BiK F Biodiversity and Climate Research Centre

Genome Skimming of Symbiotic Communities – an *in silico* Evaluation of Assembly Approaches

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

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

Lichens, an association of a filamentous fungus and one to several algal or cyanobacterial photobionts, are a hallmark for the success of mutualistic symbioses involving eukaryotes. They can colonize extreme ecological niches, frequently act as pioneering organisms, and are a promising resource for novel bioactive substances of medical and economical relevance. Still, the full potential of lichens for evolutionary and biotechnological studies has not been tapped, mainly since comprehensive genomic data are lacking. Extending the collection of lichen genomes is not trivial as a separate sequencing of the closely interacting symbionts is often not possible. Genome skimming of the lichen metagenome is an obvious and cost-effective solution to rapidly broaden the data basis for genomic research on lichens. Here we address the questions how to best assemble genome skimming data from a eukaryotic species mixture, what pitfalls can occur, and at what quality one can expect to reconstruct the individual genome sequences from a given experiment.

1. *in silico* Generation of Simulated Twin Sets

Our *in silico* analysis mirrors a real world metagenome skimming of DNA extracted from a thallus of the lichen *Lasallia pustulata*. We generated 15 million Whole Genome Shotgun read pairs of length 250 bp using the Illumina MiSeq technology. To estimate the insert size distribution we merged overlapping read pairs using *FLASH* [1] and fitted the parameters of a censored Weibull distribution to the observed insert sizes (Figure 1).

The draft genomes of the lichenized fungus *Cladonia grayi* [2] and its photobiont *Asterochloris sp.* [3] served as basis for the simulations. For each organism we created a pseudogenome, consisting of a single chromosome, by concatenating all scaffolds and removing ambiguous positions. A *RepeatMasker* [4] analysis (Box I), together with a dot plot analysis to display self similarities (Figure 2), revealed a markedly higher genomic repeat content for *Cladonia grayi* compared to *Asterochloris sp*.

We simulated Whole Genome Shotgun reads with *ART* [5], using the two pseudogenomes as templates. Read number, length and insert size distribution resembled that of the *Lasallia pustulata* genome skimming. From the simulated reads we compiled 11 twin sets (Table 1). To investigate the influence of different extents of data mixture on the genome reconstructions, we varied the coverage ratios for the fungus and the alga, from 0:10 to 10:0 in steps of one.

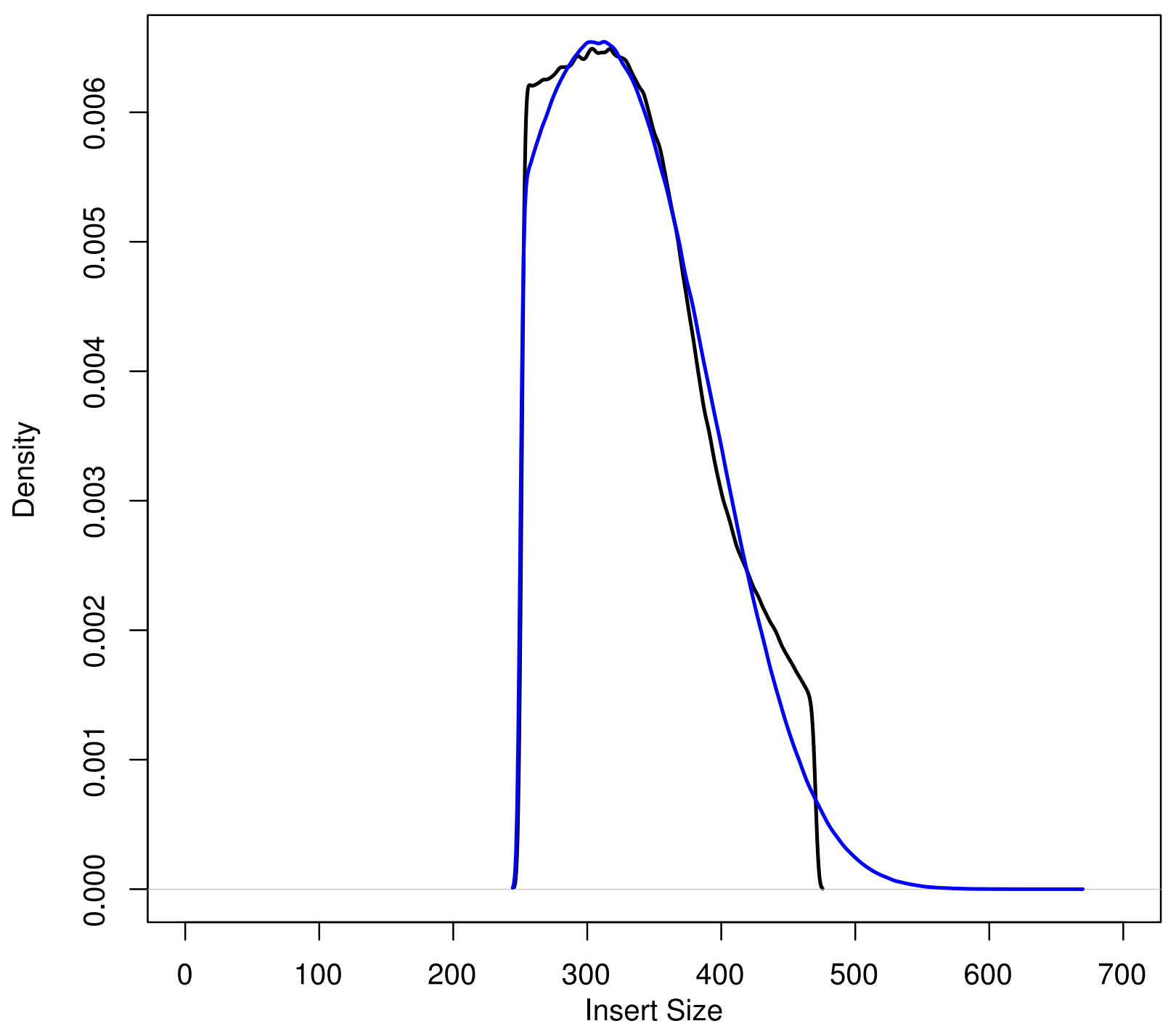


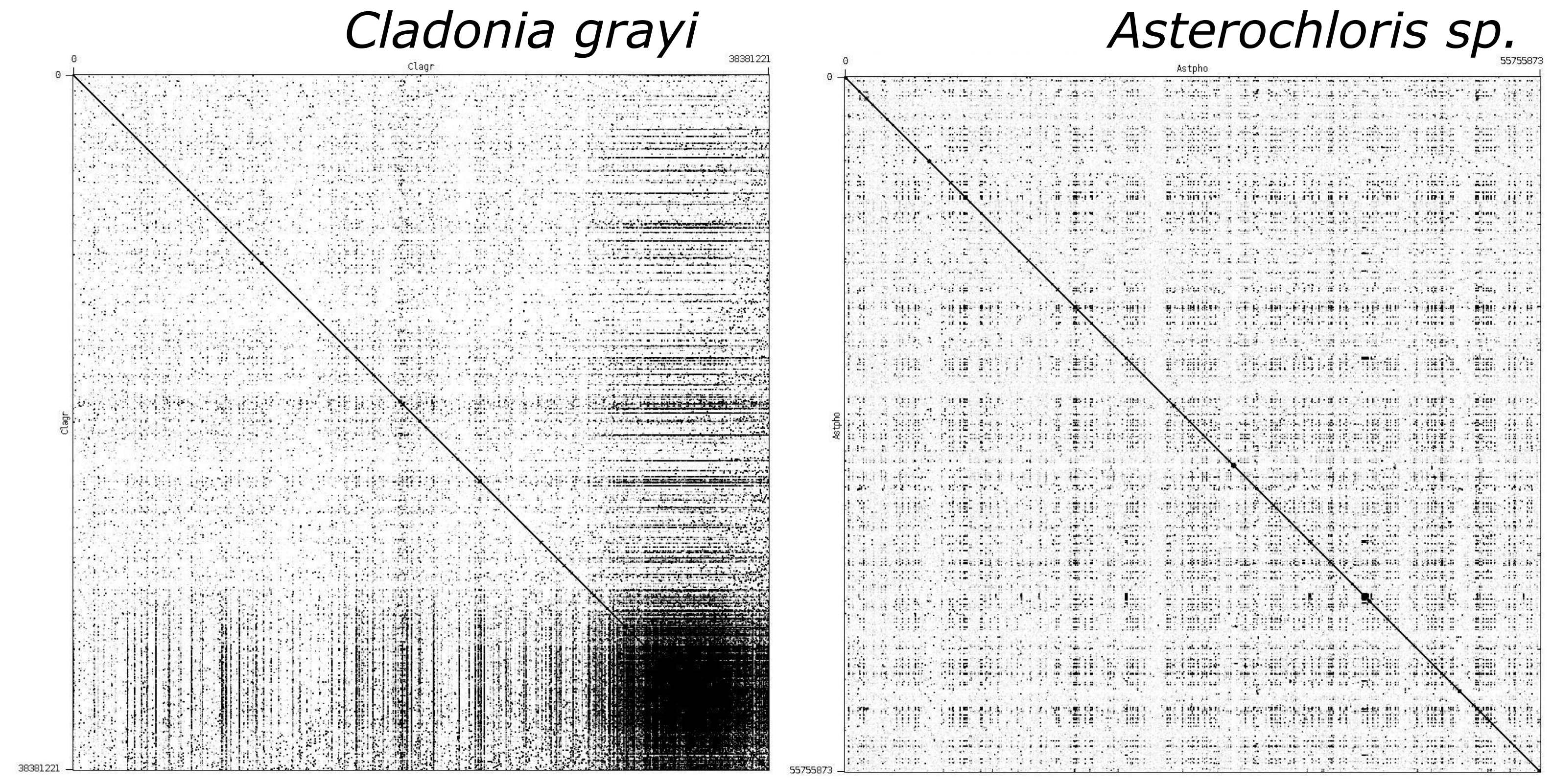
Figure 1: Insert Size distribution of the *L. pustulata* whole genome shotgun library (black: observed, blue: fitted Weibull distribution)

	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**

Table 1: Relative and 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


Cladonia grayi

Asterochloris sp.

Figure 2: Self-similarities of the pseudo-genomes of *C. grayi* and *Asterochloris sp.* visualised by a dot plot analysis.

2. Assembler Selection & Optimisation

De Bruijn Graph based

- **Velvet** [6] Standard de Bruijn Graph
- **MetaVelvet** [7] Metagenome DBG Assembler
- **SPAdes** [8] Multisized de Bruijn Graph

Overlap Layout Consensus based

- **MIRA** [9] Overlap Layout Graph Based
- **Omega** [10] Metagenome OLC Assembler
- **sga** [11] String Graph Assembler

Box II **Assembler Choice**

For *Omega*, *sga*, *Velvet* & *MetaVelvet* we explored the parameter space (overlap size and k-mer size respectively), using the maximisation of the N50 size as the acceptance criterion.

3. Assembly Results

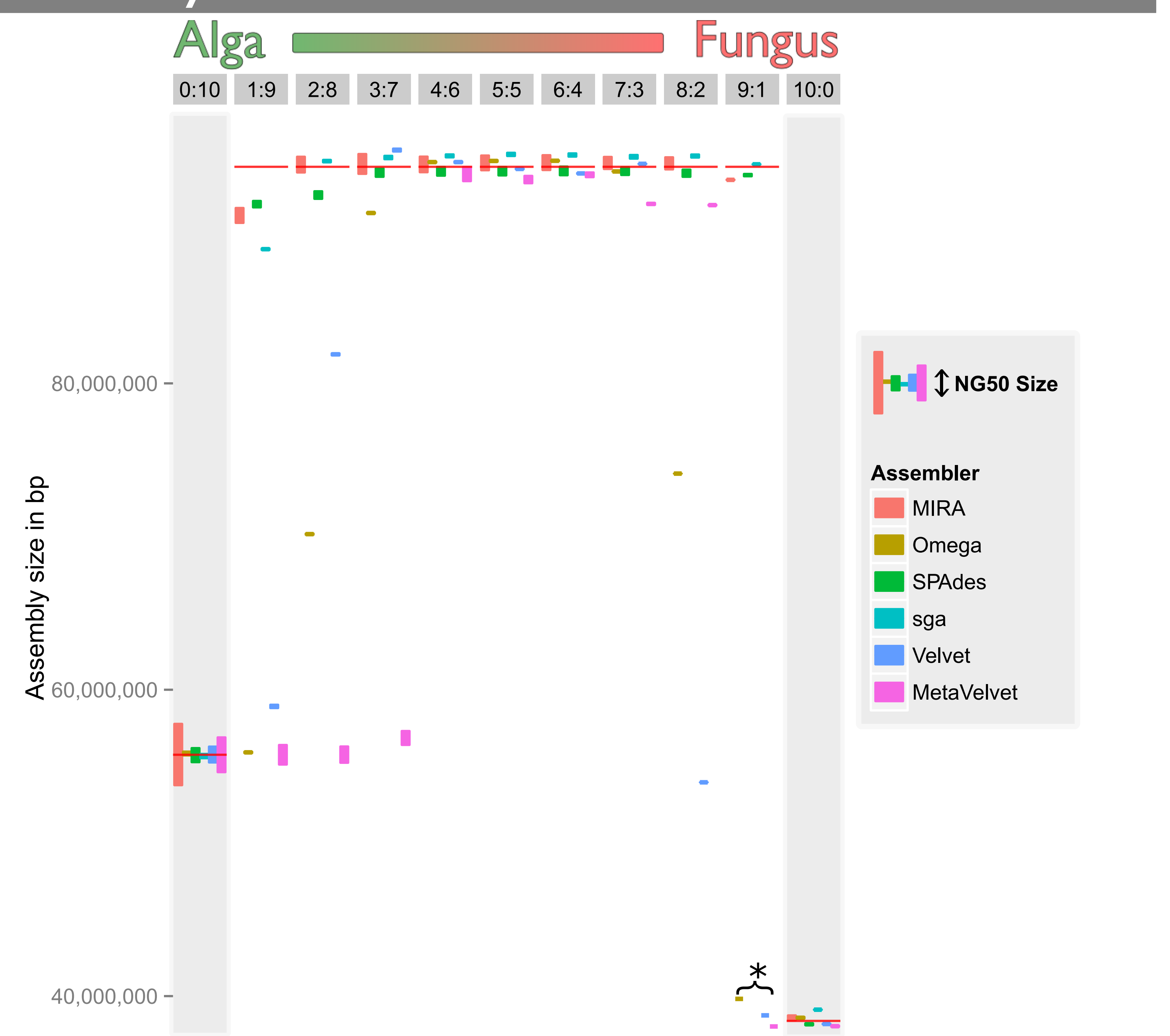
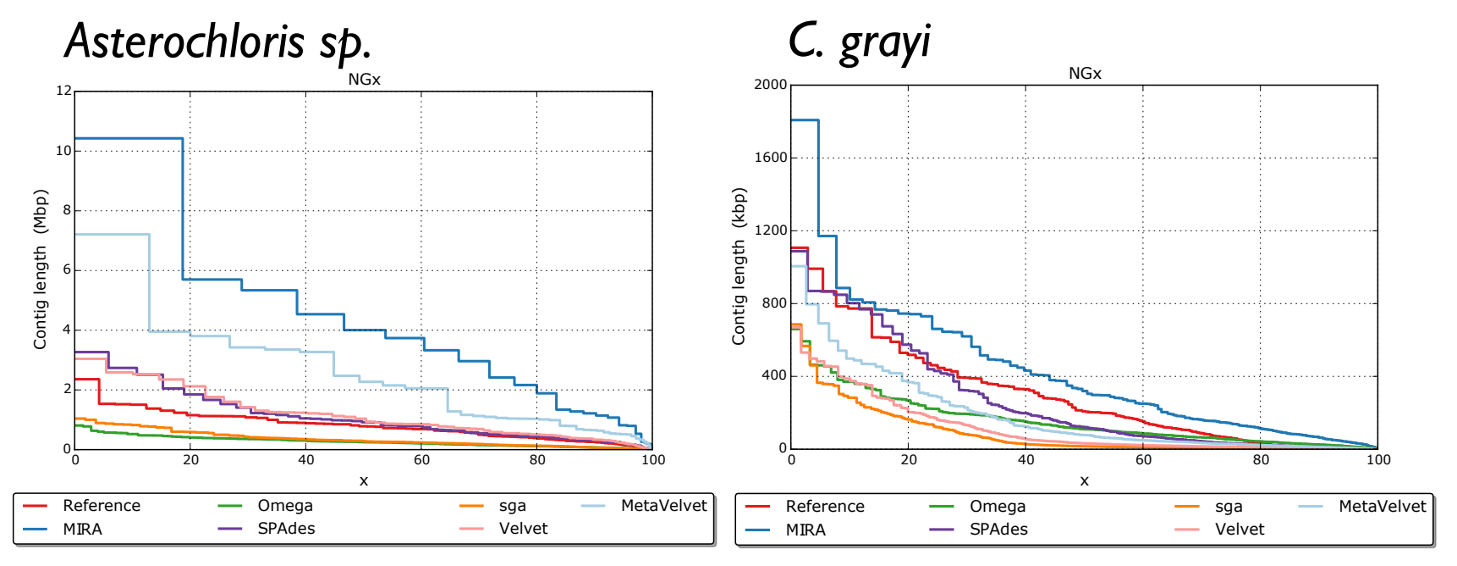


Figure 3: Assembly results over the 11 data sets. Bars are centered at total assembly length (red lines mark the reference lengths). Height of the bars represents the NG50 size. The asterisk marks assemblies covering less than 50% of the reference length. A default height was used in these instances.

For the 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. The repeat-poor alga yields on average much larger contigs compared to the repeat-rich fungus. For the alga many assemblers were able to exceed the NG50 size of the original draft genome. In case of the fungus, repeats hindered such an extension with the present WGS library layout (Figure 4).


Asterochloris sp.

C. grayi

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

For the the mixed species data, completeness of the genome reconstructions depends heavily on assembler choice and coverage ratios. *MIRA* and *SPAdes* perform best across all data sets, even when correcting for the higher number of assembly errors (Fig. 5). In contrast, *Omega*, *Velvet* and in particular *MetaVelvet* fail to assemble large parts of the low coverage genome once coverage ratios become extreme (Fig. 3, 1:9 - 3:7, 9:1).

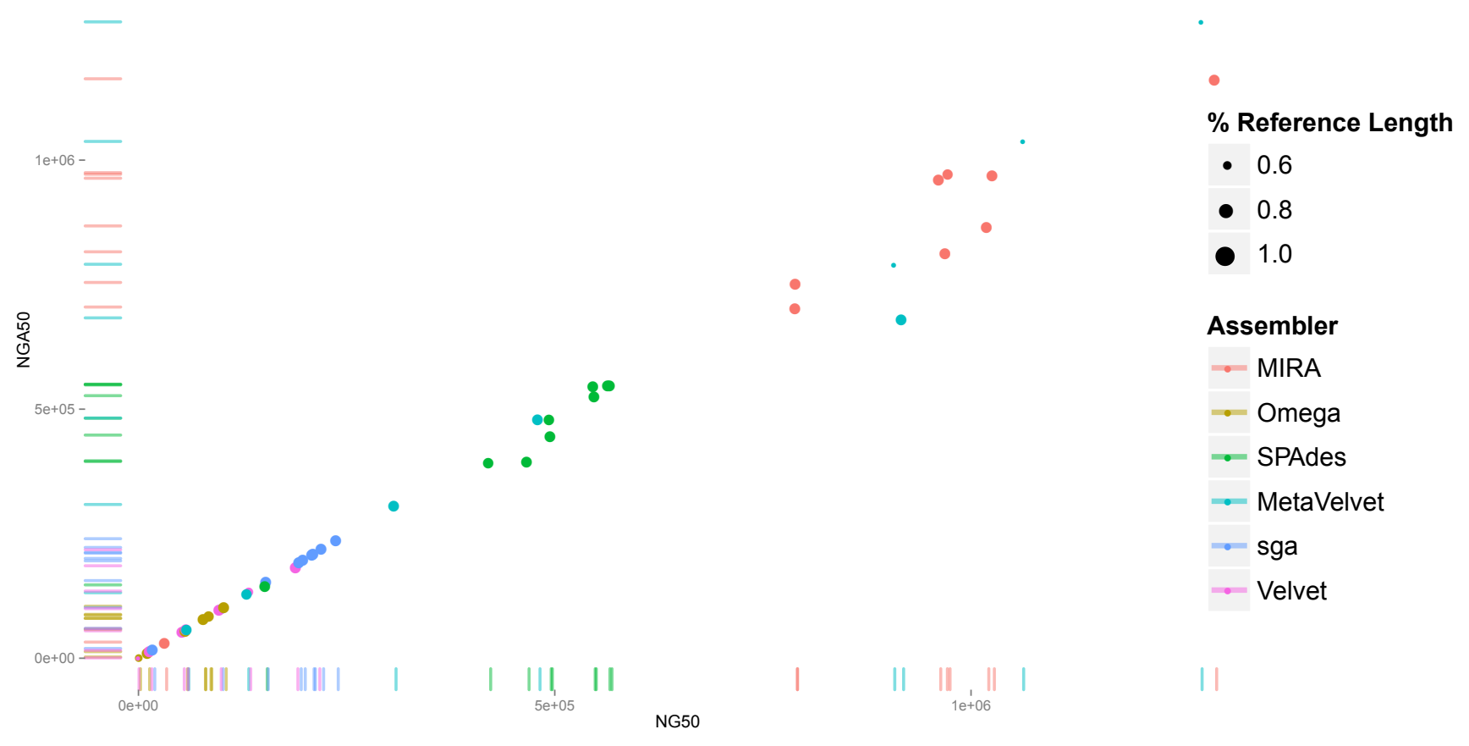


Figure 5: NG50 vs NGA50 (calculated using blocks aligned to the reference genome instead of the contigs) for the assemblies from the simulated twin sets.

The k-mer coverage plots provide an explanation for the sensitivity of some assemblers to biased coverage ratios. Increasing the value of k reduces the frequency of all k-mers (Fig. 6). A high k precludes k-mers from the low coverage genome from the assembly, since their frequency overlaps with that of k-mers introduced by sequencing errors. This prevents the formation of typically short contigs, thus optimizing the N50 size.

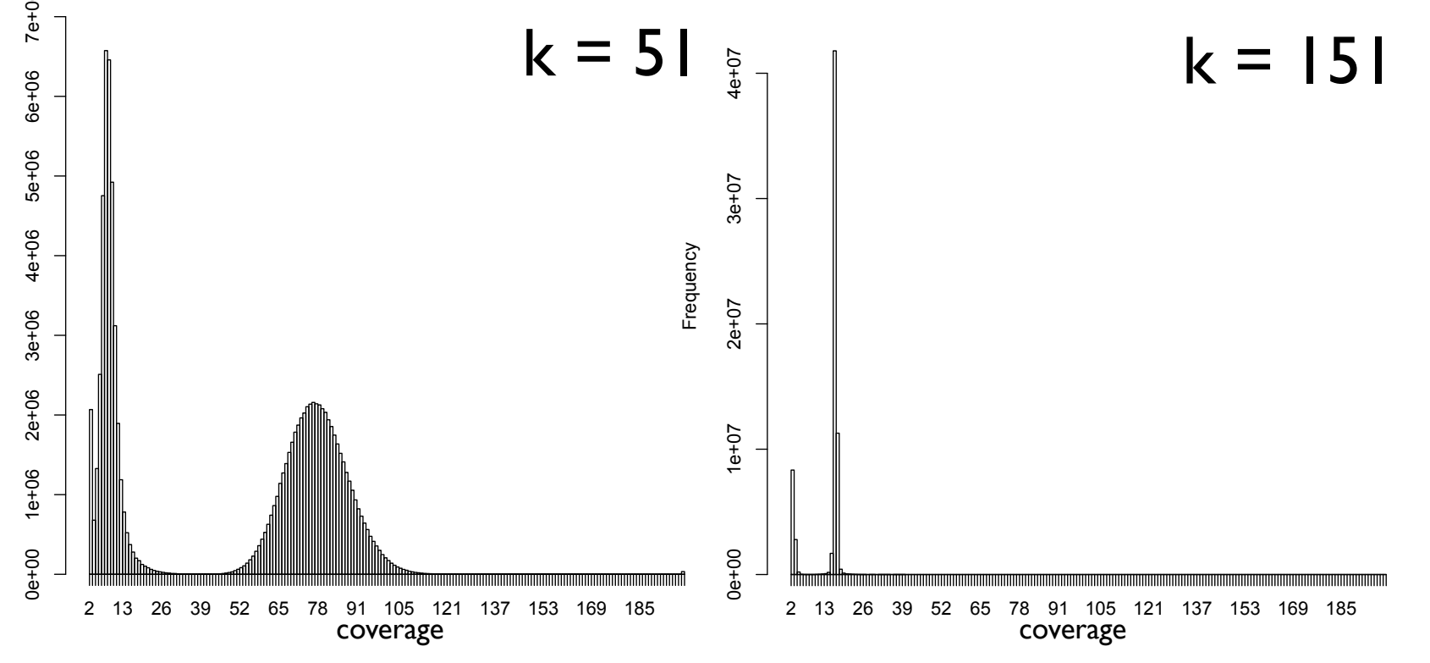


Figure 6: kmer-coverage frequencies for the 1:9 data set.

Similarly, the completeness of gene predictions varies substantially with coverage ratio and assembler choice (Table 2). Nearly all reference genes of the low coverage genome can be recovered from the *SPAdes* and *MIRA* assemblies. For *MetaVelvet*, *Velvet* and *Omega*, this number depends to a large extent on the coverage ratio.

Table 2: Number of fungal gene predictions mapping to the 10740 reference genes.

Assembler	1:9	2:8	3:7
MIRA	10348	10715	10718
Omega	72	5825	10302
SPAdes	10656	10666	10683
sga	10100	10674	10675
Velvet	2845	8817	10530
MetaVelvet	4	66	1657

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* performing best.
- Assembler performance in our data is not driven by misassemblies.
- Optimising the N50 can lead to the preclusion of sequences representing the low-coverage genome.
- Assembler choice has a large impact on the completeness of gene predictions.

doi:10.6084/m9.figshare.1337845

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] 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