Aidan Foo[1], Louise Cerdeira[2], Grant L. Hughes[1], Eva Heinz[3*]

Author affiliations

[1]Departments of Vector Biology and Tropical Disease Biology, Centre for Neglected Tropical Disease, Liverpool School of Tropical Medicine, Liverpool, UK

[2]Department of Vector Biology, Liverpool School of Tropical Medicine, Liverpool, UK

[3]Departments of Vector Biology and Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK

*Corresponding author: eva.heinz@lstmed.ac.uk

Title

Recovery of Metagenomic Data from the *Aedes aegypti* Microbiome using a Reproducible Snakemake Pipeline: MINUUR

Abstract

Background

Ongoing research of the mosquito microbiome aims to uncover novel strategies to reduce pathogen transmission. Sequencing costs, especially for metagenomics, are however still significant. A resource that is increasingly used to gain insights into host-associated microbiomes is the large amount of publicly available genomic data based on whole organisms like mosquitoes, which includes sequencing reads of the host-associated microbes and provides the opportunity to gain additional value of these initially host-focused sequencing projects.

Methods

To analyse non-host reads from existing genomic data, we developed a snakemake workflow called MINUUR (**M**icrobial **IN**sights **U**sing **U**nmapped **R**eads). Within MINUUR, reads derived from the host-associated microbiome were extracted and characterised using taxonomic classifications and metagenome assembly followed by binning and quality assessment. We applied this pipeline to five publicly available *Aedes aegypti* genomic datasets, consisting of 62 samples with a broad range of sequencing depths.

Results

We demonstrate that MINUUR recovers previously identified phyla and genera and is able to extract bacterial metagenome assembled genomes (MAGs) associated to the microbiome. Of these MAGS, 42 are high-quality representatives with >90% completeness and <5% contamination. These MAGs improve the genomic representation of the mosquito microbiome and can be used to facilitate genomic investigation of key genes of interest. Furthermore, we show that samples with a high number of KRAKEN2 assigned reads produce more MAGs.

Conclusions

Our metagenomics workflow, MINUUR, was applied to a range of *Aedes aegypti* genomic samples to characterise microbiome-associated reads. We confirm the presence of key mosquito-associated symbionts that have previously been identified in other studies and recovered high-

quality bacterial MAGs. In addition, MINUUR and its associated documentation are freely available on GitHub (https://github.com/aidanfoo96/MINUUR) and provide researchers with a convenient workflow to investigate microbiome data included in the sequencing data for any applicable host genome of interest.

Keywords

Mosquito, *Aedes aegypti,* Microbiome, Metagenomics, Snakemake, MAGs, Read Classification.

54 **Introduction**

55 *Aedes aegypti* is an important vector of human pathogens including dengue virus (DENV), yellow
56 fever virus, chikungunya virus and Zika virus. DENV cases alone are estimated to cause 10,000
57 deaths and 100 million infections per year, contributing to significant burden of human morbidity
58 and mortality worldwide (1). Interest in the mosquito microbiome has emerged due to evidence of
59 its influence in vectorial capacity (2,3), offering potential for novel approaches to reduce pathogen
60 transmission from mosquitoes to vertebrate hosts (2,4–7).

61 Mosquito microbiomes are highly variable, dependent on multiple deterministic processes such
62 as the environment (8–12), season (13), host factors (14,15), microbial interactions (16–18) and
63 mosquito-microbe interactions (19–23). These findings from mosquito microbiome studies are
64 largely driven by amplicon based 16S rRNA sequencing approaches (18,24,25). Metagenomic
65 approaches for mosquito microbiome characterisation are limited in number (26,27), but would
66 facilitate our understanding by providing the genomic context of symbionts (28). An attractive
67 resource for gaining additional insights into microbiomes is to make use of the microbiome reads
68 derived from whole genome sequencing (WGS), especially in cases where preparation of the host
69 for sequencing included its associated microbiome. Studies in *Drosophila*, bumble bees, moths
70 and nematodes have shown existing WGS data is a rich source to characterize associated
71 symbionts (29–34). Whilst some studies include specific enrichment of non-host with bait
72 sequences targeting a specific taxon of interest (29,32); it is also possible to assess microbes
73 present in the sequencing experiment without prior enrichment, taking into account that in the
74 latter case the microbiome recovered is likely a biased representation and lack of presence does
75 not prove absence.

76 Whole genome shotgun sequencing is commonly used to study mosquito genomics at the
77 individual (35,36) and population (37,38) level; meaning non-mosquito sequence data (we refer
78 to these as non-reference reads, since they do not map to the reference genome of interest) are

a source to identify mosquito microbiome members using metagenomics. Genomic surveillance programs such as the *Anopheles gambiae* 1000 Genomes Project contain a large number of genomic samples with each release (39) and, at time of writing, currently 100,514 *Aedes aegypti* whole-genome sequencing runs are deposited on the European Nucleotide Archive. As such, there is great potential to leverage existing mosquito WGS data to explore members of their microbiomes from non-reference reads.

To make use of this resource and streamline future large-scale analysis of mosquito metagenomes, we developed a Snakemake pipeline to provide **M**icrobial **IN**sight **U**sing **U**nmapped **R**eads from WGS data. MINUUR uses short read, whole genome sequencing data as input and performs an analysis of non-reference reads derived from a host sequencing experiment. We used MINUUR to study five published *Aedes aegypti* sequencing projects (in total 124 FASTQ files) currently available on the European Nucleotide Archive. We gained insight of associated microbes based on taxonomic read classifications and recover high-quality metagenome assembled genomes (MAGs) of mosquito-associated bacteria. To assess the suitability of samples for MAG recovery, we also investigated patterns between KRAKEN2 assigned reads and the number of MAGs within samples. We show samples with taxa assigned high numbers of KRAKEN2 classified reads produce relatively more high and medium quality MAGs. The pipeline is open source and available on GitHub (https://github.com/aidanfoo96/MINUUR) with an accompanying JupyterBooks page (https://aidanfoo96.github.io/MINUUR/).

Materials and Methods

Specifications

102 To undertake this analysis, we developed a metagenomics workflow called MINUUR (**M**icrobial
103 **IN**sights Using Unmapped Reads), using the workflow manager Snakemake (40) (Figure 1). This
104 pipeline was developed to facilitate the following analysis and future studies aiming to
105 characterise non-reference reads in mosquitoes or other organisms. A breakdown of the pipeline
106 that produced each result of this study is provided in the following section, with details of how
107 each step was configured in the final section.
108

109 **Database Setup**

110 MINUUR requires several databases to perform the analysis. This includes a high quality bowtie2-
111 indexed reference genome (41) to separate host and non-host reads, and a KRAKEN2 (42),
112 BRACKEN (43) and MetaPhlAn3 (44) database for taxonomic read classifications. All databases
113 are available in their respective GitHub repositories. The databases used in this study are the
114 default MetaPhlAn3 marker gene database, KRAKEN2 and BRAKEN indexes from the Ben
115 Langmead repository located here: https://benlangmead.github.io/aws-indexes/k2. For our study,
116 we downloaded and compiled these default databases on the 08/09/2022.

117 **Data Preparation**

118 MINUUR accepts either BAM or paired FASTQ inputs. For FASTQ inputs, MINUUR performs
119 quality control (QC) using FASTQC (v0.11.9) (45), providing a QC report per sample. MINUUR
120 does not use the FASTQC report in subsequent steps, but only as a quality assurance metric for
121 the user and to estimate if read trimming is required. Read trimming can be performed within
122 MINUUR using Cutadapt (v1.5) (46) with user defined parameters for minimum read length, base
123 quality and adapter content (default: minimum base length = 50, average base quality = 30). To
124 separate host and non-host, reads are aligned to a user defined indexed reference genome (the
125 relevant host genome) using Bowtie2 (v2.4.4) (41). Alignment sensitivity and type can be adjusted

126 within the pipeline at the user's discretion. A high quality, chromosome level assembled, reference
127 genome is recommended if available. In situations where this is not possible, users should be
128 aware that read alignment will likely result in mismatches between the reference and target
129 sequence and produce alignments with poor coverage (47). As a result, non-reference reads used
130 in subsequent steps are likely to contain a substantial number of host data. In this instance, we
131 included a feature within MINUUR to extract KRAKEN2 assigned reads pertaining to known
132 microbes and potentially improve metagenome assemblies. Non-reference reads within the
133 coordinate sorted binary alignment (BAM) file are extracted using samtools using the command
134 "samtools -view -f 4" (v.1.14) (48) and converted to FASTQ format using bedtools (v2.3.0) (49)
135 ("–bamToFastq"). As users might already have their data in BAM format mapped against the host
136 reference (e.g. in large-scale sequencing projects like the Ag1000G), we also included the option
137 of a BAM input. Here, any non-reference reads within the BAM file will be extracted, converted to
138 FASTQ and used in downstream steps.

139 For this study, we used five published genome datasets of *Aedes aegypti* publicly available on
140 the European Nucleotide Archive (35,37,50–53). We selected the data sets to cover a range of
141 sequencing depths, DNA extraction method and sequencing platform (Figure 2C). All raw FASTQ
142 files of published sequencing data were retrieved from the European Nucleotide Archive (ENA)
143 under the project accession numbers PRJEB33044 (37), PRJNA255893 (50), PRJNA385349
144 (51), PRJNA718905 (35), PRJNA776956 (52) and PRJNA992905 (53).

145 **Read Classification**

146 MINUUR uses two read classification approaches to infer taxonomy. KRAKEN2 (v2.1.2) (42),
147 which uses a k-mer based approach to map read fragments of k-length against a taxonomic
148 genome library of k-mer sequences, and MetaPhlAn3 (v3.0.13) (44) to align reads against a
149 library of marker genes using Bowtie2 (41). MINUUR also provides the option to use KRAKEN2

classified reads, parsed from KrakenTools (v1.2), to select a specific set of reads (for example bacterial) for metagenome assembly. To estimate the relative taxonomic abundance from KRAKEN2 classifications, MINUUR will parse KRAKEN2 read classifications to BRACKEN (v2.6.2) (43) which uses a Bayesian probability approach to redistribute reads assigned at higher taxonomic levels to lower (species) taxonomic levels.

MINUUR outputs classified and unclassified reads to paired FASTQ files and generates BRACKEN estimated taxonomic abundance profiles for further analysis. An additional feature we included within MINUUR is the option to extract a specific taxon or group of taxa from KRAKEN2, using the program KrakenTools (42). This option is useful in situations where a specific group of taxa are of interest or to exclude groups of taxa such as viral or archaeal reads. Alternatively, if alignment to a reference is poor, this option can be used to remove host reads that did not map to its reference.

Metagenome Assembly, Binning and Quality Assurance

MINUUR uses the non-reference reads to perform *de novo* metagenome assemblies (the same reads used for KRAKEN2 and MetaPhlAn3 taxonomic classifications). Reads are parsed to MEGAHIT (v1.2.9) (54), a rapid and memory efficient metagenome assembler, for *de novo* metagenome assembly. Assembled contigs are quality checked using QUAST (v5.0.2) (55) to assess contig N50 and L50 scores. The resultant contigs, which are fasta files with sequences pertaining to genomic regions of a microbe, need to be placed within defined taxonomic groups - referred to as a bin. For this, contigs are indexed using the Burrows Wheeler Aligner (BWA) (v0.7.17) (56), and the original non-reference reads are aligned to the indexed contigs using ”–bwa-mem”. The subsequent coordinate sorted BAM file is parsed to the “jgi_summarize_bam_contig_depth” script from MetaBAT2 (v2.12.1) (57) to produce a depth file of contig coverage. The depth file and assembled contigs are input to the metagenome binner

175 MetaBAT2 (v2.12.1) (57), to group contigs within defined genomic bins. Each bin is a predicted
176 metagenome assembled genome (MAG). CheckM (v1.1.3) (58) is then used for quality assurance
177 of each bin by identifying single copy core genes. Specifically, bin contamination is assessed by
178 looking for one single copy core gene within each bin, and completeness by calculating a required
179 set of single copy core genes.

180 **Pipeline Configuration**

181 All sample names were listed in the samples.tsv file of the configuration directory and paths to
182 the FASTQ directories given in the samples_table.tsv file. To implement the pipeline, the
183 configuration file was set to the following parameters: FASTQ = True, QC = True,
184 CutadaptParams = ”–minimum-length 50 -q 30”, RemoveHostFromFastqGz = True,
185 AlignmentSensitivity = ”–sensitive-local”, ProcessBam = True, From-Fastq = True,
186 KrakenClassification = True, ConfidenceScore = 0, KrakenSummaries = True,
187 GenusReadThreshold = 1000, SpeciesReadThreshold = 30000, MetagenomeAssm = True,
188 MetagenomeBinning = True, MinimumContigLength = 1500, CheckmBinQA = True. All databases
189 were installed from their respective repositories on 08/09/22. The pipeline was run on an Ubuntu
190 Linux system with 660gb of available memory and 128 CPUs. For our analysis, with the above
191 settings and 10 cores available, runtime was two weeks; the maximum Resident Set Size (RSS)
192 of an individual sample during this run was 9771 RSS (occurring during metagenome assembly);
193 and total storage used (including temporary files) was 5.18Tb (terabytes) across all samples used
194 in this study.

195 **Taxonomic Classification of MAGs with GTDB-Tk**

196 Separate from MINUUR, all bins produced from MetaBAT2 were taxonomically classified with
197 GTDB-Tk (59) (v2.1.1) using ”–classify-wf” against the Genome Taxonomy Database (GTDB)

198 (release 07-R207 08/04/22, downloaded on 08/09/22). GTDB-Tk assigns genes to MAGs using
199 Prodigal (v2.6.3) (60) and ranks the taxonomic domain of each MAG using a database of 120
200 bacteria and 122 archaea marker genes (61) using HMMER3 (62). With this information, MAGs
201 are then placed into domain specific reference trees with pplacer (v1.1) (63). Taxonomic
202 classifications with GTDB-Tk are based on placement within the GTDB reference tree, relative
203 evolutionary divergence, and average nucleotide identity (ANI) scores with its closest reference
204 genome. The relative evolutionary divergence score is used to refine ambiguous taxonomic rank
205 assignments and ANI scores are used to define species classifications (59).

206 **Results**

207 **Extraction of Non-Reference Reads Post-Alignment to *Aedes aegypti***

208 In total, we retrieved 62 samples (124 FASTQ files) across six sequencing experiments (Figure
209 2A) and parsed them through MINUUR. After alignment to the *Aedes aegypti* reference genome
210 (AaegL5.3) (64) with Bowtie2, the proportion of mapped and non-reference reads were calculated.
211 Of our initial 62 samples, on average, 176,310,722 reads aligned to the AaegL5.3 reference
212 genome, ranging between 20,655,064 and 543,378,958 reads (Figure 2B). To estimate the
213 number of reads associated to the microbiome among the non-reference reads, the overall
214 number of KRAKEN2 classifications were counted. The average proportion of KRAKEN2
215 classified reads from all non-reference reads was 17.9%. The number and proportion of
216 KRAKEN2 classifications per sample is given in Supplementary Table 1 within the extended data.
217

218 **Taxonomic Classifications Produced from Non-Reference Reads of *Aedes aegypti***

219 The KRAKEN2 classifications of non-reference reads were counted from each sequencing
220 project. Four out of the six sequencing projects showed a much lower number of classifications
221 compared to sequencing projects PRJNA776596 and PRJEB33044 (Figure 3A). PRJEB33044

222 gave the highest average proportion of taxonomic classification (81.3%), while PRJNA255893
223 was the lowest (0.893%).
224
225 Multiple phyla were identified across sequencing projects. *Bacteroidetes, Proteobacteria,*
226 *Firmicutes and Actinobacteria* were the most common phylum present across all samples, with
227 varied relative abundance between projects (Figure 3C). For example, *Bacteroidetes* and
228 *Proteobacteria* were dominant in PRJEB33044 and PRJNA882905, *Firmicutes* and
229 *Proteobacteria* in PRJNA255893, and *Bacteroidetes, Actinobacteria* and *Proteobacteria* in
230 PRJNA385349 and PRJNA718905. At the genera level, there are several dominant members
231 across samples, including *Wolbachia, Pseudomonas, Serratia* and *Elizabethkingia.* Generally,
232 however, there is considerable variation at the genus level within and between sequencing
233 experiments (Figure 3C). We summarised all KRAKEN2 classifications across 62 samples. The
234 most common genera, identified in >50% of all samples, were *Pseudomonas, Elizabethkingia,*
235 *Clostridium, Bacillus* and *Chryseobacterium.* Genera identified in 25-50% of samples were
236 *Acinetobacter*, *Enterobacter, Serratia* and *Delftia* (Figure 3B). All KRAKEN2 taxonomic
237 classifications are given in Supplementary Table 2.
238
239 **Metagenome Assembly and Binning**
240 We used MINUUR to further parse non-reference reads through assembly, binning and quality
241 assurance steps, with the aim to recover MAGs associated to the *Aedes aegypti* microbiome.
242 Assembly was conducted with MEGAHIT and binning with MetaBat2. We used CheckM to assess
243 MAG completeness and contamination through the presence and copy number of single copy
244 core genes, and recovered 105 MAGs (Figure 4A). Using the standards of MAG quality set by the
245 genome standards consortium (GSC) (65), 42 MAGs met the criteria of high quality with
246 completeness >90% and contamination <5%; and 20 MAGs classify as medium quality with
247 completeness >50% and contamination <5%. Overall, 62 high and medium quality MAGs were

248 recovered. The remaining MAGs were low-quality (<50% complete and <10% contamination) or
249 contaminated >10%, and therefore excluded from MAG classification. Of the high and medium-
250 quality MAGs, the mean N50 (the minimum contig length of an assembled contig that covers 50%
251 of the genome) was 177.2 kilobase pairs (KBP), ranging between 4.72KBP and 471.6KBP (Figure
252 4B). The average genome size was 3.36 megabase pairs (MBP), ranging between 1.07MBP and
253 6.19MBP (Figure 4C), while low-quality MAGs showed a wider range of genome sizes between
254 0.24MBP and 106MBP (Supplementary Figure 1).
255
256 **Relationship Between KRAKEN2 Taxonomic Classifications and MAG Recovery**
257 The suitability of a sample for MAG recovery would be of interest to estimate in advance, and we
258 investigated if there was a correlation between the proportion of KRAKEN2 classifications from
259 non-reference reads and MAG recovery. In one project, PRJNA255893, we recovered no high-
260 quality MAGs, (Figure 5A) and the reads used for assembly contained no taxa exceeding 100,000
261 KRAKEN2 assigned reads (Figure 5B). In contrast, PRJNA33044, PRJNA776596 and
262 PRJNA882905 allowed retrieval of more high and medium quality MAGs, and within these
263 projects, multiple samples contained >100,000 reads assigned to a taxon (Figure 5A, B). As such,
264 MAG recovery is likely linked to a sufficient number of classified reads to a taxon (Figure 5A,
265 Figure 5B), rather than overall number of KRAKEN2 assigned reads within a sample
266 (Supplementary Figure 2). In accordance with this, the total number of classified taxa totaling or
267 greater than 100,000 KRAKEN2 assigned reads for PRJEB33044 is 33 taxa (summed across all
268 samples) and 22 high and medium quality MAGs were recovered from this sample. Similarly,
269 PRJNA776956 shows 19 taxa with associated reads totaling or greater than 100,000 KRAKEN2
270 assigned reads, resulting in 16 high and medium-quality MAGs. Furthermore, taxonomic
271 classifications with a high number of KRAKEN2 assigned reads (>100,000) are akin to the
272 taxonomic classifications of medium and high-quality MAGs (Supplementary Figure 3). There are,
273 however, exceptions to these patterns; PRJNA385349 contained five taxa with over 100,000

274 KRAKEN2 assigned reads, yet no high-quality MAGs could be recovered from this project. As
275 such, applicable for future application of this approach, the number of KRAKEN2 classifications
276 assigned to a specific taxon is one factor that can help estimate high-quality MAG recovery.
277
278 **Taxonomic Classification of MAGs with GTDB-Tk**
279 Following MAG recovery using MINUUR, we used the taxonomic classifier GTBD-Tk to classify
280 high and medium quality MAGs against the Genome Taxonomy Database (GTDB). We compared
281 the genome size of each MAG to its closest reference genome in the GTDB (Figure 6A, Figure
282 6B). Of the high-quality MAGs, these were larger than their reference genome by mean = 183kb,
283 whereas medium quality MAGs deviated from their reference genomes by 1768kb (Figure 6A).
284 Congruent with pairwise size differences between MAG and reference genome, we found the
285 overall distribution of high-quality MAGs compared to their reference genome size to be similar,
286 but significantly different between medium-quality MAGs (Figure 6B). Of these, 48 MAGs were
287 classified to the species level with a mean FastANI score of 98.4%, ranging between 95.6% and
288 100%. No MAGs were identified <95% ANI to a known species, indicating no undescribed *Aedes*
289 *aegypti* associated bacterial species were present within these MAGs (Figure 6C).
290
291 **Discussion**
292 Metagenomic datasets of mosquito microbiomes are so far limited in number (66). In this study,
293 we developed a metagenomics like workflow called MINUUR to facilitate recovery and use of
294 host-associated bacterial sequences using non-reference reads from existing host WGS projects.
295 We demonstrate that MINUUR can be used to recover genus level taxonomic classifications and
296 draft high and medium quality MAGs from host WGS projects. The recovery of metagenomic
297 information, such as MAGs, are applied in large scale metagenomic studies from chickens (67),
298 humans (68) to cows (69–71), with these studies yielding between 400 to 92,000 MAGs per study.
299 We apply a similar approach with non-reference *Aedes aegypti* sequencing reads across a range

300 of different studies and can demonstrate that using MINUUR expands the genomic representation
301 of known mosquito-associated bacterial symbionts. Overall, these provide a valuable resource for
302 researchers in the field and can be used in further work such as facilitating biosynthetic gene
303 cluster discovery (71) or to identify genetic targets in symbiont pathogen blocking approaches (2).
304
305 The data retrieved in this study agrees with published insights; the phylum level classifications
306 are consistent with findings from other mosquito microbiome studies (8,9,11,18,72,73), showing
307 that *Proteobacteria, Bacteroidetes* and *Firmicutes* are dominant phyla of the mosquito
308 microbiome; and our taxonomic classifications highlight the inherent variability of the *Aedes*
309 *aegypti* microbiome (18,24,74). These findings give us confidence that taxonomic classifications
310 with KRAKEN2, within the pipeline, can accurately predict the presence of microbes associated
311 to the *Aedes aegypti* microbiome from non-reference sequences. At the genus level, we also find
312 consistent observations of taxa previously identified in other studies (16,19,23,75,76). The
313 KRAKEN2 classifications show within the two most common phyla, *Proteobacteria* and
314 *Bacteroidetes*; *Elizabethkingia, Pseudomonas* and *Serratia* are the most common. All three of
315 these symbionts are documented to play key roles in either blood digestion (19,77), iron-
316 acquisition (75) and microbial interactions (16). Notably *Elizabethkingia* has previously been
317 implicated in responses to iron fluxes in *Anopheles gambiae* (75), and blood meals in *Aedes*
318 *albopictus* (25) and *Aedes aegypti* (24). Similarly, *Pseudomonas* has shown to interact with
319 *Elizabethkingia*, triggering the expression of *hemS* to break down heme into biliverdin catabolites
320 (17). It is encouraging to note the presence of these two bacteria within our taxonomic
321 classifications. We recovered high and medium quality MAGs associated to these taxa (*Serratia,*
322 *Elizabethkingia* and *Pseudomonas*), which should allow further interrogation of genes associated
323 to these biological processes. To note, the presence of *Wolbachia* in the projects we have
324 analysed is expected since these mosquitoes were transinfected with high titers of this bacterium,
325 further validating the pipeline results (37).

326 Whilst the nature of the samples can act as confounder given differences in sample handling
327 preparation, we note the relative abundance of the key phyla, and constituent genera, are varied
328 across these projects. Putative biological causes of this variation are likely multifactorial,
329 supported by studies showing environment (9,15), host genetics (23) and competitive
330 mechanisms amongst bacteria (16) to be influential for bacterial colonization in the mosquito.

331 **Conclusion**

332 In summary, we present a reproducible workflow to analyse host-associated microbial sequence
333 data derived from host WGS experiments, leveraging a vast resource of data for additional
334 insights. Our work focuses on the mosquito microbiome, where future considerations and
335 prospects were recently established by the Mosquito Microbiome Consortium (66). A key point
336 highlighted in this statement was the need for (meta)genomics approaches with solid
337 reproducibility for data analysis within the field. Our pipeline provides a workflow to assess non-
338 host reads from existing mosquito genome sequence data and increases our knowledge of
339 mosquito-associated bacterial genomes. This approach and accompanying workflow will facilitate
340 more analyses of existing WGS data within *Aedes aegypti* and other organisms of interest for the
341 scientific community.

342

343 **Data and Software Availability**

344 All original sequencing projects can be accessed under the following accession numbers:
345 PRJEB33044 (https://www.ebi.ac.uk/ena/browser/view/PRJEB33044), PRJNA255893
346 (https://www.ebi.ac.uk/ena/browser/view/PRJNA255893), PRJNA385349
347 (https://www.ebi.ac.uk/ena/browser/view/PRJNA385349), PRJNA718905
348 (https://www.ebi.ac.uk/ena/browser/view/PRJNA718905), PRJNA776956

(https://www.ebi.ac.uk/ena/browser/view/PRJNA776956), PRJNA882905 (https://www.ebi.ac.uk/ena/browser/view/PRJNA882905). The source code for our workflow, MINUUR, is available here: https://github.com/aidanfoo96/MINUUR with an accompanying jupyter books page to run the analysis available here: https://aidanfoo96.github.io/MINUUR/.

Extended Data

Supplementary Figures and Tables are available in the FigShare repository under the title "Recovery of Metagenomic Data from the *Aedes aegypti* Microbiome Using a Reproducible Snakemake Pipeline" and can be accessed using the following URL: https://figshare.com/projects/Recovery_of_Metagenomic_Data_from_the_Aedes_aegypti_Microbiome_using_a_Reproducible_Snakemake_Pipeline/158210. This repository contains the following data:

Supplementary Table 1: KRAKEN2 Summary Report

Supplementary Table 2: KRAKEN2 Per Sample Genera Classifications

Supplementary Figure 1: Genome Size Comparison Between High, Medium and Low-Quality MAGs

Supplementary Figure 2: Correlation Between Classified Read Number and Number of MAGs

Supplementary Figure 3: GTDB Classifications vs KRAKEN2 Assigned Taxa > 100,000 Reads

Author Contributions

AF, EH, GLH conceived the project. AF and EH designed the methodology. AF performed the analyses and wrote the pipeline. LC provided technical expertise. EH and GLH provided oversight throughout the project. AF and EH drafted the manuscript, and all authors contributed to and approved the final version.

Grant Information

AF was supported by the DTP scholarship (Medical Research Council MR/N013514/1). GLH was supported by the BBSRC (BB/T001240/1, BB/V011278/1, and BB/W018446/1), the UKRI (20197 and 85336), the EPSRC (V043811/1), a Royal Society Wolfson Fellowship (RSWF\R1\180013), the NIHR (NIHR2000907) and the Bill and Melinda Gates Foundation (INV-048598). EH acknowledges funding from Wellcome (217303/Z/19/Z) and the BBSRC (BB/V011278/1).

References

1. Messina JP, Brady OJ, Pigott DM, Brownstein JS, Hoen AG, Hay SI. A global compendium of human dengue virus occurrence. Sci Data. 2014 Dec;1(1):140004.

2. Cansado-Utrilla C, Zhao SY, McCall PJ, Coon KL, Hughes GL. The microbiome and mosquito vectorial capacity: rich potential for discovery and translation. Microbiome. 2021 Dec;9(1):111.

3. Utarini A, Indriani C, Ahmad RA, Tantowijoyo W, Arguni E, Ansari MR, et al. Efficacy of *Wolbachia*-Infected Mosquito Deployments for the Control of Dengue. N Engl J Med. 2021 Jun 10;384(23):2177–86.

4. Scolari F, Casiraghi M, Bonizzoni M. *Aedes* spp. and Their Microbiota: A Review. Front Microbiol. 2019 Sep 4;10:2036.

5. Steven B, Hyde J, LaReau JC, Brackney DE. The Axenic and Gnotobiotic Mosquito: Emerging Models for Microbiome Host Interactions. Front Microbiol. 2021 Jul 12;12:714222.

6. Bongio NJ, Lampe DJ. Inhibition of *Plasmodium berghei* Development in Mosquitoes by Effector Proteins Secreted from *Asaia* sp. Bacteria Using a Novel Native Secretion Signal. PLOS ONE. 2015 Dec 4;10(12):e0143541.

7. Romoli O, Gendrin M. The tripartite interactions between the mosquito, its microbiota and *Plasmodium*. Parasit Vectors. 2018 Dec;11(1):200.

8. Dada N, Jumas-Bilak E, Manguin S, Seidu R, Stenström TA, Overgaard HJ. Comparative assessment of the bacterial communities associated with *Aedes aegypti* larvae and water from domestic water storage containers. Parasit Vectors. 2014;7(1):391.

9. David MR, Santos LMB dos, Vicente ACP, Maciel-de-Freitas R. Effects of environment, dietary regime and ageing on the dengue vector microbiota: evidence of a core microbiota throughout *Aedes aegypti* lifespan. Mem Inst Oswaldo Cruz. 2016 Aug 25;111(9):577–87.

10. Saab SA, Dohna H zu, Nilsson LKJ, Onorati P, Nakhleh J, Terenius O, et al. The environment and species affect gut bacteria composition in laboratory co-cultured *Anopheles gambiae* and *Aedes albopictus* mosquitoes. Sci Rep. 2020 Dec;10(1):3352.

11. Onyango GM, Bialosuknia MS, Payne FA, Mathias N, Ciota TA, Kramer DL. Increase in temperature enriches heat tolerant taxa in *Aedes aegypti* midguts. Sci Rep. 2020 Dec;10(1):19135.

12. Wang Y, Gilbreath TM, Kukutla P, Yan G, Xu J. Dynamic Gut Microbiome across Life History of the Malaria Mosquito *Anopheles gambiae* in Kenya. PLoS ONE. 2011 Sep 21;6(9):e24767.

13. Sandeu MM, Maffo CGT, Dada N, Njiokou F, Hughes GL, Wondji CS. Seasonal variation of microbiota composition in *Anopheles gambiae* and *Anopheles coluzzii* in two different eco-geographical localities in Cameroon. Med Vet Entomol. 2022 Sep;36(3):269–82.

14. Kakani P, Gupta L, Kumar S. Heme-Peroxidase 2, a Peroxinectin-Like Gene, Regulates Bacterial Homeostasis in *Anopheles stephensi* Midgut. Front Physiol. 2020 Sep 8;11:572340.

15. Minard G, Tran FH, Tran Van V, Fournier C, Potier P, Roiz D, et al. Shared larval rearing environment, sex, female size and genetic diversity shape *Ae. albopictus* bacterial microbiota. PLOS ONE. 2018 Apr 11;13(4):e0194521.

16. Kozlova EV, Hegde S, Roundy CM, Golovko G, Saldaña MA, Hart CE, et al. Microbial interactions in the mosquito gut determine *Serratia* colonization and blood-feeding propensity. ISME J. 2021 Jan;15(1):93–108.

17. Ganley JG, D'Ambrosio HK, Shieh M, Derbyshire ER. Coculturing of Mosquito-Microbiome Bacteria Promotes Heme Degradation in *Elizabethkingia anophelis*. ChemBioChem. 2020 May 4;21(9):1279–84.

18. Hegde S, Khanipov K, Albayrak L, Golovko G, Pimenova M, Saldaña MA, et al. Microbiome Interaction Networks and Community Structure From Laboratory-Reared and Field-Collected *Aedes aegypti*, *Aedes albopictus*, and *Culex quinquefasciatus* Mosquito Vectors. Front Microbiol. 2018 Sep 10;9:2160.

19. Heu K, Romoli O, Schönbeck JC, Ajenoe R, Epelboin Y, Kircher V, et al. The Effect of Secondary Metabolites Produced by *Serratia marcescens* on *Aedes aegypti* and Its Microbiota. Front Microbiol. 2021 Jul 7;12:645701.

20. Mitri C, Bischoff E, Belda Cuesta E, Volant S, Ghozlane A, Eiglmeier K, et al. Leucine-Rich Immune Factor APL1 Is Associated With Specific Modulation of Enteric Microbiome Taxa in the Asian Malaria Mosquito *Anopheles stephensi*. Front Microbiol. 2020 Feb 26;11:306.

21. Vogel KJ, Valzania L, Coon KL, Brown MR, Strand MR. Transcriptome Sequencing Reveals Large-Scale Changes in Axenic *Aedes aegypti* Larvae. PLoS Negl Trop Dis. 2017 Jan 6;11(1):e0005273.

22. Short SM, Mongodin EF, MacLeod HJ, Talyuli OAC, Dimopoulos G. Amino acid metabolic signaling influences *Aedes aegypti* midgut microbiome variability. PLoS Negl Trop Dis. 2017 Jul 28;11(7):e0005677.

23. Stathopoulos S, Neafsey DE, Lawniczak MKN, Muskavitch MAT, Christophides GK. Genetic Dissection of *Anopheles gambiae* Gut Epithelial Responses to *Serratia marcescens*. PLoS Pathog. 2014 Mar 6;10(3):e1003897.

24. Muturi EJ, Dunlap C, Ramirez JL, Rooney AP, Kim CH. Host blood meal source has a strong impact on gut microbiota of *Aedes aegypti*. FEMS Microbiol Ecol. 2019 Jan 1;95(1).

25. Chen S, Zhang D, Augustinos A, Doudoumis V, Bel Mokhtar N, Maiga H, et al. Multiple Factors Determine the Structure of Bacterial Communities Associated With *Aedes albopictus* Under Artificial Rearing Conditions. Front Microbiol. 2020 Apr 15;11:605.

26. Zhao T, Li B qi, Gao H ting, Xing D, Li M jin, Dang Y qi, et al. Metagenome Sequencing Reveals the Microbiome of *Aedes albopictus* and Its Possible Relationship With Dengue Virus Susceptibility. Front Microbiol. 2022 May 11;13:891151.

27. Wang Y ting, Shen R xin, Xing D, Zhao C pei, Gao H ting, Wu J hong, et al. Metagenome Sequencing Reveals the Midgut Microbiota Makeup of *Culex pipiens quinquefasciatus* and Its Possible Relationship With Insecticide Resistance. Front Microbiol. 2021 Feb 25;12:625539.

28. Pérez-Cobas AE, Gomez-Valero L, Buchrieser C. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. Microb Genomics. 2020 Aug 1;6(8).

29. Hooper R, Brealey JC, Valk T, Alberdi A, Durban JW, Fearnbach H, et al. Host-derived population genomics data provides insights into bacterial and diatom composition of the killer whale skin. Mol Ecol. 2019 Jan;28(2):484–502.

30. Ghanavi HR, Twort VG, Duplouy A. Exploring bycatch diversity of organisms in whole genome sequencing of *Erebidae* moths (*Lepidoptera*). Sci Rep. 2021 Dec;11(1):24499.

31. LaBonte NR, Jacobs J, Ebrahimi A, Lawson S, Woeste K. Data mining for discovery of endophytic and epiphytic fungal diversity in short-read genomic data from deciduous trees. Fungal Ecol. 2018 Oct;35:1–9.

481 32. Salzberg SL, Hotopp JCD, Delcher AL, Pop M, Smith DR, Eisen MB, et al.
482 Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila* species.
483 Genome Biol. 2005;8.

484 33. Martinson VG, Magoc T, Koch H, Salzberg SL, Moran NA. Genomic Features of a
485 Bumble Bee Symbiont Reflect Its Host Environment. Appl Environ Microbiol. 2014
486 Jul;80(13):3793–803.

487 34. Fierst JL, Murdock DA, Thanthiriwatte C, Willis JH, Phillips PC. Metagenome-
488 Assembled Draft Genome Sequence of a Novel Microbial *Stenotrophomonas*
489 *maltophilia* Strain Isolated from *Caenorhabditis remanei* Tissue. Genome Announc.
490 2017 Feb 16;5(7).

491 35. Chen C, Compton A, Nikolouli K, Wang A, Aryan A, Sharma A, et al. Marker-
492 assisted mapping enables effective forward genetic analysis in the arboviral vector
493 *Aedes aegypti*, a species with vast recombination deserts. Genetics; 2022 Nov
494 1;222(3).

495 36. Crava C, Varghese FS, Pischedda E, Halbach R, Palatini U, Marconcini M, et al.
496 Immunity to infections in arboviral vectors by integrated viral sequences: an
497 evolutionary perspective. http://biorxiv.org/lookup/doi/10.1101/2020.04.02.022509

498 37. Ford SA, Allen SL, Ohm JR, Sigle LT, Sebastian A, Albert I, et al. Selection on
499 *Aedes aegypti* alters *Wolbachia*-mediated dengue virus blocking and fitness. Nat
500 Microbiol. 2019 Nov;4(11):1832–9.

501 38. Faucon F, Dusfour I, Gaude T, Navratil V, Boyer F, Chandre F, et al. Identifying
502 genomic changes associated with insecticide resistance in the dengue mosquito
503 *Aedes aegypti* by deep targeted sequencing. Genome Res. 2015 Sep;25(9):1347–
504 59.

505 39. The Anopheles gambiae 1000 Genomes Consortium. Genome variation and
506 population structure among 1142 mosquitoes of the African malaria vector species
507 *Anopheles gambiae* and *Anopheles coluzzii*. Genome Res. 2020 Oct;30(10):1533–
508 46.

509 40. Köster M. Sustainable data analysis with Snakemake. F1000Research. 2022 Apr
510 21;10(33).

511 41. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods.
512 2012 Apr;9(4):357–9.

513 42. Lu J, Salzberg SL. Ultrafast and accurate 16S rRNA microbial community analysis
514 using Kraken 2. Microbiome. 2020 Dec;8(1):124.

515 43. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species
516 abundance in metagenomics data. PeerJ Comput Sci. 2017 Jan 2;3:e104.

517 44. Truong D. T, Franzosa A. E, Tickle T.L, Scholz M, Weingart G, et al. MetaPhlAn2 for
518 enhanced metagenomic taxonomic profiling. Nat Methods. 2015 Oct;12(10):902-3.

519 45. Andrew, S A S. FASTQC: A Quality Control Tool for High Throughput Sequence
520 Data. www.bioinformatics.babraham.ac.uk/projects/fastqc/

521 46. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing
522 reads. EMBnet. 2011 May; 17(1):10-12.

523 47. Valiente-Mullor C, Beamud B, Ansari I, Francés-Cuesta C, García-González N,
524 Mejía L, et al. One is not enough: On the effects of reference genome for the
525 mapping and subsequent analyses of short-reads. PLOS Comput Biol. 2021 Jan
526 27;17(1):e1008678.

527 48. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
528 Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

529 49. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
530 features. Bioinformatics. 2010 Mar 15;26(6):841–2.

531 50. Hall AB, Basu S, Jiang X, Qi Y, Timoshevskiy VA, Biedler JK, et al. A male-
532 determining factor in the mosquito *Aedes aegypti*. Science. 2015 Jun
533 12;348(6240):1268-1270.

534 51. Lee Y, Schmidt H, Collier TC, Conner WR, Hanemaaijer MJ, Slatkin M, et al.
535 Genome-wide divergence among invasive populations of *Aedes aegypti* in
536 California. BMC Genomics. 2019 Dec;20(1):204.

537 52. Lau MJ, Schmidt TL, Yang Q, Chung J, Sankey L, Ross PA, et al. Genetic stability
538 of *Aedes aegypti* populations following invasion by wMel *Wolbachia*. BMC
539 Genomics. 2021 Dec 14;22(1):894.

540 53. Rose NH, Dabo S, da Veiga Leal S, Sylla M, Diagne CT, Faye O, et al. Enhanced
541 mosquito vectorial capacity underlies the Cape Verde Zika epidemic. PLOS Biol.
542 2022 Oct 26;20(10):e3001864.

543 54. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node
544 solution for large and complex metagenomics assembly via succinct de Bruijn
545 graph. Bioinformatics. 2015 May 15;31(10):1674–6.

546 55. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for
547 genome assemblies. Bioinformatics. 2013 Apr 15;29(8):1072–5.

548 56. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler
549 transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

550 57. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive
551 binning algorithm for robust and efficient genome reconstruction from metagenome
552 assemblies. PeerJ. 2019 Jul 26;7:e7359.

553 58. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM:
554 assessing the quality of microbial genomes recovered from isolates, single cells,
555 and metagenomes. Genome Res. 2015 Jul;25(7):1043–55.

556 59. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify
557 genomes with the Genome Taxonomy Database. Bioinformatics. 2019 Nov
558 15;btz848.

559 60. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal:
560 prokaryotic gene recognition and translation initiation site identification. BMC
561 Bioinformatics. 2010 Dec;11(1):119.

562 61. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, et al.
563 A standardized bacterial taxonomy based on genome phylogeny substantially
564 revises the tree of life. Nat Biotechnol. 2018 Nov;36(10):996–1004.

565 62. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity
566 searching. Nucleic Acids Res. 2011 Jul 1;39:W29–37.

567 63. Matsen FA, Kodner RB, Armbrust EV. pplacer: linear time maximum-likelihood and
568 Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC
569 Bioinformatics. 2010 Dec;11(1):538.

570 64. Matthews BJ. Improved reference genome of *Aedes aegypti* informs arbovirus
571 vector control. 2018 Nov;563(7732):501-507.

572 65. The Genome Standards Consortium, Bowers RM, Kyrpides NC, Stepanauskas R,
573 Harmon-Smith M, Doud D, et al. Minimum information about a single amplified
574 genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and
575 archaea. Nat Biotechnol. 2017 Aug;35(8):725–31.

576 66. Dada N, Jupatanakul N, Minard G, Short SM, Akorli J, Villegas LM. Considerations
577 for mosquito microbiome research from the Mosquito Microbiome Consortium.
578 Microbiome. 2021 Feb 1;9(1):36.

579 67. Glendinning L, Stewart RD, Pallen MJ, Watson KA, Watson M. Assembly of
580 hundreds of novel bacterial genomes from the chicken caecum. Genome Biol. 2020
581 Dec;21(1):34.

582 68. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new
583 genomic blueprint of the human gut microbiota. Nature. 2019 Apr
584 25;568(7753):499–504.

585 69. Wilkinson T, Korir D, Ogugo M, Stewart RD, Watson M, Paxton E, et al. 1200 high-
586 quality metagenome-assembled genomes from the rumen of African cattle and their
587 relevance in the context of sub-optimal feeding. Genome Biol. 2020 Dec;21(1):229.

588 70. Watson M. New insights from 33,813 publicly available metagenome-assembled-
589 genomes (MAGs) assembled from the rumen microbiome. Microbiology; 2021 Apr.
590 http://biorxiv.org/lookup/doi/10.1101/2021.04.02.438222

591 71. Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, et al.
592 Assembly of 913 microbial genomes from metagenomic sequencing of the cow
593 rumen. Nat Commun. 2018 Dec;9(1):870.

594 72. Mancini MV, Damiani C, Accoti A, Tallarita M, Nunzi E, Cappelli A, et al. Estimating
595 bacteria diversity in different organs of nine species of mosquito by next generation
596 sequencing. BMC Microbiol. 2018 Dec;18(1):126.

597 73. Coon KL, Hegde S, Hughes GL. Inter-species microbiota transplantation
598 recapitulates microbial acquisition and persistence in mosquitoes. Microbiology.
599 2022 Apr 11;10(1):58.

600 74. Muturi EJ, Njoroge TM, Dunlap C, Cáceres CE. Blood meal source and mixed
601 blood-feeding influence gut bacterial community composition in *Aedes aegypti*.
602 Parasit Vectors. 2021 Dec;14(1):83.

603 75. Chen S, Johnson BK, Yu T, Nelson BN, Walker ED. *Elizabethkingia anophelis*:
604 Physiologic and Transcriptomic Responses to Iron Stress. Front Microbiol. 2020
605 May 7;11:804.

606 76. Onyango MG, Lange R, Bialosuknia S, Payne A, Mathias N, Kuo L, et al. Zika virus
607 and temperature modulate *Elizabethkingia anophelis* in *Aedes albopictus*. Parasit
608 Vectors. 2021 Dec;14(1):573.

609 77. Gaio A de O, Gusmão DS, Santos AV, Berbert-Molina MA, Pimenta PF, Lemos FJ.
610 Contribution of midgut bacteria to blood digestion and egg production in *Aedes*
611 *aegypti*. Parasit Vectors. 2011 Dec;4(1):105.

612 78. Dahl EM, Neer E, Bowie KR, Leung ET, Karstens L. microshades: An R Package for
613 Improving Color Accessibility and Organization of Microbiome Data. 2022;11(11).

614

615

616

617

618

Figures

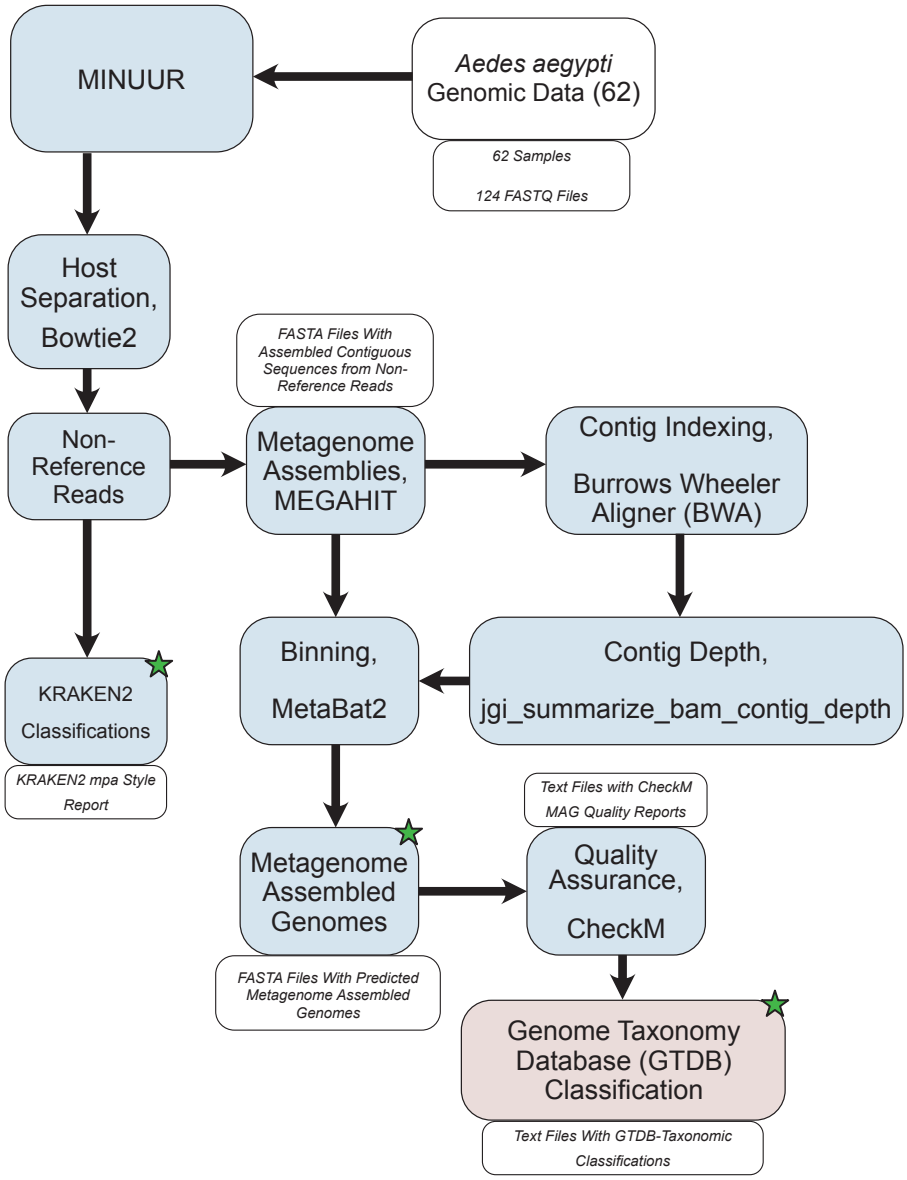
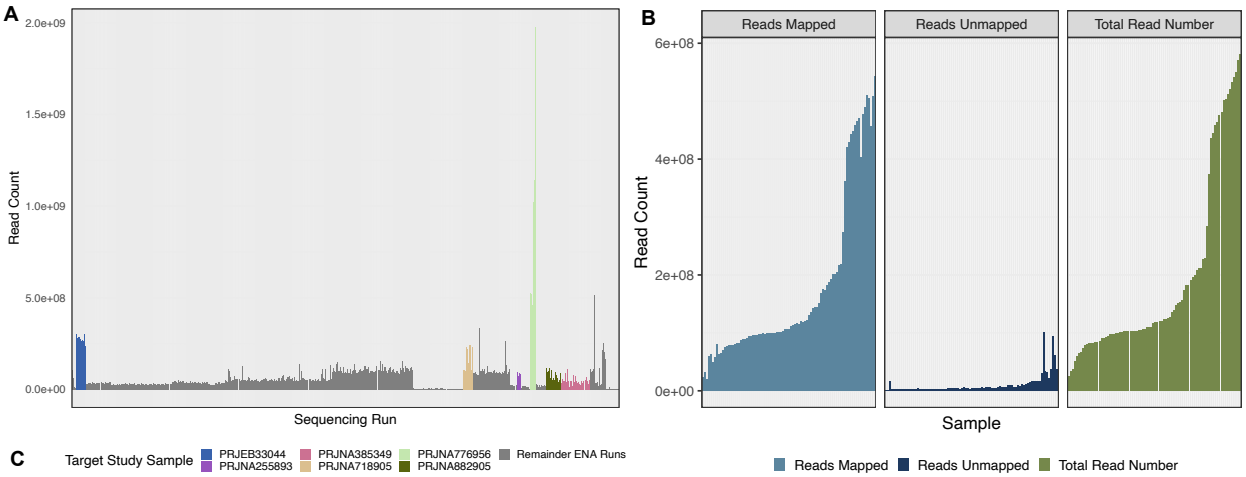


Figure 1: Study Workflow. Workflow describing the main steps of characterising non-reference reads from *Aedes aegypti* whole genome sequencing data. 62 samples (124 FASTQ files) were input to our metagenomics workflow MINUUR (**M**icrobial **IN**sight **U**sing **U**nmapped **R**eads). Reads were mapped to the *Aedes aegypti* reference genome (AaegL5.3) using Bowtie2 (v2.4.4), and unmapped (=non-reference) reads extracted, parsed for taxonomic classification with KRAKEN2

(v2.1.2) and metagenome assembly with MEGAHIT (v1.2.9). The resulting contigs are indexed with burrows wheeler aligner (BWA) (v0.7.17) and used to produce a depth file generated using the jgi_summarize_bam_contig_depth script from MetaBAT2 (v2.12.1). Bins are produced using the assembled contigs and depth file with MetaBAT2. The resulting metagenome assembled genomes (MAGs) are quality checked using CheckM (v1.1.3). High and medium quality MAGs, based on definitions set by the genome standards consortium (65), are taxonomically classified using GTDB-Tk (v1.5.0). Blue boxes denote the steps within MINUUR, red boxes denote steps outside of MINUUR. Green starts denote key outputs of the analyses.



Study Accession	TaxID	Mosquito Species	Sequencing Platform	Isolation Source	Study Date	Location	DNA Extraction Method	Number of Samples
PRJEB33044	7159	aedes_aegypti	Hiseq_3000	lab	6/14/19	Pennsylvania	TRIzol_Reagent	10
PRJNA255893	7159	aedes_aegypti	HiSeq 2000	lab	28/06/16	Virginia	QiaGEN_blood_tissue_kit	4
PRJNA385349	7159	aedes_aegypti	HiSeq 4000	field	7/1/17	California	MagMAX_protocol	28
PRJNA718905	7159	aedes_aegypti	HiSeq X Ten	lab	4/13/21	Austria	QIAGEN_GenomicTip_DNA	6
PRJNA718905	7159	aedes_aegypti	Illumina NovaSeq 6000	lab	4/13/21	Austria	QIAGEN_GenomicTip_DNA	4
PRJNA776956	7159	aedes_aegypti	Hiseq_2500	field	11/6/21	Australia	QiaGEN_blood_tissue_kit	6
PRJNA882905	7159	aedes_aegypti	Novaseq_6000	lab	10/7/22	Cape_Verde	QIAGEN_Dneasy	15

Figure 2: Sequencing Project Selection and Alignment. **A.** Bar chart depicting the total number of whole genome sequencing projects referred to as "*Aedes aegypti"* on the European Nucleotide Archive. Colour bars show the target sequencing samples we used in our study, colour coded in the legend. **B.** Facetted bar charts showing the total number of reads (right, green), aligned reads (left, blue) and non-reference reads (middle, dark blue) of our selected samples. Mapped and non-reference reads result from the alignment to the *AaegL5.3* reference genome using Bowtie2 (v2.4.4) (41) within our metagenomics workflow MINUUR. **C.** Table showing (left to right) the original sequencing project, taxonomic id, mosquito species, sequencing platform, isolation source, study date, location of the sequencing project, DNA extraction method and number of samples.

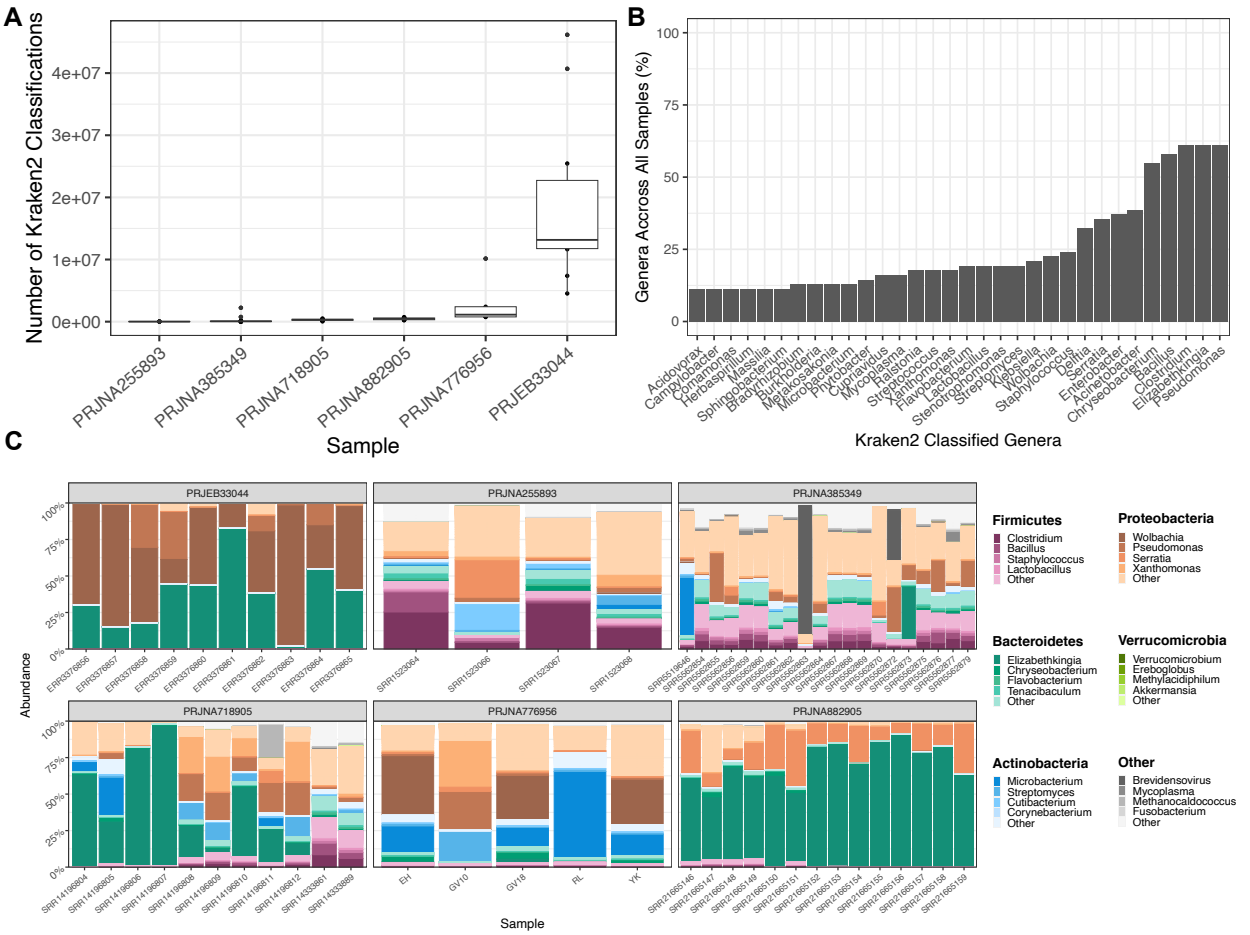


Figure 3: KRAKEN2 (v1.2) Genera Classifications of Non-Reference Reads After Alignment to the *AaegL5.3* Reference Genome. **A.** Box plot depicting the number of KRAKEN2 (42) classifications from non-reference reads across 6 publicly available *Aedes aegypti* sequencing projects. Each dot represents one sample within the sequencing project. **B.** Bar chart showing the proportion of genera summed across all samples. **C.** Facetted heatmaps showing phylum and genus level KRAKEN2 assignments of taxa identified from non-reference reads. Colour of each phylum is denoted in the right-hand legend panel, with the most abundant genera depicted by a colour gradient corresponding to their respective phylum. The graph was generated using the microshades R package (78).

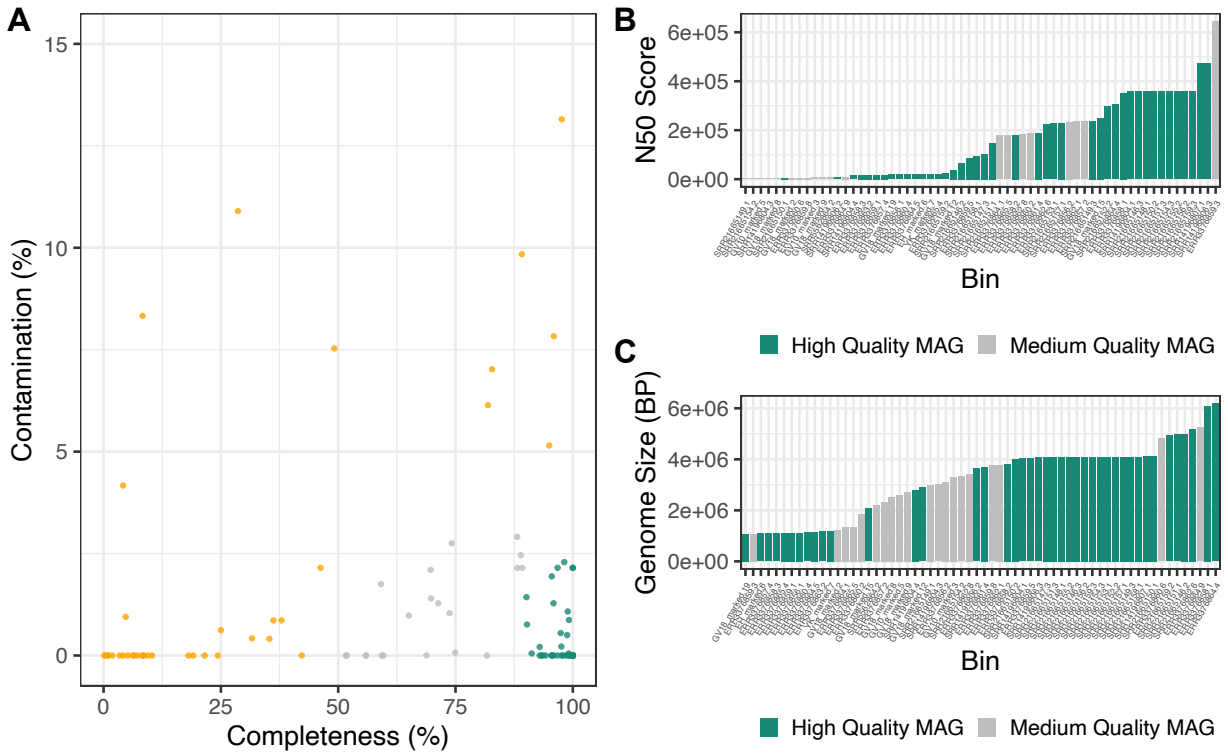
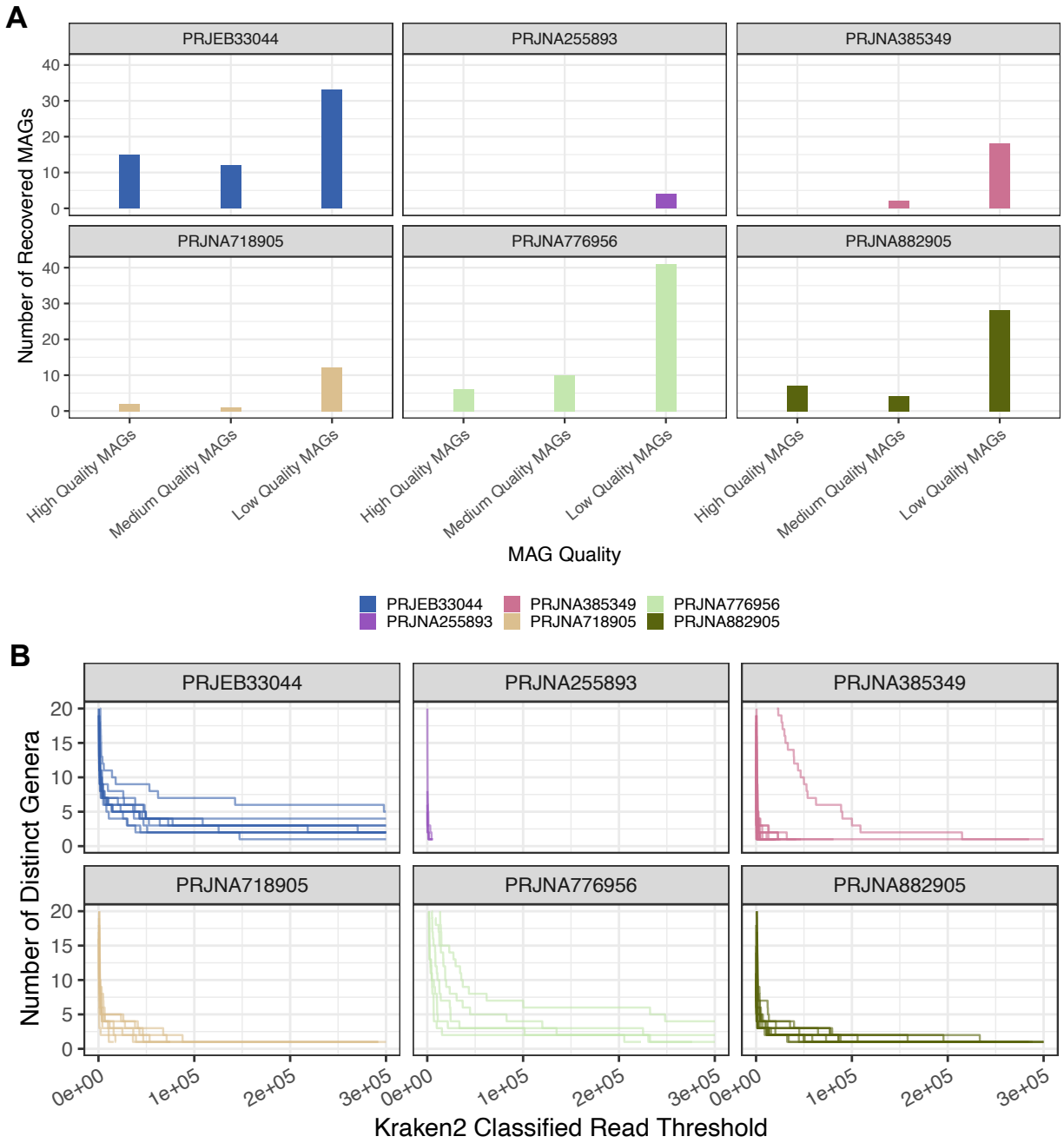


Figure 4: Recovered *Aedes aegypti* Associated Bacterial Metagenome Assembled Genomes (MAGs). A. MAGs assembled from non-reference reads using MEGAHIT (v1.2.9) and binned with MetaBAT2 (v2.12.1). Colours denote MAG genome standards consortium (GSC) high, medium and low MAG quality; green = high-quality MAGs (>90% completeness, <5% contamination), grey = medium-quality MAGs (>50% completeness, <5% contamination) and orange = low-quality MAGs. **B.** Bar graph depicting N50 score of medium and high-quality MAGs. Green = high-quality MAG and grey = medium-quality MAGs. **C.** Bar graph depicting genome size in base pairs of medium and high-quality MAGs. Green = high-quality MAG, grey = medium-quality MAG.



680

681 **Figure 5: Relationship Between KRAKEN2 classifications and MAG Recovery**. **A.** Bar chart
682 depicting the recovery of high, medium and low-quality MAGs from *Aedes aegypti* non-
683 reference reads across each sequencing project. Y-axis = number of recovered MAGs, X-axis =
684 MAG quality ranked using CheckM completeness and contamination. **B.** KRAKEN2
685 classifications depicted through increased stringency of a KRAKEN2 classification threshold.
686 Each line represents a sample within the sequencing project, Y-axis shows the number of
687 distinct KRAKEN2 classified genera and X-axis represents the number of KRAKEN2
688 classifications. Each graph represents the filtering of samples towards taxa with a high number
689 of assigned reads.

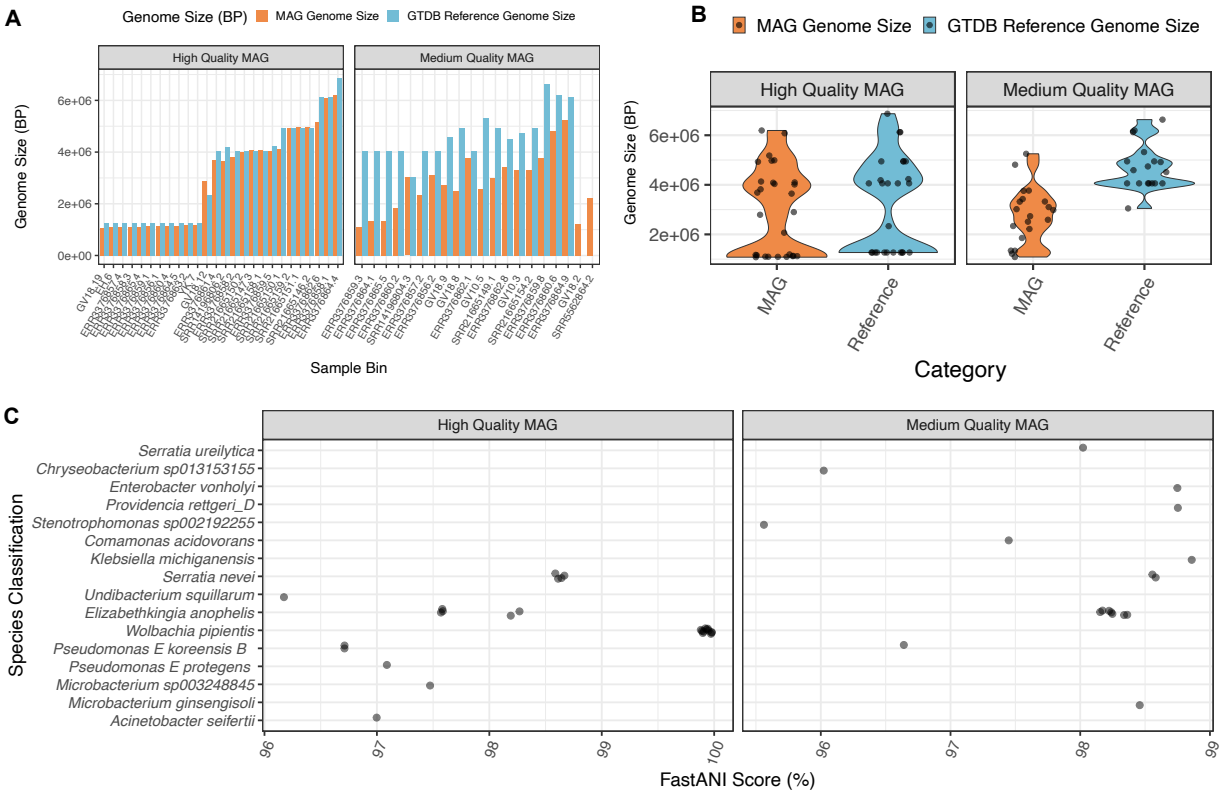


Figure 6: *Aedes aegypti* Associated Bacterial MAG Classifications with GTDB-Tk (v2.1.1). **A.** Genome size (base pairs) of *Aedes aegypti* associated MAGs (orange), split between high-quality (CheckM completeness > 90%, contamination <5%) and medium-quality (CheckM completeness >50%, contamination <5%) rankings, compared to their closest GTDB-Tk reference genome (blue). Each bar denotes the origin sample and the bin number (sample.bin). **B.** Violin plot showing the distribution of *Aedes aegypti* MAG genome size (orange) compared to their GTDB-Tk classified reference genome. Each point denotes a MAG and graphs are facetted between high and medium quality MAGs (see above). **C.** Taxonomic classifications from the Genome Taxonomy Database (GTDB) of MAGs obtained from non-reference *Aedes aegypti* reads. X-axis = FastANI (Average Nucleotide Identity) (%) score to the closest related reference genome in the GTDB, Y-axis = GTDB-Tk species classification.