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

Whole genome shotgun sequencing of multi species communities using only a single library layout is commonly used to assess taxonomic and functional complexities of large and diverse microbial communities. We investigate to what extent such metagenome skimming approaches are applicable for in-depth genomic characterizations of obligate symbiotic communities involving eukaryotes, e.g. lichens. It is still unclear how a eukaryotic species mixture 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 investigate how those paradigms perform on eukaryotic species mixtures, we performed an *in silico* study based on a lichen system, mimicking a genome skimming study. 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. Reconciling the sexpectations from the *in silico* study with the outcome of a real-world metagenome skimming of the lichen *Lasallia pustulata* indicates methodological problems causing the underrepresentation of one symbiont in the shotgun library.

1. *in silico* Sequencing

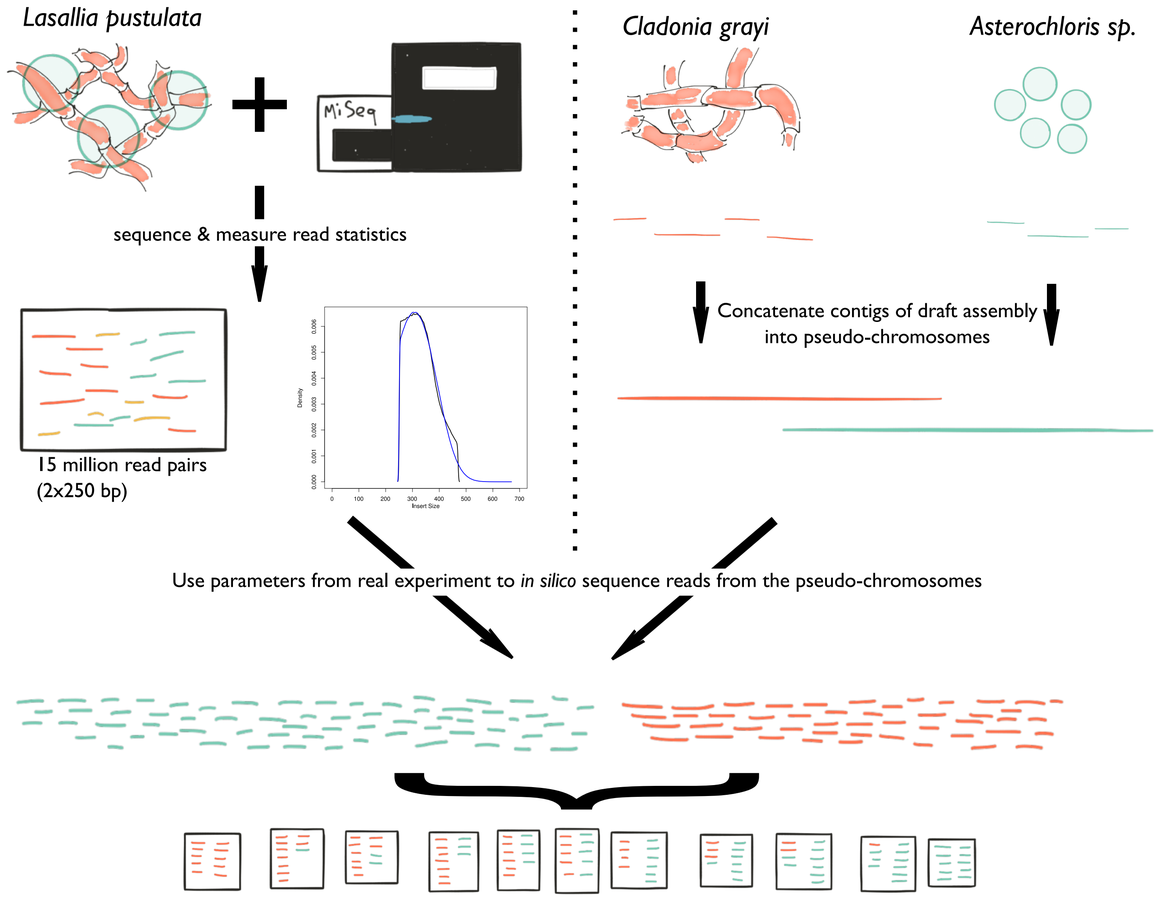

Figure 1: Workflow for generating twin data sets, which resemble a real sequencing data set in terms of insert size distribution, read number and read length.

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 sizes (Figure 1, left).

The scaffolds of the draft genomes of *Cladonia grayi* [2] and *Asterochloris sp.* [3] were concatenated to create contiguous pseudo-chromosomes (Figure 1, right). These were analysed using Repeatmasker [4] (Box I) and Gepard [5] (Fig 2.)

From the pseudo-chromosomes reads were simulated with the parameters estimated from *L. pustulata,* using ART [6], and mixed into 11 twin data sets of varying composition (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

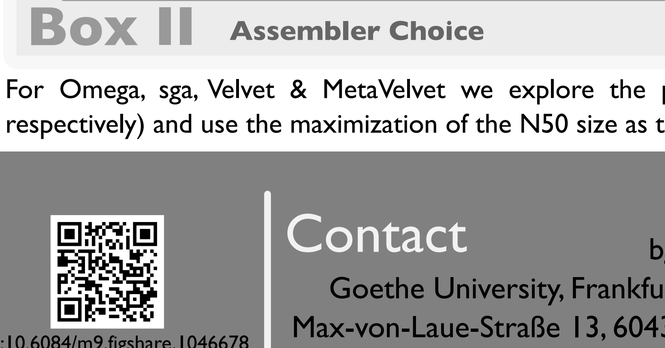

Figure 2: Dotplot of the pseudo-chromosomes of *Cladonia grayi* and *Asterochloris sp.* The end (~ 1/4th) of the chromosome of *Cladonia grayi* shows high self-similarity, while this effect is not observed for Asterochloris, which shows little self similarity.

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 Bruijn Graph* based**	Velvet [7] Standard de Bruijn Graph	MetaVelvet [8] Metagenome DBG Assembler	SPAdes [9] Multisized de Bruijn Graph
***Overlap Layout Consensus* based**	MIRA [10] Regular Overlap Layout Consensus	Omega [11] Metagenome OLC Assembler	sga [12] Sting Graph Assembler

Box II Assembler Choice

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

3. Assembly Results

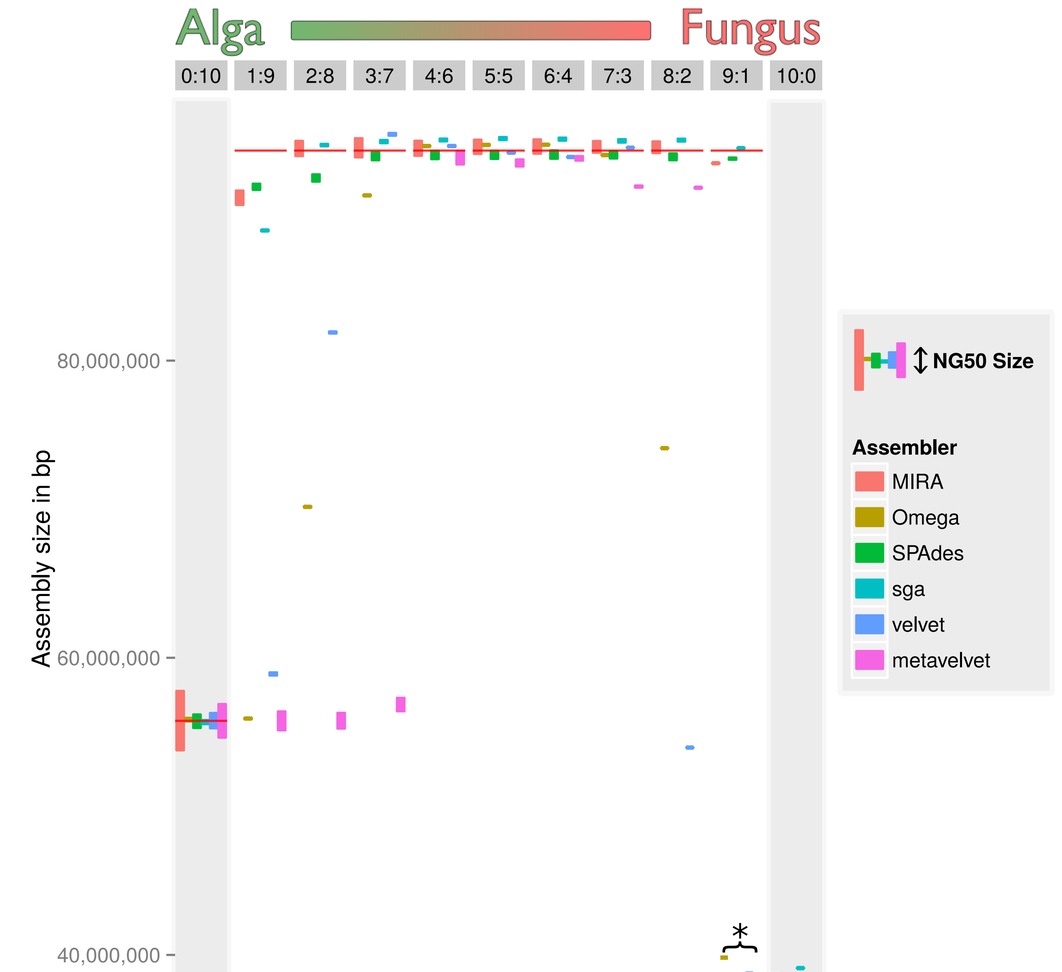

Figure 3: Assembly results over the 11 data sets. Bars are centered at total assembly length (red lines are reference lengths). Height of the bars shows the NG50 size. For the assemblies with the asterisk the total assembly length was less than 50% of the reference length. A default height was used in those 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. For the alga many assemblers were able to exceed the NG50 size of the original draft genome. For the fungus, repeats hindered such an extension with the present WGS library layout (Figure 4).

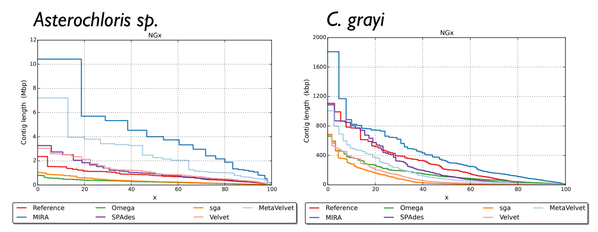

Figure 4: NGx distributions for *Asterochloris sp.* & *C. grayi*

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. 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).

The k-mer coverage plots provide an explanation for the sensitivity 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.

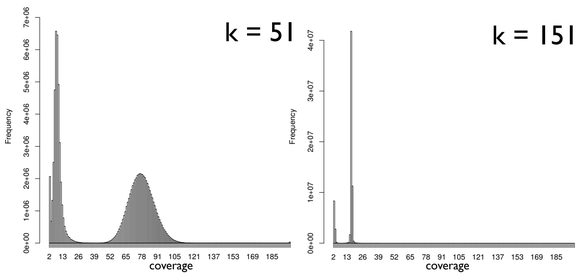

Figure 6: kmer-coverage frequencies for the 1:9 data set.

	Number of Contigs		
	64,180		
	Total Length		
	119 Mbp		
	N50		
Whole Assembly	3.3 Kbp		
	6977	8872	19,371
	Total Length		
	37 Mbp	14 Mbp	34 Mbp
	N50		
	19 kbp	2 kbp	3 kbp

Box III Assembly Thallus

The WGS data of the *L. pustulata* thallus was assembled using *MIRA* and the contigs taxonomically assigned using *MEGAN* [13]. Both, fungal and algal, assemblies are much more fragmented than expected given the simulations. This is a result of the bacterial diversity also present in the thallus: nearly 1/3 of the total assembly is of bacterial origin (Box III, right column).

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.
- Optimising the N50 can lead to the preclusion of sequences representing the low-coverage genome.
- Real world data sets from lichen thalli not only contain fungal & algal DNA but also a large bacterial proportion.

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