Do we reference Matt's Oyster River paper?

Quality Metric	**Higher in NCGR**	**Higher in DIB**
Transrate score, “cds”	44	583
Transrate score, “nt”	495	143
Mean ORF %	42	596
Percentage of references with CRBB	100	538
Number of contigs	12	626

Table 1. Number of assemblies with higher values in NCGR or DIB assemblies for each quality metric.

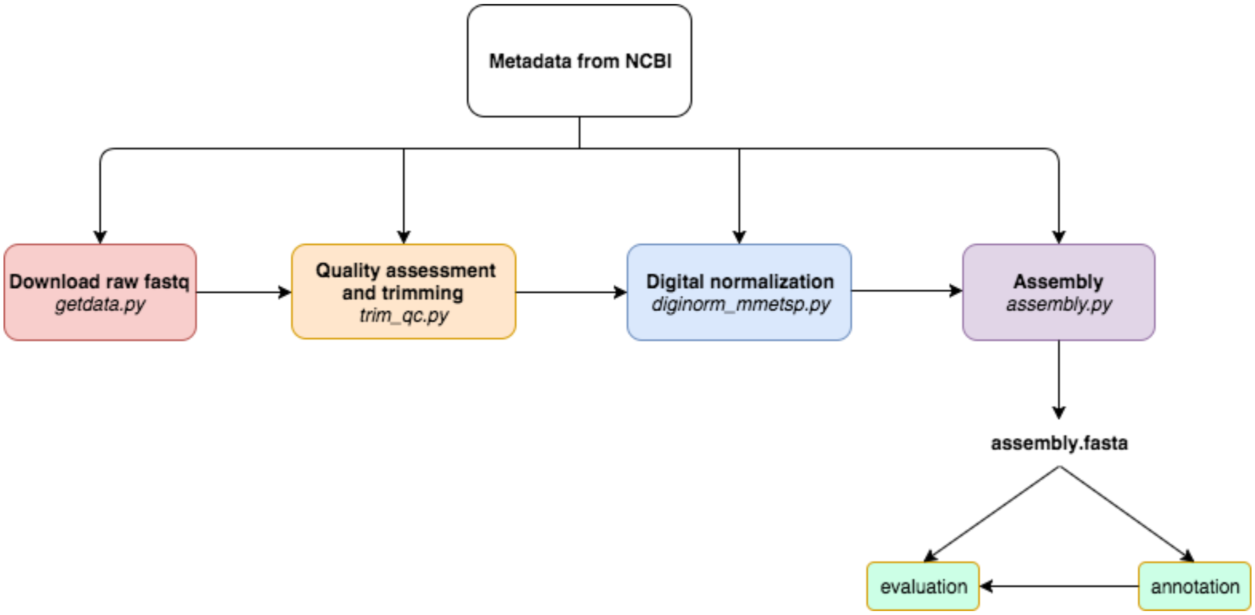


Figure 1. A modularized *de novo* transcriptome assembly pipeline workflow where tools used for each step can be substituted such that output assembly files can be used to test the effects of the individual tools on the overall pipelines. Metadata in the SraRunInfo.csv file downloaded from NCBI was used as input to the pipeline. The steps of the pipeline are as follows: download raw fastq data with the fastq-dump script in the SRA Toolkit [1], quality assessment with FastQC [40] and trimming residual Illumina adapters and low quality bases (Q<2) with Trimmomatic [42], digital normalization with khmer version 2.0 [3], and assembly with Trinity [18]. Each script in the pipeline uses a metadata file obtained from the SRA as input. If a process is terminated, the automated nature of this pipeline allows

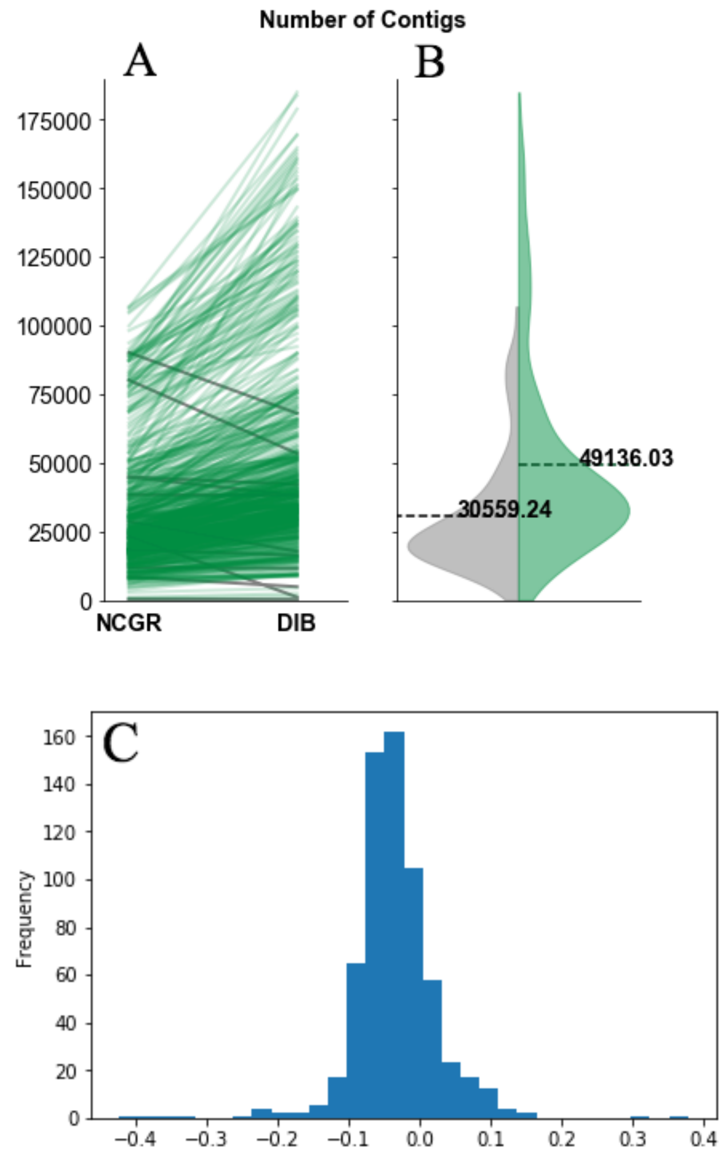


Figure 2. Comparing quality metrics between each assembly. (A) The line plot compares contig numbers between the DIB and the NCGR assemblies. Gray lines represent values where NCGR was higher than DIB and green lines represent values where DIB was higher than NCGR. and ~~split violin plots of the number of assembled contigs.~~ (B) Split violin plots comparing distributions of contig numbers. In the green (right side of B) are the DIB re-assemblies and in gray (left side of B) are the original assemblies from NCGR. (C) Histogram showing the frequency distribution of transrate score differences between the NCGR 'nt' version and DIB re-assemblies. Negative values on the x-axis indicate that NCGR had a higher Transrate score and positive values indicate that the DIB had a higher Transrate score.

?? why?
(will look in paper)

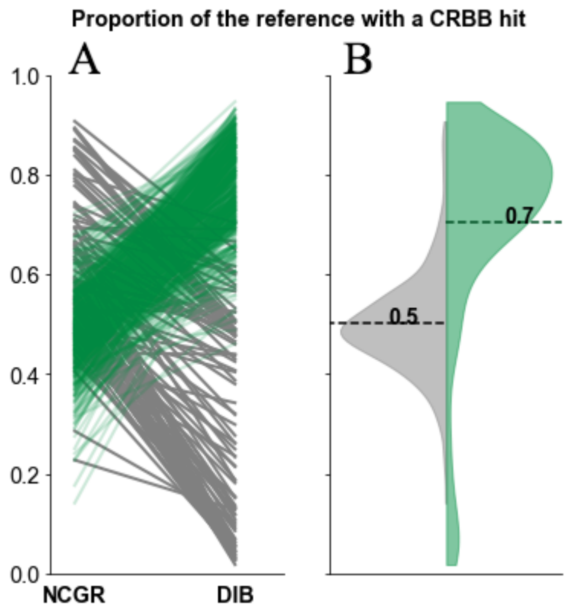
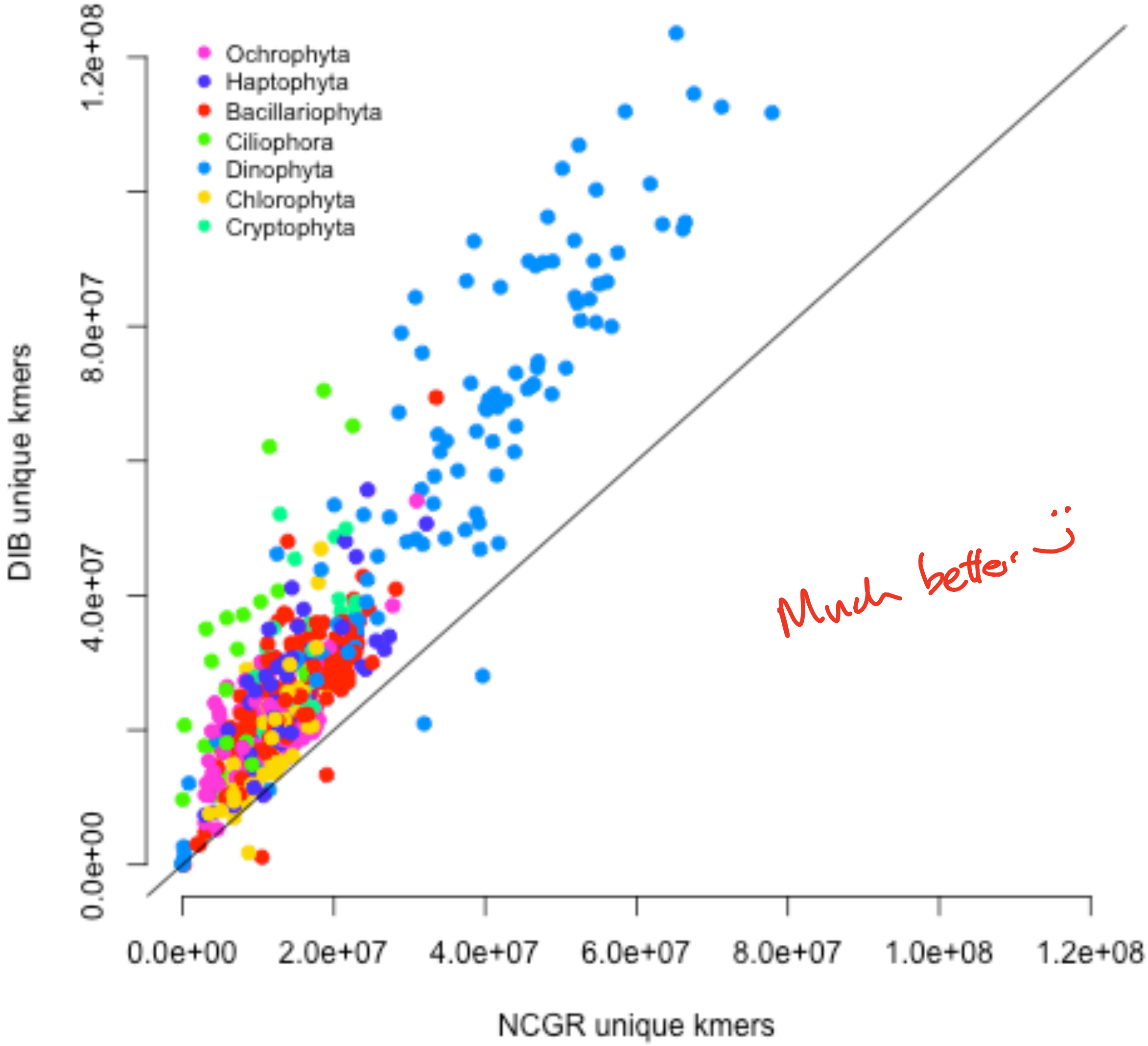


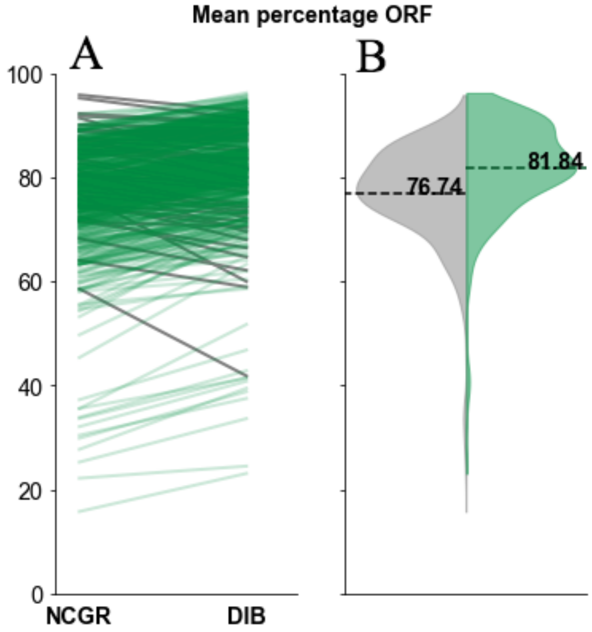
Figure 3. When the DIB re-assemblies were assigned as references, there was a higher proportion of conditional reciprocal best BLAST (CRBB) hits compared to when NCGR ‘nt’ assemblies were assigned as references. This indicated that the DIB re-assemblies contained more of the sequence content in the NCGR assemblies than the NCGR assemblies contained the DIB re-assemblies. (A) Line plot comparing proportion of CRBB hits between NCGR ‘nt’ assemblies and DIB between the same samples. (B) Violin plots showing the distribution of the proportion of NCGR transcripts with reciprocal BLAST hits to DIB (grey) and the proportion of DIB transcripts with reciprocal BLAST hits to NCGR (green).



Much better :)

Figure 4. Unique numbers of *k*-mers (*k*=25), calculated with the HyperLogLog function in the khmer software package [43,51], in each of the 678 DIB re-assemblies compared to the NCGR 'nt' assemblies along a 1:1 line. Samples are colored based on their phylum level affiliation. There were 536 samples that had higher unique *k*-mer values in the DIB re-assemblies than in the NCGR assemblies whereas 99 of the samples had higher unique *k*-mer content in the original NCGR assemblies.

really?
are you sure?



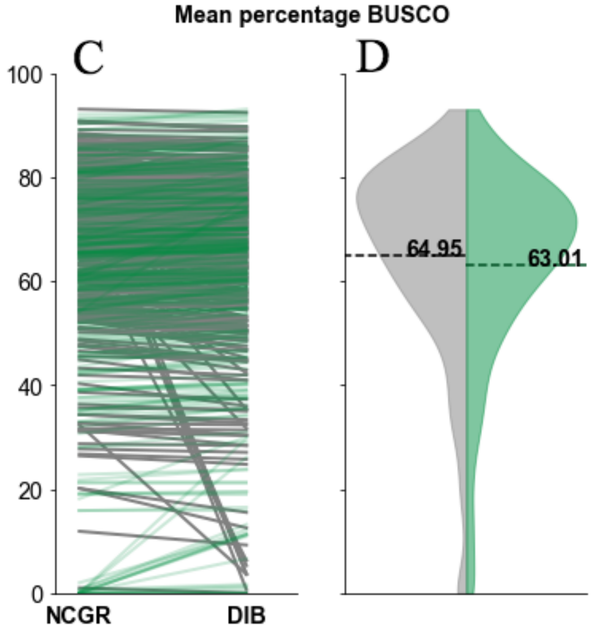


Figure 5. The percentage of contigs with predicted open reading frame (ORF) (A, B) and the percentage of complete protistan benchmarking universal single-copy orthologs (BUSCO) recovered in each assembly (C, D). In the green (right side B, D) are the “DIB” re-assemblies and in gray (left side of B, D) are the original assemblies from NCGR ‘nt’ assemblies. Line plots (A,C) compare values between the DIB and the NCGR ‘nt’ assemblies. Gray lines represent values where NCGR was higher than DIB and green lines represent values where DIB was higher than NCGR.

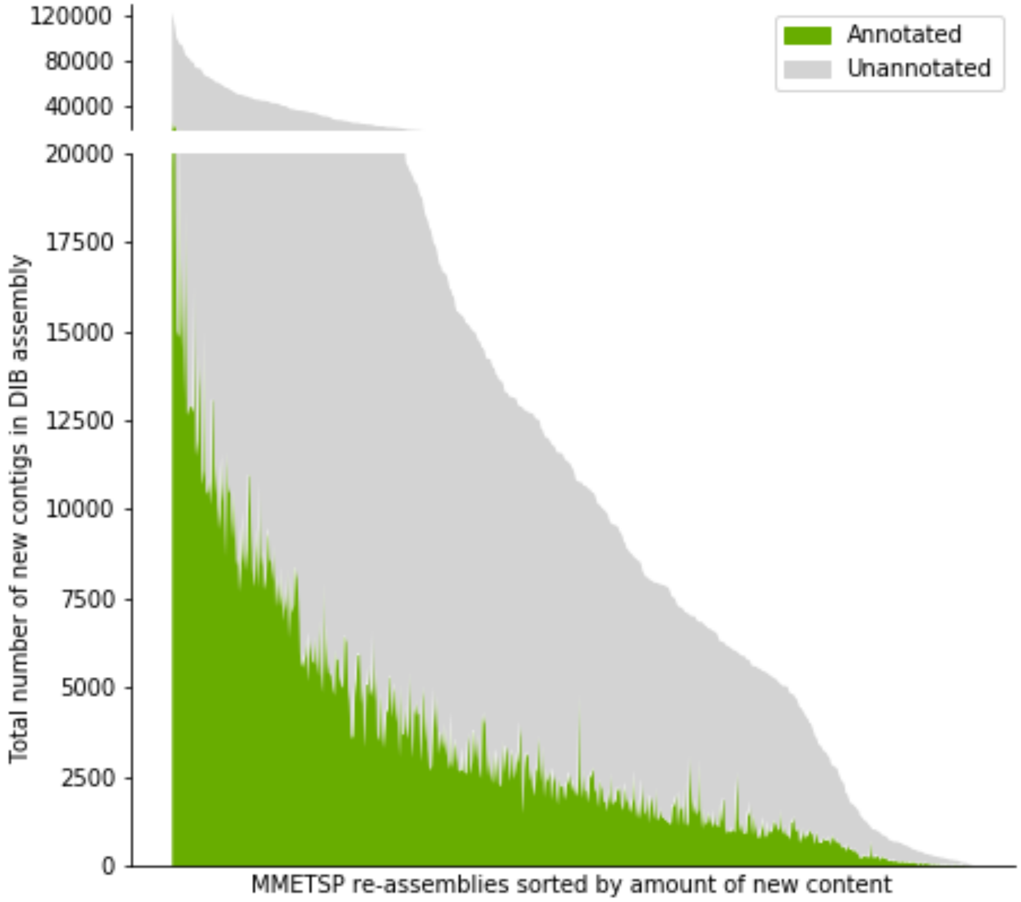


Figure 6. A histogram across MMETSP samples depicting the number of contigs identified as novel in DIB assemblies. These contigs were absent in the NCGR assemblies, based on negative conditional reciprocal best BLAST (CRBB) results. Samples are sorted from highest to lowest number of 'new' contigs. The region in gray indicates the number of unannotated contigs present in the DIB re-assemblies, absent from NCGR 'nt' assemblies. Highlighted in green are contigs that were annotated with dammit [44] to a gene name in the Pfam, Rfam, or OrthoDB databases, representing the number of contigs unique to the DIB re-assemblies with an annotation.

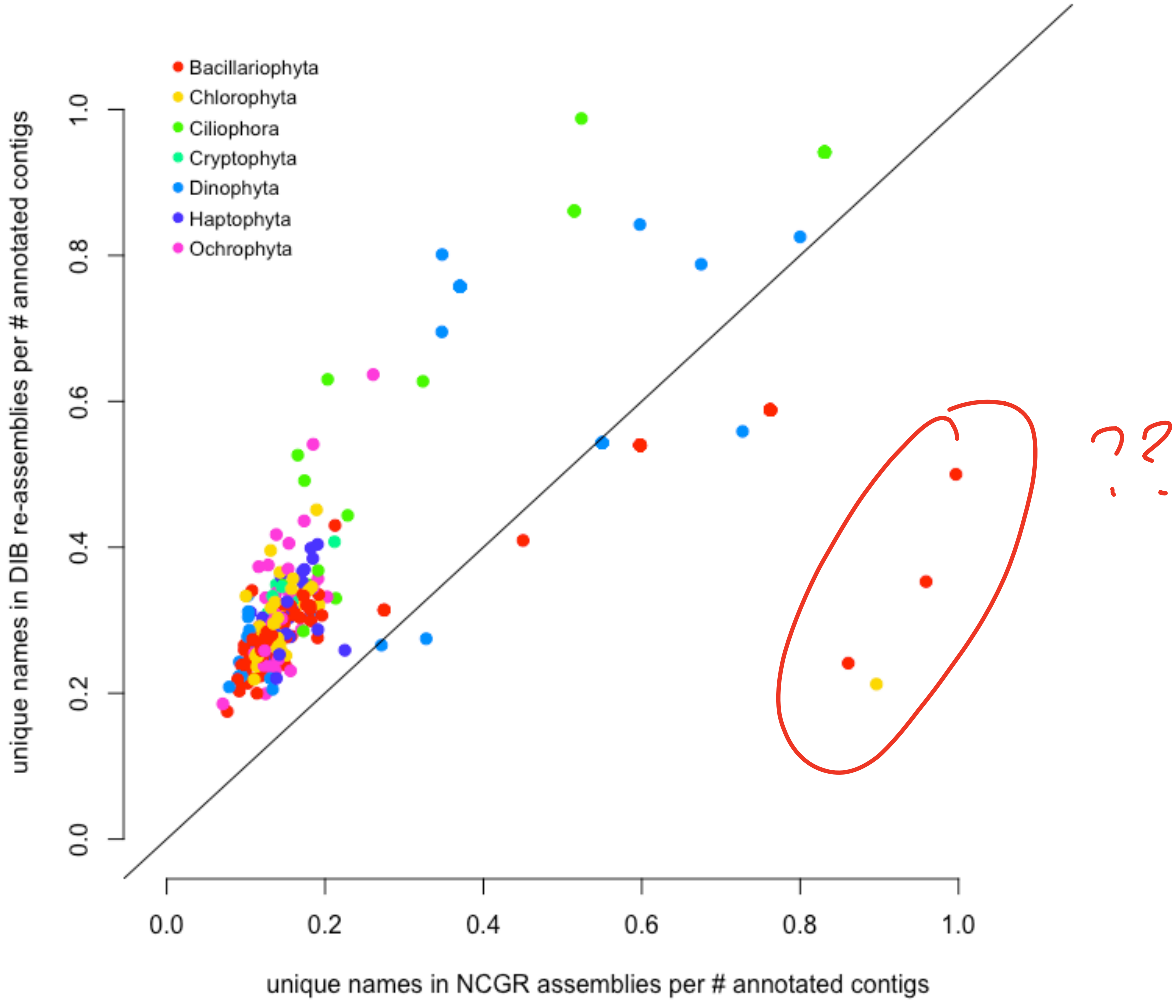
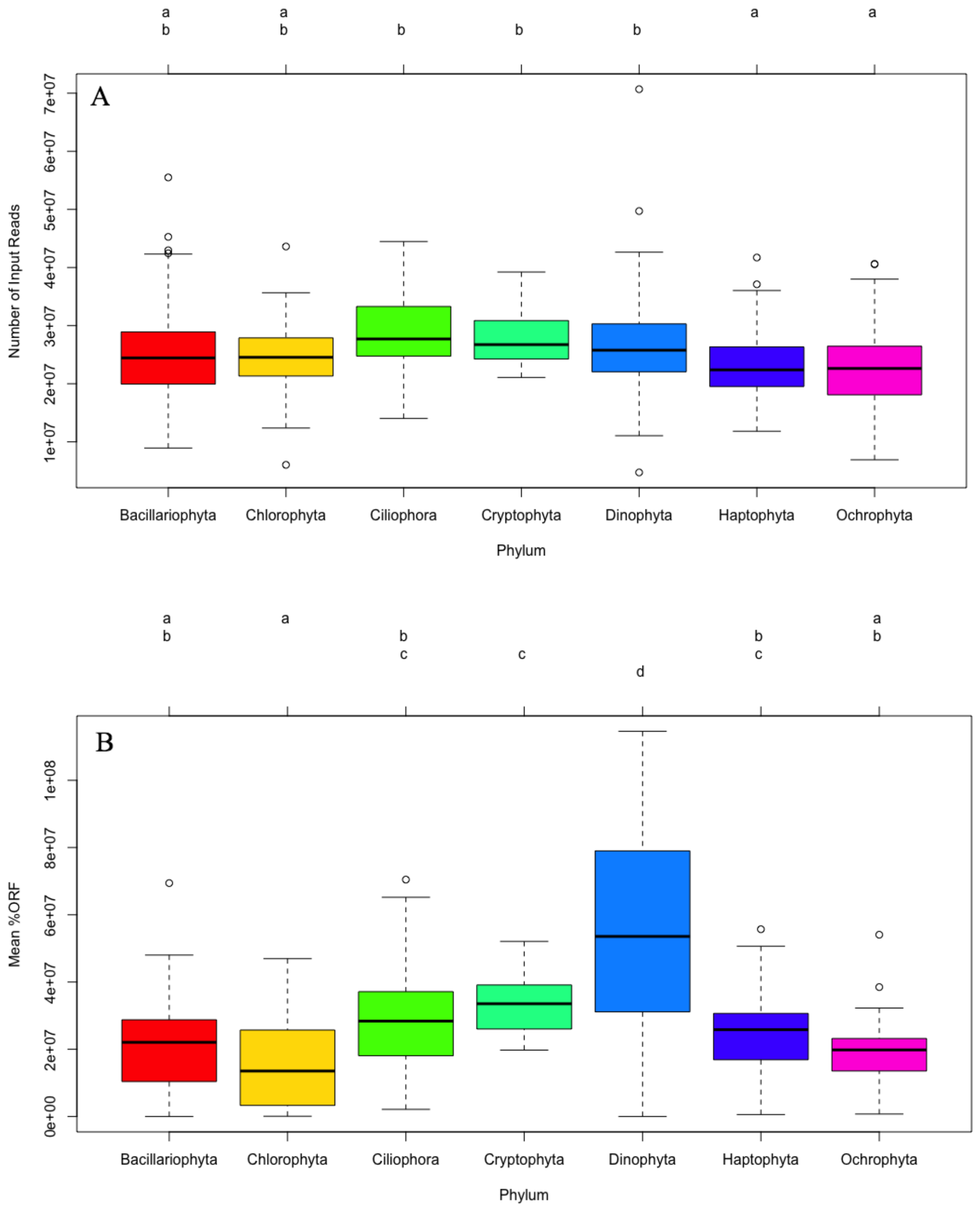


Figure 7. Unique gene names found in either NCGR ‘nt’ assemblies or DIB re-assemblies but not found in the other assembly, normalized to the number of annotated contigs in each assembly. The line indicates a 1:1 relationship between the number of unique gene names in DIB and NCGR. DIB assemblies had the highest number of unique names not found in NCGR assemblies. Several NCGR assemblies had gene names not found DIB assemblies.

what does this sentence mean?


A
a b
a b
b
b
b
a
a
Number of Input Reads
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
7e+07
Bacillariophyta
Chlorophyta
Ciliophora
Cryptophyta
Dinophyta
Haptophyta
Ochrophyta
Phylum
B
a b
a
b c
c
d
b c
a b
Mean %ORF
0e+00
2e+07
4e+07
6e+07
8e+07
1e+08
Bacillariophyta
Chlorophyta
Ciliophora
Cryptophyta
Dinophyta
Haptophyta
Ochrophyta
Phylum

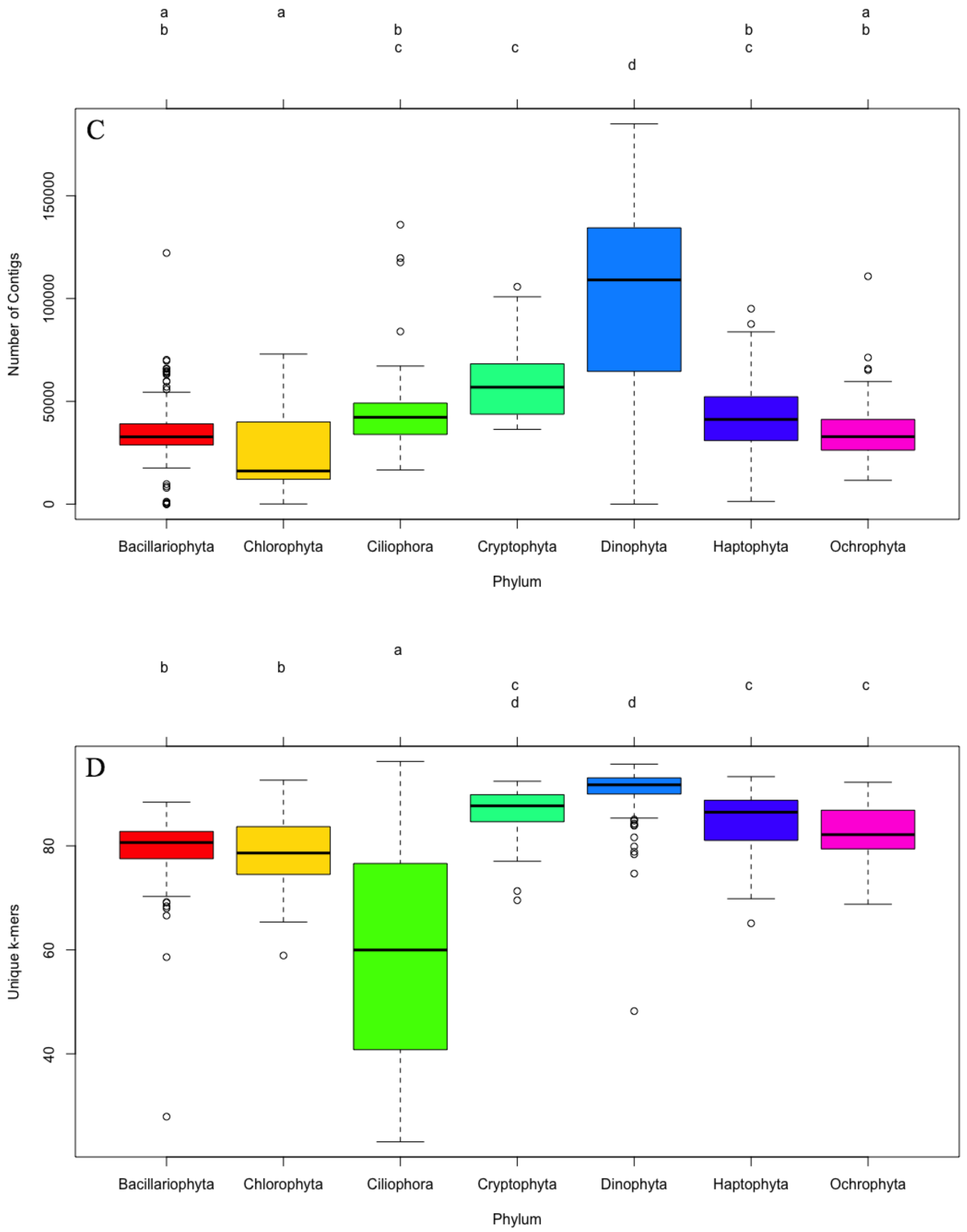


Figure 8. Box-and-whisker plots for the seven most common phyla in the MMETSP dataset. Clockwise from the top left (A) number of input reads, (B) mean percentage open reading frame (ORF) content, (C) number of contigs in the assembly, (D) unique *k*-mers (*k*=25) in the assembly. Groups sharing a letter in the top margin are not significantly different at the 5% level. While there do not appear to be differences in the number of input reads between these phyla (A), the Dinophyta phylum has higher percentage of ORF and number of contigs and the Ciliophora phylum has lower unique *k*-mer content.

a higher

(the highest?)

ORFs

Supplemental Files:

Data Table 1. MMETSP_all_evaluation_matrix.csv
Data Table 2. MMETSP_all_evaluation_matrix_METADATA.csv
Supplemental Figure 1. Transrate scores comparisons between NCGR ‘cds’ and ‘nt’ versions vs. DIB.
Supplemental Figure 2. Transrate score differences colored by taxonomic grouping.
Supplemental Figure 3. BUSCO scores with the Protista database, NCGR ‘nt’ vs. DIB.
Data notebook. Different Trinity versions