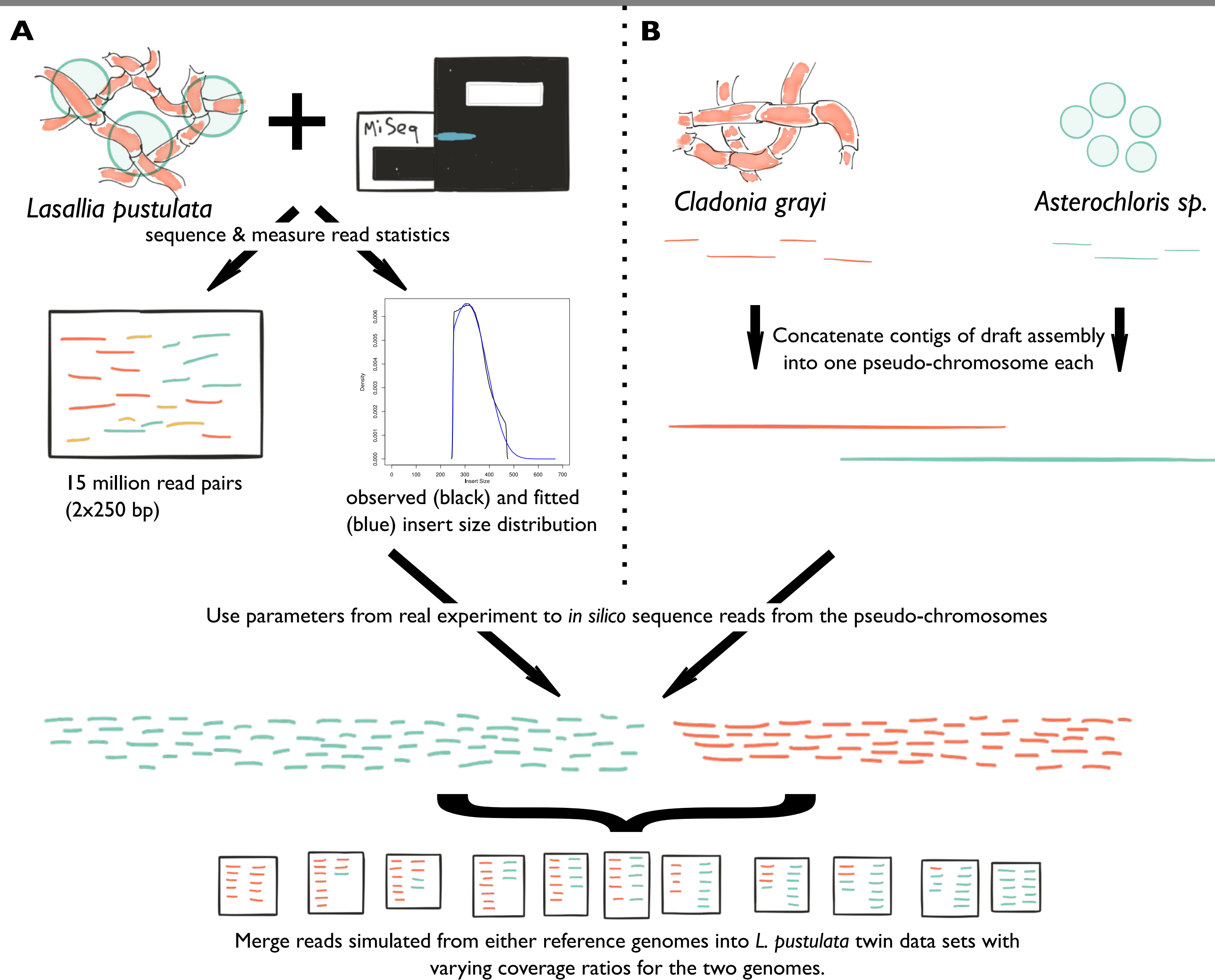


# Potential and pitfalls of eukaryotic metagenome skimming:

## 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 Bruijn Graph* based methods or *Overlap Layout* based assemblers and how to optimize assembly parameters as k-mer or overlap sizes.

## I. *in silico* Sequencing

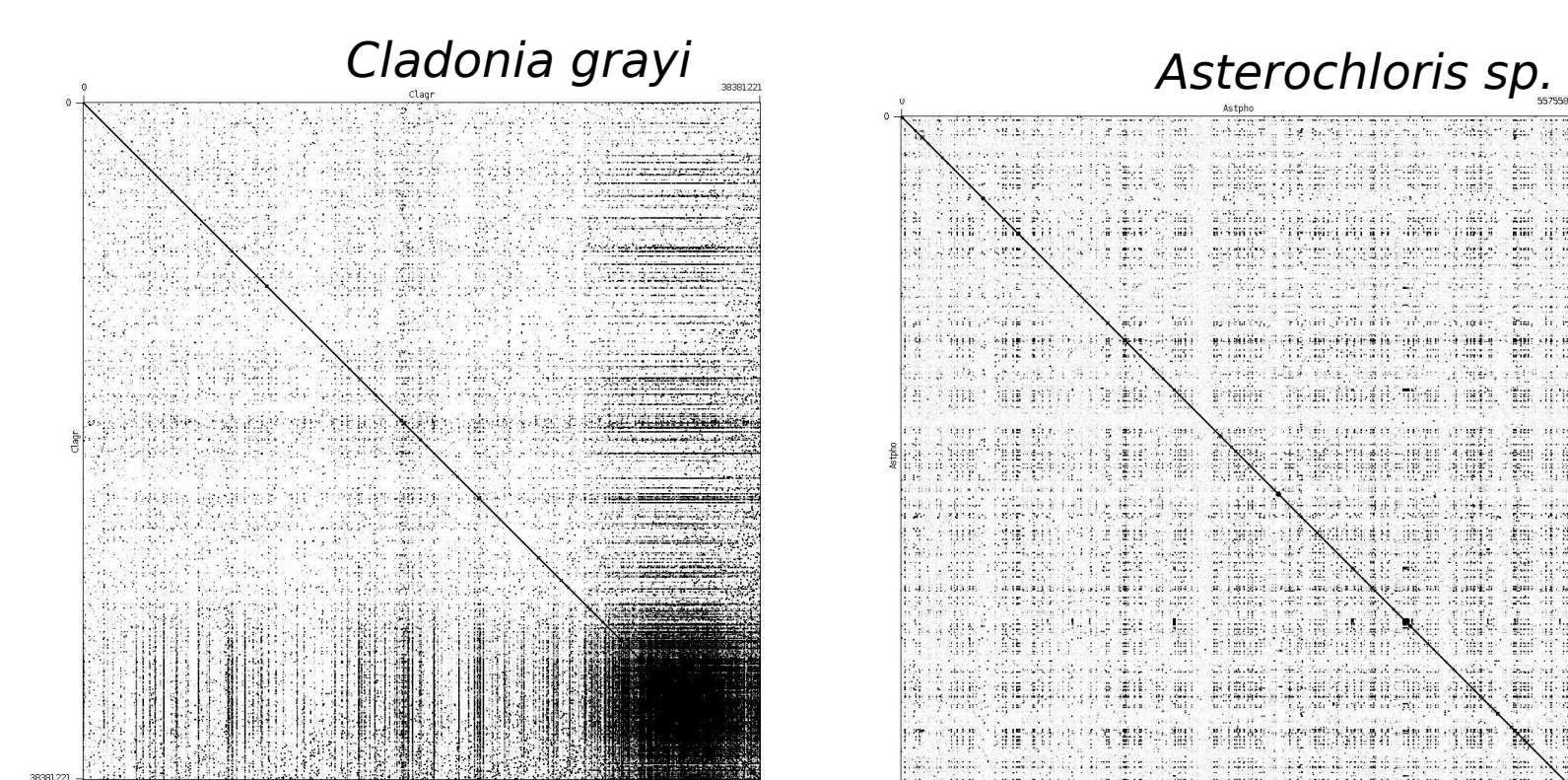


**Figure 1:** Workflow for generating twin sets, resembling a real sequencing data set with respect to insert size distribution, read number and read length.

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).

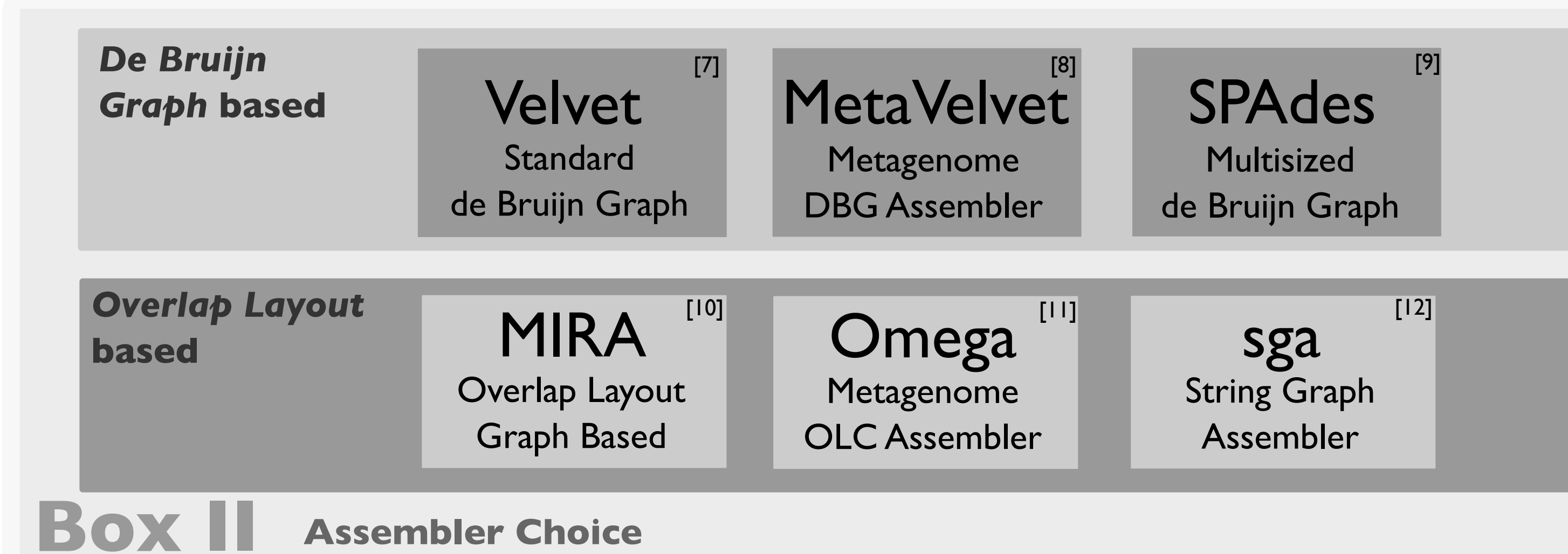
The scaffolds of the genomes of *Cladonia grayi* [2] and *Asterochloris sp.* [3] were each concatenated to create a contiguous pseudo-chromosome, respectively (Figure 1, B). Both were checked for repeat content & self-similarity using Repeatmasker [4] (Box I) and Gepard [5] (Fig 2).

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 I).



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

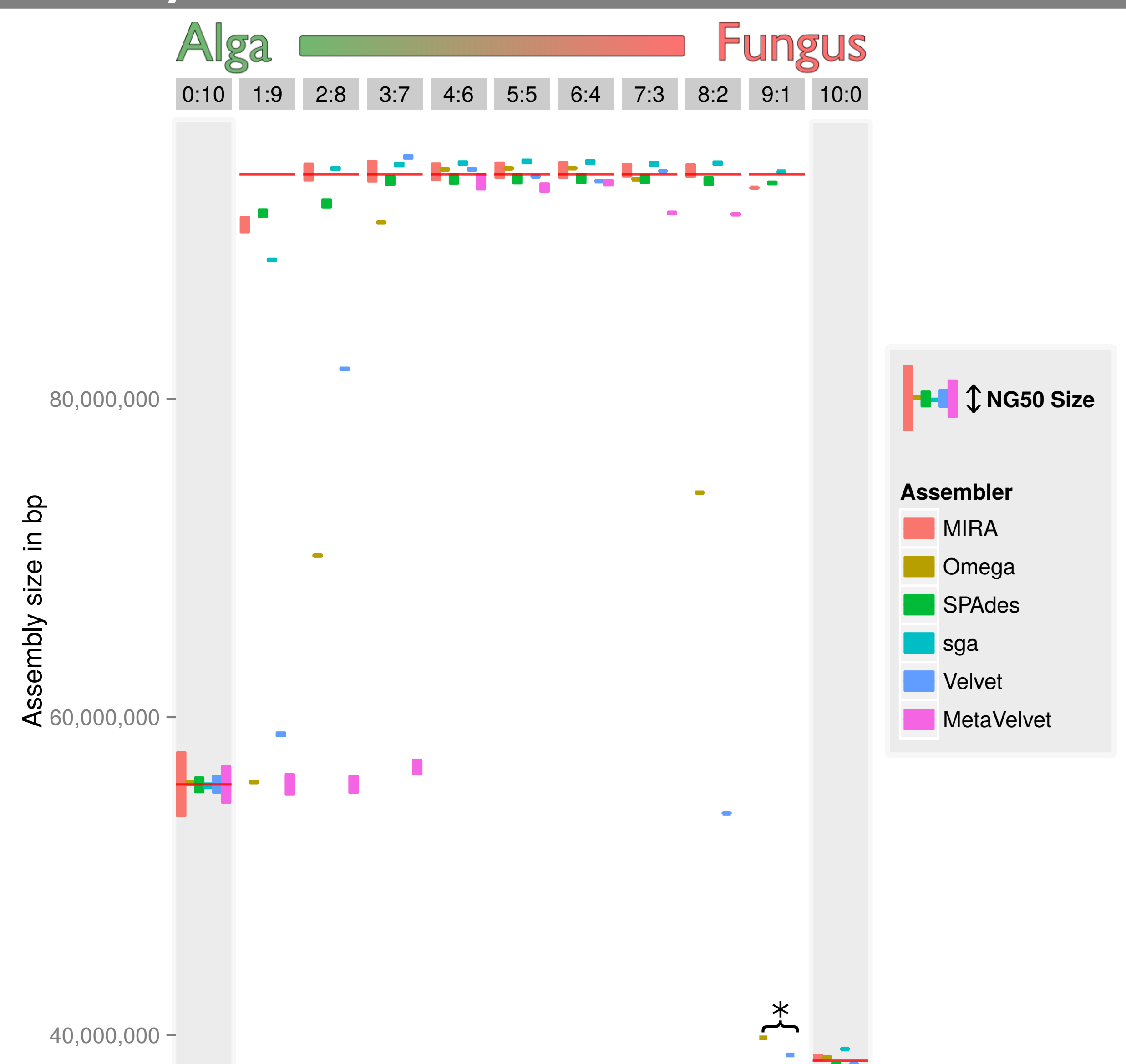
## 2. Assembler Selection & Parameter Selection



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.

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. Transferring the results of the simulations to a real-world metagenome skimming experiment of the lichen *Lasallia pustulata*, we not only see a larger species diversity, but also hints to biased sequencing coverage for the algal genome.

## 3. Assembly Results



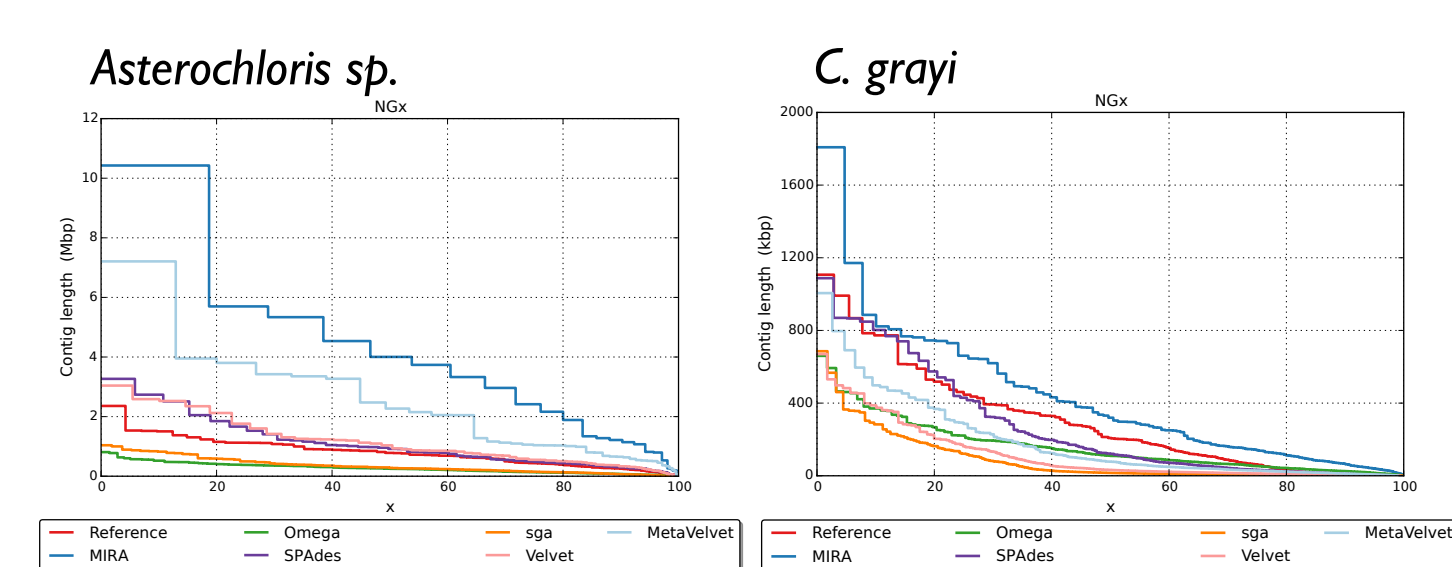
**Figure 3:** Assembly results for the 11 data sets and the different assemblers. Bars are centered at total assembly length, red lines are reference lengths. Height of the bars shows the NG50 size. For the assemblies with the asterisk the total assembly length was below 50% of the reference length.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

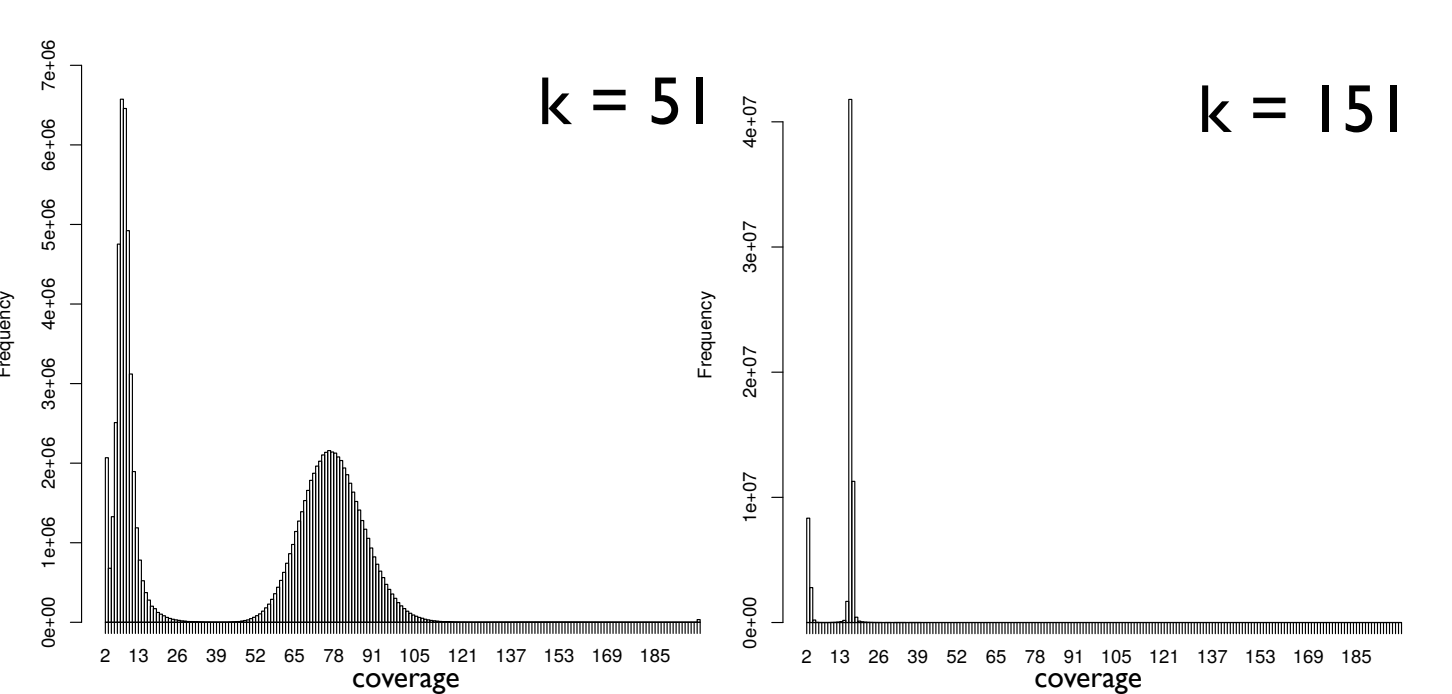
**Mixed species data** Completeness of the genome reconstructions depends heavily on assembler choice and coverage ratios. *MIRA* and *SPAdes* perform best across all data sets. In contrast *Velvet* and the metagenome assemblers fail to assemble large parts of the low coverage genome once coverage ratios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

**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 during 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.



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



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



## 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 measure can 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.

