Quality Metric	Higher in NCGR	Higher in DIB
Transrate score	52	575
Mean ORF %	45	606
Percentage of references with CRBB	106	545
Number of contigs	107	544

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

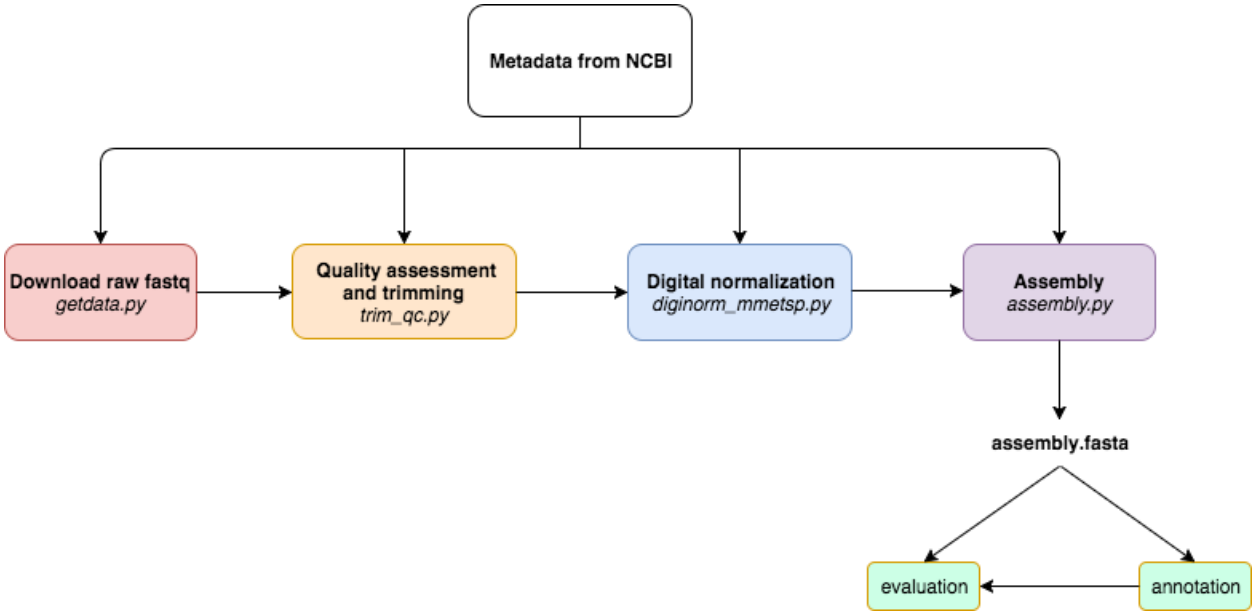

Figure 1. A modularized *de novo* transcriptome assembly pipeline workflow where tools used for each step can be substituted such that output assembly files can be used to test the effects of the individual tools on the overall pipelines. Metadata in the SraRunInfo.csv file downloaded from NCBI was used as input to the pipeline. The steps of the pipeline are as follows: 1) download data with the fastq-dump script in the SRA Toolkit [25], 2) quality assessment with FastQC (Andrews 2010) and trimming residual adapters and low quality bases (Q<2) with Trimmomatic [28], 3) digital normalization with khmer version 2.0 [29], 4) assembly with Trinity [8]. Each script in the pipeline uses the metadata (SraRunInfo.csv) file obtained from the SRA as input.

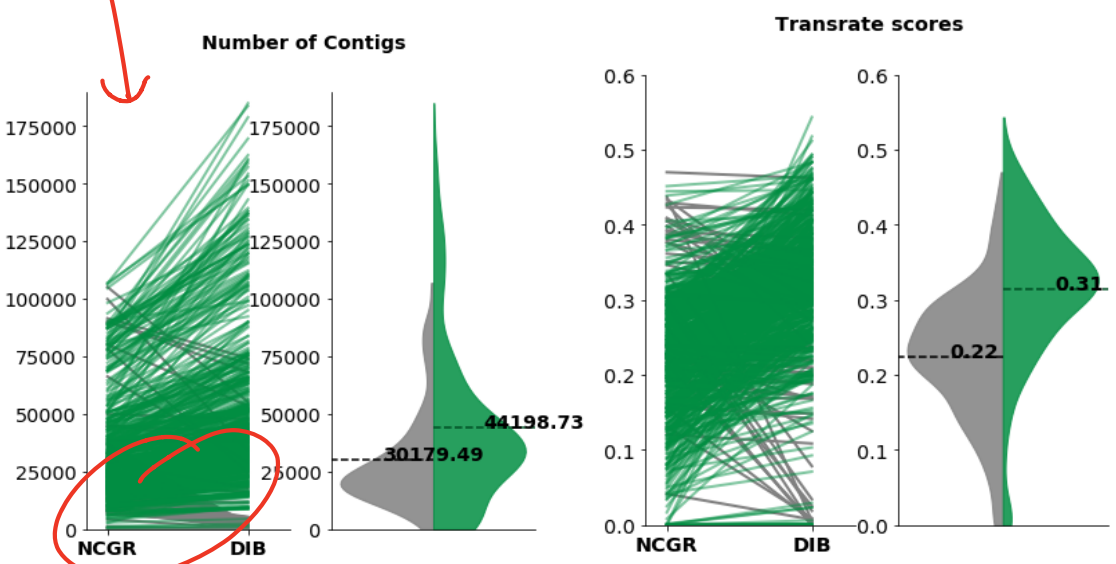

Figure 2. Split violin plots of the number of assembled contigs (left) and Transrate quality scores (right) of each pipeline. In the green (right side of each plot) are the DIB re-assemblies and in gray (left side of each plot) are the original assemblies from NCGR.

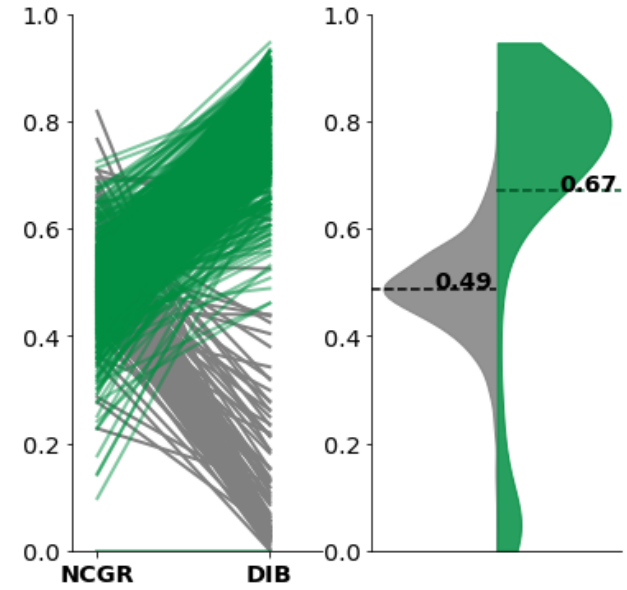

Figure 3. Split violin plot of the proportion of references with a conditional reciprocal best BLAST (CRBB) hit. The distribution plot on the left (grey) contains data were NCGR assemblies were aligned against the DIB assemblies as reference and on the left (green), the DIB assemblies were aligned against the NCGR assembly as reference.

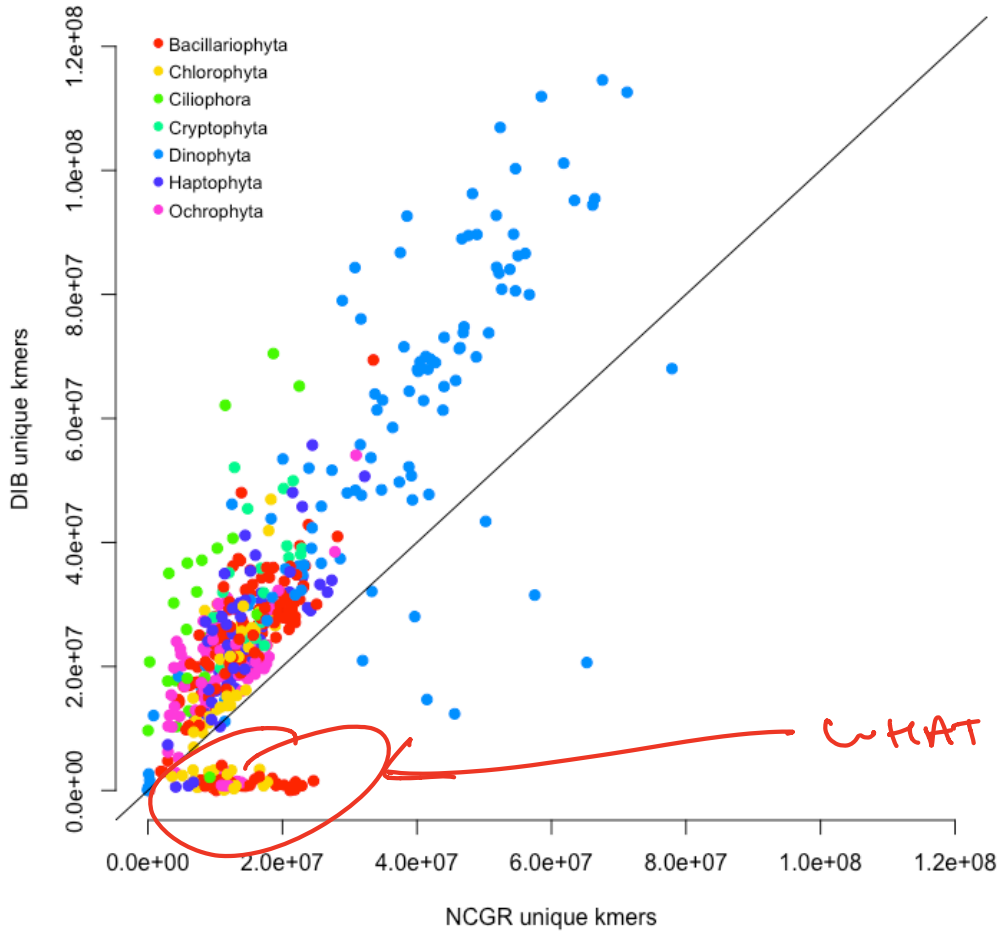

Figure 4. Unique *k*-mers (*k*=25) from the DIB re-assembly compared NCGR assembly for each of the 678 samples. The line ~~indicates a 1:1 relationship between DIB re-assemblies and NCGR assemblies.~~ There were 536 samples that had higher unique *k*-mer values in the DIB re-assemblies than in the NCGR assemblies whereas 99 of the samples had higher unique *k*-mer content in the original NCGR assemblies.

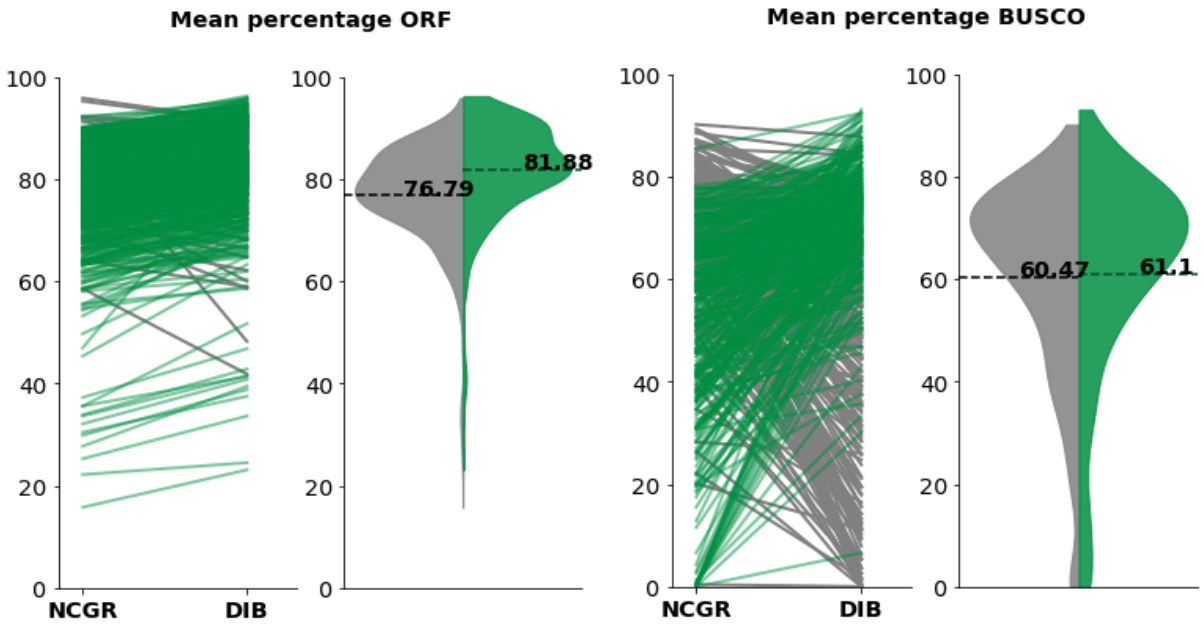

Figure 5. Split violin plots of the percentage of contigs with predicted open reading frame (ORF) in each assembly (left) and the percentage of complete protistan benchmarking universal single-copy orthologs (BUSCO) recovered in each assembly (right). In the green (right side of each plot) are the "DIB" re-assemblies and in gray (left side of each plot) are the original assemblies from NCGR.

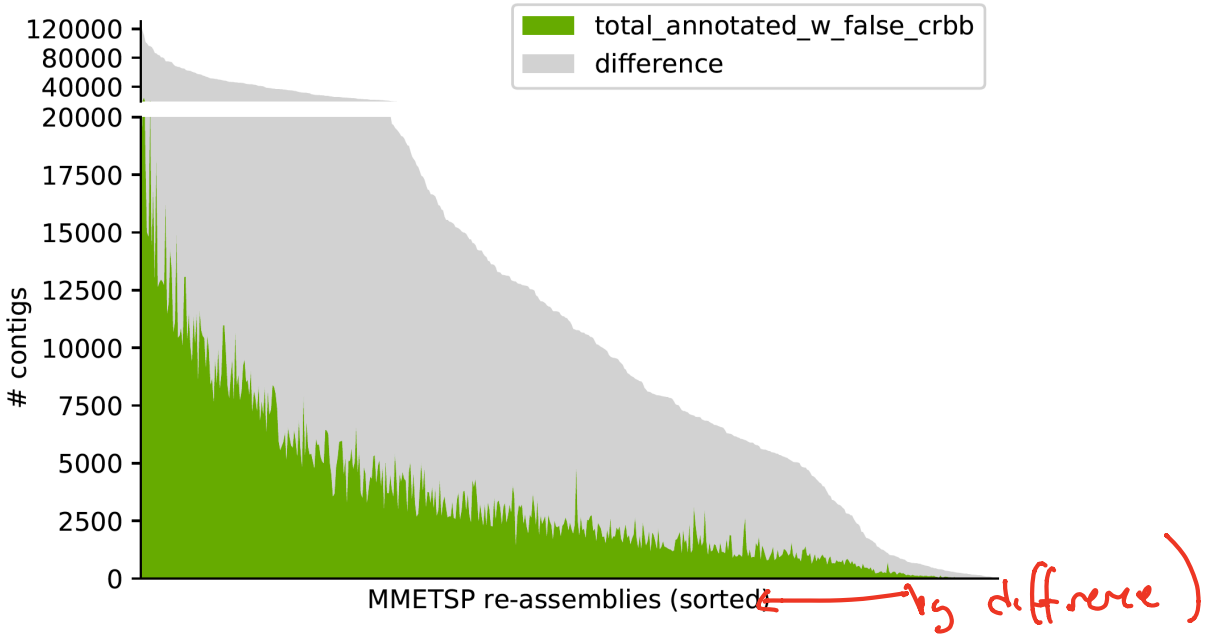

Figure 6. A histogram across assembled samples depicting the number of contigs identified as present in DIB assembly but absent in the NCGR assembly as measured with a "false" CRBB result. Samples are sorted from highest to lowest with the number of contigs. Highlighted in green are the contigs that could be annotated with a known gene name from the Pfam, Rfam, or OrthoDB databases. The region in gray indicates the number of contigs that are present in the DIB re-assemblies, absent from NCGR but with no annotation result. The region in green represents the number of contigs that are unique to the DIB re-assemblies that did have a positive annotation result.
have an

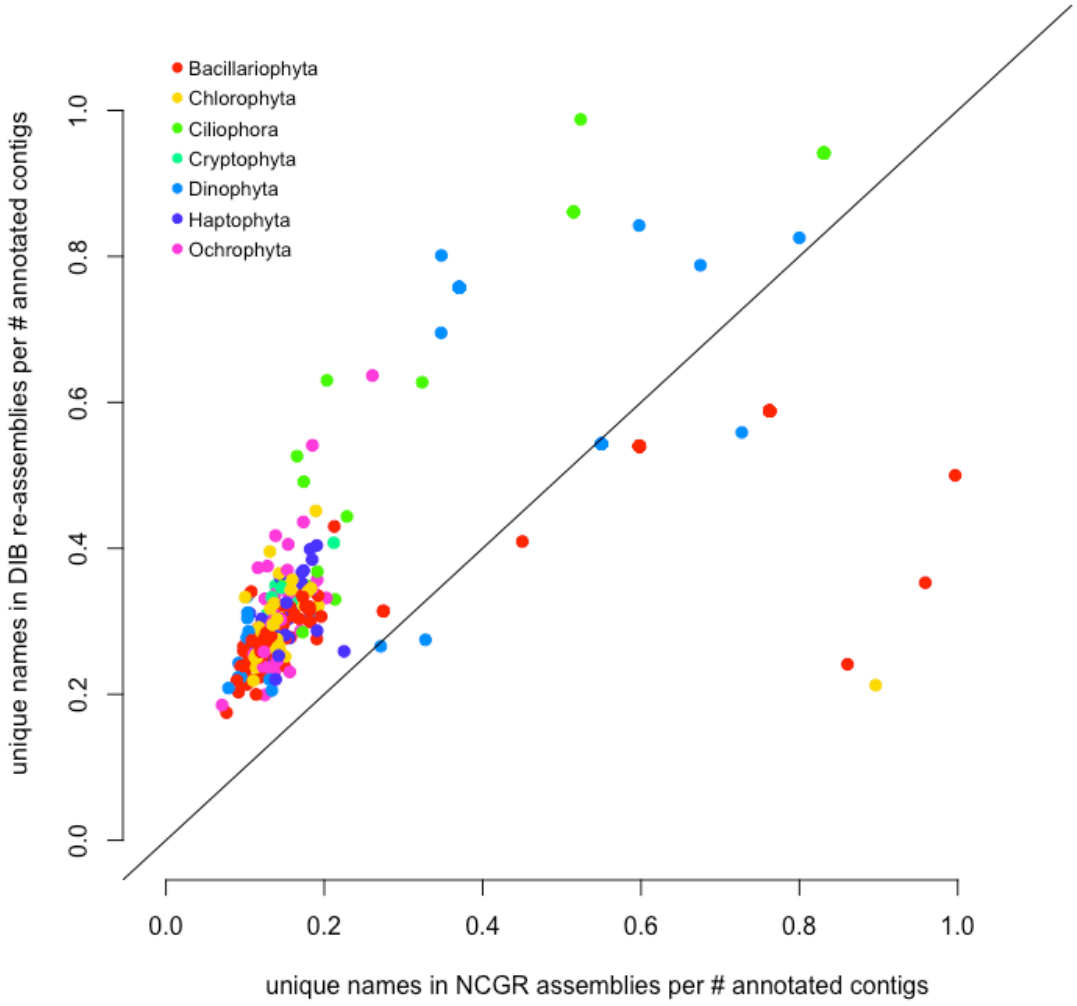

Figure 7. Unique gene names found in either NCGR or DIB assemblies but not found in the other assembly. The line indicates a 1:1 relationship between the number of unique gene names in DIB and NCGR. Numbers of unique gene names are normalized to the number of annotated contigs in each assembly. DIB assemblies had the highest number of unique names not found in NCGR assemblies. Several NCGR assemblies had gene names not found DIB assemblies. Most MMETSP samples had gene names that were unique in both NCGR and DIB assemblies.

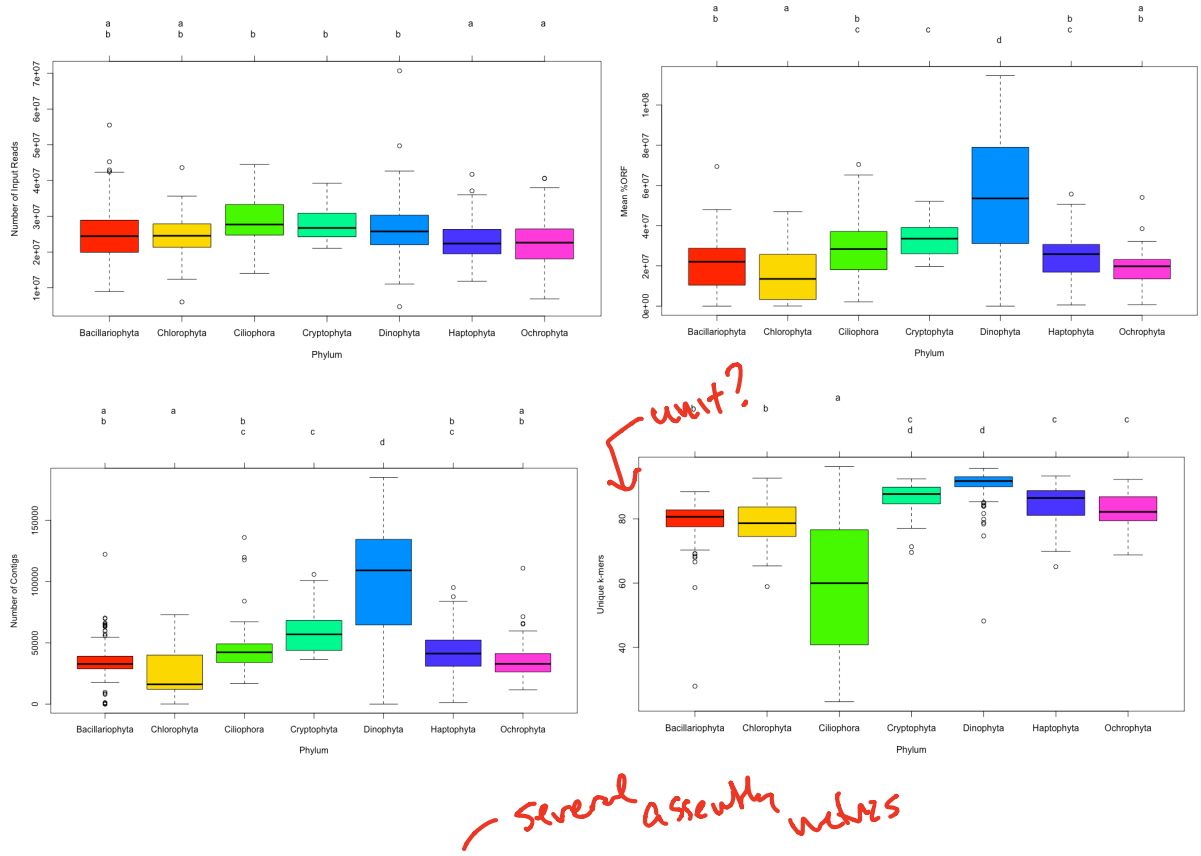

Figure 8. Box-and-whisker plots for the seven most common phyla in the MMETSP dataset. Clockwise from the top left (A) number of input reads, (B) mean percentage open reading frame (ORF) content, (C) number of contigs in the assembly, (D) unique *k*-mers (*k*=25) in the assembly. Groups sharing a letter 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.