Building an integrated plant microbiome MAG catalogue

Using an SSU linkage based approach

A Thesis presented for the degree of Master of Science.

Wageningen University & Research
Department of Bio-informatics

Author: Martijn de Vries
Supervisors: prof.DR. M.H Medema & DR. A Kupczok & L.J.U Pronk

Building an integrated plant microbiome MAG catalogue

Using an SSU linkage based approach

Martijn de Vries

Abstract

The plant associated microbiome is one of the main pillars of a plants general health and success. Metagenomics is an indispensable tool for the study of this microbiome, this technique is defined by directly sequencing environmental samples. This allows for the construction of MAG catalogues, which facilitate the future annotation of sequencing projects, without the computation power needed to generate the catalogue. In this paper, different strategies for the construction of a tomato plant rhizosphere MAG catalogue are explored. Namely, an all in one co-assembly of pooled samples and individual sample assemblies, in order to determine the optimal approach for this particular dataset. In addition, normalisation of the short read data is explored with the aim of improving the assembly quality and increasing the assembly speed. Furthermore, a recently published method to link SSU genes to MAGs is used to assign taxonomy to the MAGs in this catalogue. An open source snakemake workflow was designed specifically for this project. The MAG catalogue that was constructed consisted of 99 high quality MAGs with a species level taxonomic resolution. Including many species that remain uncultured as of to date.

Contents

Chapter 1
Introduction

Over the course of millions of years, plants have evolved alongside associated microorganisms. This led to the formation of a complex network of plant-microbiome interactions. It is currently well understood, that the plant-associated microbiome is one of the main pillars of a plants general health and success [1]. However, studying the plant microbiome is challenging. This is mainly, because many species in the microbiome operate in specialised niches under specific conditions. This means that many species can not be cultured and studied using traditional culture-based analysis [2]. Furthermore, if an organism can be cultured in the lab, the observed behaviour in this setting can be different compared to natural conditions [3].

Therefore, culture-independent methods. Such as shotgun metagenomic sequencing, provide a useful tool for studying the microbiome. This technique is defined by directly sequencing the genetic content of environmental samples [4]. This allows for the construction of metagenome-assembled genome (MAG) catalogues [5]. These catalogues facilitate the taxonomic and functional annotation of future sequencing projects, and can maximize the information that can be obtained from such projects [6], without the substantial computational power required to generate the catalogue.

The construction of MAG catalogues has greatly increased our understanding of species that remained uncultured before. This includes, for example, the recent elucidation of the bacterial super-phyla referred to as the candidate phyla radiation. Which may include up to 35 individual candidate phyla, many of which, until recently, had no genomic representation [6]. In fact, even the vast recourse of reference genomes that is available as of to date, does not represent the phylogenetic diversity in nature that is currently known [7]. The construction of comprehensive MAG catalogues from various environments can help fill in this gap and aid in completing the complex picture that is the tree of life.

For this particular project, a MAG catalogue of the tomato plant rhizosphere is constructed. The rhizosphere is is a thin region of soil that is directly influenced by the root exudates of the plant [8]. This is a distinct ecosystem that, in turn, has a distinct microbiome. In general, this microbiome is found to be dominated by Proteobacteria, and other highly abundant phyla include Actinobacteria and Firmicutes, among others. In contrast, this highly complex microbiome is also inhabited by various bacteria, fungi and unicellular eukaryotes that are relatively low in abundance [9]. This combination makes the recovery of full length MAGs from this microbiome difficult. The presence of many closely related species poses the risk of strain collapse during the assembly stage, potentially fragmenting assemblies [10]. The low abundance of the other organisms, on the other hand, may lead to the failure to recover genomes altogether [11].

There are various strategies for both the metagenome assembly and the construction of MAG catalogues [10]. For this project, two different strategies are used, with the aim to determine the optimal choice for this dataset. Namely, a sample specific assembly approach and an all in one co-assembly approach of multiple pooled samples.

The sample specific approach implies that both the assembly and the MAG construction is done for each sample individually. After this, the MAGs are combined and clustered to produce a non redundant MAG catalogue. In addition to reducing the redundancy, this clustering involves the selection of the best representative genomes, which is important for subsequent downstream analyses [12]. One of the main arguments that can be made for a sample specific approach, as opposed to the co-assembly approach, is that individual samples are less complex than a pool of samples. In general, lower sample complexity leads to higher quality assemblies [13]. There are however, some disadvantages to using this strategy. When MAGs are dereplicated, valuable information on single nucleotide variants (SNVs) and auxiliary gene content among representatives from the same species can be lost in the process [12]. Therefore, a strong case could be made for not dereplicating the MAGs.

The co-assembly approach implies that the reads of multiple samples are pooled and assembled together. A MAG catalogue is then directly constructed from this assembly. This avoids the problems associated with the dereplication [12]. In addition, the pooling of reads can artificially inflate the read depth of low abundance organisms. This can potentially increase the recovery rate of low abundance MAGs [11]. However, one of the potential downsides of this strategy, is that the pooling of reads may amplify the strain collapse problem that was highlighted earlier [10]. Therefore, the ideal solution depends on the specific dataset and or study goal.

The construction of a MAG catalogue from a complex metagenomic sample, highlights another specific problem large scale metagenomics faces currently. In order to properly assign a taxonomy to the MAGs in this catalogue, taxonomic information of a high resolution, preferably species level is required. As mentioned before, many species found in natural microbiomes have poor genomic representation in current databases. Therefore, this information is often best obtained by analysing universal marker genes such as the short sub unit (SSU) ribosomal ribonucleic acids (rRNA) genes [14]. Currently, there is a vast resource of sequenced SSU genes including organisms that have never been observed in labs before. However, because it is notoriously difficult to reconstruct full length SSU sequences in MAGs [6], The integration of MAGs with this resource is challenging [11].

In the past, various attempts have been made to link SSU genes to MAGs, usually resulting in poor linkage rates [6]. In a recent study by Lesker et al [11] a framework for a linkage strategy was proposed. This strategy was able to link 70% of MAGs in a mouse gut microbiome MAG catalogue. This strategy will also be applied to the MAG catalogue that is created for this project. This will potentially allow the high resolution taxonomic annotation of MAGs with poor genomic representation in databases. This will allow researchers in the future to compare short read data to this catalogue and analyse that data at the same resolution. And, naturally, this resource adds to the growing amount of MAGs that is available for research in general. Finally, a dedicated open source workflow is designed for this project that allows others to create MAG catalogues in the same way, in the hopes this encourages others to further improve upon the strategy that is described here, in order to build more MAG catalogues integrated with SSU gene taxonomic information.

Chapter 2
Methods

In this chapter, the (computational) methods that were used for this research project are described. Most tools mentioned in this chapter are part of a Snakemake (v7.20) [15] workflow that was designed for this study. In order to improve the compatibility of this pipeline with various systems, the package management is handled with anaconda (v22.9.0) [16]. This workflow is open source and publicly available on Github as the Plant Microbiome Catalogue (PMC) tool [17]. An overview of this workflow is presented in appendix F.

2.1 Downloading input data

The input data consisted exclusively of short read (150 bp) paired end Illumina [18] shotgun sequencing data. The sequencing was performed on the NovaSeq paired-end platform. The data that was used to construct the MAG catalogue was previously generated for the study of Oyserman et al [19]. As part of that study, shotgun metagenome sequencing of the tomato plant rhizosphere was performed. Specifically, this was performed for a recombinant inbred line of *Solanum Lycopersicum* and *Solanum pimpinellifolium*, and the respective parental lines. 20 samples were randomly selected (Appendix A) and downloaded for a total size of approximately 158 gigabases (Gb) hereafter referred to as the total data set. This data was downloaded from the NCBI Sequence Read Archive (SRA) using the SRA toolkit (v3.03) [20]. The data can be found using the BioProject ID PRJNA789467. In order to decrease the memory footprint of the total data set, the resulting FASTQ files were immediately compressed using pigz (v2.4) [21].

2.2 Input data quality control

The quality of the input FASTQ files was assessed, and improved upon, based on two criteria: (1) the overall quality of the reads in the FASTQ file; (2) the amount of contamination that is present in each sample.

In order to determine and improve the quality of the input reads fastp (v0.22) [22] was used. This tool can perform quality control, adapter trimming and quality filtering of FASTQ files. Fastp was run with the following adjusted settings: -q 20, -n 0, -e 20. These settings are slightly stricter than the default settings, further increasing the strictness of the quality control led to more than 2% of reads being removed from the samples. Therefore, the choice was made to settle for these parameters. This also means N characters that indicate a failed base-call are removed. Other than that the default settings were used. For each sample, quality reports were generated and the reads that passed the quality filters were written to dedicated compressed FASTQ files.

The reads that passed the previously mentioned quality control were mapped against the reference genomes of possible contaminants. In the context of this project this implies the host plant (*Solanum Lycopersicum*) and the ϕX174 bacteriophage (Potential positive spike in sequencing control) genomes. Both genomes were downloaded from the NCBI genome browser using the SL3.1 and ASM3342559v1 assem-

bly accessions respectively. The reads were mapped to the reference genomes with minimap2 (v2.22) [23] on default settings. The resulting alignments were converted to sorted BAM files and extracted using samtools (v1.9) [24]. Reads that could not be mapped to either reference genome were written to FASTQ files that were subsequently compressed. Hereafter referred to as quality-controlled samples.

2.3 Individual assemblies

A random selection of 5 quality-controlled samples (appendix A) was selected for individual denovo assemblies using metaSPAdes (v3.15.5) [25]. MetaSPAdes was run with the default settings but was given the total system RAM (3TB) memory to perform each assembly on 128 cores. After the assemblies were completed, quality reports were generated using MetaQUAST (v5.0.2) [26]. MetaQUAST was run without determining the taxonomy of contigs by supplying the -max-ref-number 0 argument. This choice was made as this parameter is prone to crashing the workflow. Other than that the default settings were used. In addition, the reads of the quality controlled samples were mapped back to their respective assemblies using minimap2 on default settings. The results were converted into sorted BAM files with samtools. Subsequently, samtools was also used to calculate mapping statistics using the "flagstat" option. In addition, a prediction of open reading frames was made using the default Prokka (v1.13.4) [27] workflow.

2.4 Read pooling and normalisation

The quality controlled samples were pooled using a custom python script. This script can be adjusted to split input samples to create pools of a total desired read count. For this particular project the desired read pool depth was adjusted to include all 20 quality controlled samples in one single pool with approximately 1 billion read pairs. From this point onward this will be referred to as the pooled samples.

Subsequently, the pooled samples were normalised using BBnorm (v39.01) [28]. This software package is designed mainly for normalising coverage, by K-mer based downsampling of reads in areas with relatively high read depth (>100). This leads to a reduction in the total amount of data which can significantly speed up the assembly process. In addition, this software package can also exclude reads based on a minimum required k-mer depth. Therefore, removing reads with many low depth k-mers (<5) potentially indicating sequencing errors. In that regard, pooling multiple samples into one assembly also means that all the sequencing errors are pooled in one assembly. This normalisation can possibly also correct for that. This can also potentially improve the overall quality of assemblies [29]. In total, approximately 786 million read pairs passed the normalisation procedure and were written to separate FASTQ files. From this point onward referred to as the normalised pooled samples.

2.5 Co-assemblies

Co-assemblies were performed on the pooled samples and the normalised pooled samples with MEGAHIT (v1.2.9) [30]. It is worth noting, that metaSPAdes has been shown to routinely outperform MEGAHIT for single sample assemblies. However, metaSPAdes can not handle the memory requirements of a co-assembly [11]. Therefore, metaSPAdes can not be used for this purpose. The assemblies were performed with the following adjusted settings: -f, -m 0.9, −−presets meta-large.

These settings are recommended for the assembly of complex metagenomes, such as those of soil by the original authors of MEGAHIT [31]. Other settings were kept as default. After the assemblies had concluded, the same post assembly analyses that are mentioned in the "Individual assemblies" section were performed.

2.6 Binning contigs to MAGs

For this project, MAGscoT (v1.0.0) [32] was chosen as the tool to perform the binning of contigs to MAGs. In short, this tool relies on two sets of marker genes from the Genome Taxonomy Database Toolkit [33], that are stored as Hidden Markov Models (HMMs). The presence profiles of these markers gene sets in bins produced by different binning tools is compared, and new hybrid candidate bins are created, that are scored alongside the original bins in the same way that was introduced by the binning tool DASTool [34]. MAGscoT is a fast tool that is able to outperform or at least compete with many current popular binning tools. Furthermore, this tool is not limited by the amount of binning tools that can be used as input. Therefore, this tool can be updated with newer and better binning tools in the future.

Three different binning tools were used as input for MAGscoT; (1) MetaBAT2 (v2:2.15) [35], (2) MaxBin2 (v2.2.7) [36], (3) CONCOCT (v1.1.0) [37]. All three binning tools were run with the default settings, a custom python script was used to parse the output bins so that they could be used by MAGscoT. MAGscoT was run with the default settings. A custom python script was made that parses the output file of the MAGscoT run to bin FASTA files. After the binning process was completed, the quality of the MAGs was assessed using the default CheckM (v1.2.2) [38] lineage workflow. In addition, the original pooled reads were mapped to the MAG catalogues using minimap2 on default settings.

2.7 MAG dereplication

Dereplication of the MAG sets was performed with the tool dRep (v2.2.3) [13]. This is a tool that can rapidly cluster genomes based on average nucleotide identity (ANI). Dereplication was performed firstly, to asses the redundancy of the constructed MAG catalogues. Secondly, the number of distinct MAGs, and the clustering of the MAGs, is used to make a comparison between the different assembly strategies. Dereplication was performed using the default dRep settings.

2.8 Reconstruction of SSU genes

The reconstruction of full length SSU sequences was performed as part of this project using the tool Phyloflash (v3.4) [39]. This tool takes short read data as input, and attempts to reconstruct, and classify full length SSU gene sequences. It does so by mapping the input reads to an SSU gene reference database, the reads only require 70% identity to be mapped. This allows reads from SSU genes that are not in the database to be mapped as well. This allows the discovery of novel SSU genes and this is the main reason this tool is included in the workflow. After the mapping, the mapped reads are extracted and a targeted assembly of SSU genes is performed only using those reads. The taxonomy of the SSU genes is then inferred based on the closest SSU gene in the database. Phyloflash was run using the default workflow and the reads of the pooled samples were used as input. The full lenght SSU sequences were stored in FASTA format, from this point onward referred to as

the reconstructed SSU sequences.

2.9 Linkage of SSU genes to MAGs

The linkage strategy that is described here follows the original approach described by Lesker et al. However, a few things are worth noting here; The original code that was provided by the authors did not work out of the box. Therefore, the code that is needed to perform the linkage procedure was rewritten for this project. This includes bash scripts that are used to parse the required data. The output of the new parsing scripts was compared to example output of the original scripts that was provided by the authors. Furthermore, the provided R script that performs the actual linkage had some associated problems. This includes hard coded directories not present on the system this project was performed on, and in some parts, missing code or functions that seem to have been deprecated in the meantime. In that regard, it is worth noting that the linkage R script itself was originally designed in 2019. Nevertheless, this script was rewritten following the approach that is described in original paper of Lesker et al. A detailed explanation of how the linkage score is calculated is presented in appendix B. The general strategy is briefly explained in the remainder of this section.

Firstly, one of the MAG catalogues that was created was selected for the linkage procedure. The choice was made, based on the results of the assemblies and binning, to perform the linkage procedure on the MAG catalogue that was obtained after the original pooled samples assembly (See discussion). Before the linkage, a selection of high quality MAGs was made. The quality cut-off for high quality MAGs for this project was completeness - contamination of at least 70%. This choice was made so that a minimal quality of MAGs in the final catalogue, at least based on these criteria, can be guaranteed. In total 99 MAGs were selected for the linkage procedure.

The linkage of SSU genes to MAGs is a three step process that involves; (1) Blast (v2.13.0) [40] search of SSU genes in MAGs, (2) Split mapping of reads to both SSU genes and MAGs using BBsplit (v38.90) [28], (3) Statistical correlation based linkage. Each step gives a score that indicates the connection between each MAG-SSU pair in the data set. All three scores are combined into an integrated score that is the final score indicating the likelihood a MAG and SSU gene are linked (appendix B). For this project, the taxonomy of a MAG was determined by the SSU gene that had the highest integrated scores. After that, manual curation was performed. The taxonomy inferred by the SSU gene linkage was compared to the taxonomic classification obtained by the analyses described in the next section. If there was taxonomic disagreement the second best SSU gene based on linkage score was compared. If this led to an agreement in taxonomic classification, the second best SSU hit was used to infer the taxonomy of a MAG (See discussion).

2.10 Additional taxonomic assignment / SSU prediction

To allow the validation of the results of the linkage procedure, additional analyses were performed on the high quality MAGs. This includes a prediction of MAG taxonomies using the Bin Annotation Tool (BAT) (v4.6) [41]. This taxonomic as-

signment provides a frame of reference against which the SSU linkage procedure can be compared.

In addition, a prediction of SSU genes in the high quality MAGs was done with Barrnap (v0.9) [42]. This tool can predict likely locations of SSU genes, It was used to give an indication of both the amount and the distribution of SSU genes in the high quality MAGs. This result is also compared with the linkage approach.

Chapter 3
Results

3.1 Input data quality control:

In this section the results of the input data quality control are summarised. A comprehensive table with the results of the quality control and contamination removal can be found in appendix C. On average, 98,62% (98,06% - 99,02%) of reads in all samples passed the quality control. Additionally, on average 9,67% (5.65% - 15,71%) of reads were removed from the samples after they were mapped to the genomes of possible contaminants. There was no observed ϕX174 contamination in any of the samples.

3.2 Individual assemblies:

Individual MetaSPAdes assemblies were performed for 5 randomly selected samples. In this section a summary is provided of those assemblies. A more detailed table with the results for each sample and additional statistics is presented in appendix D. On average, the assemblies produced 352.656 contigs (180.649 - 503.483). With a total length of on average 368.978.315 base pairs (172.969.935 - 511.353.652). After mapping the reads of all samples back to the respective assemblies, an average mapping rate of 66.03% (58.53% - 73.15%) was observed. A prediction of the total amount of ORFs in the individual assemblies is shown in figure 3.1.

3.3 Pooled assemblies:

Pooled assemblies with MEGAHIT were performed for the original and the normalised pooled samples. The resulting assemblies had 5.803.284 and 5.159.371 contigs for total sizes of 7.070.626.611 and 6.504.501.359 base pairs respectively. The reads of the original pooled samples were mapped to both assemblies, resulting in an observed mapping rates of 88.99% and 87.58% respectively. More detailed assembly statistics are presented in appendix D. The results of the ORF predictions for both assemblies are shown in figure 3.1.

3.4 Binning contigs to MAGs:

After running MAGScoT on the assembly of the original pooled samples, 211 MAGs were recovered. On average the MAGs had a size of 4.225.739 (470.944 - 11.335.146) base pairs. Binning the assembly of the normalised pooled samples yielded 186 MAGs in total with an average size of 4.219.538 (643.389 - 9.350.779) base pairs. The individual sample assemblies combined in total produced 80 MAGs after binning. With an average size of 3.571.757 (498.682 - 8.702.377) base pairs. The binning efficiency was determined by comparing the ORFs in MAGs to the total ORFs of the assembly, the results are shown in figure 3.1. In addition, The quality and completeness scores of the MAGs was determined, the results are shown in figure 3.2. The original pooled reads were mapped to the catalogues, for the individual assemblies MAG set a mapping rate of 21,40% was observed. For the original pooled assembly MAGs this was 16,82% and the lowest observed rate was 15,57% for the normalised assembly MAGs.

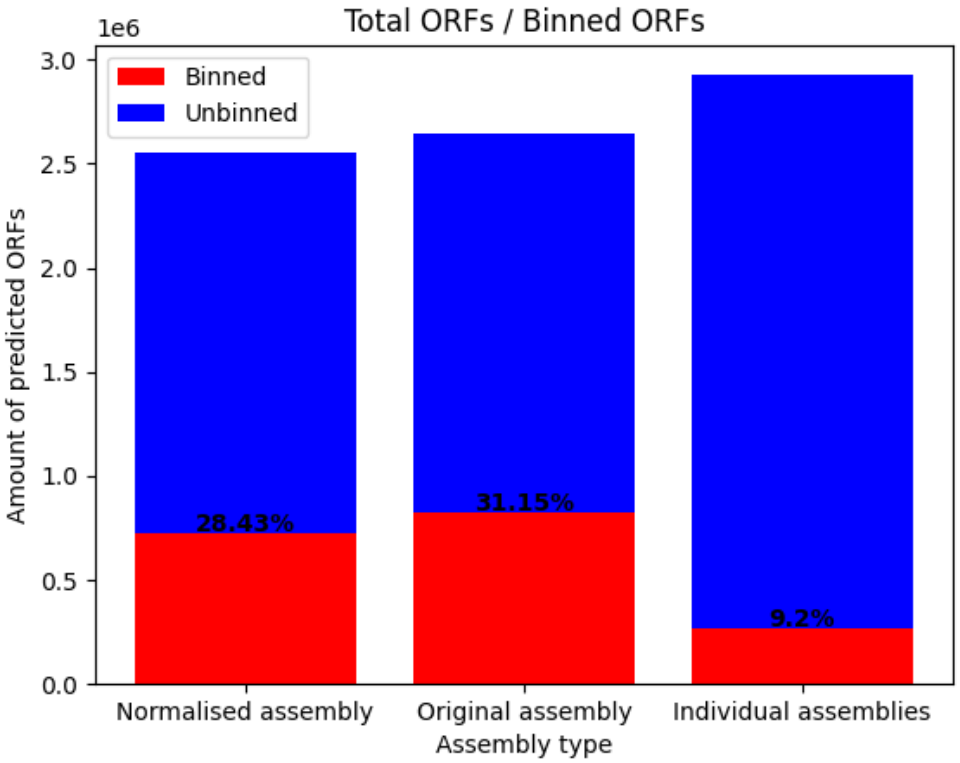


Figure 3.1: Bar plot that indicates the total ORFs predicted and the percentage of ORFs that were binned for each assembly strategy.

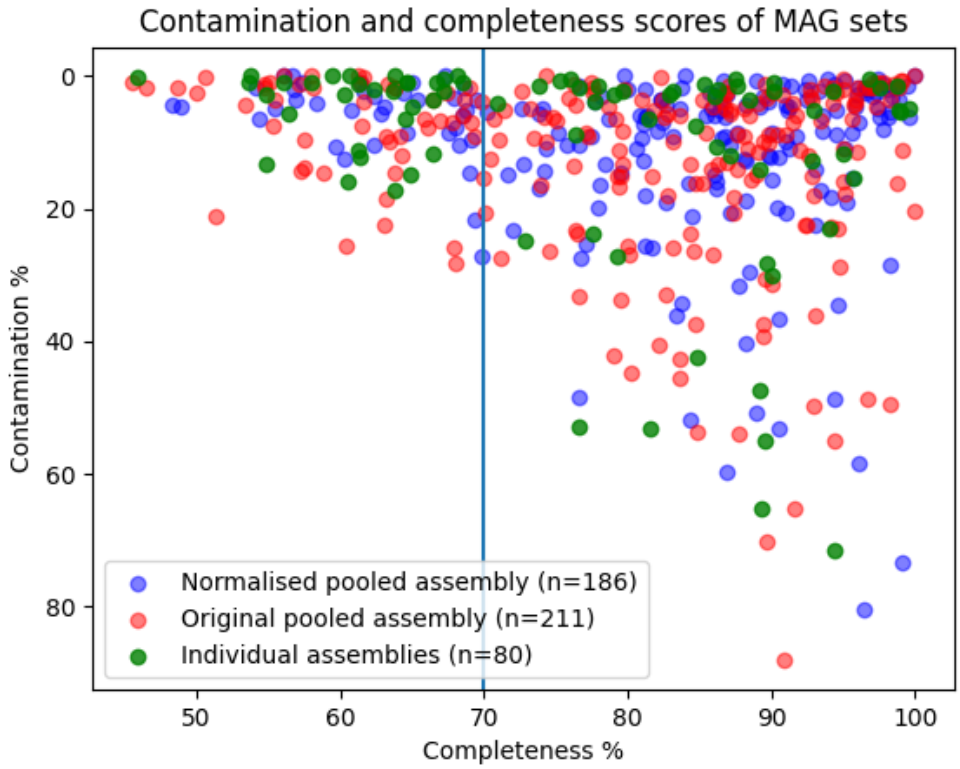


Figure 3.2: Plot showing CheckM contamination and completion scores for the produced MAG sets. The blue line indicates the minimal required completeness score to be included in the catalogue.

3.5 Dereplication of MAGs

The results of the dereplication are discussed in this section, the dendrograms that were obtained after clustering each MAG set are shown in appendix E. It is worth noting, that deRep uses an in-build quality filtering function that selects MAGs of a certain quality for clustering. This selection is slightly different, and less strict then the selection that was made for high quality MAGs in the catalogue build for this project (Completeness minimum 75 and contamination maximum 25). Therefore, slightly more MAGs are included in the clustering than would be in the catalogue. Firstly, the pooled assemblies are discussed. In both cases, the amount of output clusters equals the amount of input genomes. In other words, species level dereplication has not led to the reduction in the total number of genomes. In both cases, a comparable number of distinct MAGs were recovered after dereplication, and quality filtering. With 118 and 116 MAGs for the normalised and original assembly MAG sets respectively. The dereplication of the single sample assemblies shows a different picture. Dereplication of this dataset led to a 37,5% reduction in the number of genomes (from 32 to 20). When compared to the pooled assemblies, the individual sample approach yielded less distinct MAGs overall, but the co-assembly approach does not allow recovery closely related but distinct MAGs, with average nucleotide identity between MAGs never exceeding at least 90%. In addition, after dereplication of the individual assembly MAG sets, the mapping rate of the original pooled samples was re-evaluated at 12,91% a reduction of 8,49% after dereplication.

3.6 Reconstruction of SSU genes:

The reconstruction of SSU genes with Phyloflash produced a total of 238 (2% Eukaryotic, 97% Prokaryotic, 1% Archeal) distinct reconstructed SSU sequences. An overview of Phyloflash results is presented in plots a-d of figure 3.3. Plot a shows the identity (%) of reads mapped to the SSU database. The distribution is heavily skewed towards high identities of over 90%. This is an indication the most of the SSU genes that were detected trough mapping have a good representation in the SSU gene database. In plot b, the percentage of reads that mapped to the SSU database is shown. The SSU reads have a very low abundance of 0,101%. In most cases the reads could be mapped properly as a pair. The observed insert sizes for read pairs are shown in plot c. The uni-modal distribution of this plot indicates that there are no contaminating libraries present in the analysis. This was expected beforehand since all the samples are from the same sequencing experiments. Lastly, plot d shows the proportion of assembled reads, after mapping them to the SSU database. From this plot, the observation can be made that most reads were not assembled. This can either be caused by a general failure of assembly, however, it can also be caused by the presence of highly diverse and low abundance organisms [39]. In addition to the reconstruction of the SSU genes, a prediction of SSU gene location in the high quality MAGs used for the linkage was performed. This analysis gives an indication of the distribution of SSU sequences over MAGs and the results are shown in plot e of figure 3.4. In 15% of the MAGs, multiple SSU genes were detected. It is possible that these are multiple copies of the same SSU gene, however, it is also possible that these are distinct SSU genes and even false positive detections by barrnap [42]. For the other MAGs, there seems to be a roughly equal split of MAGs with no SSU genes (43%) and MAGs with an unique SSU gene predicted (41%).

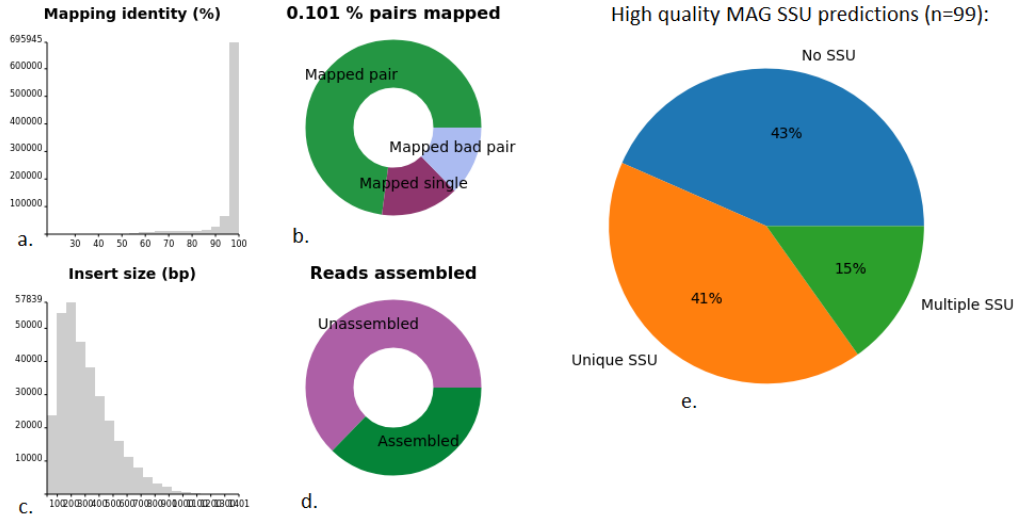


Figure 3.3: Plots showing the results of the SSU reconstruction (a-e) and the SSU gene predictions in high quality MAGs (e).

3.7 Taxonomic classification of MAGs with BAT

In figure 3.4 a Krona chart of the taxonomy of high quality MAGs, as defined by BAT, is shown. This taxonomic classification was performed on the MAG set of the original pooled assembly, as these were used for the linkage procedure. Strikingly, 23% of the MAGs could only be classified to a kingdom specific level. In addition, the overall taxonomic resolution of this classification is low. The majority of the classifications are only specific to a phylum level.

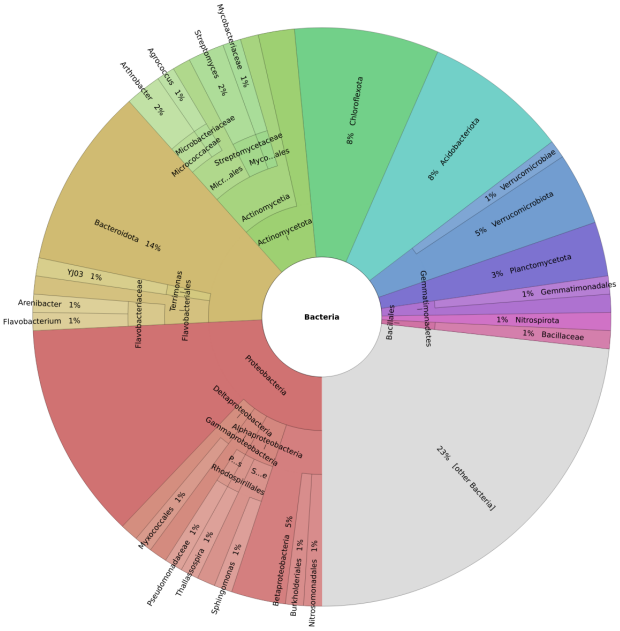


Figure 3.4: Krona plot that shows the taxonomic classifications of high quality MAGs used for the linkage procedure according to BAT.

3.8 Linkage of SSU genes to MAGs:

As mentioned in the introduction, the linkage of SSU genes to MAGs was mainly performed in order to assign taxonomy to MAGs in the catalogue at a high resolution. After the linkage procedure, the Krona plot shown in figure 3.5 was made. This Krona plot indicates the taxonomic classifications of the high quality MAGs obtained after the co-assembly of the original pooled samples. The catalogue contains 99 MAGs in total. The taxonomic resolution of the MAG catalogue has improved significantly as compared to the BAT classifications. All MAGs have a taxonomy assigned to a species level. Strikingly, in most cases, the taxonomy that was inferred trough the SSU linkage suggest that the MAGs belong to uncultured species. In this context this implies the species is only known trough various environmental 16S sequencing projects. The inferred composition of the microbiome is along the expectations for a rhizosphere sample [9]. for instance, proteobacteria is among the most dominant phyla. Interestingly, the microbiome also contains members of the earlier mentioned "Candidate phyla radiation" that generally have poor genomic representation.

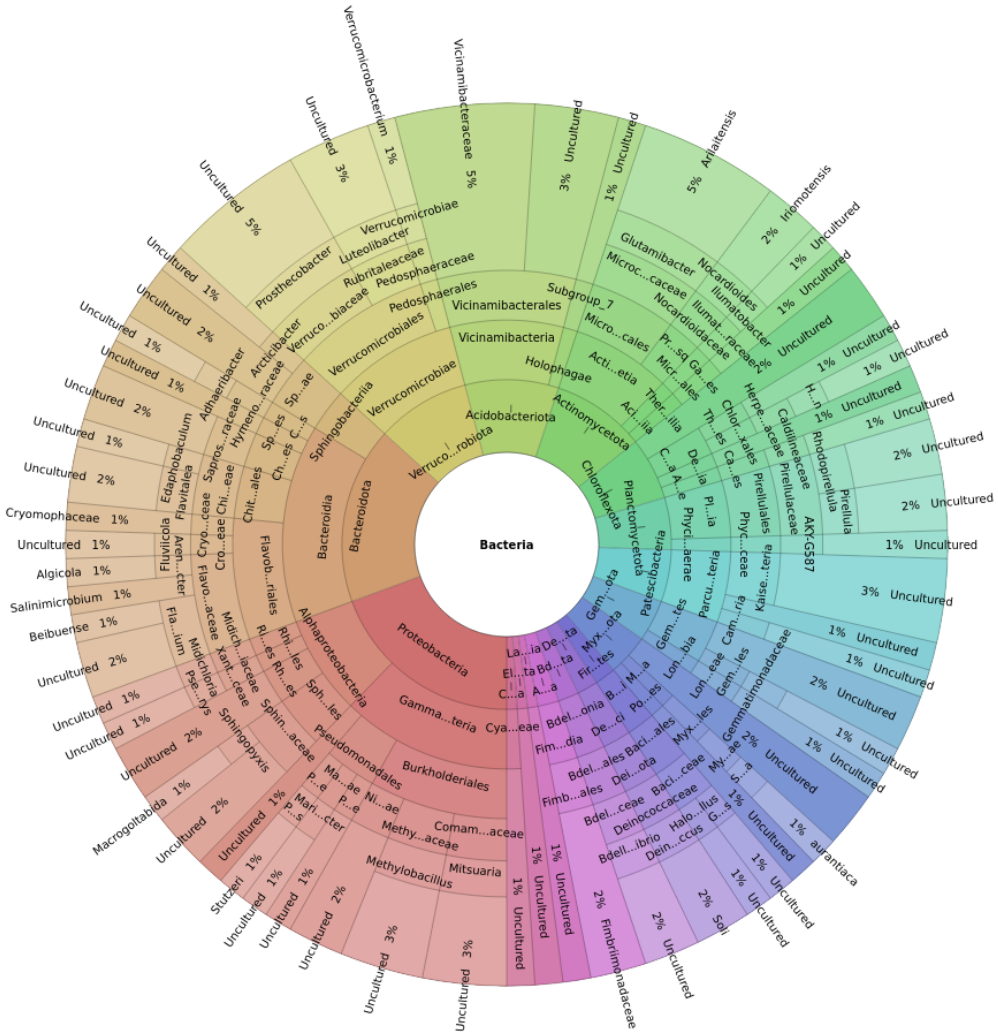


Figure 3.5: Krona plot showing the taxonomic assignment of high quality MAGs according to the SSU linkage approach.

Chapter 4
Discussion

Firstly, a brief discussion on the dataset that was used to construct the MAG catalogue. In general, the samples that were chosen for this project, all had high observed quality scores and low contamination. This is taking into account that these are metagenomic samples that are notorious for high contamination [3]. The samples that were used for this project both include the recombinant inbred line and the respective parental lines. The goal being to make the MAG catalogue as comprehensive as possible. That being said, the relative abundance of parental line samples in the dataset is comparatively low, this is a result of the experimental setup during the study of Oyserman et al. Therefore, this potentially reduces the recovery of MAGs from organisms exclusively associated with the microbiome of the parental lines.

When comparing the results of the assemblies, some interesting observations can be made. Both pooled assemblies have a substantially larger total assembly size and larger contig sizes compared to the individual assemblies. In addition, the mapping rates of the pooled assemblies are higher than the individual assemblies. Furthermore, the assembly sizes and the mapping rates vary substantially for the individual assemblies. This can, of course, be partly attributed to the fact that the amount of input reads varies per sample. However, in that regard, it is worth noting that the largest individual assembly was not obtained from the largest input sample. In the context of this project, this variation is most likely caused by the complexity of the input sample. In an earlier study on denovo assemblies, it was shown that an increase in sample complexity leads to more assembly fragmentation and generally reduces the quality of assemblies, also, the variation between assemblies was observed to increase along with sample complexity [43]. Strikingly, when comparing the size of the largest individual assembly to the co-assembly of the original samples, an increase of 1282% was observed favoring the co-assembly. Based on the assembly results, the argument could be made that the co-assemblies performed better for this particular dataset, in the sense that a larger portion of the microbiome was assembled. In the study of Lesker et al, the co-assembly strategy was also explored. The main point that was put forward in favour of the co-assembly by that study, was that it could recover organisms with a relatively low abundance that the individual assemblies could not [11]. This can partly explain the large difference in assembly sizes. This point is further supported by additional data that will be discussed in this chapter. Normalisation of the assemblies did have an effect on the outcome. The original pooled assembly has the largest total size, however, the normalised assembly has the longest contig. Other than that the assemblies are highly comparable. When the original pooled reads were mapped to both assemblies, a difference in mapping rate of 1,5% was observed in favour of the original assembly. That being said, the assembly time was reduced by approximately 30% for the normalised assembly (CPU time). While the normalisation for this dataset took approximately 4 hours. Therefore, a strong argument can be made for the normalisation of metagenomic co-assemblies for saving time. The effect of normalisation on the construction of the MAG catalogue is also touched upon later in this chapter. It is worth noting, that the individual assemblies combined did contain more ORFs in total than

either of the co-assemblies. However, as will be discussed, there is some amount of redundancy present.

After the assemblies, MAG sets were constructed. For all three MAG sets, the observed quality and contamination scores were comparable. In that regard, the assembly strategy does not seem to have a significant impact. The individual assemblies combined produced less bins than either of the co-assemblies. However, the individual assemblies were only performed for 5 samples. Therefore, a meaningful comparison of the amount of MAGs can not be made at this point. The binning efficiency of each assembly strategy was determined by the percentage of open reading frames in the assemblies that were binned. This is were the differences between the co-assemblies and the individual assemblies becomes more clear. The binning of the individual assemblies is far less efficient than binning the contigs of the co-assemblies, with less than 10% of predicted open reading frames ending up in the MAGs. A variety of factors may have potentially attributed to this low binning efficiency [10]. However, since the binning strategy was the same for each assembly, and both pooled assemblies show higher binning efficiency. This is most likely caused by the differences in the assemblies. The individual assemblies are shorter in general and contain more short contigs compared to the co-assemblies. Prior to the binning, contigs smaller then 1000 base pairs were removed as part of the default binning procedure. In addition, a minimal MAG size of 50.000 base pairs was required at minimum for a MAG to be included in the output, this is also the default of MAGscoT. It is worth noting, that for some individual assemblies, this means over half of the contigs were not even considered during the binning (N50 less then 1000). In addition, longer contigs logically contain more information that can be used during the binning stage [10]. Therefore, all things considered, the small assembly and contig sizes of the individual assemblies are the most likely cause of the difference in binning efficiency.

Dereplication provided some additional insight into the differences between the assembly strategies. When the MAGs of the co-assemblies were dereplicated, no species level clustering was observed. This is however, highly unlikely. For example, in a recent study it has been shown that the rhizosphere metagenome also shows a high phylogenetic diversity even at the species level [44]. As mentioned in the introduction, co-assemblies have a high potential for strain collapse during the assembly stage [45], part of this problem is that closely related strains collapse onto the same contigs or become highly fragmented. Comparing the results of the recent study in species level diversity with this observation, makes this the most likely explanation for the absence of dereplication. In contrast, the dereplication of the individual assemblies MAG set provided a different picture. In this case, there was a reduction in the total number of MAGs after dereplication. Because the MAGs of several assemblies were pooled, some redundancy was expected beforehand. However, the most striking difference between the MAGs of the different assembly strategies, is the total number of distinct MAGs obtained after dereplication. In that regard, the co-assemblies have substantially outperformed the individual assemblies, at least when capturing a wide range of distinct MAGs is concerned. This can partly be explained by the point that was made earlier. In the study of Lesker et al, it was observed that individual assemblies have difficulties with assembling the genomes of relatively low abundance organisms. In that study a similar observation was made

as is the case here. The individual assembly MAGs showed a very high redundancy and an inability to rival the co-assembly in total MAGs recovered. In that study, the TPM of MAGs was determined for both assembly strategies. And it was concluded that a large number of co-assembly MAGs had abundances well below those of the single sample assembly. This particular analysis was not performed for this project, mainly to save time. However, it is expected that this is also the case here. In conclusion, there does not seem to be an optimal choice in this case when it comes to single sample assembly versus co-assemblies. In general, a co-assembly seems to be more suited for a broad screening of the microbiome, picking up many representative genomes, while losing taxonomic resolution in the process. In contrast, individual sample assemblies trade the large scope for a higher resolution, meaning the ability to recover closely related genomes from different samples. Because the goal of this project is the construction of a MAG catalogue, the argument was made to construct the most comprehensive catalogue and thus, use the larger co-assembly MAG set. Because the original sample MAG set was already made, and was slightly better then the normalised sample MAG set, taking the read mapping rates into account, this set was chosen for the linkage. For future projects, normalisation is still highly recommended for saving assembly time, since even after dereplication, the differences between the assemblies are small, and a similar number of distinct MAGs was recovered from both co-assemblies.

In the following paragraph, the results of the SSU reconstruction are discussed. A low read abundance of around 0,1% was observed, this low amount is in line with a complex metagenomic sample [46]. It was observed that the majority of reads that mapped to the SSU database could not be assembled. In general, it is hard to reconstruct full SSU sequences because they share highly conserved regions and in general share a lot of similarity, making the assembly of full length sequences difficult [39]. The fact that a large proportion of reads could not be assembled can be further attributed to the low abundance of some included species [39]. Nevertheless, the reconstruction of SSU genes was successful and yielded Prokaryotic, Eukaryotic and Archeal SSU genes. All reconstructed sequences were included for the linkage procedure. This was done for a specific reason. At this point, it was clear that all the MAGs in the MAG catalogue are of bacterial origin. This was confirmed by both checkM and BAT. However, in the paper of Lesker et al, it was mentioned that in some cases, false positive SSU linkages were observed in the results. In that study, this was confirmed by comparing the SSU inferred taxonomy to the known taxonomy of the linked genomes. It was concluded that a taxonomic disagreement was found in around 30% of the MAG classifications. In the study of Lesker et al, disagreement at the family level meant that the MAGs were filtered out of the catalogue. However, because for this project, the reference taxonomy provided by BAT is of a much lower resolution, SSU genes of different kingdoms were included to measure the false positive rate after the linkage of SSU genes to MAGs.

As mentioned, the taxonomic classification of the MAGs with BAT produced a low taxonomic resolution, with most MAG classifications only specific to the phylum level. There are several factors that can explain the low resolution that was observed. Firstly, this could be result of the binning procedure. If, during the binning, contigs belonging to different species are binned together. This leads to difficulties with the classification [47]. Specifically for this reason, the quality of the

MAGs was assessed with checkM, and MAGs were filtered based on a minimal completeness and maximum contamination. In addition, an observation that was made during the study of Lesker at al, may also apply here. In that study, the genomes of bacteria that were known to be present in the microbiome were blasted against the MAGs obtained after co-assembly. It was observed that the genomes were mostly contained withing distinct MAGs. It is worth noting, that metabat2 was used for the binning in that study, MAGscoT has been shown to outperform metbat2 in general [32]. It is therefore argued that bin contamination is not the major factor in the low resolution taxonomic classification. Mainly because, based on what is mentioned, it seems unlikely that the bins are contaminated to a degree that they only allow kingdom/phyla level classification in general. Another explanation for this observation is the poor genomic representation of many species that are present in the MAG catalogue. If the query organism is highly divergent from organisms that are present in the database, BAT by default decreases the level of taxonomic classification until a hit of sufficient score is achieved [47]. Various metagenomic studies, have already shown, that many of the MAGs that are recovered belong to species that are currently not well studied and catalogued [6], [11]. This can also explain the low taxonomic resolution that was obtained after the MAG classification with BAT. This is further touched upon in the next paragraph. Most likely, the low resolution is the result of a combination of the factors mentioned here, possibly also including the strain collapse problems mentioned with the assembly before.

The linkage of SSU genes to the MAGs was mainly performed to assign taxonomy to the MAG catalogue that was constructed. The linkage procedure itself, followed the approach of Lesker et al. However, a different strategy was used for the manual curation of the MAG catalogue than is described in that paper. In the original study, MAGs were excluded from the catalogue if the taxonomy of the SSU gene and the additional classifications differed at the family level. In that study, it was argued that these are false positive classifications. This was argued for the following reason; the score that is calculated depends on two factors. This includes the direct correlation between SSU genes and MAGs. This correlation is determined by blasting and mapping. However, as the results of the SSU gene prediction suggests, in a large portion of the MAGs the SSU genes have not been reconstructed to a sufficient degree. Therefore, the association for these MAGs is mainly based on the Pearson and Spearman correlations. This leaves a chance that MAGs and SSU genes are linked purely based on correlation and not causation, likely leading to a taxonomic disagreement with other methods. However, as mentioned, the reference taxonomy available for this project was relatively low. Therefore, using family level agreement as a baseline was not an option. This is the reason that the SSU genes of other kingdoms were included in the analysis, in order to get an indication of the false positive rate. Furthermore, during this project, if taxonomic disagreement was observed at any level, the second best SSU gene hit was evaluated in that case. It was observed that in every case were this was needed (Approximately 10% of MAGs), this led to a taxonomic agreement between the BAT classification and the SSU classification, at the level that was available for that particular MAG. Therefore, the choice was made to include these MAGs in the catalogue so that it is as comprehensive as possible. In the study of Lesker et al, it was shown that a false classification rate of around 30% is to be expected (At the family level) with this approach. However, because of the resolution of the reference taxonomy, filtering

the false positives was not possible to that degree for this project. However, it is worth noting, that these disagreements are mostly class level and below for this catalogue. Therefore, the general structure of the microbiome can confidently be determined using this approach. At the very least to a substantially higher degree than the BAT classifications.

In conclusion, a MAG catalogue of the tomato plant rhizosphere was constructed. This MAG catalogue holds 99 representative high quality genomes. This MAG catalogue has a species level taxonomic resolution due to an SSU linkage based approach. This MAG catalogue includes many species that are currently uncultured and only known trough SSU gene sequencing. In order to construct this catalogue, both a single sample and co-assembly approach were tested. This has given valuable insight into the strategies that can be used for the future construction of MAG catalogues. Finally, an open source workflow was constructed, in the hope that others further improve upon the MAG catalogue construction framework that is explored and described in this report.

Bibliography

[1] Stéphane Compant, Abdul Samad, Hanna Faist, and Angela Sessitsch. A review on the plant microbiome: ecology, functions, and emerging trends in microbial application. *Journal of advanced research*, 19:29–37, 2019.

[2] Philip Hugenholtz, Brett M Goebel, and Norman R Pace. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of bacteriology*, 180(18):4765–4774, 1998.

[3] John C Wooley and Yuzhen Ye. Metagenomics: facts and artifacts, and computational challenges. *Journal of computer science and technology*, 25(1):71–81, 2010.

[4] Thomas J Sharpton. An introduction to the analysis of shotgun metagenomic data. *Frontiers in plant science*, 5:209, 2014.

[5] Jessica D Forbes, Natalie C Knox, Jennifer Ronholm, Franco Pagotto, and Aleisha Reimer. Metagenomics: the next culture-independent game changer. *Frontiers in microbiology*, 8:1069, 2017.

[6] Donovan H Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J Woodcroft, Paul N Evans, Philip Hugenholtz, and Gene W Tyson. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature microbiology*, 2(11):1533–1542, 2017.

[7] Christian Rinke, Patrick Schwientek, Alexander Sczyrba, Natalia N Ivanova, Iain J Anderson, Jan-Fang Cheng, Aaron Darling, Stephanie Malfatti, Brandon K Swan, Esther A Gies, et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–437, 2013.

[8] James M Lynch, Melissa J Brimecombe, and Frans AAM De Leij. Rhizosphere. *e LS*, 2001.

[9] Thomas R Turner, Euan K James, and Philip S Poole. The plant microbiome. *Genome biology*, 14(6):1–10, 2013.

[10] Luis Fernando Delgado and Anders F Andersson. Evaluating metagenomic assembly approaches for biome-specific gene catalogues. *Microbiome*, 10(1):1–11, 2022.

[11] Till R Lesker, Abilash C Durairaj, Eric JC Gálvez, Ilias Lagkouvardos, John F Baines, Thomas Clavel, Alexander Sczyrba, Alice C McHardy, and Till Strowig. An integrated metagenome catalog reveals new insights into the murine gut microbiome. *Cell reports*, 30(9):2909–2922, 2020.

[12] Jacob T Evans and Vincent J Denef. To dereplicate or not to dereplicate? *Msphere*, 5(3):e00971–19, 2020.

[13] Matthew R Olm, Christopher T Brown, Brandon Brooks, and Jillian F Banfield. drep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME journal*, 11(12):2864–2868, 2017.

[14] F Maia Da Silva, H Noyes, M Campaner, ACV Junqueira, JR Coura, N Añez, Jeffrey Jon Shaw, JR Stevens, and Marta Maria Geraldes Teixeira. Phylogeny, taxonomy and grouping of trypanosoma rangeli isolates from man, triatomines and sylvatic mammals from widespread geographical origin based on ssu and its ribosomal sequences. *Parasitology*, 129(5):549–561, 2004.

[15] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[16] Anaconda software distribution, 2020.

[17] de Vries Martijn. PMC-tool. `https://github.com/mvries/PMC-tool`, 4 2023.

[18] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *nature*, 456(7218):53–59, 2008.

[19] Ben O Oyserman, Stalin Sarango Flores, Thom Griffioen, Xinya Pan, Elmar van der Wijk, Lotte Pronk, Wouter Lokhorst, Azkia Nurfikari, Joseph N Paulson, Mercedeh Movassagh, et al. Disentangling the genetic basis of rhizosphere microbiome assembly in tomato. *Nature Communications*, 13(1):3228, 2022.

[20] SRA toolkit development team. Sra toolkit. `https://github.com/ncbi/sra-tools/wiki/01.-Downloading-SRA-Toolkit`.

[21] Senthilkumar Palani. Pigz – compress and decompress files in parallel in linux. `https://ostechnix.com/pigz-compress-and-decompress-files-in-parallel-in-linux/`.

[22] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018.

[23] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[24] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, et al. Twelve years of samtools and bcftools. *Gigascience*, 10(2):giab008, 2021.

[25] Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel A Pevzner. metaspades: a new versatile metagenomic assembler. *Genome research*, 27(5):824–834, 2017.

[26] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. Quast: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 2013.

[27] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014.

[28] B Bushnell. Bbtools: a suite of fast, multithreaded bioinformatics tools designed for analysis of dna and rna sequence data. *Joint Genome Institute*, 2018.

[29] Dilip A Durai and Marcel H Schulz. Improving in-silico normalization using read weights. *Scientific Reports*, 9(1):1–10, 2019.

[30] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676, 2015.

[31] aquaskyline. megahit. `https://github.com/voutcn/megahit`, 10 2019.

[32] Malte Christoph Rühlemann, Eike Matthias Wacker, David Ellinghaus, and Andre Franke. Magscot: a fast, lightweight and accurate bin-refinement tool. *Bioinformatics*, 38(24):5430–5433, 2022.

[33] Pierre-Alain Chaumeil, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. Gtdb-tk: a toolkit to classify genomes with the genome taxonomy database, 2020.

[34] CHRISTIAN SIEBER. Dereplication, aggregation and scoring tool (das tool) v1. 0. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2017.

[35] Dongwan D Kang, Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359, 2019.

[36] Yu-Wei Wu, Blake A Simmons, and Steven W Singer. Maxbin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4):605–607, 2016.

[37] Johannes Alneberg, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Nicholas J Loman, Anders F Andersson, and Christopher Quince. Concoct: clustering contigs on coverage and composition. *arXiv preprint arXiv:1312.4038*, 2013.

[38] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, 25(7):1043–1055, 2015.

[39] Harald R Gruber-Vodicka, Brandon KB Seah, and Elmar Pruesse. phyloflash: rapid small-subunit rrna profiling and targeted assembly from metagenomes. *Msystems*, 5(5):e00920–20, 2020.

[40] Stephen F Altschul. Blast algorithm. *e LS*, 2001.

[41] Diego D Cambuy, Felipe H Coutinho, and Bas E Dutilh. Contig annotation tool cat robustly classifies assembled metagenomic contigs and long sequences. *BioRxiv*, page 072868, 2016.

[42] Torsten Seemann. barrnap 0.9: rapid ribosomal rna prediction. *Google Scholar*, 2013.

[43] Bin Yang, Yu Peng, Henry CM Leung, Siu-Ming Yiu, Jing-Chi Chen, and Francis YL Chin. Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers. In *Proceedings of the third international workshop on Data and text mining in bioinformatics*, pages 3–10, 2009.

[44] Jiajia Zuo, Mengting Zu, Lei Liu, Xiaomei Song, and Yingdan Yuan. Composition and diversity of bacterial communities in the rhizosphere of the chinese medicinal herb dendrobium. *BMC Plant Biology*, 21(1):1–13, 2021.

[45] Jurgen F Nijkamp, Marcel A van den Broek, Jan-Maarten A Geertman, Marcel JT Reinders, Jean-Marc G Daran, and Dick de Ridder. De novo detection of copy number variation by co-assembly. *Bioinformatics*, 28(24):3195–3202, 2012.

[46] Bin Zou, JieFu Li, Quan Zhou, and Zhe-Xue Quan. Mipe: A metagenome-based community structure explorer and ssu primer evaluation tool. *PLoS One*, 12(3):e0174609, 2017.

[47] FA Bastiaan Von Meijenfeldt, Ksenia Arkhipova, Diego D Cambuy, Felipe H Coutinho, and Bas E Dutilh. Robust taxonomic classification of uncharted microbial sequences and bins with cat and bat. *Genome biology*, 20:1–14, 2019.

Appendix A
Input sample accessions

The following samples were randomly selected from SRA bioproject ID PRJNA789467:

1. SRR17255001 (*Solanum Lycopersicum X Solanum Pimpinellifolium* #262)
2. SRR17255002 (*Solanum L. X Solanum P.* #210)
3. SRR17255003 (*Solanum L. X Solanum P.* #261)
4. SRR17255004 (*Solanum L. X Solanum P.* #260)
5. SRR17255025 (*Solanum L. X Solanum P.* #240)
6. SRR17255037 (*Solanum L. X Solanum P.* #289)
7. SRR17255040 (*Solanum L. X Solanum P.* #286)
8. SRR17255041 (*Solanum L. X Solanum P.* #285)
9. SRR17255042 (*Solanum L. X Solanum P.* #284)
10. SRR17255043 (*Solanum L. X Solanum P.* #283)
11. SRR17255044 (*Solanum L. X Solanum P.* #212)
12. SRR17255045 (*Solanum L. X Solanum P.* #282)
13. SRR17255046 (*Solanum L. X Solanum P.* #281)
14. SRR17255047 (*Solanum L. X Solanum P.* #280)
15. SRR17255048 (*Solanum L. X Solanum P.* #278)
16. SRR17255049 (*Solanum L. X Solanum P.* #277)
17. SRR17255076 (*Solanum P.* #5)
18. SRR17255083 (*Solanum L.* #5)
19. SRR17255087 (*Solanum L.* #1)
20. SRR17255092 (*Solanum L. X Solanum P.* #206)

The following samples were randomly selected for individual assemblies with metaSPAdes:

1. SRR17255001
2. SRR17255002
3. SRR17255025
4. SRR17255037
5. SRR17255087

Appendix B
Mathematical equations

The mathematical equations used to perform the linkage of SSU genes to bins are displayed here and explained in further detail. Three approaches were used in order to calculate an integrated score:

(I) Indirect association: Uses normalised abundance values of SSU genes and MAGs in samples to calculate a correlation (Pearson and Spearman). A regularized interdependence score is than calculated based on these correlations. This is done by multiplying the correlation values and removing the negative sign if needed:

$$Vreg(x,y) = Value[Ireg(x,y)] = abs(P(x,y)) * abs(S(x,y))$$

$$Sg(x,y) = sign[(I(x,y))] = \begin{cases} \text{-, any [P(x,y), S(x,y)]} < 0 \\ +, else \end{cases}$$

$$Ireg(x,y) = Vreg(x,y) * Sg(x,y)$$

Where:
P(x,y) = Pearsons correlation between SSU gene (x) and MAG (y).
S(x,y) = Spearmans correlation between SSU gene (x) and MAG (y).
I(x,y) = Integrated correlation between SSU gene (x) and MAG (y).

(II) Direct correlation with split mapping [M(x,y)]: Reads were mapped to both MAG and SSU sequences: This quantifies the fraction of reads in a MAG (y) that were aligned with an SSU sequence (x). The number of uniquely mapped reads in a MAG (y) that were also mapped to an SSU gene (x) [m(x,y)] was normalised by the total number of SSU reads mapping to MAG (y):

$$\sum_{i=1}^{n} m(x,i)$$

$$M(x,y) = \frac{m(x,y)}{\sum_{i=1}^{n} m(x,i)}$$

Where:
n = number of MAGs

(III) Blasting bin sequences to to SSU sequences [B(x,y)]: This quantifies the fraction of a MAG (y) that is present in an SSU gene (x). The total length of a MAG uniquely blasted to an SSU gene 'x'b [(x,y)] was normalized by the longest length of a MAG (y) blasted to an SSU gene (x), note that only blast hits of at least 100bp and 95% identity were included:

$$\max_{0<i\leq n} b(x,i)$$

The final score is then calculated with:

$$B(x,y) = \frac{b(x,y)}{\max_{0<i\leq n}^{b(x,i)}}$$

Where:
n = number of MAGs

The frequency of observations differs for each of the calculated scores. For instance, correlation based linkage is always possible to an extent but blasting may not yield a score in some cases. Therefore, it is needed to integrate the scores in a way that does not allow one of the scores to dominate the result. i.e taking the geometric mean of the scores:

$$Freg = 1 - \sqrt[3]{(1 - Ireg) * (1 - B) * (1 - M)}$$

Where:
Freg = combined score of MAG to SSU relationship.

Finally, The probability of a MAG (y) being linked to an SSU gene (x) is calculated, The Freg score is normalised against the background distribution of all possible MAG to SSU links:.

$$Pr(x,y) = \frac{Freg(x,y)}{\max_{0<i\leq m} Freg(i,y)} * \frac{Freg(x,y)}{\max_{0<j\leq n} Freg(x,j)}$$

Where:
n = number of mags.
m = number of SSU genes.

A score is calculated for each possible MAG to SSU pair. The highest scoring MAG to SSU pairs are used to assign a taxonomy to a MAG.

Appendix C

Quality control results table

Table C.1: Input quality control results.

Sample:	Pre qc reads:	Post qc reads:	% Passed:	Post CR reads:	% Phix DNA:	% Plant DNA:
SRR17255001	32943206	32511472	98.68945967	29122308	0	13.12017578
SRR17255002	35995004	35367350	98.25627468	32078500	0	12.20912449
SRR17255003	37958628	37396566	98.51927736	35713906	0	6.285288425
SRR17255004	47669098	47073402	98.75035185	41765360	0	14.13548931
SRR17255025	88338498	87127464	98.62909827	77331124	0	14.23407993
SRR17255037	76941788	75893182	98.63714371	69119310	0	11.31735545
SRR17255040	36492860	35786722	98.06499682	31536250	0	15.71718261
SRR17255041	45237464	44740796	98.902087	41451814	0	9.132652192
SRR17255042	50811204	50193206	98.78373675	48018866	0	5.815085263
SRR17255043	69394132	68718134	99.02585711	62209876	0	11.54841717
SRR17255044	43193442	42597210	98.61962378	39549326	0	9.214103927
SRR17255045	49633798	48947888	98.61805861	45683572	0	8.646928922
SRR17255046	43802082	43316782	98.8920618	39974514	0	9.575020724
SRR17255047	48603920	47980684	98.71772483	45946562	0	5.783583982
SRR17255048	42051996	41462684	98.59861111	39186332	0	7.312917167
SRR17255049	38610826	38115142	98.71620462	35747776	0	8.009029709
SRR17255076	78904264	77915538	98.74692957	71342530	0	10.59919518
SRR17255083	74254924	73211184	98.59438278	70278394	0	5.658253944
SRR17255087	77290856	76093054	98.45026687	73064520	0	5.784388921
SRR17255092	75606074	74276486	98.241427	69169894	0	9.304886314

Appendix D

Assembly result tables:

Table D.1: Results of the individual MetaSPAdes assemblies

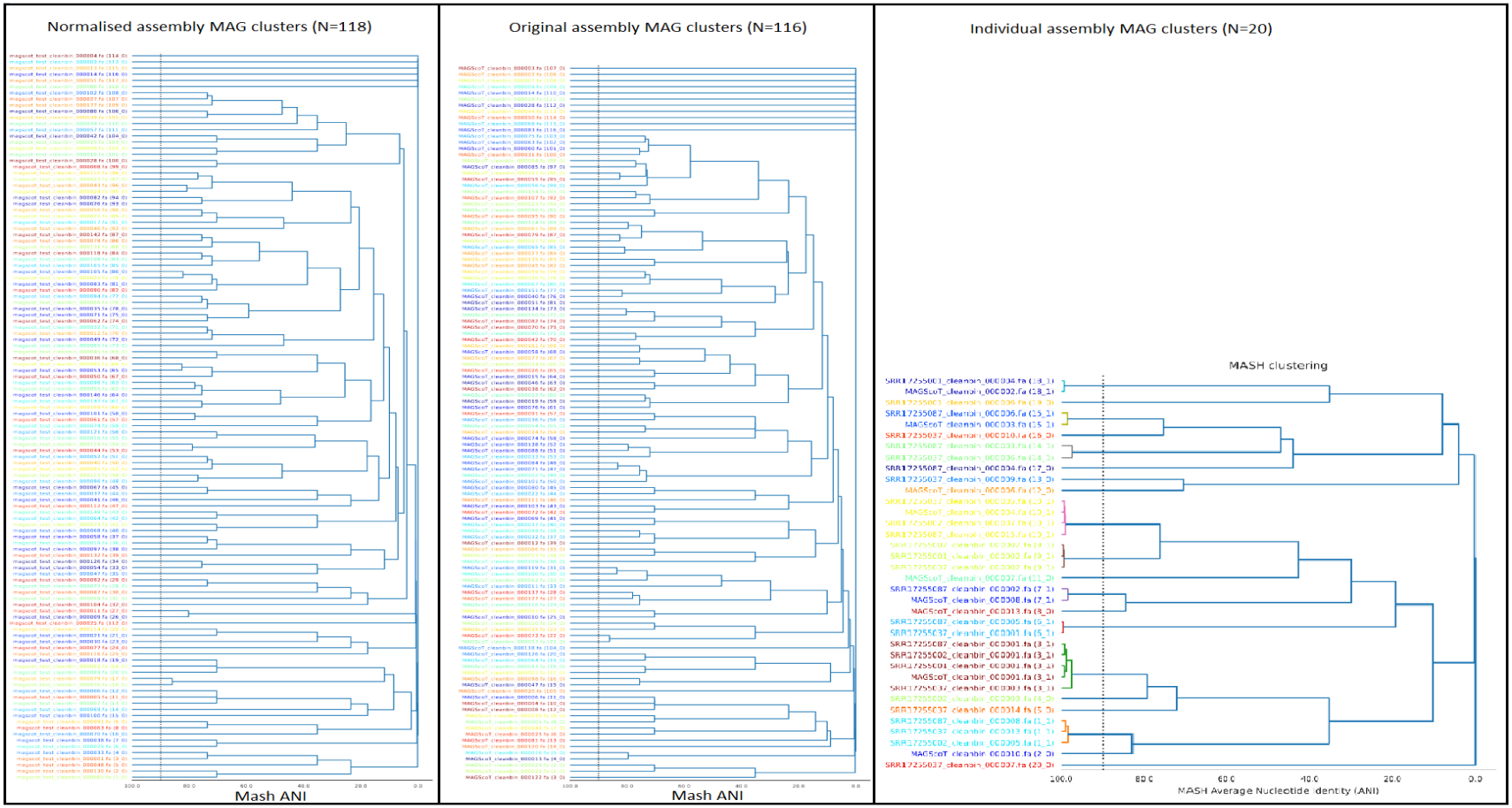
Sample	# Contigs:	Largest contig:	Total length:	% GC:	N50:	L50:	% Reads mapped:
SRR17255001	180649	211905	172969935	62.78	915	44953	58.53
SRR17255002	189675	211887	188628708	63.47	974	45361	59.20
SRR17255025	503483	335562	511353652	61.6	1026	121645	67.60
SRR17255037	452041	106322	484348725	61.55	1113	101181	71.70
SRR17255087	437436	236800	487590555	62	1187	91977	73.15

Table D.2: Results of the pooled MEGAHIT assemblies

Sample:	#Contigs:	Largest contig:	Total length:	% GC:	N50:	L50:	% Reads mapped:
Original pooled samples	5803284	779038	7070626611	62.80	1368	1104264	88.99
Normalised pooled samples	5159371	895074	6504501359	62.82	1451	950824	87.58

Appendix E

Dereplication cluster dendrograms



Appendix F
Workflow flowchart