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.

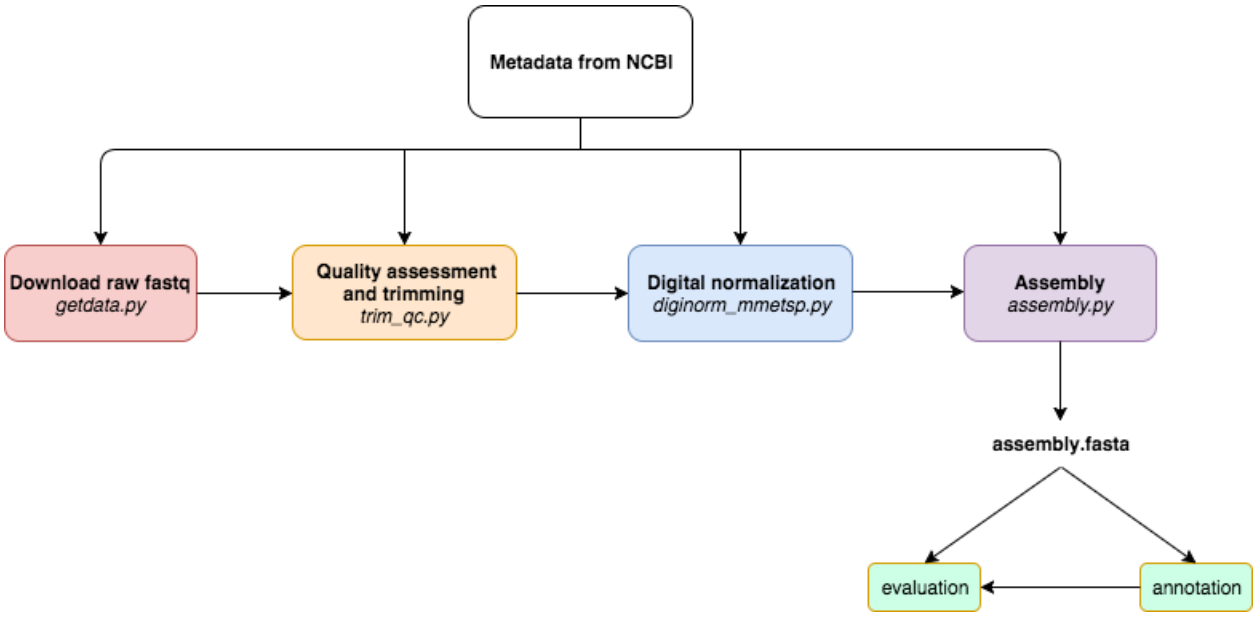
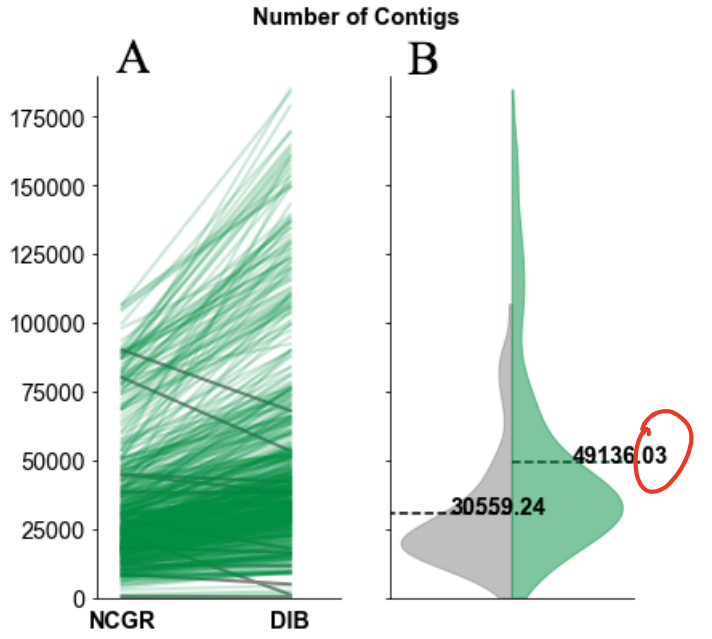


Figure 1. A programmatically automated *de novo* transcriptome assembly pipeline was developed for this study. Metadata in the SraRunInfo.csv file downloaded from NCBI was used as input for each step of the pipeline to indicate which samples were processed. The steps of the pipeline are as follows: download raw fastq data with the fastq-dump script in the SRA Toolkit, quality assessment with FastQC and trimming residual Illumina adapters and low quality bases (Q<2) with Trimmomatic, digital normalization with khmer version 2.0, and *de novo* transcriptome assembly with Trinity. If a process was terminated, the programmatically automated nature of this pipeline allowed for the last process to be run again without starting the pipeline over again. In the future, if a new sample is added, the pipeline can be run from beginning to end with just new samples, without having to repeat the processing of all samples in the dataset as one batch. If a new tool becomes available, for example a new assembler, it can be substituted in lieu of the original Trinity tool used by the assembly.



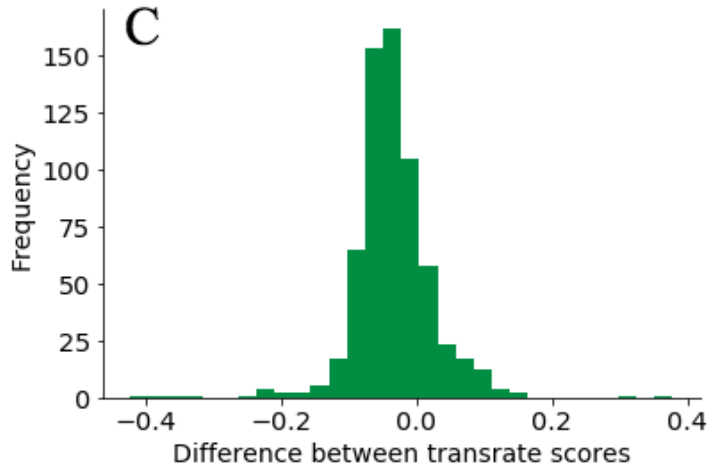


Figure 2. The number of contigs and Transrate quality score varied between DIB and NCGR assemblies. (A) Slopegraphs show shifts in the number of contigs for each individual sample between the DIB and the NCGR assembly pipelines. Gray lines represent values where NCGR was higher than DIB and green lines represent values where DIB was higher than NCGR. (B) Split violin plots show the distribution of the number of contigs in each assembly with the original assemblies from NCGR in gray (left) and the DIB re-assemblies and in green (right side of B). (C) The difference in Transrate score between the DIB and NCGR assemblies is shown as a histogram. Negative values on the x-axis indicate that the NCGR assembly had a higher Transrate score and positive values indicate that the DIB assembly had a higher Transrate score.

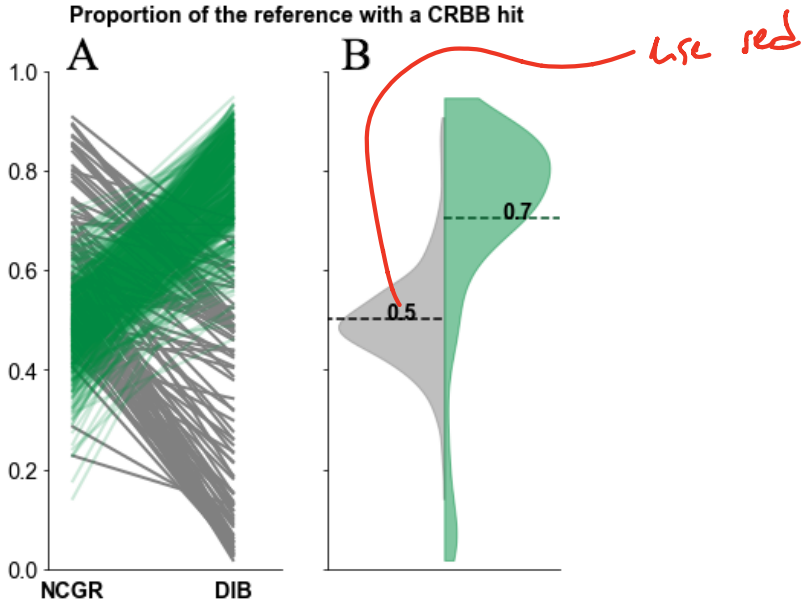


Figure 3. (A) Line plot comparing proportion of CRBB hits between NCGR 'nt' assemblies and DIB between the same samples. (B) Violin plots showing the distribution of the proportion of NCGR transcripts with reciprocal BLAST hits to DIB (grey) and the proportion of DIB transcripts with reciprocal BLAST hits to NCGR (green).

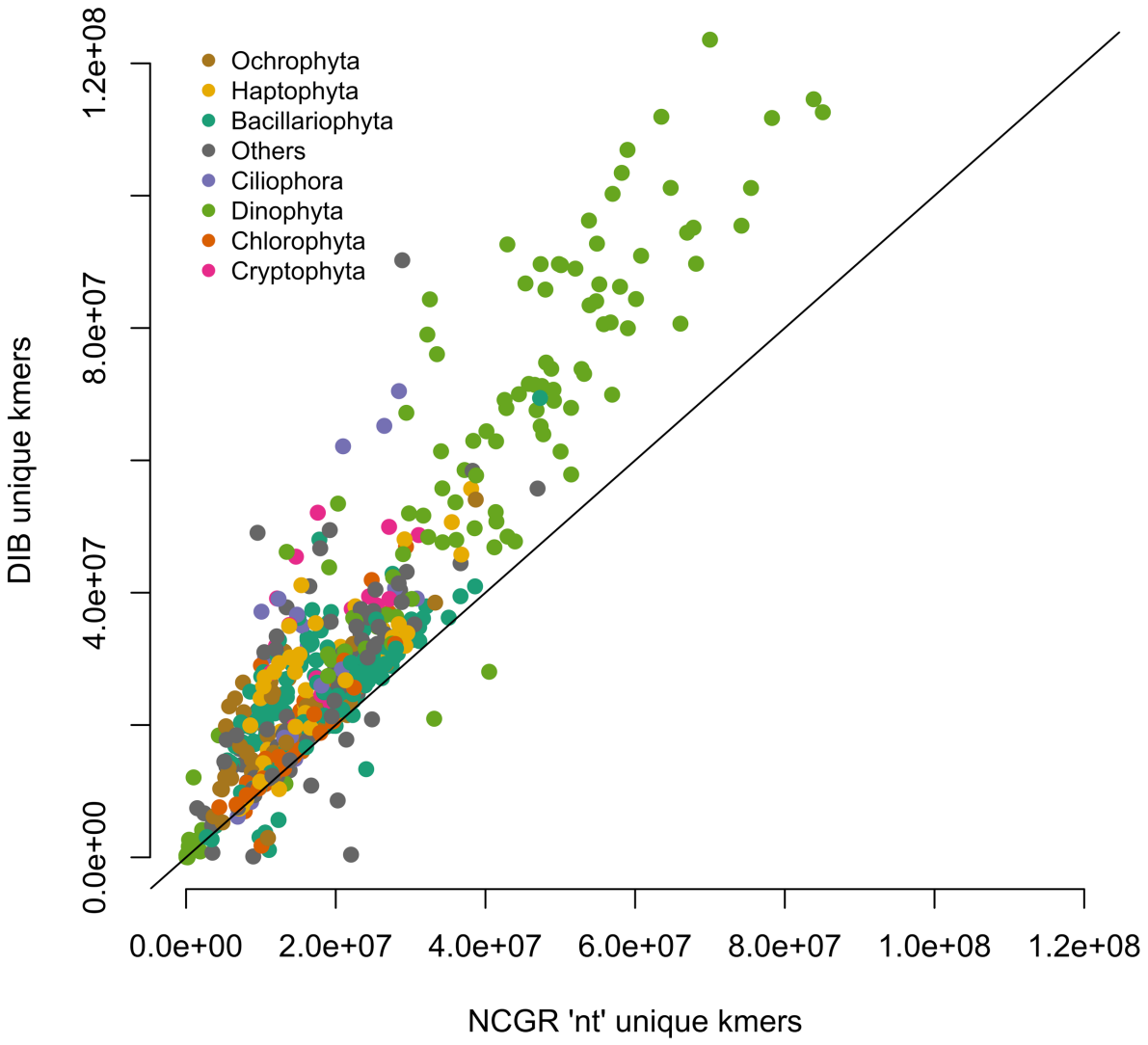


Figure 4. Unique numbers of *k*-mers (*k*=25) in seven most represented phyla, calculated with the HyperLogLog function in the khmer software package. DIB re-assemblies were compared to the NCGR 'nt' assemblies along a 1:1 line. Samples are colored based on their phylum level affiliation. More than 95% of the DIB re-assemblies had ~~higher~~ unique *k*-mers compared ~~to~~ the NCGR assembly of the same sample.

more

than

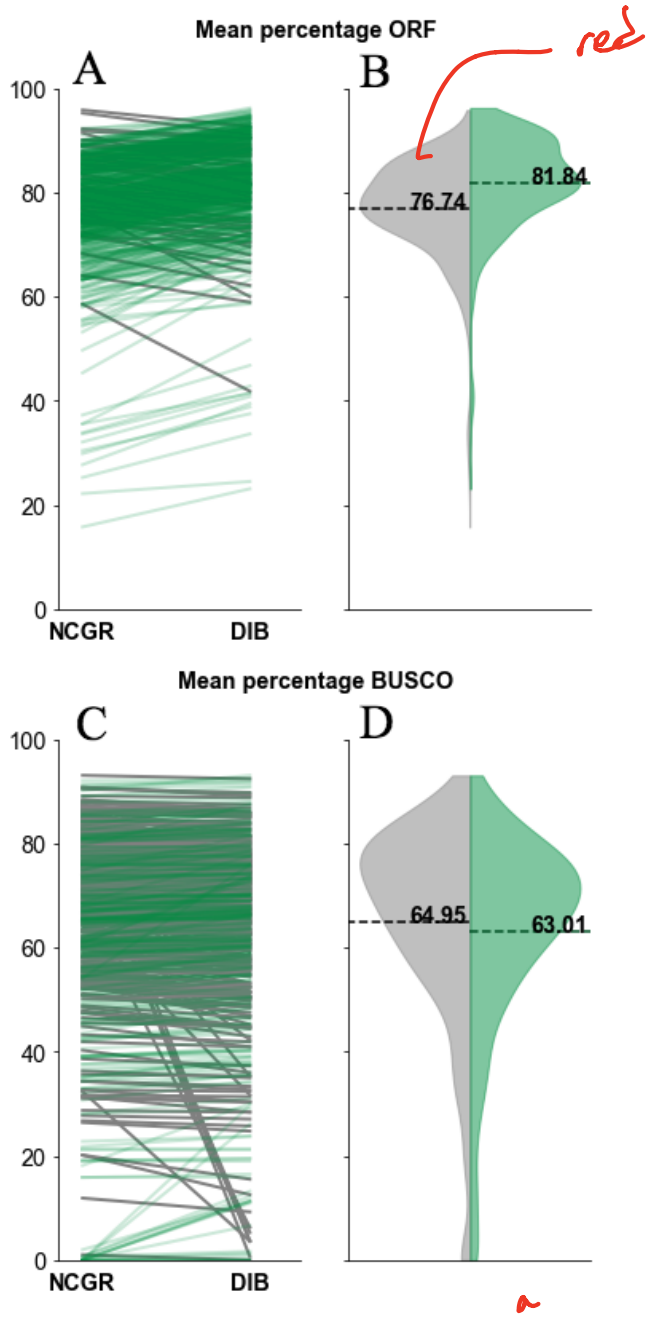


Figure 5. The percentage of contigs with predicted open reading frame (ORF) (A, B) and the percentage of complete protistan benchmarking universal single-copy orthologs (BUSCO) recovered in each assembly (C, D). In the green (right side B, D) are the "DIB" re-assemblies and in gray (left side of B, D) 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.

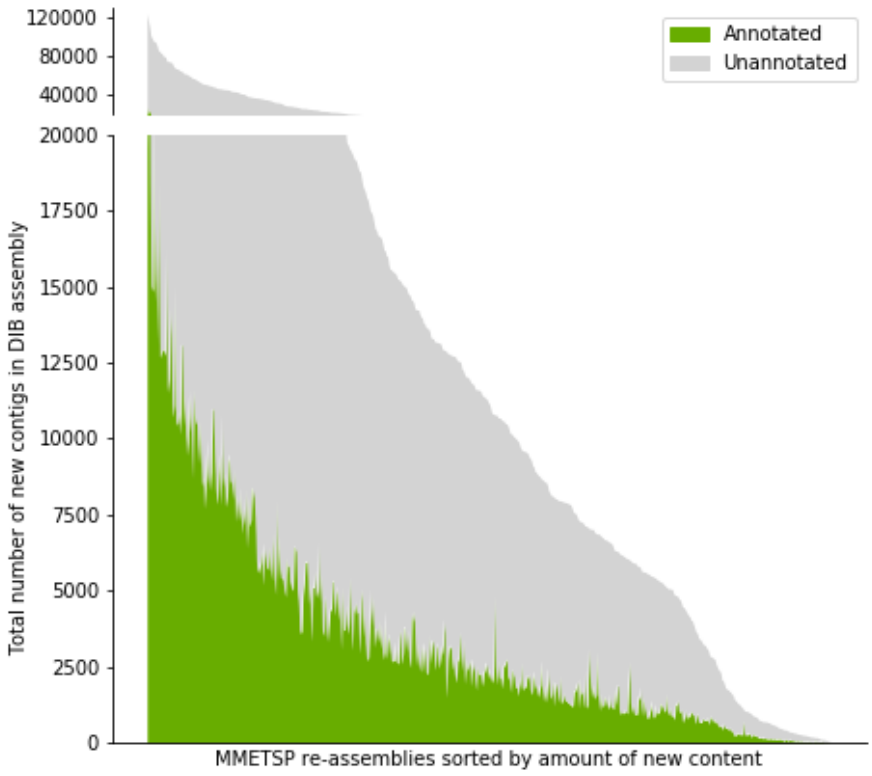


Figure 6. A histogram across MMETSP samples depicting the number of contigs identified as novel in DIB assemblies. These contigs were absent in the NCGR assemblies, based on negative conditional reciprocal best BLAST (CRBB) results. Samples are sorted from highest to lowest number of 'new' contigs. The 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.

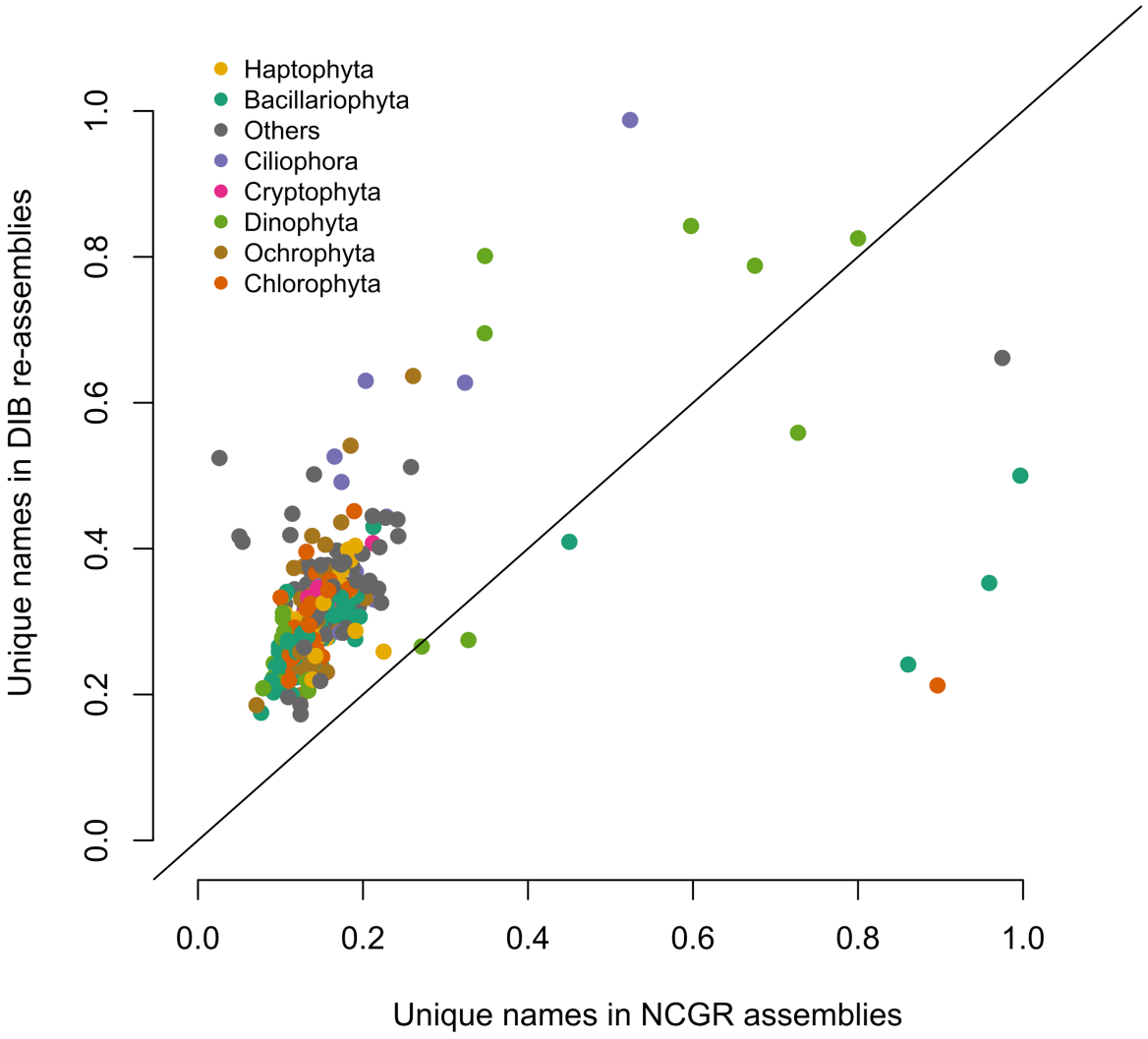


Figure 7. Unique gene names found in a subset (296 samples) of either NCGR 'nt' assemblies or DIB re-assemblies but not found in the other assembly, normalized to the number of annotated contigs in each assembly. The line indicates a 1:1 relationship between the unique gene names in DIB and NCGR. More than 97% of the DIB assemblies had more unique gene names than in NCGR assemblies of the same sample.

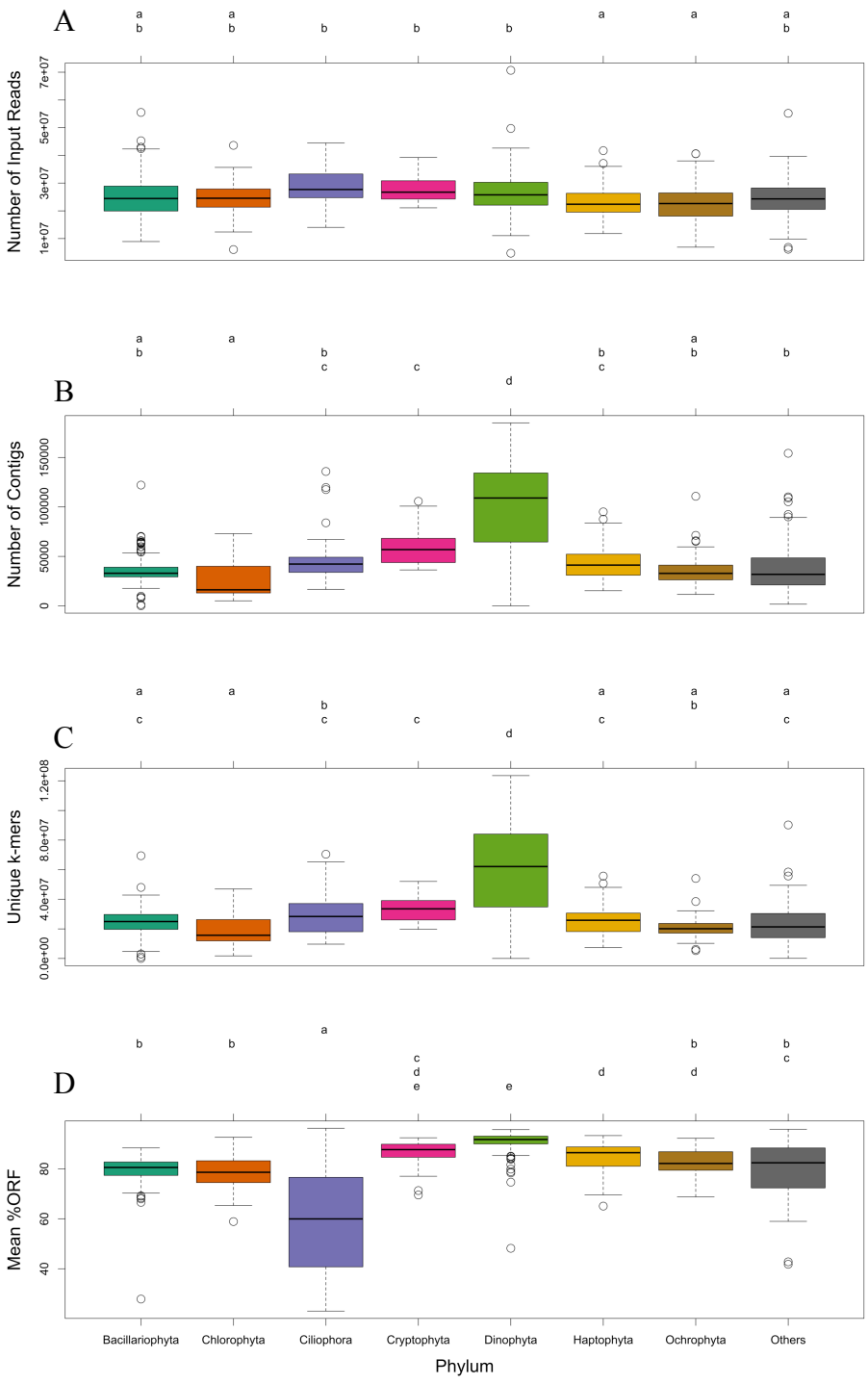


Figure 8. Box-and-whisker plots for the seven most common phyla in the MMETSP dataset. Clockwise from the top left (A) number of input reads, (B) number of contigs in the assembly, (C) unique *k*-mers ($k = 25$) in the assembly, (D) mean percentage open reading frame (ORF). Groups sharing a letter in the top margin were compared using Tukey's HSD post-hoc range test of multiple pairwise comparisons was used in conjunction with an ANOVA.

fix sentence structure.

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