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

Summary

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. Here we investigate to what extent such metagenome skimming approaches are applicable for the in-depth characterizations of genomes represented 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 twin sets of genome skimming experiments of a lichen. We show that the quality of genome reconstructions from such data depends on assembler choice, but more importantly also on the parameter optimisation strategy and the ratio of the taxa in the metagenome. Optimising for assembly metrics such as N50 can in extreme cases lead to the exclusion of complete genomes. Transfering the results of the simulations to a real-world metagenome skimming experiment of the lichen *Lasallia pustulata*, we not only shows a larger species diversity, but also hints to biased sequencing coverage for the algal genome.

1. *in silico* Sequencing

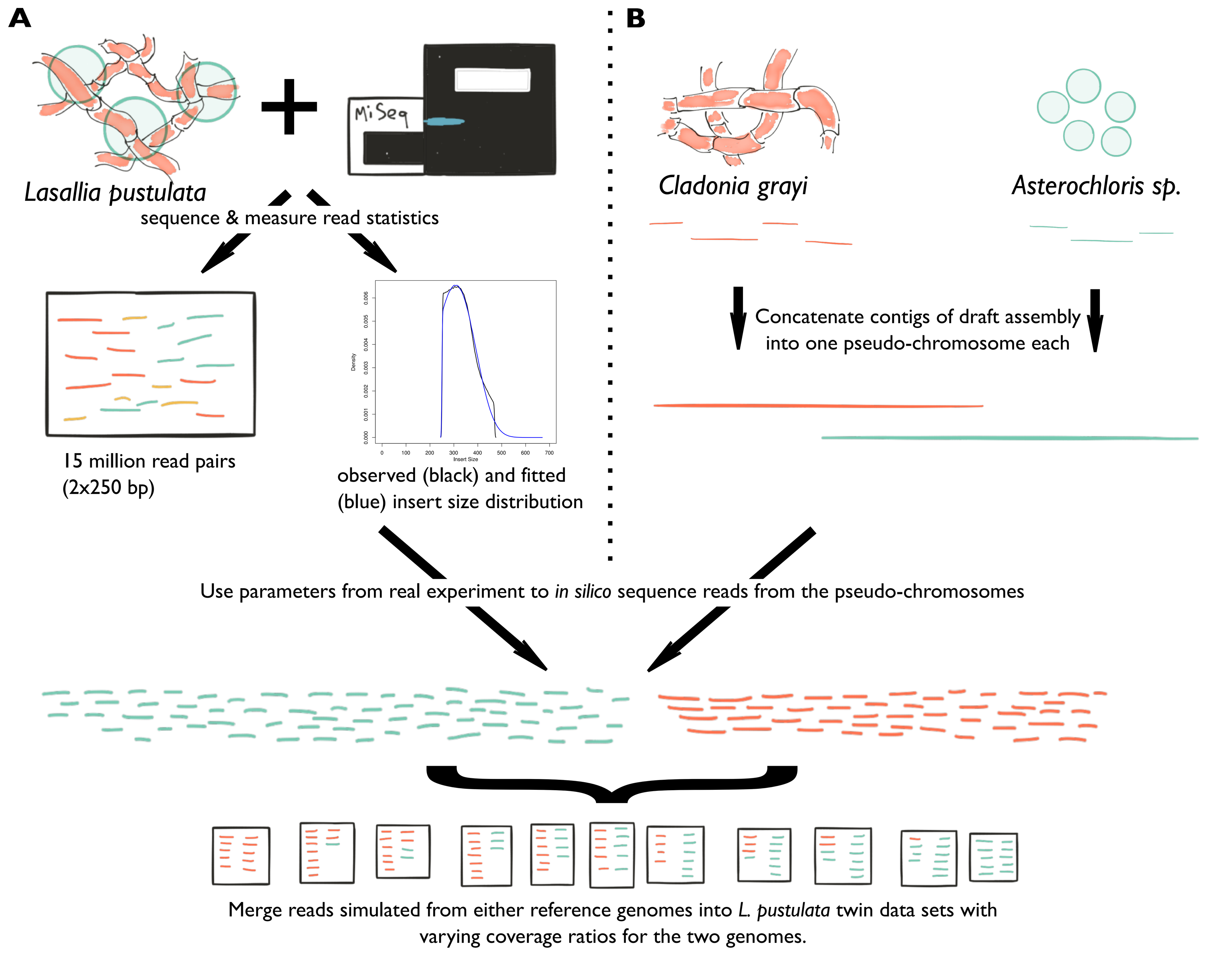

Figure 1: Workflow for generating twin sets, resembling a real sequencing data set with respect to insert size distribution, read number and read length.

DNA from a thallus of *Lasallia pustulata* was sequenced using Illumina MiSeq technology, yielding 15 million read pairs with a read length of 250 bp. To estimate the insert size distribution, we joined overlapping 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.)

	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

Using the pseudo-chromosomes as templates, we simulated reads using ART [6], parameterized with the values estimated from the *L. pustulata* data . The reads were used to compile 11 twin sets by mixing fungal and algal reads at varying ratios (Table 1).

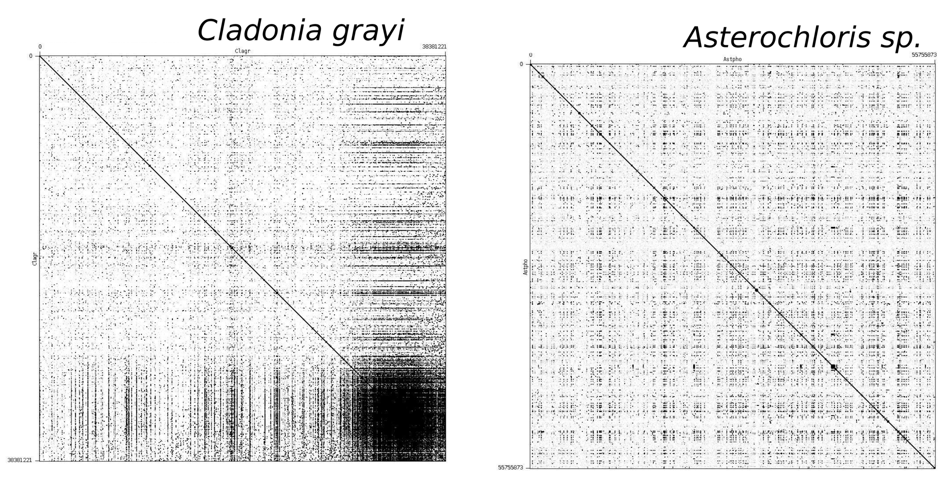

Figure 2: Dotplot of the pseudo-chromosomes of *Cladonia grayi* and *Asterochloris sp.*

Table 1: Absolute coverages for each organism per twin 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 & Parameter Selection

De Bruijn 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 used the maximization of the N50 size as the objective.

3. Assembly Results

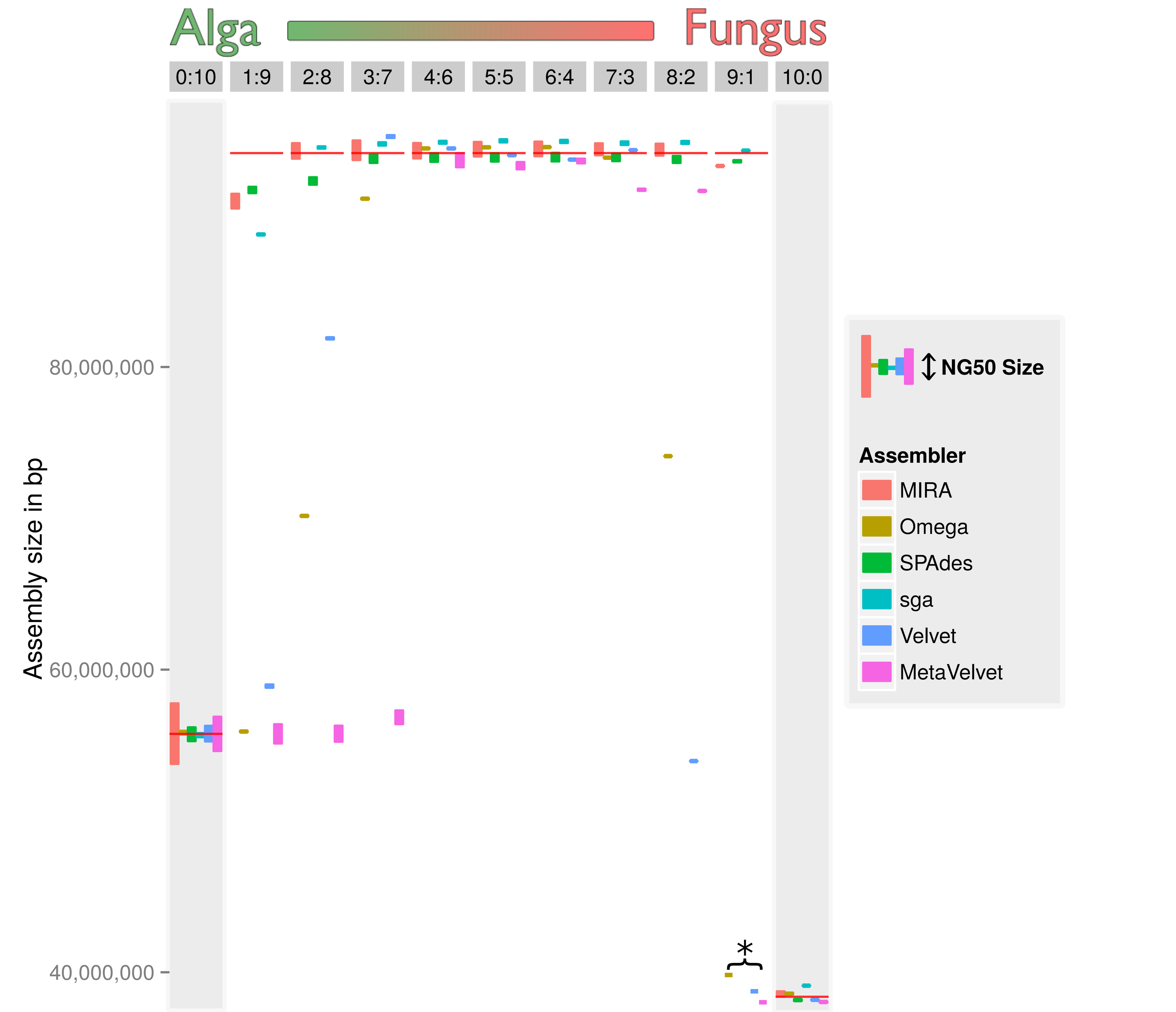

Figure 3: Assembly results for the 11 data sets and the different assemblers. Bars are centered at total assembly length, red lines are reference lengths. Height of the bars shows the NG50 size. For the assemblies with the asterisk the total assembly length was below 50% of the reference length. A default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

Single species data Almost all assemblers reconstruct the two genomes over their full length from the simulated sequences (Figure 3, column 0:10 & 10:0), however with varying contiguity. For the alga most assemblers exceeded the NG50 size of the original draft assembly. For the fungus, it appears that the repetitive nature of the genome (c.f. Fig. 2) hinders the generation of longer contigs with the present WGS library layout (Figure 4).

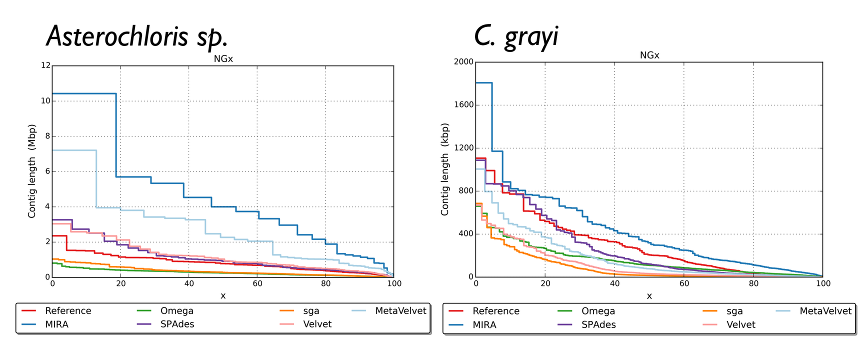

Figure 4: NGx distributions for *Asterochloris sp.* & *C. grayi*

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 *Velvet* and the metagenome assemblers fail to assemble large parts of the low coverage genome once coverage ratios become extreme (Fig. 3, 1:9 - 3:7, 9:1).

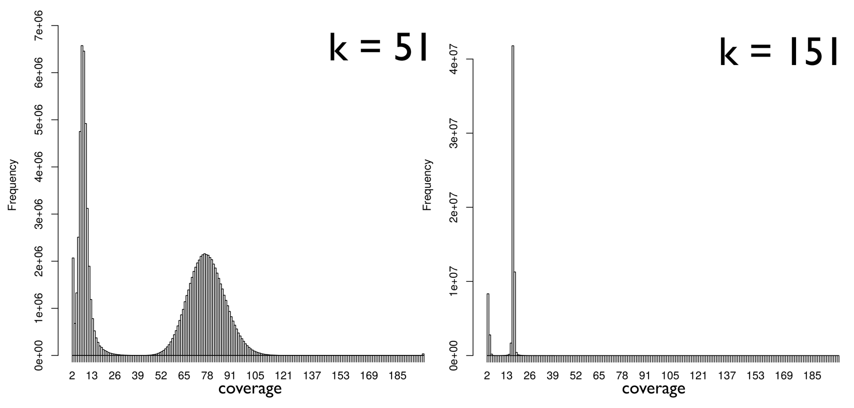

Figure 6: kmer-coverage frequencies for the 1:9 data set.

Parameter optimization for N50 impairs metagenome assembly Contigs from the low coverage genomes tend to be short, decreasing the overall N50 size. Increasing the word length k to larger values shifts the word frequencies of the low coverage genomes to overlap with those introduced by sequencing errors (Fig. 6). Thus they are ignored turing assembly, increasing the N50 size.

L. pustulata The assembly was done with *MIRA* and the contigs were taxonomically assigned using *MEGAN* [13]. The algal assembly is much more fragmented than expected given the *in silico* study. (Box III). It appears that this is a result of a biased library preparation, yielding a highly uneven read coverage for the algal genome.

Whole Assembly

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

	Fungal	Algal	Bacteria
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

Key Points

- 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.
- For metagenomes optimising the assembly parameters using the N50 size as quality measurecan lead to the preclusion of entire genomes.
- Despite a 10x coverage, the algal genome assembly is highly fragmented, representing <1/4 of the estimated genome size. This may be due to an uneven algal genome coverage, resulting from a biased library preparation.

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.1046678

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