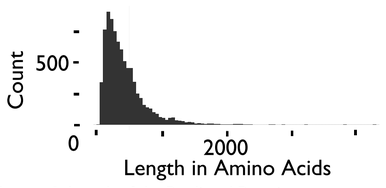


Figure 4: Length of the Predicted Proteins.

3. Tree Reconstruction

We used HaMStR [10] to identify orthologs to 162 genes, which have previously been used to resolve the pezizomycete phylogeny [11]. Orthologous sequences were aligned with *MAFFT* [12] (-linsi) and concatenated into a supermatrix with 115,155 amino acid positions. Removing columns with >50% undetermined amino acids or gaps retained a supermatrix of 45,999 amino acid positions.

Maximum likelihood tree reconstruction with *RAxML* [13], (LG+G+F) obtained the tree in Figure 5.

L. pustulata is placed into monophyletic Lecanoromycetes forming the sister clade of the Dothideomycetes.

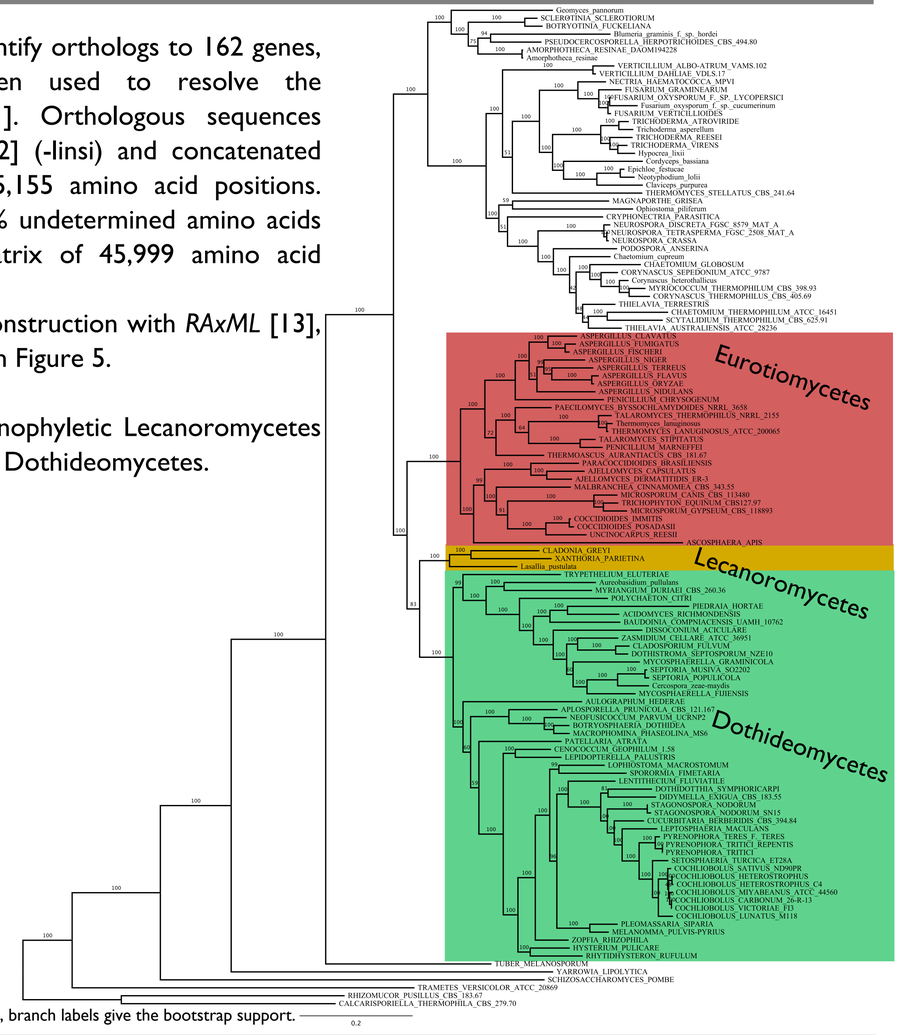


Figure 5: Phylogeny of the Pezizomycotina, branch labels give the bootstrap support.

4. Functional Annotation (Gene Ontology)

We preliminary annotated the 8,156 L. pustulata genes with Gene Ontology terms using Blast2GO [14]. About 7,500 of our genes could be annotated with GO terms (Figure 6) and 5,000 genes were assigned Enzyme Codes (Figure 7).

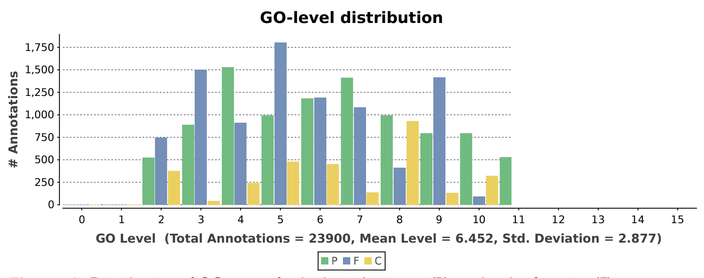


Figure 6: Distribution of GO terms for biological process (P), molecular function (F) and cellular component (C)

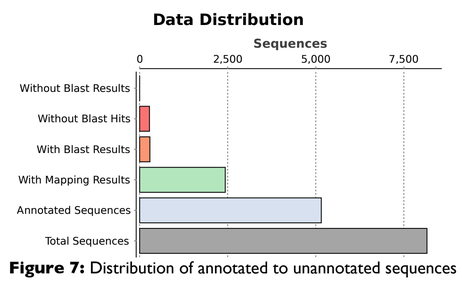
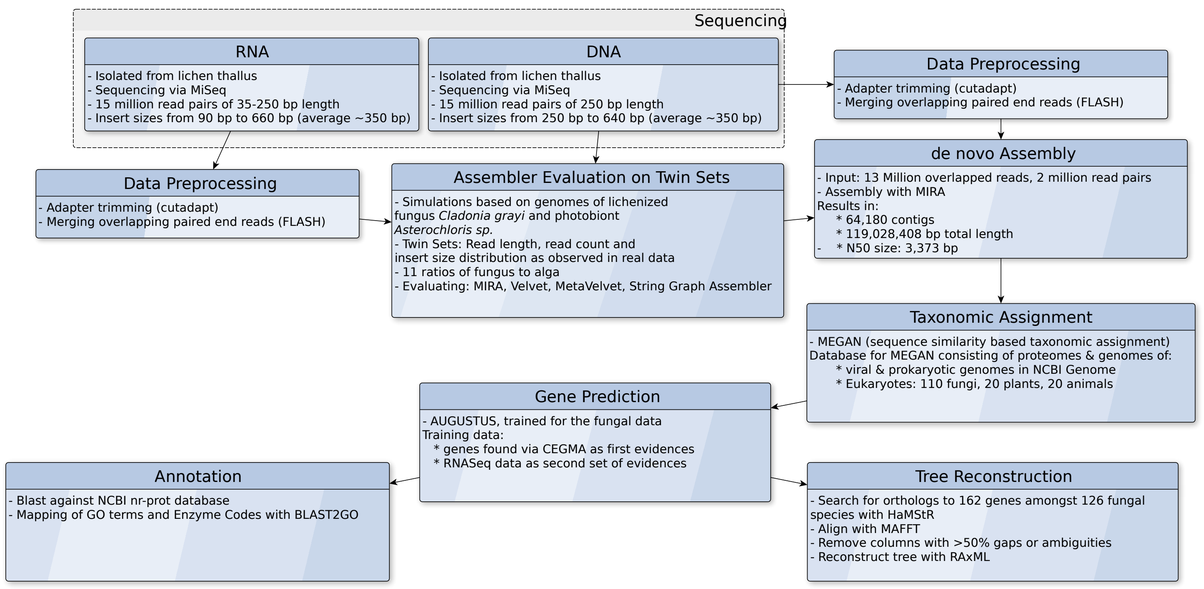


Figure 7: Distribution of annotated to unannotated sequences

Summary



- *MIRA* (Overlap Consensus-based) consistently performs best on assembling simple meta-genomes
- Using *MIRA* & *MEGAN* we were able to recover ~37.5 Mbp of the mycobiont genome of *Lasallia pustulata*
- *AUGUSTUS* trained with additional RNAseq data annotated 8,156 genes
- Phylogenomic tree reconstruction firmly places *L. pustulata* within the Lecanoromycetes and those as sister group to the dothideomycetes
- Preliminary functional annotation assigned GO terms and/or Enzyme Codes to about 7,500 of the predicted genes

Poster URL

doi:10.6084/m9.figshare.1046678

Contact

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

References

[1] Huang W, Li L, Myers JR, and Marth GT. Bioinformatics (2012) 28 (4): 593-594
[2] http://sourceforge.net/projects/mira-assembler/
[3] Zerbino DR and Birney E. Genome Research (2008) 18:821-829.
[4] Simpson JT and Durbin R. Bioinformatics (2010) 26 (12): i367-i373
[5] Namiki T, Hachiya T, Tanaka H, Sakakibara Y. Nucleic Acids Res, (2012) 40(20), e155
[6] Huson DH, Auch AF, Qi J, et al. Genome Research (2007) 17: 000
[7] Stanke M, Steinkamp R, Waack S and Morgenstern B (2004) Nucleic Acids Research, Vol. 32, W309-W312
[8] Genis Parra, Keith Bradnam and Ian Korf (2007) Bioinformatics, 23: 1061-1067
[9] Trapnell C, Pachter L, Salzberg SL. Bioinformatics (2009) 25 (9): 1105-1111.
[10] http://sourceforge.net/projects/hamstr/
[11] Ebersberger I, de Matos Simoes R, Kupczok A, Gube M, Kothe E, Voigt K, and von Haeseler A. Mol Biol Evol (2012) 29 (5): 1319-1334
[12] Katoh, Standley (2013) Molecular Biology and Evolution 30:772-780
[13] Stamatakis A. Bioinformatics (2006) 22 (21): 2688-2690
[14] Conesa A, Götz S, Garcia-Gomez JM, Terol J, Talon M and Robles M. Bioinformatics, (2005) 21: 3674-3676.