Potential and pitfalls of eukaryotic metagenome skimming: A test case for lichens

Bastian Greshake†, 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

Motivation

Metagenomic sequencing with only a single library layout is used to quickly and cheaply assess the taxonomic and functional complexity of large and diverse microbial communities. We investigate to what extent such metagenome skimming approaches are applicable for the in-depth characterizations of genomes found in obligate symbiotic communities of eukaryotes, e.g. lichens. It is still unclear how a eukaryotic species mixture, with larger and more repeat-rich genomes, influences different *de novo* assembly paradigms, such as *de Brujin Graph* based methods or *Overlap Layout* based assemblers and how to optimize assembly parameters as k-mer or overlap sizes.

To address these questions, we performed an *in silico* study, simulating a genome skimming experiment of a lichen. We show that the quality of genome reconstructions from metagenome skimming data depends essentially on assembler choice, but also on the parameter optimisation strategy used. In the worst case optimising standard assembly metrics can lead to the exclusing of complete genomes. Reconciling the expectations from the *in silico* study with the outcome of a real-world metagenome skimming of the lichen *Lasallia pustulata* indicates an even larger biodiversity, causing the underrepresentation of one symbiont in the shotgun library.

1. *in silico* Sequencing

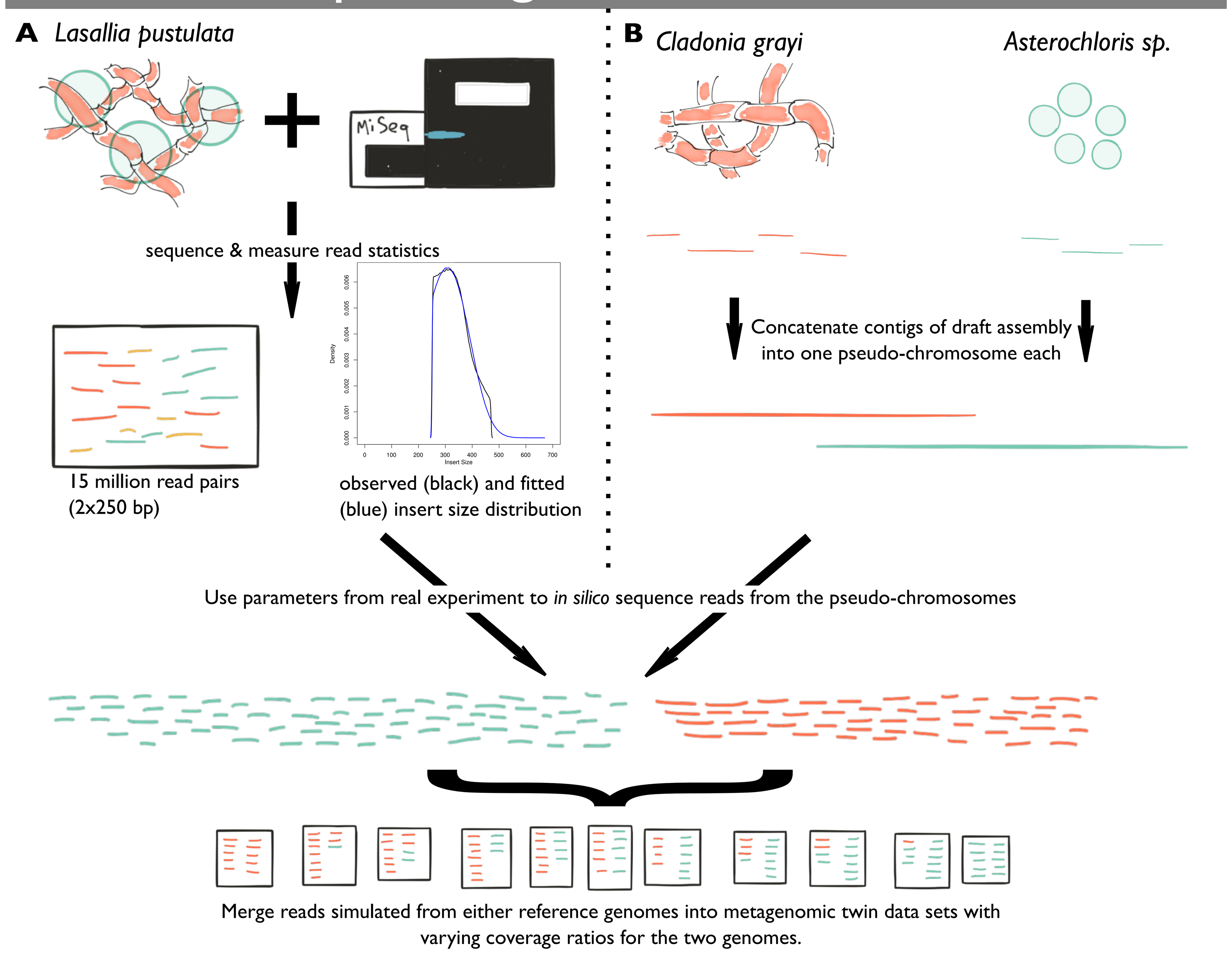

Figure 1: Workflow for generating twin data sets, resembling a real sequencing data set with respect to insert size distribution, read number and read length.

DNA from a thallus of *Lasallia pustulata* was sequenced using Illumina MiSeq technology, yielding 15 million read pairs with a length of 250 bp. To estimate the insert size distribution, we overlapped read pairs using *FLASH* [1] and fitted a censored Weibull distribution to the observed insert size distribution (Figure 1,A).

The scaffolds of the genomes of *Cladonia grayi* [2] and *Asterochloris sp.* [3] were each concatenated to create a contiguous pseudo-chromosome, respectively (Figure 1, B). Both were checked for repeat content & self-similarity using Repeatmasker [4] (Box 1) and Gepard [5] (Fig 2.)

Using the pseudo-chromosomes as templates, we simulated reads using ART [6], parameterized using the values from Fig. 1, A. The reads were mixed into 11 twin data sets by mixing fungal and algal reads in varying ratios (Table 1).

	Cladonia grayi	*Asterochloris sp.*
Number of Scaffolds	1506	153
Total Length	38 Mbp	55 Mbp
GC content	44 %	58 %
% Repetitive	5 %	2.8 %

Box I Reference Genomes

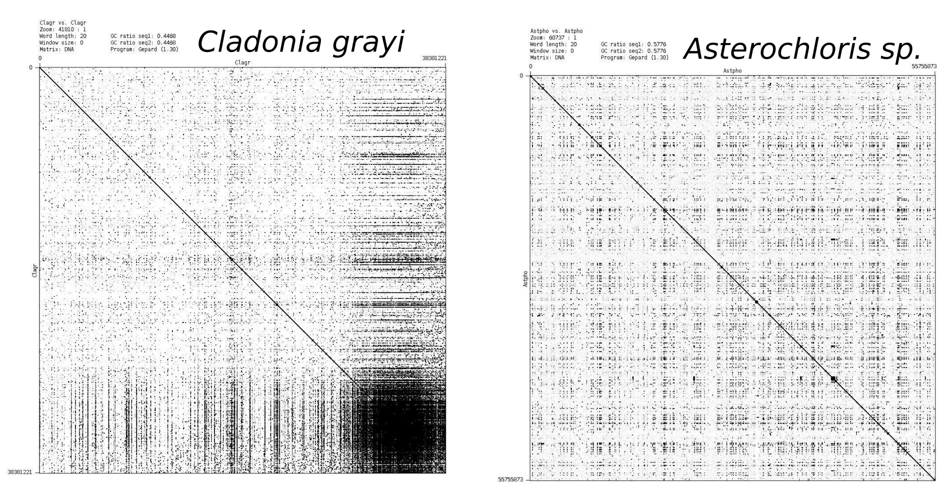

Figure 2: Dotplot of the pseudo-chromosomes of *Cladonia grayi* and *Asterochloris sp.*

Table 1: Absolute coverages for each organism per data set

Coverage Ratio *C. grayi* : *Asterochloris sp.*	Coverage *C. grayi*	Coverage *Asterochloris sp.*
10:0	182x	0x
9:1	157x	17x
8:2	134x	33x
7:3	112x	48x
6:4	92x	61x
5:5	74x	74x
4:6	56x	86x
3:7	40x	97x
2:8	26x	107x
1:9	13x	116x
0:10	0x	125x

2. Assembler Selection & Optimisation

De Brujin Graph based	Velvet [7] Standard de Bruijn Graph	MetaVelvet [8] Metagenome DBG Assembler	SPAdes [9] Multisized de Bruijn Graph
Overlap Layout based	MIRA [10] Overlap Layout Graph Based	Omega [11] Metagenome OLC Assembler	sga [12] String Graph Assembler

Box II Assembler Choice

For Omega, sga, Velvet & MetaVelvet we explored the parameter space (overlap size and k-mer size respectively) and use the maximization of the N50 size as the acceptance objective.

3. Assembly Results

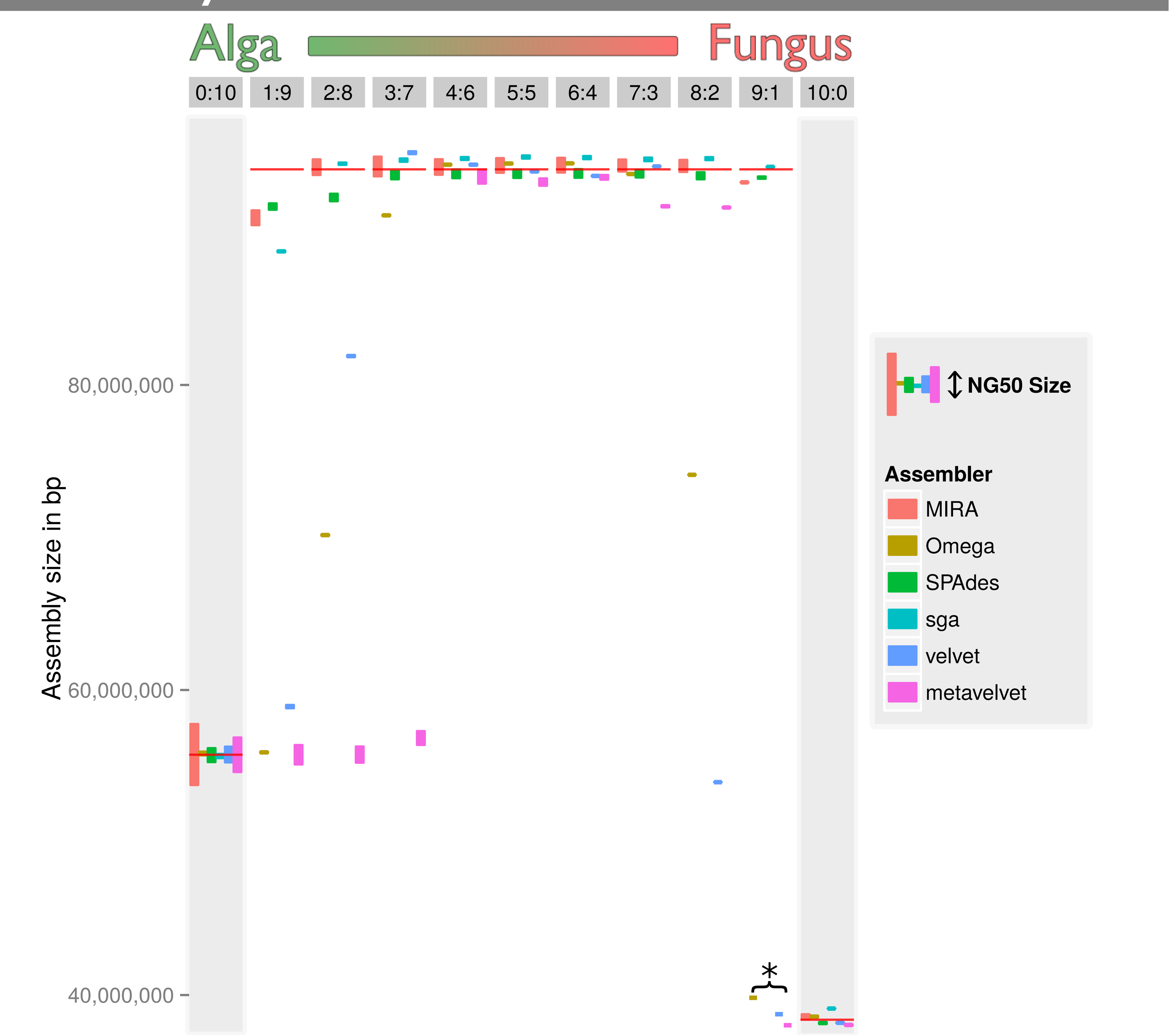

Figure 3: Assembly results for the 11 data sets and the diffferent assemblers. Bars are centered at total assembly length, red lines are reference lengths. Height of the bars shows the NG50 size. For the assemblers with the asterisk the total assembly length was less than 50% of the reference length. A default height was used in those instances.

Single Species Data Sets Almost all assemblers reconstruct the two genomes over their full length (Figure 3, column 0:10 & 10:0), however with varying NG50 sizes. For the alga many assemblers exceeded the NG50 size of the original draft assembly. For the fungus, repeats hindered such an extension with the present WGS library layout (Figure 4).

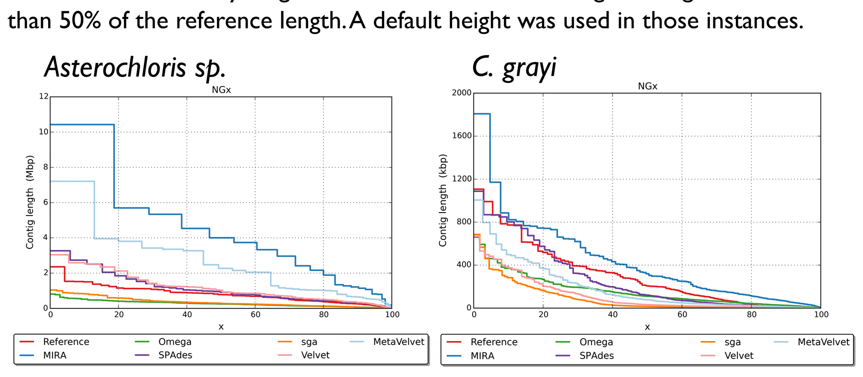

Figure 4: NGx distributions for *Asterochloris sp.* & *C. grayi*

Mixed Species Data Sets Completeness of the genome reconstructions depends heavily on assembler choice and coverage ratios. *MIRA* and *SPAdes* perform best across all data sets. In contrast, *Omega*, *Velvet* and *MetaVelvet* fail to assemble large parts of the low coverage genome once coverage ratios become extreme (Fig. 3, 1:9 - 3:7, 9:1).

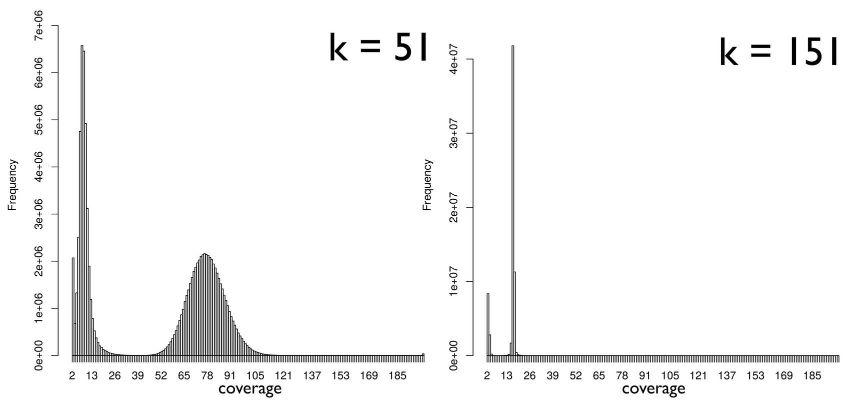

Figure 6: kmer-coverage frequencies for the 1:9 data set.

Sensitivity to Biased Coverage Ratios Increasing the value of k reduces the frequency of all k-mers (Fig. 6), causing k-mers from the low coverage genome to overlap with those introduced by the sequencing error. This prevents the formation of typically short contigs, thus increasing the N50 size by not assembling the low coverage genome.

***L. pustulata* Assembly** Done with *MIRA*, contigs were taxonomically assigned using *MEGAN* [13]. The fungal and algal genome are much more fragmented than expected given the *in silico* study. This is a result of the larger microbial diversity: nearly 1/3 of the assembly is of bacterial origin, which decreases the absolute coverage, especially for the alga (Box III, right column).

Whole Assembly

Number of Contigs	64,180
Total Length	119 Mbp
N50	3.3 Kbp

	Fungus		Alga
Number of Contigs	6977	8872	19,371
Total Length	37 Mbp	14 Mbp	34 Mbp
N50	19 kbp	2 kbp	3 kbp
Coverage	80x	10x	14x

Box III Assembly Thallus

Summary

- Twin sets are valuable for guiding strategic decisions during planning of metagenome sequencing and assembly.
- Assembler performance already varies substantially for single species data.
- Mixing data from different species inflates the assembler performance differences, with *MIRA & SPAdes* yielding the most contiguous sequences
- Optimising the N50 can lead to the preclusion of sequences representing the low-coverage genome.
- The assembly of *L. pustulata* shows that lichen thalli are even more diverse in their composition, driving down absolute coverages for all organisms.

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] 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] Stanke M, Steinkamp R, Waack S and Morgenstern B (2004) Nucleic Acids Research, Vol. 32, W309-W312