Trimming and decontamination of metagenomic data can significantly impact assembly and binning metrics, phylogenomic and functional analysis

Jason Whitham (jmwhitha@ncsu.edu)
North Carolina State University https://orcid.org/0000-0003-2623-6292
Amy M. Grunden
North Carolina State University

Research

Keywords: metrics, placement of metagenome, multiple marker, phenotypes

Posted Date: May 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-539358/v1

Abstract

Background

Investigators using metagenomic sequencing to study their microbiomes are often provided data that has been trimmed and decontaminated or do it themselves without knowing the effect these procedures can have on their downstream analyses. Here we evaluated the impact that JGI trimming and decontamination procedures had on assembly and binning metrics, placement of metagenome assembled genomes into species trees, and functional profiles of metagenome-assembled genomes (MAGs) extracted from twenty three complex rhizosphere metagenomes. We also investigated how more aggressive trimming impacts these binning metrics.

Results

We found that JGI trimmed and decontamination of input reads had some significant impacts in assembly and binning metrics compared to raw reads, and that differences in placement of MAGs in species trees increased with decreasing completeness and contamination thresholds. More aggressive trimming beyond those used by JGI were found to reduce MAG counts.

Conclusions

Mild trimming and decontamination of metagenomics reads prior to assembly can change an investigator's answer to the questions, "Who is there and what are they doing? However, mild trimming and decontamination of metagenomic reads with high quality scores is recommended for those who elect to do so.

Introduction

Metagenomics is the study of communities of microorganisms. Investigators in this field often ask the questions, "Who is there, and what are they doing?" They answer these questions by comparing microorganisms' genomes from one environment to ones found in other environments, in space and time. Differences in sequences of homologous universal genetic markers are used to construct phylogenetic (one marker) or phylogenomic (multiple marker) trees, enabling inference of identities, phenotypes, and functions of microorganisms based on those of neighboring organisms. Functional potential of community members is further evaluated by comparing non-universal genes to homologous genes of characterized enzymes, their domains, and the metabolic pathways they belong to. Leading up to these analyses, is a pipeline of trimming and decontamination, assembly, and binning steps.

Assemblies are long fragments of DNA, contigs, constructed from smaller fragments, known as reads, which are commonly acquired by shotgun sequencing of the DNA found in environmental samples. Since

assemblies are made of contigs, some key metrics for evaluating assembly quality are total length of all assembly contigs, total number of assembly contigs, the number of bases of the largest contig in an assembly, and GC content [1]. When comparing metagenome assemblies generated with the same datasets, algorithm and parameters, assemblies with longer total lengths are generally better than ones with short total lengths. Conversely, total length can be inflated with misassemblies, and misassembly frequency can vary with different assemblers and datasets [2]. Having the largest contig relative to other assemblies is typically a favorable quality since longer contigs are good starting points for generating gap-free MAGs. Unlike the first two metrics mentioned, interpreting a high number for total contigs is a little less straightforward. An assembly with more total contigs relative to other assemblies could mean that it is larger, but it could also mean that it is more fragmented. So, the total contigs metric should be interpreted in the context metrics describing the size of contigs, such as the quantity of 1k, 10k, 100k bp contigs. GC content can be useful for simple assembly sequence comparisons.

Binning contigs into MAGs is a common step in characterizing metagenomes. Depending on the study, investigators might be interested in generating as many MAGs as possible or MAGs of a certain quality threshold in accordance with accepted recommendations [3]. The quality of MAGs is often defined by their completeness and contamination; each of which can be calculated by presence and absence of lineage-specific single-copy marker genes [4]. An important consideration when evaluating MAGs with this method is estimation bias. The real completeness is increasingly overestimated, and real contamination is increasingly underestimated with lower completeness and higher contamination estimates. The estimation bias is due to contaminant markers being counted as completeness markers. However, this bias is less than 3% when MAGs are estimated to be greater than 70% complete and less than or equal to 10% contaminated [4]. In other words, at these estimate limits, the MAGs might actually be 67% complete and 13% contaminated. This level of completeness is often good enough for investigators answering the questions "Who's there, and what are they doing?" Investigators considering quality in their MAG selection often even use recommended medium quality thresholds; completeness of greater than or equal to 50% and less than 10% contamination [3]. In comparison to the low level of bias found with 70% completeness and 10% contamination, one can expect as much as 8% bias with medium quality thresholds, resulting in some MAGs with 42% completeness and 18% contamination.

Prior to assembly and binning, investigators often perform trimming and decontamination of reads with the purpose of removing low quality bases, presumed errors generated in the sequencing process, contaminants like adaptors used in sequencing library preparation, and foreign DNA introduced during DNA extraction and library construction. The conventional wisdom is that trimming and decontamination of metagenomes are important to do before assembly and binning. The United States Department of Energy (DOE) Joint Genome Institute (JGI) follows this workflow and many bioinformatic forums like BioStars and SeqAnswers advocate for some level of trimming and decontamination, offering parameter suggestions for commonly used applications. However, we have not seen any published research that addresses the consequences of trimming and decontamination on assembly and binning metrics.

In this study, we evaluated how trimming and decontamination impacted total assembly length, total number of assembly contigs, largest contig length, the number of 1k, 10k, 100k bp contigs, GC content, total counts of bins, their completeness and contamination, those of "good" and "medium quality" bins. We did this with select data from a large sequencing effort completed by JGI in 2018 to enable investigators to identify beneficial ecofunctional genes selected by rhizospheres of three important biofuel crops - corn, switchgrass, and *Miscanthus xgiganteus* [5]. Rhizosphere samples were taken at seven plot areas at the Kellogg Biological Station (KBS) in Michigan during five seasons of crop establishment. Some of the data generated early on was used to create a pipeline for faster identification of short-subunit rRNA gene fragments and enabled unsupervised operational taxonomic unit community analysis with shotgun data [6]. Shotgun sequencing was compared to amplicon sequencing, demonstrating bias in the later for specific community members. More recently, the data was repurposed and combined with data from other sequencing efforts to evaluate the prevalence and activity of trace gas oxidizers of soil microbial communities [7]. In line with our research interests, we focused on metagenomes from *Miscanthus xgiganteus*, selecting 23 metagenomes that spanned all sampling seasons and plots and clustered well in principal component analysis (PCA) by taxon.

Researchers and students alike are increasingly interested in metagenomics. So, we made our analysis highly accessible to active and new metagenomics investigators by using KBase [8], an open science data platform funded by the DOE. This platform has a variety of command-line bioinformatics applications that have been simplified into modules that can be arranged into Jupyter Notebook [9] narratives. This format removes many of the learning curves associated with working in the terminal and provides visual displays for data and analyses, making research and findings more interoperable, reproducible, and repurposable. While the platform is not perfect, it is under heavy development, with increasing modular applications, and a funded help desk, with staff that are quick to respond to bugs and inquires. In addition to research, it is also being used to teach students in the classroom and asynchronously online [10], especially since the beginning of the coronavirus pandemic.

Methods

Trimming and Decontamination

Select raw fastq files ("raw") *Miscanthus xgiganteus* rhizosphere metagenomes from the JGI Project "Metagenomic analysis of the rhizosphere of three biofuel crops at the KBS intensive site" (Proposal ID: 1296) [5] were imported from JGI's Integrated Microbial Genomes and Microbiomes (IMG/M) website [11–12]. Selection of metagenomes was based on clustering in PCA plots by taxon, accomplished with the genome clustering tool on the JGI IMG/M website. Metadata for these metagenomes is provided in Table 1.

JGI trimmed and decontaminated reads (referred to as "qc" in this study) were imported directly into KBase for subsequent assembly and binning. JGI reports indicated that raw fastq file used in this study contained contaminant reads including PhiX (< 0.3%), DNA Spike-ins (< 3%), *E. coli* (< 0.7%), Mitocondria

(< 4%), Chloroplast (< 4%), and rRNA (< 3%). The following parameters for trimming and decontamination were used to generate qc libraries from raw reads. BBDuk [13] adapter trimming (mink = 11, k = 23, hdist = 1, hdist2 = 1, ktrim = r, tpe, tbo, minlen = 40, minlenfraction = 0.6, ftm = 5) was used to remove known Illumina adapters. The reads were then processed using BBDuk filtering and trimming (maq = 8, maxns = 1, minlen = 40, minlenfraction = 0.6, k = 27, hdist = 1, trimq = 12, qtrim = rl). At this stage reads ends were trimmed where quality values were less than 12. Read pairs containing more than three 'N', or with quality scores (before trimming) averaging less than 3 over the read, or length under 51bp after trimming, as well as reads matching illumina artifact, spike-ins and phiX were discarded. Remaining reads were mapped to a masked version of human HG19 with BBMap (fast local minratio = 0.84 maxindel = 6 tipsearch = 4 bw = 18 bwr = 0.18 usemodulo printunmappedcount idtag minhits = 1), discarding all hits over 93% identity.

Additional trimming and decontamination were performed on six randomly chosen raw fastq files with BBDuk (version 37.62) on Henry2, the high-performance computing cluster at North Carolina State University, because BBDuk was not available on KBase at the time of this study. Six files were processed fourteen different ways with varying aggressiveness. Parameters used for each step were derived from author recommendations provided on the JGI website [14] and bioinformatics forums, SeqAnswers [15] and BioStars [16–17]. BBDuk trimmed adaptors of paired 151 bp raw reads; targeted kmers sized between 11 (mink = 11) and 23 (k = 23) and ones with a single substitution (hdist = 1), all bases to the right of 3' adapters (ktrim = r), and paired reads to the same length in events where the adapter kmer was only detected in one of the paired reads (tpe). Paired overlapping reads were also merged (tpo). Reads that were quality trimmed were either trimmed within the recommend Q8-Q12 range (trimq = 10) or to Q20 (trimq = 20) from the left and right (qtrim = rl); reads shorter than 100 base pair were discarded (minlen = 100). Illumina artifacts, spike-ins and phiX were also discarded from libraries that were decontaminated. Force trimming was also performed on some libraries since the last two (ftr = 149) and first six (ftl = 7) were found to be low quality. Six files were chosen for additional trimming and decontamination in accordance with the minimum five to six groups recommended for mixed effects models [18].

Table 1
Data sources for this study. Starred(*) fastq files were randomly selected for BBDuk trimming and decontamination.

IMG/M GENOME #	Raw Fastq File	M-Tag (Plot-Season)
3300025938	9672.8.140931.GTAGAG.fastq.gz	M1-2
3300025940	9117.4.122649.GTAGAG.fastq.gz*	M1-3
3300025899	9041.8.119154.GTCCGC.fastq.gz	M2-2
3300025960	9117.5.122651.GTCCGC.fastq.gz	M2-3
3300025934	10158.6.150237.GTGAAA.fastq.gz*	M2-4
3300013296	11260.5.198617.GTGAAA.fastq.gz	M2-5
3300026089	9053.2.119381.GTGAAA.fastq.gz	M3-2
3300025937	9117.6.122653.GTGAAA.fastq.gz	M3-3
3300025935	10186.3.150267.ACAGTG.fastq.gz	M3-4
3300013308	11263.1.198788.GTGGCC.fastq.gz	M3-5
3300025901	7333.1.69391.CGTACG.fastq.gz	M4
3300026023	9053.3.119383.GTGGCC.fastq.gz	M4-2
3300025926	9117.7.122655.GTGGCC.fastq.gz*	M4-3
3300014969	11306.3.200370.TGACTGA-GTCAGTC.fastq.gz*	M4-5
3300025942	9053.4.119385.GTTTCG.fastq.gz	M5-2
3300025907	9117.8.122657.GTTTCG.fastq.gz*	M5-3
3300025927	10186.4.150269.GCCAAT.fastq.gz	M5-4
3300014745	11306.1.200366.CCTTCCT-AAGGAAG.fastq.gz	M5-5
3300025908	9053.5.119387.CGTACG.fastq.gz	M6-2
3300025893	9108.1.122552.CGTACG.fastq.gz	M6-3
3300011119	10158.8.150241.GTTTCG.fastq.gz	M6-4
3300013297	11260.6.198619.GTTTCG.fastq.gz	M6-5
3300026121	9108.2.122554.GAGTGG.fastq.gz*	M7-3

Assembly and Binning

Assembly and binning were performed in KBase. Currently, KBase has the leading metagenome assemblers, MEGAHIT [19], HipMER [20], SPAdes [21], and IDBA-UD [22] for assembling metagenomes.

We used MEGAHIT in this study because it was the fastest, required the least amount of RAM, and had the longest assemblies [23]. KBase also currently has two of the leading binning tools, MetaBAT2 [24] and MaxBin2 [25]. Of these, we chose MetaBAT2 to bin all of our assemblies because it recently outperformed MaxBin2, generating more bins from the high-complexity Critical Assessment of Metagenome Interpretation dataset [26]. We binned some assemblies though with MaxBin2 to demonstrate similar results regardless of binning algorithm or min contig length.

Reads were assembled using MEGAHIT v1.2.9 with the meta-large parameter preset and a minimum contig length of 1000 bp. Assembly quality metrics including number of contigs, length of contigs, longest contig, and the number of contigs with sizes greater than 1k, 10k, and 100k bp were assessed with QUAST v4.4. MetaBAT2 v1.7 was used to bin minimum length contigs of 2500 bp. This minimum length was chosen because the MetBAT2 application occasionally failed when specifying shorter minimum contig lengths. Quality metrics including number of bins, bin completeness, bin contamination, single-copy and multi-copy markers were evaluated with CheckM v1.0.18.

Phylogenomic and Functional Analyses

Good quality MAGs that were greater than 70% complete and less than or equal to 10% contaminated were extracted from select raw and qc libraries as assembly sets with the KBase "Extract Bins as Assemblies from BinnedContigs" utility v.1.0.2. Assembly sets were annotated with RASTtk [27] v1.073. Annotated genome sets were inserted into species trees based on multiple universal phylogenetic markers with the KBase Species Tree application v2.2.0 [28]. Refseq (GCF) IDs were removed to condense tree figures, and the final trees visualized in ETE toolkit [29].

Functional assessment including COG [30], Pfam [31], and TIGRFAM [32] annotations of MAG genes was performed with the "View Function Profile for Genomes" application v1.4.0 MAGs and "Annotate Domains in a GenomeSet" utility v1.4.0. Defaults were used except that the lower limit threshold for TIGRFAM annotations was decreased from 10–9% of genes for the 9053.2 (M3-2) medium quality raw and qc genomesets to accommodate one pair of poorly annotated MAGs. Categories common to genomesets were statistically compared. Less than 0.2% of genes were not included in the analysis because they were absent in some genomesets.

Statistical Analyses, Data and Code Availability

Trimming, decontamination, assembly, binning, and functional analysis data used in this study was saved to Excel files. Python was used to combine Excel worksheets into Pandas dataframes, converted to lists, and then copy and pasted into KBase for statistical analysis in narrative code cells. Python coding was used for statistical analyses with the exception of calculating effect sizes, which were performed with the R package effsize [33]. The Excel files, data wrangling code, effect size and power calculation code are hosted at GitHub [34], and the code for all other statistical analyses are in the narratives [35–36] for complete transparency and reproducibility.

Paired Student's t-test and the nonparametric Wilcoxon signed-rank test were used to determine if there were significant differences in the means and mean signed-ranks of average and median assembly lengths, total contigs, the largest contig length, GC%, and the number contigs greater than or equal to 1k, 10k, 100k bp of raw and qc assemblies ($\alpha = 0.05$). These tests were also used for binning metrics- MAG completeness, contamination, single-copy and multi-copy marker counts, and total MAG counts. Ordinary least squares and linear mixed effects models were used to determine significant correlations in read and base counts of raw and trimmed and decontaminated reads with MAG counts, completeness and contamination.

Results

JGI trimming and decontamination methods significantly impacted assembly metrics, reducing total contigs counts, counts of 10k + contigs, and total contig length

Trimming and decontamination can impact the quality and quantity of MAGs if assembled contigs used for binning are substantially altered in length, count, or sequence. JGI trimming and decontamination removed 0.6–7.7% of reads in fastq files selected for this study due to detected artifacts or low quality. Performing statistical tests on the assembly metrics, we did not detect significant differences in average GC% or average counts of contigs greater than 100k bp in length. QC assemblies had fewer total contigs ($p = 1.227e\text{-}05$, raw avg = 1,058,433 bp, qc avg = 1,031,110 bp, Cliff's delta = 0.0850), contigs greater than 10k bp in length ($p = 0.0053$, raw avg = 6129, qc avg = 6071, Hedges's g = -0.0225), and smaller total lengths ($p = 1.227e\text{-}05$, raw avg = 1,952,563,350 bp, qc avg = 1,929,639,607 bp, Cliff's delta = 0.0888) though than raw assemblies.

MAG counts of binned raw and qc assemblies were similar regardless of binning application or minimum contig length

Minimum contig length is a key parameter that significantly impacts binning metrics. The MetaBAT2 application on KBase allows users to select a minimum contig length as low as 1500 bp (default: 2500 bp). In some preliminary experiments, we found that the application failed more frequently when minimum contig lengths less than 2500 bp were used, which lead us to choose this default minimum contig length for this study.

MaxBin2 is another high-performance automated binning algorithm in KBase's suite of applications, and does not have a minimum contig length. We tested MaxBin2 with default 1000 bp and 2500 bp minimum contig length, and binned three randomly chosen raw and qc assembly pairs to see if there was reason for additional testing with alternative binning applications or minimum contig lengths. Figure 1 shows that total raw and qc MAG counts were not substantially different when a minimum contig length of 2500 bp was used. When 1000 bp minimum contig length was applied, application runtime increased by multiple days and the number of total MAGs substantially increased relative to MAGs generated with 2500 bp minimum contig length, but the total raw and qc MAG counts remained even with each other. Good quality raw and qc MAG counts were also close to each other when the same binning algorithm and

minimum contig lengths were applied. Since neither these factors contributed to differences in raw and qc MAG counts, we did not see a reason to perform additional testing with MaxBin2 or alternative contig lengths.

JGI trimming and decontamination methods did not significantly impact most binning metrics

The average total counts of MAGs, and the averages of the completeness means, single-copy marker count means, and multi-copy marker count means were not significantly different between binned raw and qc assemblies. Mean contaminations were on average 2.0% higher for binned raw assemblies compared to binned QC assemblies ($p = 0.0391$, Raw = 68.7%, QC = 66.6%, Hedges's $g = -0.0225$). Since means can be strongly influenced by outliers, and these distributions have outliers (Fig. 2), we also tested for significant differences in the averages of contamination medians. Averages of contamination medians were much lower than the contamination means, and were not significantly different at $\alpha = 0.05$ ($p = 0.4749$, Raw = 3.9%, QC = 3.8%). For the sake of thoroughness, averages of completeness medians, single-copy marker count medians and multi-copy marker count medians were found not to be significantly different at $\alpha = 0.05$. Finally, we tested whether trimming and decontamination impacted these binning metrics when only good or medium quality bins were considered. They did not.

Medium quality raw and qc MAGs paired less frequently in species trees than other quality MAGs

A major concern for the practicality of this study was whether trimming and decontamination of reads ultimately impacted the placement of MAGs in phylogenomic trees. Placement of MAGs is dependent upon differences in sequences of multiple universal genetic markers found in the binned fragments, and these markers might be altered by trimming or removed by decontamination. Alternatively, contaminant markers could also cause misplacement of MAGs. To test the impact of trimming and decontamination, we annotated medium and good quality MAGs generated from raw and qc assemblies and inserted them into species trees. Readsets from plots three and five and seasons two and four were chosen because binning of the assemblies yielded a similar number of medium and good quality MAGs.

Figure 3 shows that good quality raw and qc MAGs typically paired together. 9053.2 (M3-2) had 19 pairs out of a possible 21; two raw and three qc MAGs were discretely placed in the tree. 10186.3 (M3-4) had 15 pairs out of a possible 18; three raw and three qc MAGs were discretely placed in the tree. 9053.4 (M5-2) had 15 pairs out of a possible 17; three raw and two qc MAGs were discretely placed in the tree. 10186.4 (M5-4) tree had 14 pairs out of a possible 15; six raw and one qc MAGs were discretely placed in the tree. Thus, there was consistent placement of good quality MAGs in the phylogenomic tree regardless of whether the reads that MAGs were derived from were trimmed and decontaminated or not.

Medium quality raw and qc MAGs were also mostly paired in phylogenomic trees (Supplemental_Fig.1), but to a consistently lower percentage than good quality MAGs. Medium quality MAGs were paired 30/34 (88.2%), 23/28 (82.1%), 23/26 (88.5%), and 25/29 (86.2%) compared to 19/21 (90.5%), 15/18 (83.3%), 15/17 (88.2%), and 14/15 (93.3%) for M3-2, M3-4, M5-2, and M5-4 good quality MAGs. Instead of five to

seven discreetly placed MAGs as seen with the good quality MAGs, there were seven to 10 for medium quality MAGs.

Trends of higher pairing and fewer discreet placements emerged when MAGs with high quality completeness and contamination thresholds (> 90%, < 5%, respectively) were also compared. Pairing was 100% for all but 10186.3 (M3-4, 5/6 or 83.3%), and only zero to four MAGs were discretely placed.

Phylogenomically paired raw and qc MAGs have similar functional analyses

MAGs that are phylogenomically unrelated typically do not have similar functional profiles. However, we questioned whether raw and qc MAGs that paired in species trees based on a relatively short list of genes would have similar functional profiles. Functional analysis could be impacted by domain alterations or removal from trimming or decontamination, or binning of contaminant gene domains because they were not removed. As with the phylogenomic analysis, we considered raw and qc MAGs with high, good, and medium quality completeness and contamination thresholds. Less than 5% of domain namespaces of protein-coding genes common to good quality raw and qc MAGs were differentially present. One out of 24 COG, 12 out of 359 PFAMs, and four out of 95 TIGRFAM domains had a significantly different percentage of annotated protein-coding genes in raw compared to qc MAGs. While these percentage differences were statistically significant, these differences were practically minor (Table 2). Average percentages were higher in raw MAGs for seven of differentially represented domains and were higher in qc MAGs for ten of the differentially represented domains.

Even fewer domain namespaces were differentially present in raw and qc MAGs with medium and high quality thresholds compared to good quality MAGs. Just four out of 101 TIGRFAM domains, no COGs or PFAMs, were differentially present in medium quality raw and qc MAGs. One COG, one PFAM, and two TIGRFAMs were differentially present in raw and qc MAGs with high quality thresholds. Significant differences were practically minor though (Table 2).

Table 2
Averages and standard deviations of percentages of differentially represented annotated protein-coding genes

MAG Quality	Domain Namespace	Category	Category Description	Raw Avg %	Raw Std %	QC Avg %	QC Std %
High	COG	Q	Secondary metabolites biosynthesis, transport, and catabolism	1.65	0.7	1.59	0.69
Good	COG	R	General function prediction only	8.27	1.4	8.36	1.44
High	PFAM	CL0127	ClpP_crotonase	0.52	0.15	0.54	0.16
Good	PFAM	CL0037	Lysozyme	0.15	0.11	0.16	0.11
Good	PFAM	CL0040	tRNA_synt_II	0.47	0.24	0.46	0.24
Good	PFAM	CL0051	NTF2	0.4	0.22	0.41	0.23
Good	PFAM	CL0052	NTN	0.21	0.09	0.22	0.09
Good	PFAM	CL0071	His_phosphatase	0.1	0.08	0.11	0.09
Good	PFAM	CL0113	GT-B	0.95	0.34	0.97	0.35
Good	PFAM	CL0228	Acetyltransferase	0.17	0.1	0.18	0.1
Good	PFAM	CL0254	THDP-binding	0.54	0.2	0.53	0.2
Good	PFAM	CL0265	HIT	0.09	0.06	0.09	0.06
Good	PFAM	CL0331	EpsM	0.05	0.06	0.05	0.05
Good	PFAM	CL0380	IDO-like	0.02	0.02	0.02	0.02
Good	PFAM	CL0401	AsmA-like	0.12	0.1	0.11	0.09
High	TIGRFAM	role:11040	Amino Acid Biosynthesis Pyruvate Family	0.94	0.54	0.99	0.54
High	TIGRFAM	role:17060	Energy Metabolism Sugars	0.43	0.31	0.46	0.33
Good	TIGRFAM	role:13020	Fatty Acid and Phospholipid Metabolism Degradation	0.07	0.1	0.06	0.09
Good	TIGRFAM	role:17050	Energy Metabolism Pentose Phosphate Pathway	0.8	0.29	0.84	0.3
Good	TIGRFAM	role:17080	Energy Metabolism Other	0.55	0.36	0.58	0.37
Good	TIGRFAM	role:25510	Unknown Function General	3.5	0.57	3.43	0.55
Medium	TIGRFAM	role:11060	Amino Acid Biosynthesis Histidine Family	0.71	0.46	0.79	0.5

MAG Quality	Domain Namespace	Category	Category Description	Raw Avg %	Raw Std %	QC Avg %	QC Std %
Medium	TIGRFAM	role:13020	Fatty Acid and Phospholipid Metabolism Degradation	0.06	0.1	0.05	0.08
Medium	TIGRFAM	role:14050	Riboflavin, FMN, and FAD	0.53	0.27	0.58	0.31
Medium	TIGRFAM	role:17080	Energy Metabolism Other	0.59	0.42	0.64	0.5

Trimming and decontamination removed as much as tens of millions of reads and tens of billions of bases from read files

Eighty-four trimmed and/or decontaminated fastq were generated from raw fastq files to evaluate the effects of these methods on assembly and binning metrics. The creation of these eighty-four read files involved sequentially force trimming, kmer trimming, quality trimming, decontaminating, or some combination of these steps based on recommendations posted in online bioinformatics forums. The number and percentage of reads and bases removed are provided in the Supplemental Files. When used, force trimming did not impact reads but removed about five percent of bases. Kmer trimming removed about four percent or fewer reads and as much as five percent of bases. Decontamination of raw fastq files removed between zero and seven percent of reads and bases; less than three percent for files that were force, quality, and/or kmer trimmed prior to decontamination. Quality trimming to Q10 removed about two to four percent of reads and about two to four percent of bases, while quality trimming to Q20 removed about eight to 15% of reads and about nine to 16% of bases. Generalizations could not be made about which steps had the greatest impact on reads or bases with the exception that quality trimming to Q20 consistently removed the most of each. Total reads removed by all combinations of steps tested ranged from about zero to 16%. Similarly, total bases removed were one to 22%. The greatest change in reads and bases was from about 399M to 334M and from about 60.3B to 46.7B, respectively.

In addition to the read files generated, the raw and JGI processed reads were included in the subsequent analyses, making a total of 96 read files. These ranged from 245M to 399M reads, a span of 154M reads, and from 34.4 to 60.3B bases, a span of 25.9B.

Total MAG counts correlated with bases and reads

MAGs are binned contigs assembled from metagenomic reads. Therefore, it is no surprise that we found MAG counts were correlated with input reads (p_raw = 0.040, 1.664 MAGs/tMreads, 95% CI [0.080, 3.249], adjusted Pearson's r = 0.382), and their base counts (p_raw = 0.041, 1.099 MAGs/Bbases, 95% CI [0.338, 1.392], adjusted Pearson's r = 0.382). Read and base counts were also correlated with medium (0.648 medium MAGs/tMreads, 95% CI [0.129, 1.166], adjusted Pearson's r = 0.455; 0.428 medium MAGs/Bbases, 95% CI [0.085, 0.772], adjusted Pearson's r = 0.455) and good MAGs from raw reads (p_raw = 0.004, 0.529 good MAGs/tMreads, 95% CI [0.187, 0.872], adjusted Pearson's r = 0.545; p_raw =

0.004, 0.350 good MAGs/Bbases, 95% CI [0.123, 0.577], adjusted Pearson's r = 0.544). We wanted to know though if reduction of reads and bases due to trimming and decontamination also reduced MAG counts.

Since trimmed and decontaminated reads were observations dependent upon the original raw files, we used mixed linear effects models to avoid violating the ordinary least squares model assumption that observations are independent [37]. We found that MAG counts were correlated with read and base counts of trimmed and decontaminated reads (p_trim_decon = 0.000, 2.095 MAGs/tMreads, 95% CI [1.435, 2.755]; p_trim_decon = 0.000, 1.320 MAGs/Bbases 95% CI [1.018, 1.622]). Read and base counts of trimmed and decontaminated reads were also correlated with medium MAGs (p_trim_decon = 0.000, 0.883 medium MAGs/tMreads, 95% CI [0.492, 1.275]; p_trim_decon = 0.000, 0.610 medium MAGs/Bbases, 95% CI [0.423, 0.797]) and good MAGs (p_trim_decon = 0.003, 0.399 good MAGs/tMreads, 95% CI [0.160, 0.638]; p_trim_decon = 0.000, 0.309 good MAGs/Bbases, 95% CI [0.196, 0.421]). No significant correlations were found with average MAG completeness or contamination and read or base counts for raw or trimmed and decontaminated reads.

Discussion

In this study, we demonstrated that JGI trimming and decontamination procedures had little impact on the quantity or quality of MAGs from complex rhizosphere metagenomes, or the functional profiling of raw and qc MAGs that were phylogenomically paired (Table 2). However, we did observe that the number of raw and qc MAGs discretely placed in species trees increased from zero to four MAGs to five to seven MAGs to seven to 10 MAGs as quality thresholds for completeness and contamination were decreased from high to good to medium quality (Fig. 3, Supplemental_Fig.1). Phylogenomic differences of MAGs may be explained by differences in binning and assembly metrics including the 2.0% lower average contamination of qc MAGs compared to raw MAGs, and significantly higher total contig counts, contigs greater than 10k bp in length, and larger total lengths of raw assemblies compared to qc assemblies. Since choosing JGI trimmed and decontaminated or raw reads means reporting and depositing a similar quantity and quality of MAGs, some with phylogenomic differences, researchers may choose to assemble each and retain the union of discreet and paired MAGs to increase the total number in their analysis and avoid missing functionally important community members. We believe our methods were appropriate for the questions we were asking, but there are other ways of analyzing the data. To illustrate this point, consider that binning of a single assembly generates multiple MAGs, 77 MAGs for example. Binning multiple assemblies generates multiple MAG counts (e.g. assembly01 = 77 MAGs, assembly02 = 66 MAGs, assembly03 = 91 MAGs, ... assembly24 = 73 MAGs). So, a distribution of MAG counts can be generated from a set of assemblies, which can subsequently be compared to another distribution of MAG counts from an alternative assembly set (e.g. raw vs qc assembly MAG counts). However, each MAG has its own completeness percentage (e.g. bin001 = 14.9%, bin002 = 8.7%, bin003 = 93.4%, ... bin077 = 23.1%), contamination percentage, and counts of single-copy and multi-copy markers, used to calculate the completeness and contamination percentages. Since each assembly has multiple MAGs, each assembly set contains multiple distributions for these other metrics. To evaluate differences in binning metrics besides MAG counts, we elected to average MAGs single-copy marker counts, multi-copy marker counts,

completeness scores, and contamination scores for each assembly. Distributions used for statistical testing were therefore average values. The consequence of this is that we tested differences in the averages of averages. A possible alternative method could be to make distributions by combining all values for each MAG metric for all assemblies generated with the same trimming and decontamination procedure, disregarding intuitive assembly-level groupings. We believe our method is more relevant to the researcher who wants to know if their assembly, when binned, is going to have better or worse binning metrics than if they used an assembly prepared a different way (raw vs qc).

We failed to reject the null hypotheses that there were no significant differences in several key binning metrics for assemblies that were JGI trimmed and decontaminated compared to raw assemblies. These include total counts of MAGs, and completeness averages, single-copy marker count averages, and multi-copy marker count averages of assembly MAGs. However, our study was unpowered, comparing 23 assembly pairs. It is expected that differences in these metrics could be found significant given a much higher sample size. Based on the small effect sizes of less than 0.1 found for the significant difference in average contamination, it is also expected though that significance would have a small practical effect. We calculate that a powered study (power = 0.8) would need a sample size of greater than 824 assembly pairs (801 more pairs than what we used) for an effect size less than or equal to 0.1 and $\alpha = 0.05$. Then again, an effect size may be greater for low quality data, and some JGI datasets are worse quality than the ones used in this study. Therefore, in addition to sample size, future studies should consider using average Q scores as a factor or filter in experiment designs.

We also found that more aggressive trimming reduced MAG counts, including good quality MAGs, with small to medium effect. While JGIs methods of trimming and decontamination removed between 0.6–7.7% of reads in the fastq files, we removed as much as 16% of reads. Parameters that were overly aggressive included quality trimming to Q20 and discarding reads that were trimmed to less than 100 bp. More mild parameters such as trimming to Q8 - Q12 and discarding reads that are less than 40 bp are recommended for those who elect to trim their reads to avoid loss of MAGs.

Conclusions

Mild trimming and decontamination of metagenomics reads can change the way an investigator answers the questions "Who is there and what are they doing?" This is because some MAGs assembled with JGI trimming and decontamination are phylogenomically distinct from ones assembled with raw reads. Phylogenomics informs investigators of MAG identities and functions through relatedness to other organisms, and phylogenomically distinct microbes also have differing COG, PFAM, and TIGRFAM functional profiles. Since the number of MAGs discretely placed in species trees increases with inclusion of MAGs with lower qualities, the discrepancy will be more substantial with medium quality MAGs compared to high quality MAGs. While mild JGI trimming and decontamination can impact MAG identities and functions, it does not appear to impact how many are assembled. However, aggressive trimming should be avoided for this reason.

Abbreviations

IMG/M
Integrated Microbial Genomes and Microbiomes
DOE
United States Department of Energy
JGI
Joint Genome Institute
KBS
Kellogg Biological Station
MAGs
metagenome assembled genomes
PCA
principal component analysis
qc
JGI trimmed and decontaminated fastq files or reads
raw
raw fastq files or reads

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

All data and code generated and analyzed during this study are included in this published article, JGI IMG/M (Proposal ID: 1296, [5]), in the KBase narratives [35-36], and in the GitHub repository [34].

Competing interests

The authors declare that they have no competing interests.

Funding

Funding was provided by the United States Department of Energy, Award No. DE-EE0008523.

Authors' contributions

JMW was responsible for experimental design, data acquisition, wrangling, statistical analyses, creating figures and tables, depositing code and generated data into repositories, and drafted the manuscript.

AMG contributed to manuscript edits.

Acknowledgements

We acknowledge the computing resources provided on Henry2, a high-performance computing cluster operated by North Carolina State University, and acknowledge Lisa L. Lowe for her assistance with adding software packages to Henry2, which was provided through the Office of Information Technology High Performance Computing services at NC State University.

References

1. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies, Bioinformatics. 2013;29:1072-1075. doi:10.1093/bioinformatics/btt086.
2. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics. 2016;32:1088-90. doi:10.1093/bioinformatics/btv697.
3. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TB, *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nature biotechnology. 2017;35:725-31. doi:10.1038/nbt.3893.
4. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043-1055. doi:10.1101/gr.186072.114.
5. Tiedje JM. Metagenomic analysis of the rhizosphere of three biofuel crops at the KBS intensive site. United States: N. p. 2013. doi:10.25585/1488010.
6. Guo J, Cole JR, Zhang Q, Brown CT, Tiedje JM. Microbial community analysis with ribosomal gene fragments from shotgun metagenomes. Appl. Environ. Microbiol. 2016;82:157-166.
7. Bay SK, Dong X, Bradley JA, Leung PM, Grinter R, Jirapanjawat T, *et al.*. Trace gas oxidizers are widespread and active members of soil microbial communities. Nat. Microbiology. 2021:1-11.
8. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, *et al.*. KBase: The United States Department of Energy Systems Biology Knowledgebase. Nat. Biotechnology. 2018;36:566. doi:10.1038/nbt.4163.
9. Kluyver T, Ragan-Kelley B, Pérez F, Granger BE, Bussonnier M, Frederic J, *et al.*. Jupyter Notebooks - a publishing format for reproducible computational workflows. ELPUB. 2016.
10. Dow E, Wood-Charlson E, Biller S, Paustian T, Schimer A, Sheik C, Whitham J, Krebs R, Goller C, Allen B, Crockett Z, and Arkin A. Bioinformatic teaching resources - for educators, by educators - using KBase, a free, user-friendly, open source platform. United States: N. p., 2021. Web. doi:10.25982/90997.49/1783189

11. Chen IM, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, *et al.*. The IMG/M data management and analysis system v. 6.0: new tools and advanced capabilities. Nucleic Acids Res. 2021;49:D751-63. doi.org/10.1093/nar/gkaa939.
12. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Sundaramurthi JC, Lee J, *et al.*. Genomes OnLine Database (GOLD) v. 8: overview and updates. Nucleic Acids Res. 2021;49:D723-33. doi:10.1093/nar/gkaa983.
13. Bushnell B: BBTools Software Package. 2017. http://sourceforge.net/projects/bbmap. Accessed 15 Oct 2020.
14. BBDuk Guide. https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/. Accessed 15 Oct 2020.
15. SeqAnswers BBDuk. http://seqanswers.com/forums/showthread.php?t=96593&goto=nextnewest. Accessed 15 Oct 2020.
16. BioStars BBDuk 1. https://www.biostars.org/p/237714/#237745. Accessed 15 Oct 2020.
17. BioStars BBDuk 2. https://www.biostars.org/p/237931/. Accessed 15 Oct 2020.
18. Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press. 2006.
19. Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, *et al.*. MEGAHIT v1. 0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods. 2016;102:3-11.
20. Azad A, Pavlopoulos GA, Ouzounis CA, Kyrpides NC, Buluç A. HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. Nucleic Acids Res. 2018;46:e33.
21. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes De Novo Assembler. Curr. Protoc. Bioinform. 2020;70:e102.
22. Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28:1420-8. doi.org/10.1093/bioinformatics/bts174.
23. Whitham JM. KBase Silver Case Study: Determining Media Formulation Requirements for Isolation of Microbiome Constituents. United States: N. p. 2021. doi:10.25982/68579.143/1766297.
24. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7:e7359. doi.org/10.7717/peerj.7359.
25. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32:605-607.
26. Yue Y, Huang H, Qi Z, Dou HM, Liu XY, Han TF, *et al.* Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. BMC Bioinformatics. 2020;21:334. doi.org/10.1186/s12859-020-03667-3.

27. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, *et al.* RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci. Rep. 2015;5:1-6.
28. Price MN, Dehal PS, Arkin AP. FastTree 2 Approximately Maximum-Likelihood Trees for Large Alignments. PLoS One. 2010;5. doi:10.1371/journal.pone.0009490
29. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 2016;33:1635-8.
30. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. 2021;49:D274-81.
31. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, *et al.* Pfam: The protein families database in 2021. Nucleic Acids Research. 2021;49:D412-9.
32. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O. TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res. 2001;29:41-3.
33. Torchiano M. effsize: Efficient Effect Size Computation. 2020. doi:10.5281/zenodo.1480624.
34. GitHub. https://github.com/jmwhitha/Trimming_and_decon. Accessed 22 April 2021.
35. Whitham, Jason. JGI QC impact on assembly, binning, phylogenomics, and functional analysis. United States: N. p., 2021. Web. doi:10.25982/62657.1515/1779219.
36. Whitham, Jason. Impact of BBDuk metagenomic read trimming and decontamination. United States: N. p., 2021. Web. doi:10.25982/77705.1341/1779218.
37. Sainani K. The importance of accounting for correlated observations. PM&R. 2010;2:858-861.

Figures

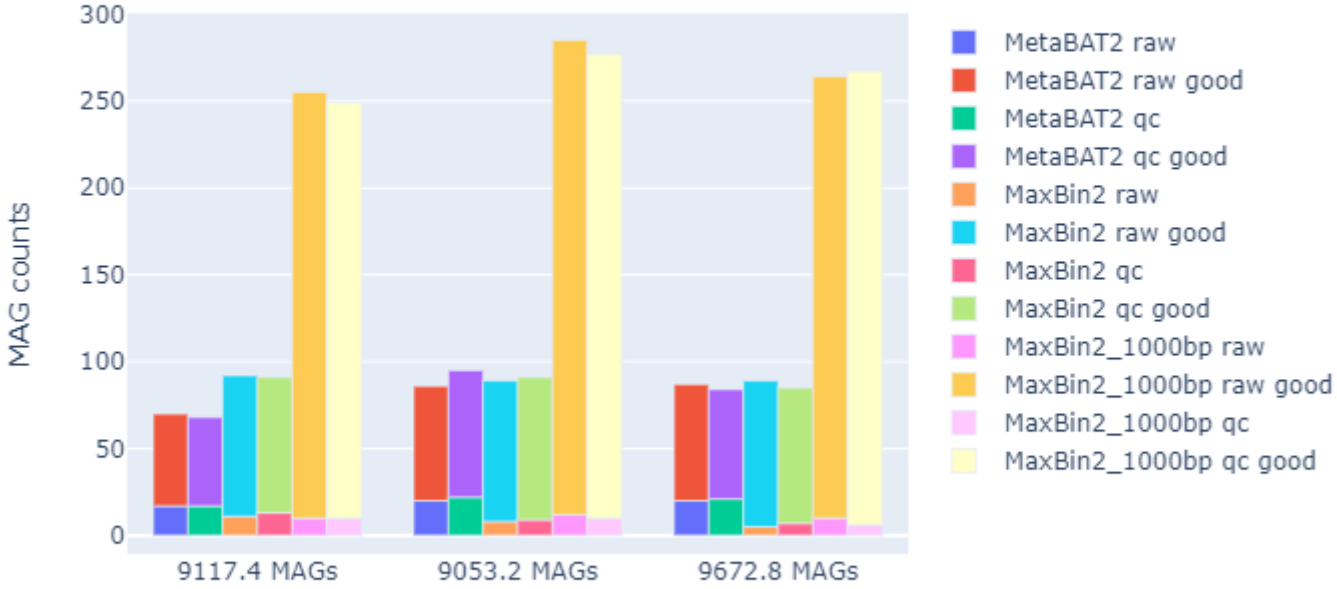

Figure 1

MAG counts of raw and qc assemblies binned with different applications and min contig lengths

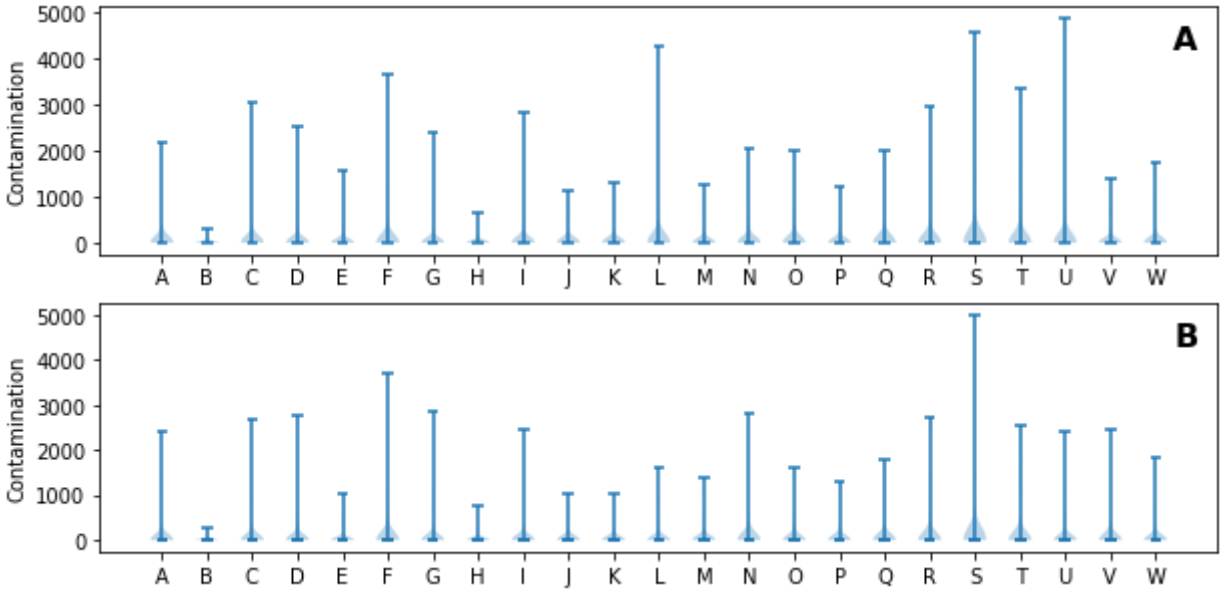

Figure 2

Distributions of percentage of MAG contamination per assembly. Raw (A) and QC (B) assemblies are ordered A-W: 10158.6, 10158.8, 10186.3, 10186.4, 11260.5, 11260.6, 11263.1, 11306.1, 11306.3, 7331.1,

9041.8, 9053.2, 9053.3, 9053.4, 9053.5, 9108.1, 9108.2, 9117.4, 9117.5, 9117.6, 9117.7, 9117.8, and 9672.8. Percentages can be higher than 100% because of multi-copy markers (e.g. 400% indicates that each lineage marker for a MAG has an average of 4 copies).

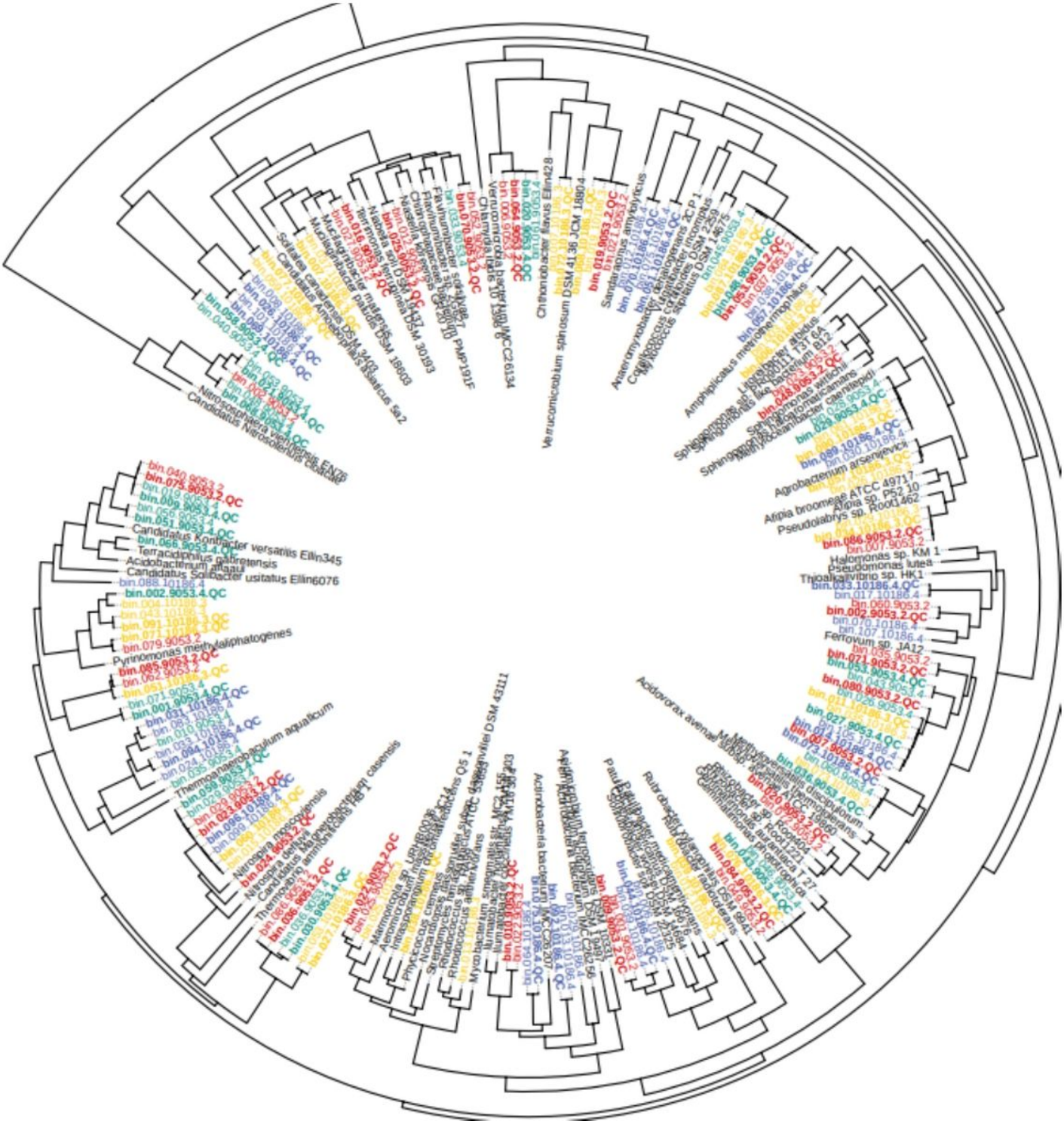

Figure 3

Phylogenomic tree of good quality raw and qc MAGs, colored by Plot-Season M3-2 (red), M3-4 (green), M5-2 (yellow), and M5-4 (blue).