METHOD ARTICLE

REVISED Recovery of metagenomic data from the *Aedes aegypti* microbiome using a reproducible snakemake pipeline: MINUUR [version 2; peer review: 1 approved, 2 approved with reservations]

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

[1]Vector Biology and Tropical Disease Biology, Liverpool School of Tropical Medicine, Liverpool, L3 5QA, UK
[2]Vector Biology, Liverpool School of Tropical Medicine, Liverpool, L3 5QA, UK
[3]Vector Biology and Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, L3 5QA, UK

v2 **First published:** 23 Mar 2023, **8**:131
https://doi.org/10.12688/wellcomeopenres.19155.1
Latest published: 26 May 2023, **8**:131
https://doi.org/10.12688/wellcomeopenres.19155.2

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 from these initially host-focused sequencing projects.
Methods: To analyse non-host reads from existing genomic data, we developed a snakemake workflow called MINUUR (Microbial INsights Using Unmapped Reads). 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-

Open Peer Review

Approval Status ? ? ✓

	1	2	3
version 2 (revision) 26 May 2023		? view	✓ view
version 1 23 Mar 2023	? view		

1. **Taylor Reiter**, Arcadia Sciences, Berkeley, USA

2. **Ellen Cameron**, Wellcome Trust Genome Campus, Hinxton, UK
Wellcome Sanger Institute, Hinxton, UK

3. **Mariana Rocha David**, Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, Brazil

Any reports and responses or comments on the article can be found at the end of the article.

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 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

Corresponding author: Eva Heinz (eva.heinz@lstmed.ac.uk)

Author roles: Foo A: Conceptualization, Formal Analysis, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Cerdeira L**: Software, Writing – Review & Editing; **Hughes GL**: Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing; **Heinz E**: Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome (217303). 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).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Foo A, Cerdeira L, Hughes GL and Heinz E. **Recovery of metagenomic data from the *Aedes aegypti* microbiome using a reproducible snakemake pipeline: MINUUR [version 2; peer review: 1 approved, 2 approved with reservations]** Wellcome Open Research 2023, **8**:131 https://doi.org/10.12688/wellcomeopenres.19155.2

First published: 23 Mar 2023, **8**:131 https://doi.org/10.12688/wellcomeopenres.19155.1

REVISED **Amendments from Version 1**

In this revised version, we have made changes to address reviewer comments. In the text, we have rephrased the reporting of mapped and unmapped reads from absolute numbers to percentages. The discussion now addresses a limitation of this workflow, which is the inference of a microbe as true symbiont verses contamination. We highlight that our results strongly suggest the recovery of mosquito-symbionts in our study, as evidenced by their matching taxonomic assignments in previous studies, and we also highlight a potential source of viral contamination. Additionally, we expand on comparisons between KRAKEN2 and GTDB-Tk classifications, giving examples of matching taxonomic assignments in both of these results. Figure 4 has been updated to improve legibility of the X-axis labels.

Within the methods, we include an additional genome quality assurance metric using BUSCO, and in the results section, we highlight the minimal detection of eukaryotic marker genes within our high and medium quality MAGs. A figure for this has now been included in Supplementary Figure 2. All extended data and Zenodo DOI reflecting changes to the code (inclusion of BUSCO and updated software versions) have been updated accordingly.

Any further responses from the reviewers can be found at the end of the article

Introduction

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

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

Whole genome shotgun sequencing is commonly used to study mosquito genomics at the individual[35,36] and population[37,38] level; meaning non-mosquito sequence data (we refer 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 "Microbial INsight Using Unmapped Reads" (MINUUR) 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 a total of 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 with an accompanying JupyterBooks page.

Methods

Specifications

To undertake this analysis, we developed a metagenomics workflow called MINUUR, using the workflow manager Snakemake[40] (Figure 1). This pipeline was developed to facilitate the following analysis and future studies aiming to characterise non-reference reads in mosquitoes or other organisms. A breakdown of the pipeline that produced each result of this study is provided in the following section, with details of how each step was configured in the final section.

Database setup

MINUUR requires several databases to perform the analysis. This includes a high quality bowtie2-indexed reference genome[41] to separate host and non-host reads, and a KRAKEN2[42], BRACKEN[43] and MetaPhlAn3[44] database for taxonomic read classifications. All databases are available in their respective GitHub repositories. The databases used in this study are the default MetaPhlAn3 marker gene database, KRAKEN2 and BRAKEN indexes from the Ben Langmead repository located here. For our study, we downloaded and compiled these default databases on 8 September 2022.

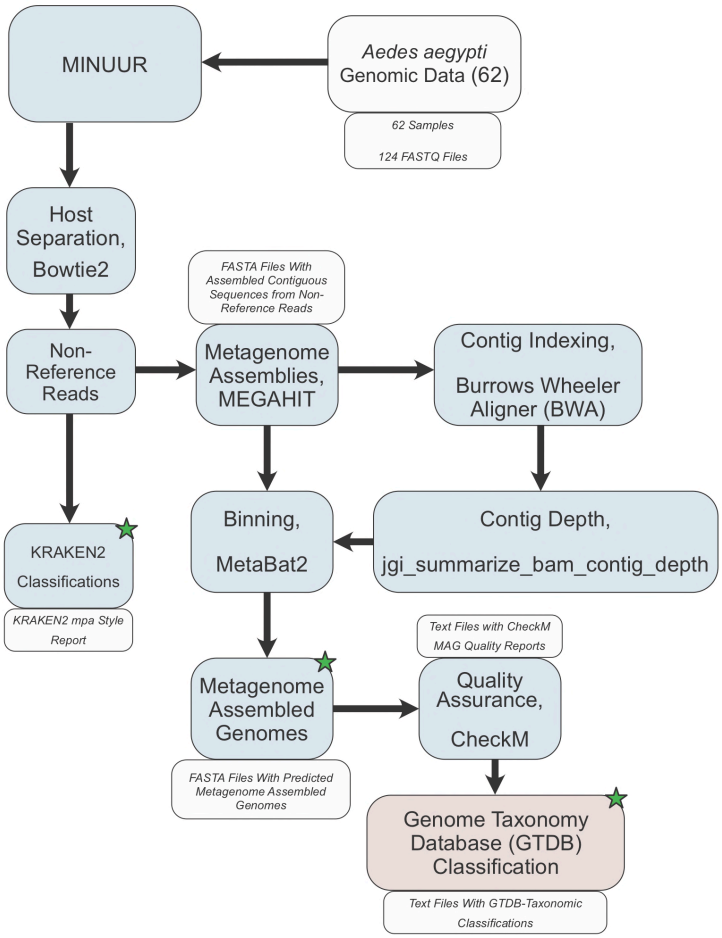


Figure 1. Study Workflow. Workflow describing the main steps of characterising non-reference reads from *Aedes aegypti* whole genome sequencing data. 62 samples (124 FASTQ files) were input to our metagenomics workflow MINUUR (**M**icrobial **IN**sight **U**sing **U**nmapped **R**eads). Reads were mapped 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[45], 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.

Data preparation

MINUUR accepts either BAM or paired FASTQ inputs. For FASTQ inputs, MINUUR performs quality control (QC) using FASTQC (v0.11.9)[46], providing a QC report per sample. MINUUR does not use the FASTQC report in subsequent steps, but only as a quality assurance metric for the user and to estimate if read trimming is required. Read trimming can be performed within MINUUR using Cutadapt (v1.5)[47] with user defined parameters for minimum read length, base quality and adapter content (default: minimum base length = 50, average base quality = 30). To separate host and non-host, reads are aligned to a user defined indexed reference genome (the relevant host genome) using Bowtie2 (v2.4.4)[41]. Alignment sensitivity and type can be adjusted within the pipeline at the user's discretion. A high-quality, chromosome level-assembled reference genome is recommended if available. In situations where this is not possible, users should be aware that read alignment will likely result in mismatches between the reference and target sequence and produce alignments with poor coverage[48]. As a result, non-reference reads used in subsequent steps are likely to contain a substantial number of host data. In this instance, we included a feature within MINUUR to extract KRAKEN2 assigned reads pertaining to known microbes and potentially improve metagenome assemblies. Non-reference reads within the coordinate sorted binary alignment (BAM) file are extracted using samtools using the command "samtools -view -f 4" (v.1.14)[49] and converted to FASTQ format using bedtools (v2.3.0)[50] ("–bamToFastq"). As users might already have their data in BAM format

mapped against the host reference (*e.g.* in large-scale sequencing projects like the Ag1000G), we also included the option of a BAM input. Here, any non-reference reads within the BAM file will be extracted, converted to FASTQ and used in downstream steps.

For this study, we used five published genome datasets of *Aedes aegypti* publicly available on the European Nucleotide Archive[35,37,51–54]. We selected the data sets to cover a range of sequencing depths, DNA extraction method and sequencing platform (Figure 2C). All raw FASTQ files of published sequencing data were retrieved from the European Nucleotide Archive (ENA) under the project accession numbers PRJEB33044[37], PRJNA255893[51], PRJNA385349[52], PRJNA718905[35], PRJNA776956[53] and PRJNA992905[54].

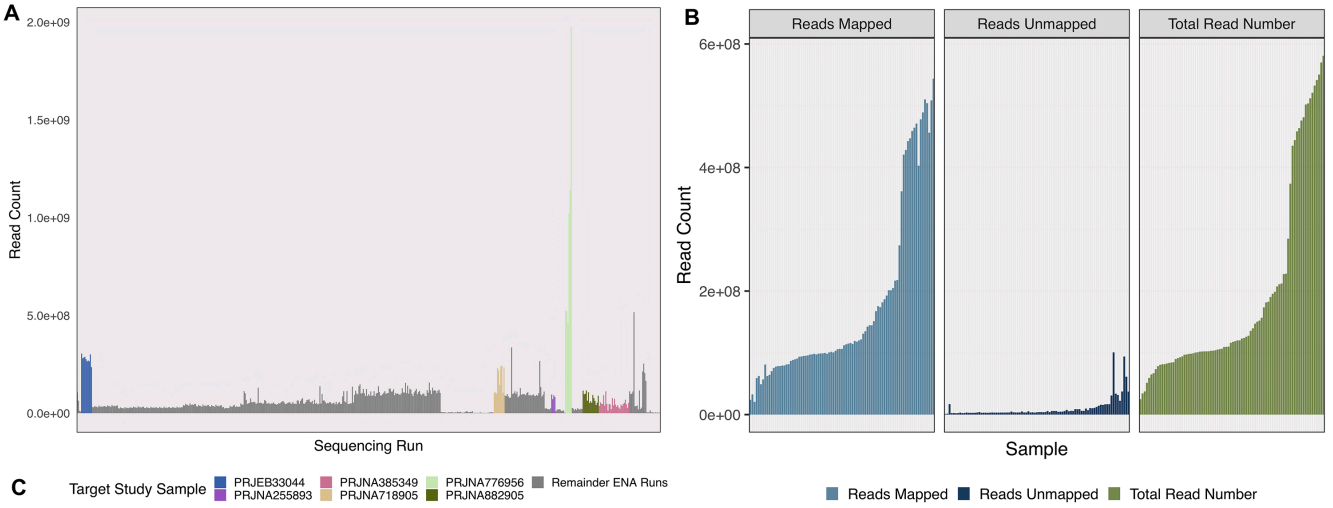
Read classification

MINUUR uses two read classification approaches to infer taxonomy. KRAKEN2 (v2.1.2)[42], which uses a k-mer based approach to map read fragments of k-length against a taxonomic genome library of k-mer sequences, and MetaPhlAn3 (v3.0.13)[44] to align reads against a 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)[55], a rapid and memory efficient metagenome assembler, for *de novo* metagenome assembly. Assembled contigs are quality-checked using QUAST (v5.0.2)[56] to assess contig N50 and L50 scores. The resultant contigs, which are fasta files with sequences pertaining to genomic



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.

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)[57], 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)[58] to produce a depth file of contig coverage. The depth file and assembled contigs are input to the metagenome binner MetaBAT2 (v2.12.1)[58], to group contigs within defined genomic bins. Each bin is a predicted metagenome assembled genome (MAG). CheckM (v1.1.3)[59] is then used for quality assurance of each bin by identifying single copy core genes. Specifically, bin contamination is assessed by looking for one single copy core gene within each bin, and completeness by calculating a required set of single copy core genes. In addition, BUSCO (v5.4.7)[60] can be included to search for eukaryotic or prokaryotic specific genes in the final MAGs as an additional quality assurance metric.

Pipeline configuration

All sample names were listed in the samples.tsv file of the configuration directory and paths to the FASTQ directories given in the samples_table.tsv file. To implement the pipeline, the configuration file was set to the following parameters: FASTQ = True, QC = True, CutadaptParams = "–minimum-length 50 -q 30", RemoveHostFromFastqGz = True, AlignmentSensitivity = "–sensitive-local", ProcessBam = True, From-Fastq = True, KrakenClassification = True, ConfidenceScore = 0, KrakenSummaries = True, GenusReadThreshold = 1000, SpeciesReadThreshold = 30000, MetagenomeAssm = True, MetagenomeBinning = True, MinimumContigLength = 1500, CheckmBinQA = True, BUSCO=True. All databases were installed from their respective repositories on 8 September 2022. The pipeline was run on an Ubuntu Linux system with 660gb of available memory and 128 CPUs. For our analysis, with the above settings and 10 cores available, runtime was two weeks; the maximum Resident Set Size (RSS) of an individual sample during this run was 9771 RSS (occurring during metagenome assembly); and total storage used (including temporary files) was 5.18Tb (terabytes) across all samples used in this study.

Taxonomic classification of MAGs with GTDB-Tk

Separate from MINUUR, all bins produced from MetaBAT2 were taxonomically classified with GTDB-Tk[61] (v2.1.1) using "–classify-wf" against the Genome Taxonomy Database (GTDB) (release 07-R207 8 April 2022, downloaded on 8 September 2022). GTDB-Tk assigns genes to MAGs using Prodigal (v2.6.3)[62] and ranks the taxonomic domain of each MAG using a database of 120 bacteria and 122 archaea marker genes[63] using HMMER3[64]. With this information, MAGs are then placed into domain specific reference trees with pplacer (v1.1)[65]. Taxonomic classifications with GTDB-Tk are based on placement within the GTDB reference tree, relative evolutionary divergence, and average nucleotide identity (ANI) scores with its closest reference genome. The relative evolutionary divergence score is used to refine ambiguous taxonomic rank assignments and ANI scores are used to define species classifications[61].

Results

Extraction of non-reference reads post-alignment to *Aedes aegypti*

In total, we retrieved 62 samples (124 FASTQ files) across six sequencing experiments (Figure 2A) and parsed them through MINUUR. After alignment to the *Aedes aegypti* reference genome (AaegL5.3)[66] with Bowtie2, the proportion of mapped and non-reference reads were calculated. Of our initial 62 samples, on average, 91.9% of reads aligned to the AaegL5.3 reference genome, while 4.6% of reads did not map to the AaegL5.3 reference genome (Figure 2B). To estimate the number of reads associated to the microbiome among the non-reference reads, the overall number of KRAKEN2 classifications were counted. The average proportion of KRAKEN2 classified reads from all non-reference reads was 17.9%. The number and proportion of KRAKEN2 classifications per sample is given in Supplementary Table 1 within the Extended data.

Taxonomic classifications produced from non-reference reads of *Aedes aegypti*

The KRAKEN2 classifications of non-reference reads were counted from each sequencing project. Four out of the six sequencing projects showed a much lower number of classifications compared to sequencing projects PRJNA776596 and PRJEB33044 (Figure 3A). PRJEB33044 gave the highest average proportion of taxonomic classification (81.3%), while PRJNA255893 was the lowest (0.893%).

Multiple phyla were identified across sequencing projects. *Bacteroidetes, Proteobacteria, Firmicutes and Actinobacteria* were the most common phylum present across all samples, with varied relative abundance between projects (Figure 3C). For example, *Bacteroidetes* and *Proteobacteria* were dominant in PRJEB33044 and PRJNA882905, *Firmicutes* and *Proteobacteria* in PRJNA255893, and *Bacteroidetes, Actinobacteria* and *Proteobacteria* in PRJNA385349 and PRJNA718905. At the genera level, there are several dominant members across samples, including *Wolbachia, Pseudomonas, Serratia* and *Elizabethkingia.* Generally, however, there is considerable variation at the genus level within and between sequencing experiments (Figure 3C). We summarised all KRAKEN2 classifications across 62 samples. The most common genera, identified in >50% of all samples, were *Pseudomonas, Elizabethkingia, Clostridium, Bacillus* and *Chryseobacterium.* Genera identified in 25–50% of samples were *Acinetobacter*, *Enterobacter, Serratia* and *Delftia* (Figure 3B). All KRAKEN2 taxonomic classifications are given in Supplementary Table 2 (*Extended data*).

Metagenome assembly and binning

We used MINUUR to further parse non-reference reads through assembly, binning and quality assurance steps, with the aim to recover MAGs associated to the *Aedes aegypti* microbiome. Assembly was conducted with MEGAHIT and binning with MetaBat2. We used CheckM to assess MAG completeness and contamination through the presence and copy number of single copy core genes, and recovered 105 MAGs (Figure 4A). Using the standards of MAG quality set by the genome standards consortium (GSC)[45], 42 MAGs met the criteria of high quality with completeness >90%

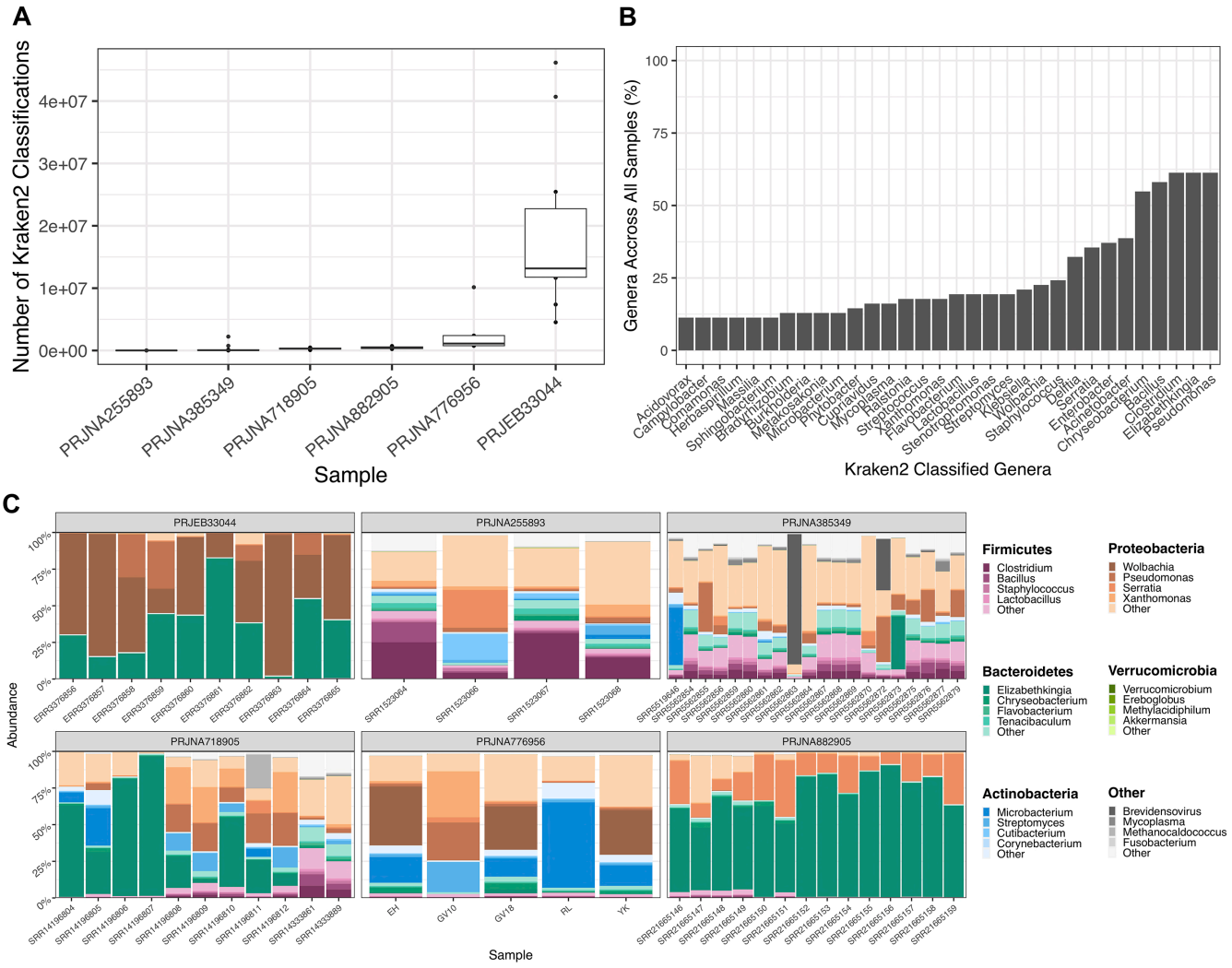


Figure 3. KRAKEN2 (v1.2) Genera Classifications of Non-Reference Reads After Alignment to the *AaegL5.3* Reference Genome. **A.** Box plot depicting the number of KRAKEN2[42] classifications from non-reference reads across 6 publicly available *Aedes aegypti* sequencing projects. Each dot represents 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[67].

and contamination <5%; and 20 MAGs classify as medium quality with completeness >50% and contamination <5%. Overall, 62 high- and medium-quality MAGs were recovered. The remaining MAGs were low-quality (<50% complete and <10% contamination) or contaminated >10%, and therefore excluded from MAG classification. Of the high- and medium-quality MAGs, the mean N50 (the minimum contig length of an assembled contig that covers 50% of the genome) was 177.2 kilobase pairs (KBP), ranging between 4.72KBP and 471.6KBP (Figure 4B). The average genome size was 3.36 megabase pairs (MBP), ranging between 1.07MBP and 6.19MBP (Figure 4C), while low-quality MAGs showed a wider range of genome sizes between 0.24MBP and 106MBP (Supplementary Figure 1, *Extended data*). We further applied BUSCO to assess eukaryotic contamination in the final assemblies. Eukaryotic contamination was 3.24% (0 – 4.47%) and 2.26% (0.4 = 4.7%) on average in high and medium quality MAGs respectively (Supplementary Figure 2, *Extended data*).

Relationship between KRAKEN2 taxonomic classifications and MAG recovery

The suitability of a sample for MAG recovery would be of interest to estimate in advance, and we investigated if there was a correlation between the proportion of KRAKEN2 classifications from non-reference reads and MAG recovery. In one project, PRJNA255893, we recovered no high-quality MAGs, (Figure 5A) and the reads used for assembly contained no taxa exceeding 100,000 KRAKEN2 assigned reads (Figure 5B). In contrast, PRJNA33044, PRJNA776596 and PRJNA882905 allowed retrieval of more high- and medium-quality MAGs, and within these projects, multiple samples contained >100,000 reads assigned to a taxon

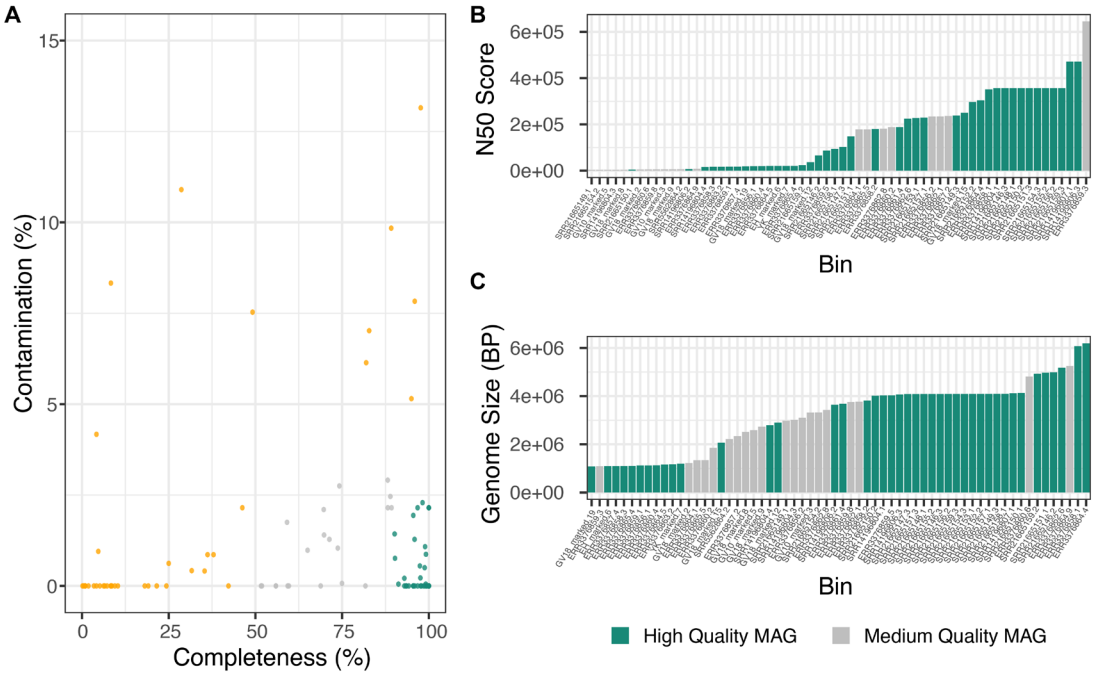


Figure 4. Recovered *Aedes aegypti* Associated Bacterial Metagenome Assembled Genomes (MAGs). A. MAGs assembled from non-reference reads using MEGAHIT (v1.2.9) and binned with MetaBAT2 (v2.12.1). Colours denote MAG genome standards consortium (GSC) high, medium and low MAG quality; green = high-quality 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.

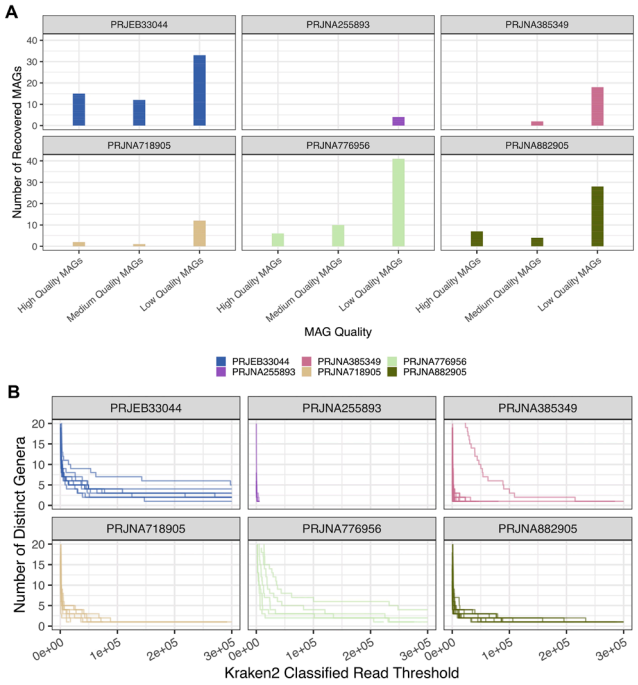


Figure 5. Relationship Between KRAKEN2 classifications and MAG Recovery. A. Bar chart depicting the recovery of high, medium and low-quality MAGs from *Aedes aegypti* non-reference reads across each sequencing project. Y-axis = number of recovered MAGs, X-axis = MAG quality ranked using CheckM completeness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

(Figure 5A, B). As such, MAG recovery is likely linked to a sufficient number of classified reads to a taxon (Figure 5A, Figure 5B), rather than overall number of KRAKEN2 assigned reads within a sample (Supplementary Figure 3, *Extended data*). In accordance with this, the total number of classified taxa totaling or greater than 100,000 KRAKEN2 assigned reads for PRJEB33044 is 33 taxa (summed across all samples) and 22 high- and medium-quality MAGs were recovered from this sample. Similarly, PRJNA776956 shows 19 taxa with associated reads totaling or greater than 100,000 KRAKEN2 assigned reads, resulting in 16 high and medium-quality MAGs. Furthermore, taxonomic classifications with a high number of KRAKEN2 assigned reads (>100,000) are akin to the taxonomic classifications of medium and high-quality MAGs (Supplementary Figure 4, *Extended data*). For example, across multiple samples *Serratia*, *Elizabethkingia* and *Wolbachia* were assigned >100,000 reads and resulted in six, 27 and 13 medium and high-quality MAGs respectively. There are, however, exceptions to these patterns; PRJNA385349 contained five taxa with over 100,000 KRAKEN2 assigned reads, yet no high-quality MAGs could be recovered from this project. As such, applicable for future application of this approach, the number of KRAKEN2 classifications assigned to a specific taxon is one factor that can help estimate high-quality MAG recovery.

Taxonomic classification of MAGs with GTDB-Tk

Following MAG recovery using MINUUR, we used the taxonomic classifier GTBD-Tk to classify high and medium quality MAGs against the Genome Taxonomy Database (GTDB). We compared the genome size of each MAG to its closest reference genome in the GTDB (Figure 6A, Figure 6B). Of the high-quality MAGs, these were larger than their reference genome by mean = 183kb, whereas medium quality MAGs deviated from their reference genomes by 1768kb (Figure 6A). Congruent with pairwise size differences between MAG and reference genome, we found the overall distribution of high-quality MAGs compared to their reference genome size to be similar, but significantly different between medium-quality MAGs (Figure 6B). Of these, 48 MAGs were classified to the species level with a mean FastANI score

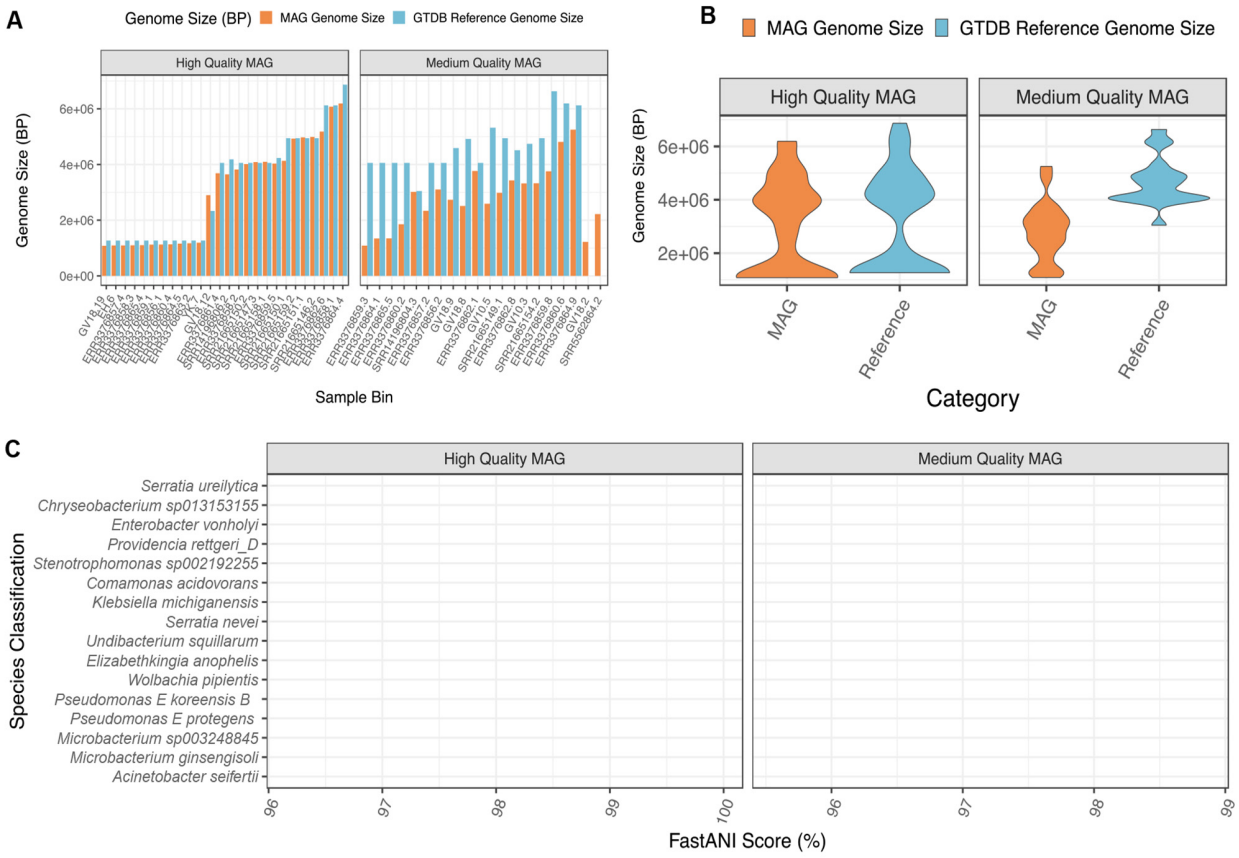


Figure 6. *Aedes aegypti* Associated Bacterial MAG Classifications with GTDB-Tk (v2.1.1). A. Genome size (base pairs) of *Aedes aegypti* associated MAGs (orange), split between high-quality (CheckM completeness > 90%, contamination <5%) and medium-quality (CheckM completeness >50%, contamination <5%) 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.

of 98.4%, ranging between 95.6% and 100%. No MAGs were identified <95% ANI to a known species, indicating no undescribed *Aedes aegypti* associated bacterial species were present within these MAGs (Figure 6C).

Discussion

Metagenomic datasets of mosquito microbiomes are so far limited in number[68]. In this study, we developed a metagenomics like workflow called MINUUR to facilitate recovery and use of host-associated bacterial sequences using non-reference reads from existing host WGS projects. We demonstrate that MINUUR can be used to recover genus level taxonomic classifications and draft high and medium quality MAGs from host WGS projects. The recovery of metagenomic information, such as MAGs, are applied in large scale metagenomic studies from chickens[69], humans[70] to cows[71–73], with these studies yielding between 400 to 92,000 MAGs per study. We apply a similar approach with non-reference *Aedes aegypti* sequencing reads across a range of different studies and can demonstrate that using MINUUR expands the genomic representation of known mosquito-associated bacterial symbionts. Overall, these provide a valuable resource for researchers in the field and can be used in further work such as facilitating biosynthetic gene cluster discovery[73] or to identify genetic targets in symbiont pathogen blocking approaches[2].

The data retrieved in this study agrees with published insights; the phylum level classifications are consistent with findings from other mosquito microbiome studies[8,9,11,18,74,75], showing that *Proteobacteria, Bacteroidetes* and *Firmicutes* are dominant phyla of the mosquito microbiome; and our taxonomic classifications highlight the inherent variability of the *Aedes aegypti* microbiome[18,24,76]. These findings give us confidence that taxonomic classifications with KRAKEN2, within the pipeline, can accurately predict the presence of microbes associated to the *Aedes aegypti* microbiome from non-reference sequences. At the genus level, we also find consistent observations of taxa previously identified in other studies[16,19,23,77,78]. The KRAKEN2 classifications show within the two most common phyla, *Proteobacteria* and *Bacteroidetes*; *Elizabethkingia, Pseudomonas* and *Serratia* are the most common. All three of these symbionts are documented to play key roles in either blood digestion[19,79], iron-acquisition[77] and microbial interactions[16]. Notably *Elizabethkingia* has previously been implicated in responses to iron fluxes in *Anopheles gambiae*[77], and blood meals in *Aedes albopictus*[25] and *Aedes aegypti*[24]. Similarly, *Pseudomonas* has shown to interact with *Elizabethkingia*, triggering the expression of *hemS* to break down heme into biliverdin catabolites[17]. It is encouraging to note the presence of these two bacteria within our taxonomic classifications. We recovered high and medium quality MAGs associated to these taxa (*Serratia, Elizabethkingia* and *Pseudomonas*), which should allow further interrogation of genes associated to these biological processes. To note, the presence of *Wolbachia* in the projects we have analysed is expected since these mosquitoes were transinfected with high titers of this bacterium, further validating the pipeline results[37].

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

A limitation of this workflow is knowing whether the data is indicative of symbionts associated to the *Aedes aegypti* microbiome, or sequencing contamination[80–82]. We believe the majority of our results support the identification of true microbial symbionts to the mosquito microbiome; our taxonomic classifications from MAGs and KRAKEN2 read assignments are congruent with previous studies[4,10,13,18,19,25,74,76]. However, we also identify likely contaminants such as *Brevidensovirus* (Figure 3), which have previously been identified as ZIKV stock contamination[83]. Discerning between symbiont or contaminant from results generated through this workflow requires further cross reference to the host-microbiome literature in question. Further analysis can also help answer questions of contamination or symbiont, such as measuring species genetic similarity with other sequenced host-symbionts.

Conclusions

In summary, we present a reproducible workflow to analyse host-associated microbial sequence data derived from host WGS experiments, leveraging a vast resource of data for additional insights. Our work focuses on the mosquito microbiome, where future considerations and prospects were recently established by the Mosquito Microbiome Consortium[68]. A key point highlighted in this statement was the need for (meta)genomics approaches with solid reproducibility for data analysis within the field. Our pipeline provides a workflow to assess non-host reads from existing mosquito genome sequence data and increases our knowledge of mosquito-associated bacterial genomes. This approach and accompanying workflow will facilitate more analyses of existing WGS data within *Aedes aegypti* and other organisms of interest for the scientific community.

Data and software availability

All original sequencing projects can be accessed under the following accession numbers: PRJEB33044 https://www.ebi.ac.uk/ena/browser/view/PRJEB33044), PRJNA255893 (https://www.ebi.ac.uk/ena/browser/view/PRJNA255893), PRJNA385349 (https://www.ebi.ac.uk/ena/browser/view/PRJNA385349), PRJNA718905 (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/ and at zenodo here under a GNU General Public License, Version 3.

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, for citation please use https://doi.org/10.5281/zenodo.7707874. 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: BUSCO Assessment of Eukaryotic Contamination in High and Medium Quality MAGs

Supplementary Figure 3: Correlation Between Classified Read Number and Number of MAGs

Supplementary Figure 4: 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.

References

1. Messina JP, Brady OJ, Pigott DM, *et al.*: **A global compendium of human dengue virus occurrence.** *Sci Data.* 2014; **1**(1): 140004.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Cansado-Utrilla C, Zhao SY, McCall PJ, *et al.*: **The microbiome and mosquito vectorial capacity: rich potential for discovery and translation.** *Microbiome.* 2021; **9**(1): 111.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Utarini A, Indriani C, Ahmad RA, *et al.*: **Efficacy of *Wolbachia*-Infected Mosquito Deployments for the Control of Dengue.** *N Engl J Med.* 2021; **384**(23): 2177–2186.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Scolari F, Casiraghi M, Bonizzoni M: ***Aedes* spp. and Their Microbiota: A Review.** *Front Microbiol.* 2019; **10**: 2036.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Steven B, Hyde J, LaReau JC, *et al.*: **The Axenic and Gnotobiotic Mosquito: Emerging Models for Microbiome Host Interactions.** *Front Microbiol.* 2021; **12**: 714222.
PubMed Abstract | Publisher Full Text | Free Full Text
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; **10**(12): e0143541.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Romoli O, Gendrin M: **The tripartite interactions between the mosquito, its microbiota and *Plasmodium*.** *Parasit Vectors.* 2018; **11**(1): 200.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Dada N, Jumas-Bilak E, Manguin S, *et al.*: **Comparative assessment of the bacterial communities associated with *Aedes aegypti* larvae and water from domestic water storage containers.** *Parasit Vectors.* 2014; **7**(1): 391.
PubMed Abstract | Publisher Full Text | Free Full Text
9. David MR, Santos LMBD, Vicente ACP, *et al.*: **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; **111**(9): 577–87.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Saab SA, Dohna HZ, Nilsson LKJ, *et al.*: **The environment and species affect gut bacteria composition in laboratory co-cultured *Anopheles gambiae* and *Aedes albopictus* mosquitoes.** *Sci Rep.* 2020; **10**(1): 3352.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Onyango GM, Bialosuknia MS, Payne FA, *et al.*: **Increase in temperature enriches heat tolerant taxa in *Aedes aegypti* midguts.** *Sci Rep.* 2020; **10**(1): 19135.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Wang Y, Gilbreath TM 3rd, Kukutla P, *et al.*: **Dynamic Gut Microbiome across Life History of the Malaria Mosquito *Anopheles gambiae* in Kenya.** Leulier F, editor. *PLoS One.* 2011; **6**(9): e24767.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Sandeu MM, Maffo CGT, Dada N, *et al.*: **Seasonal variation of microbiota composition in *Anopheles gambiae* and *Anopheles coluzzii* in two different eco-geographical localities in Cameroon.** *Med Vet Entomol.* 2022; **36**(3): 269–282.
PubMed Abstract | Publisher Full Text
14. Kakani P, Gupta L, Kumar S: **Heme-Peroxidase 2, a Peroxinectin-Like Gene, Regulates Bacterial Homeostasis in *Anopheles stephensi* Midgut.** *Front Physiol.* 2020; **11**: 572340.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Minard G, Tran FH, Tran Van V, *et al.*: **Shared larval rearing environment, sex, female size and genetic diversity shape *Ae. albopictus* bacterial microbiota.** Oliveira PL, editor. *PLoS One.* 2018; **13**(4): e0194521.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Kozlova EV, Hegde S, Roundy CM, *et al.*: **Microbial interactions in the mosquito gut determine *Serratia* colonization and blood-feeding propensity.** *ISME J.* 2021; **15**(1): 93–108.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Ganley JG, D'Ambrosio HK, Shieh M, *et al.*: **Coculturing of Mosquito-Microbiome Bacteria Promotes Heme Degradation in *Elizabethkingia anophelis*.** *ChemBioChem.* 2020; **21**(9): 1279–1284.
PubMed Abstract | Publisher Full Text
18. Hegde S, Khanipov K, Albayrak L, *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; **9**: 2160.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Heu K, Romoli O, Schönbeck JC, *et al.*: **The Effect of Secondary Metabolites Produced by *Serratia marcescens* on *Aedes aegypti* and Its Microbiota.** *Front Microbiol.* 2021; **12**: 645701.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Mitri C, Bischoff E, Cuesta EB, *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; **11**: 306.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Vogel KJ, Valzania L, Coon KL, *et al.*: **Transcriptome Sequencing Reveals Large-Scale Changes in Axenic *Aedes aegypti* Larvae.** Dimopoulos G, editor. *PLoS Negl Trop Dis.* 2017; **11**(1): e0005273.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Short SM, Mongodin EF, MacLeod HJ, *et al.*: **Amino acid metabolic signaling influences *Aedes aegypti* midgut microbiome variability.** *PLoS Negl Trop Dis.* 2017; **11**(7): e0005677.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Stathopoulos S, Neafsey DE, Lawniczak MKN, *et al.*: **Genetic Dissection of *Anopheles gambiae* Gut Epithelial Responses to *Serratia marcescens*.** Schneider DS, editor. *PLoS Pathog.* 2014; **10**(3): e1003897.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Muturi EJ, Dunlap C, Ramirez JL, *et al.*: **Host blood-meal source has a strong impact on gut microbiota of *Aedes aegypti*.** *FEMS Microbiol Ecol.* 2019; **95**(1). PubMed Abstract | Publisher Full Text

25. Chen S, Zhang D, Augustinos A, *et al.*: **Multiple Factors Determine the Structure of Bacterial Communities Associated With *Aedes albopictus* Under Artificial Rearing Conditions.** *Front Microbiol.* 2020; **11**: 605. PubMed Abstract | Publisher Full Text | Free Full Text

26. Zhao T, Li BQ, Gao HT, *et al.*: **Metagenome Sequencing Reveals the Microbiome of *Aedes albopictus* and Its Possible Relationship With Dengue Virus Susceptibility.** *Front Microbiol.* 2022; **13**: 891151. PubMed Abstract | Publisher Full Text | Free Full Text

27. Wang YT, Shen RX, Xing D, *et al.*: **Metagenome Sequencing Reveals the Midgut Microbiota Makeup of *Culex pipiens quinquefasciatus* and Its Possible Relationship With Insecticide Resistance.** *Front Microbiol.* 2021; **12**: 625539. PubMed Abstract | Publisher Full Text | Free Full Text

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 Genom.* 2020; **6**(8): mgen000409. PubMed Abstract | Publisher Full Text | Free Full Text

29. Hooper R, Brealey JC, van der Valk T, *et al.*: **Host-derived population genomics data provides insights into bacterial and diatom composition of the killer whale skin.** *Mol Ecol.* 2019; **28**(2): 484–502. PubMed Abstract | Publisher Full Text | Free Full Text

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

31. LaBonte NR, Jacobs J, Ebrahimi A, *et al.*: **Data mining for discovery of endophytic and epiphytic fungal diversity in short-read genomic data from deciduous trees.** *Fungal Ecol.* 2018; **35**: 1–9. Publisher Full Text

32. Salzberg SL, Hotopp JCD, Delcher AL, *et al.*: **Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila* species.** *Genome Biol.* 2005; **6**(3): R23. PubMed Abstract | Publisher Full Text | Free Full Text

33. Martinson VG, Magoc T, Koch H, *et al.*: **Genomic Features of a Bumble Bee Symbiont Reflect Its Host Environment.** *Appl Environ Microbiol.* 2014; **80**(13): 3793–803. PubMed Abstract | Publisher Full Text | Free Full Text

34. Fierst JL, Murdock DA, Thanthiriwatte C, *et al.*: **Metagenome-Assembled Draft Genome Sequence of a Novel Microbial *Stenotrophomonas maltophilia* Strain Isolated from *Caenorhabditis remanei* Tissue.** *Genome Announc.* 2017; **5**(7): e01646–16. PubMed Abstract | Publisher Full Text | Free Full Text

35. Chen C, Compton A, Nikolouli K, *et al.*: **Marker-assisted mapping enables forward genetic analysis in *Aedes aegypti*, an arboviral vector with vast recombination deserts.** *Genetics.* 2022; **222**(3): iyac140. PubMed Abstract | Publisher Full Text | Free Full Text

36. Crava C, Varghese FS, Pischedda E, *et al.*: **Immunity to infections in arboviral vectors by integrated viral sequences: an evolutionary perspective.** *Evol Biol.* 2020. Publisher Full Text

37. Ford SA, Allen SL, Ohm JR, *et al.*: **Selection on *Aedes aegypti* alters *Wolbachia*-mediated dengue virus blocking and fitness.** *Nat Microbiol.* 2019; **4**(11): 1832–1839. PubMed Abstract | Publisher Full Text | Free Full Text

38. Faucon F, Dusfour I, Gaude T, *et al.*: **Identifying genomic changes associated with insecticide resistance in the dengue mosquito *Aedes aegypti* by deep targeted sequencing.** *Genome Res.* 2015; **25**(9): 1347–59. PubMed Abstract | Publisher Full Text | Free Full Text

39. The Anopheles gambiae 1000 Genomes Consortium: **Genome variation and population structure among 1142 mosquitoes of the African malaria vector species *Anopheles gambiae* and *Anopheles coluzzii*.** *Genome Res.* 2020; **30**(10): 1533–1546. PubMed Abstract | Publisher Full Text | Free Full Text

40. Köster M: **Sustainable data analysis with Snakemake.** *F1000Res.* 2022; **10**(33).

41. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods.* 2012; **9**(4): 357–9. PubMed Abstract | Publisher Full Text | Free Full Text

42. Lu J, Salzberg SL: **Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2.** *Microbiome.* 2020; **8**(1): 124. PubMed Abstract | Publisher Full Text | Free Full Text

43. Lu J, Breitwieser FP, Thielen P, *et al.*: **Bracken: estimating species abundance in metagenomics data.** *PeerJ Comput Sci.* 2017; **3**: e104. Publisher Full Text

44. Truong DT, Franzosa AE, Tickle TL, *et al.*: **MetaPhlAn2 for enhanced metagenomic taxonomic profiling.** *Nat Methods.* 2015; **12**(10): 902–3. PubMed Abstract | Publisher Full Text

45. The Genome Standards Consortium, Bowers RM, Kyrpides NC, *et al.*: **Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea.** *Nat Biotechnol.* 2017; **35**(8): 725–31. PubMed Abstract | Publisher Full Text | Free Full Text

46. Andrew SAS: **FASTQC: A Quality Control Tool for High Throughput Sequence Data**. Reference Source

47. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *EMBnet.* 2011; **17**(1): 10–12. Publisher Full Text

48. Valiente-Mullor C, Beamud B, Ansari I, *et al.*: **One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads.** *PLoS Comput Biol.* 2021; **17**(1): e1008678. PubMed Abstract | Publisher Full Text | Free Full Text

49. Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–9. PubMed Abstract | Publisher Full Text | Free Full Text

50. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–2. PubMed Abstract | Publisher Full Text | Free Full Text

51. Hall AB, Basu S, Jiang X, *et al.*: **A male-determining factor in the mosquito *Aedes aegypti*.** *Science.* 2015; **348**(6240): 1268–1270. PubMed Abstract | Publisher Full Text | Free Full Text

52. Lee Y, Schmidt H, Collier TC, *et al.*: **Genome-wide divergence among invasive populations of *Aedes aegypti* in California.** *BMC Genomics.* 2019; **20**(1): 204. PubMed Abstract | Publisher Full Text | Free Full Text

53. Lau MJ, Schmidt TL, Yang Q, *et al.*: **Genetic stability of *Aedes aegypti* populations following invasion by *w*Mel *Wolbachia*.** *BMC Genomics.* 2021; **22**(1): 894. PubMed Abstract | Publisher Full Text | Free Full Text

54. Rose NH, Dabo S, da Veiga Leal S, *et al.*: **Enhanced mosquito vectorial capacity underlies the Cape Verde Zika epidemic.** *PLoS Biol.* 2022; **20**(10): e3001864. PubMed Abstract | Publisher Full Text | Free Full Text

55 Li D, Liu CM, Luo R, *et al.*: **MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph.** *Bioinformatics.* 2015; **31**(10): 1674–6. PubMed Abstract | Publisher Full Text

56. Gurevich A, Saveliev V, Vyahhi N, *et al.*: **QUAST: quality assessment tool for genome assemblies.** *Bioinformatics.* 2013; **29**(8): 1072–5. PubMed Abstract | Publisher Full Text | Free Full Text

57. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2009; **25**(14): 1754–60. PubMed Abstract | Publisher Full Text | Free Full Text

58. Kang DD, Li F, Kirton E, *et al.*: **MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies.** *PeerJ.* 2019; **7**: e7359. PubMed Abstract | Publisher Full Text | Free Full Text

59. Parks DH, Imelfort M, Skennerton CT, *et al.*: **CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.** *Genome Res.* 2015; **25**(7): 1043–55. PubMed Abstract | Publisher Full Text | Free Full Text

60. Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–2. PubMed Abstract | Publisher Full Text

61. Chaumeil PA, Mussig AJ, Hugenholtz P, *et al.*: **GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database.** *Bioinformatics.* 2019; **36**(6): 1925–1927. PubMed Abstract | Publisher Full Text | Free Full Text

62. Hyatt D, Chen GL, Locascio PF, *et al.*: **Prodigal: prokaryotic gene recognition and translation initiation site identification.** *BMC Bioinformatics.* 2010; **11**(1): 119. PubMed Abstract | Publisher Full Text | Free Full Text

63. Parks DH, Chuvochina M, Waite DW, *et al.*: **A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.** *Nat Biotechnol.* 2018; **36**(10): 996–1004. PubMed Abstract | Publisher Full Text

64. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching.** *Nucleic Acids Res.* 2011; **39**(Web Server issue): W29–37. PubMed Abstract | Publisher Full Text | Free Full Text

65. Matsen FA, Kodner RB, Armbrust EV: **pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree.** *BMC Bioinformatics.* 2010; **11**: 538. PubMed Abstract | Publisher Full Text | Free Full Text

66. Matthews BJ, Dudchenko O, Kingan SB, *et al.*: **Improved reference genome of *Aedes aegypti* informs arbovirus vector control.** *Nature.* 2018; **563**(7732): 501–507. PubMed Abstract | Publisher Full Text | Free Full Text

67. Dahl EM, Neer E, Bowie KR, *et al.*: ***microshades*: An R Package for Improving Color Accessibility and Organization of Microbiome Data.** *Microbiol Resour Announc.* 2022; **11**(11): e0079522. PubMed Abstract | Publisher Full Text | Free Full Text

68. Dada N, Jupatanakul N, Minard G, *et al.*: **Considerations for mosquito microbiome research from the Mosquito Microbiome Consortium.** *Microbiome.* 2021; **9**(1): 36.
PubMed Abstract | Publisher Full Text | Free Full Text

69. Glendinning L, Stewart RD, Pallen MJ, *et al.*: **Assembly of hundreds of novel bacterial genomes from the chicken caecum.** *Genome Biol.* 2020; **21**(1): 34.
PubMed Abstract | Publisher Full Text | Free Full Text

70. Almeida A, Mitchell AL, Boland M, *et al.*: **A new genomic blueprint of the human gut microbiota.** *Nature.* 2019; **568**(7753): 499–504.
PubMed Abstract | Publisher Full Text | Free Full Text

71. Wilkinson T, Korir D, Ogugo M, *et al.*: **1200 high-quality metagenome-assembled genomes from the rumen of African cattle and their relevance in the context of sub-optimal feeding.** *Genome Biol.* 2020; **21**(1): 229.
PubMed Abstract | Publisher Full Text | Free Full Text

72. Watson M: **New insights from 33,813 publicly available metagenome-assembled-genomes (MAGs) assembled from the rumen microbiome.** *Microbiology.* 2021.
Publisher Full Text

73. Stewart RD, Auffret MD, Warr A, *et al.*: **Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen.** *Nat Commun.* 2018; **9**(1): 870.
PubMed Abstract | Publisher Full Text | Free Full Text

74. Mancini MV, Damiani C, Accoti A, *et al.*: **Estimating bacteria diversity in different organs of nine species of mosquito by next generation sequencing.** *BMC Microbiol.* 2018; **18**(1): 126.
PubMed Abstract | Publisher Full Text | Free Full Text

75. Coon KL, Hegde S, Hughes GL: **Interspecies microbiome transplantation recapitulates microbial acquisition in mosquitoes.** *Microbiology.* 2022; **10**(1): 58.
PubMed Abstract | Publisher Full Text | Free Full Text

76. Muturi EJ, Njoroge TM, Dunlap C, *et al.*: **Blood meal source and mixed blood-feeding influence gut bacterial community composition in *Aedes aegypti.*** *Parasit Vectors.* 2021; **14**(1): 83.
PubMed Abstract | Publisher Full Text | Free Full Text

77. Chen S, Johnson BK, Yu T, *et al.*: ***Elizabethkingia anophelis*: Physiologic and Transcriptomic Responses to Iron Stress.** *Front Microbiol.* 2020; **11**: 804.
PubMed Abstract | Publisher Full Text | Free Full Text

78. Onyango MG, Lange R, Bialosuknia S, *et al.*: **Zika virus and temperature modulate *Elizabethkingia anophelis* in *Aedes albopictus*.** *Parasit Vectors.* 2021; **14**(1): 573.
PubMed Abstract | Publisher Full Text | Free Full Text

79. de O Gaio A, Gusmão DS, Santos AV, *et al.*: **Contribution of midgut bacteria to blood digestion and egg production in *Aedes aegypti* (diptera: culicidae) (L.).** *Parasit Vectors.* 2011; **4**: 105.
PubMed Abstract | Publisher Full Text | Free Full Text

80. Chrisman B, He C, Jung JY, *et al.*: **The human "contaminome": bacterial, viral, and computational contamination in whole genome sequences from 1000 families.** *Sci Rep.* 2022; **12**(1): 9863.
PubMed Abstract | Publisher Full Text | Free Full Text

81. Laurence M, Hatzis C, Brash DE: **Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes.** *PLoS One.* 2014; **9**(5): e97876.
PubMed Abstract | Publisher Full Text | Free Full Text

82. Castillo DJ, Rifkin RF, Cowan DA, *et al.*: **The Healthy Human Blood Microbiome: Fact or Fiction?** *Front Cell Infect Microbiol.* 2019; **9**: 148.
PubMed Abstract | Publisher Full Text | Free Full Text

83. Cataneo AHD, Kuczera D, Mosimann ALP, *et al.*: **Detection and clearance of a mosquito densovirus contaminant from laboratory stocks of Zika virus.** *Mem Inst Oswaldo Cruz.* 2019; **114**: e180432.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ? ? ✓

Version 2

Reviewer Report 09 August 2023

https://doi.org/10.21956/wellcomeopenres.21596.r64469

✓ **Mariana Rocha David**
Laboratório de Mosquitos Transmissores de Hematozoários, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, Brazil

The present manuscript, "Recovery of metagenomic data from the *Aedes aegypti* microbiome using a reproducible snakemake pipeline: MINUUR" is a method article that presents a workflow to analyse non-host reads from existing genomic data with the aim to extract host associated microbe sequences. Using 62 *Aedes aegypti* samples, the results indicated that it is possible to identify bacterial phyla and genera usually found in association with this mosquito species. There was a remarkable variation in taxonomic diversity between sequencing projects, as well as there was some variation between samples (which is in line with mosquito microbiota studies). This workflow can increase the knowledge about bacteria potentially associated to *Aedes aegypti*, although it is not possible to distinguish between symbionts and contaminants (as pointed by the authors). In the following text I present some questions and suggestions about the present manuscript.

Methods:

- Do the authors also consider searching for fungal, protozoan, and viral DNA sequences?

Results:

- Figure 6C: it seems to be an empty plot.

Discussion:

- What is the meaning of “constituent” in, “Whilst the nature of the samples can act as confounder given differences in sample handling preparation, we note the relative abundance of the key phyla, and constituent genera, are varied across these projects”?

- It is also important to consider that some bacteria could be on the surface of the mosquito, since the samples were not prepared for microbiota analysis (e.g. they did not have their external surface sterilized with ethanol).

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Vector ecology, symbiotic interactions, vector competence, arbovirus transmission

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 31 May 2023

https://doi.org/10.21956/wellcomeopenres.21596.r57419

?

Ellen Cameron
[1] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, UK
[2] Wellcome Sanger Institute, Hinxton, England, UK

In “Recovery of metagenomic data from the *Aedes aegypti* microbiome using a reproducible snakemake pipeline: MINUUR [version 2]”, Foo et al. describe a new Snakemake workflow, *MINUUR* . *MINUUR* is then used to characterize the *Aedes aegypti* microbiome in publicly available whole genome sequencing datasets on the European Nucleotide Archive. This workflow functions at the interface between whole genome sequencing and metagenomic sequencing to recover metagenomic assembled genomes (MAGs) of bacterial symbionts and contaminants in a host (*Aedes aegypti*) through the implementation of commonly used metagenomic tools.

Overall, *MINUUR* is rooted on the basis of re-assembly of non-mapped reads which will allow for the successful recovery of present symbionts. Numerous tools exist for metagenomic assembly and binning and currently, *MINUUR* utilizes *MEGAHIT* for assembly and *metaBAT2* for binning. On this front, I believe that there are two opportunities for further development of the workflow:

Assembly: Is there a reason why authors chose to use *MEGAHIT* and did authors also evaluate *metaSPAdes* as an assembler? *metaSPAdes* has been known to give better quality assemblies but may have higher resource requirements/run time. It would be nice for the workflow to also give users the option of using *metaSPAdes*.

Binning: Currently, the workflow only implements *metaBAT2* for binning. Including additional binners (e.g., *CONCOCT*) and a bin refinement tool (e.g., *dasTOOL*, *metaWRAP* bin refinement module, *BinSPreader, GraphBin*) may improve the recovery of metagenomic assembled genomes. I would recommend including adding in additional binning/bin-refinement to have more robust evaluation of assemblies.

It is evident that the authors have put a lot of time and effort into the development of a user-friendly Snakemake workflow. The GitHub is well organized and the Snakemake workflow structure (e.g., Snakefile, rules directory, scripts, environments) is very clean. I also appreciate their inclusion of test data which will allow for users to easily test out the workflow. The authors have also prepared detailed documentation for set-up and operation of the workflow in a Jupyter notebook. I have a few minor suggestions for the workflow which I think may make it more accessible and user friendly:

More automation in setup requirements: In the documentation, clear instructions are provided for users on what databases are required for operation. However, users are required to download these databases manually. Including the ability for databases to automatically download would make setup and operation more user-friendly. This could be done through downloading from the hosts of the databases (e.g., *wget* to link in a Snakemake rule with a “Download databases” binary True/False option in the config). Additionally, on the topic of automation in the workflow setup, the authors provide great instructions on how users can use *MINUUR* for alternative hosts by creating the bwa index as outlined in their Jupyter notebook. I believe that this could also be included as a simple option in the config as an additional optional rule in the workflow to make the tool more user-friendly and to simplify the application of this workflow in non-mosquito hosts.

Conda environments: In the instructions for *Running MINUUR*, users are instructed to use the flag *–use-conda* with the snakemake command. I noticed in the GitHub that there are currently six conda environments required for execution. Having numerous conda environments will increase the time required during the initial setup of the workflow, but also can take up a lot of storage. If there are not version conflicts, it would be great if the total number of conda environments could be reduced by including more of the dependencies in a single environment. Alternatively, I also noticed that there is a Dockerfile but no mention of this – will there be the option to run the workflow using docker/singularity in the future instead?

The authors demonstrated the utility of their workflow by applying it to numerous publicly available datasets and presented these key findings. I have a few minor comments about their figures that I believe will help to address overall readability:

X-axis labels: I noticed in several of the figures that the x-axis labels were often difficult to read because there was minimal spacing between adjacent labels. I would adjust this to ensure the labels can be clearly read. Specific instances where I noticed this include: Figure 3B, Figure 4B, Figure 4C, Figure 6A.

Axis Inconsistencies: Within some of the figures subfigures, there are inconsistencies in the axis labelling/limits. While this does not directly interfere with readability and interpretation, updating these will just make for much higher quality and cohesive visuals. Specific examples: i) Figure 3 – subfigure A and B do not expand to 0 but C does. Figure B includes (%) in the y-axis title, but figure C includes % after each number in the y-axis tick labels.

Figure 3B: I am not clear on what data is being presented here. Based on the text where this figure is referenced, I assume it is the number of samples that each genus has been identified in (e.g., a taxon that was identified in 31 samples == 50%?) but the figure legend and axis labels do not clearly communicate this. Perhaps including "detection" somewhere on the graph/figure legend text would help to clarify this?

Figure 3C: On initial look at this graph, I was confused because it looked like your bars were not adding up to 100%. However, on closer inspection, I realized that this is due to the light shade being plotted for "Other" and that it just has the appearance of being 'white'/empty space. I think that either including an outline or choosing a darker shade in place of the pale grey would improve this. I would also italicize the genera names in the legends.

Figure 4: I would include the colours for the MAG classification as a legend next to the graphs to increase readability. I think it could also be interesting to plot the genome sizes and N50 scores for the low-quality MAGs (perhaps as a supplementary figure?).

Figure 5A: Faceting the panels by the MAG quality instead could be a nice alternative to allow for cross-project comparisons? This is just personal preference though depending on what you are trying to highlight here.

Figure 6C: I cannot see any data plotted here and only am able to see the empty plots (E.g., axis labels, facet labels). This is likely just due to an issue with the geom style not cooperating when inserted as a picture in Word (I commonly see this with geom_point() and inserting as a PDF image in Microsoft Word).

I believe that the inclusion of a eukaryotic scoring also increases robustness of the workflow to not only focus on prokaryotic MAGs. However, I am interested as to how "eukaryotic contamination" is defined. Is this serving to identify instances where host (*i.e., Aedes aegypti*) reads have made it into the newly assembled reads or could this instead be the presence of eukaryotic symbionts and/or be indicative of blood meal diet source? Clarifying this in the text will help to contextualize the usage of the BUSCO scoring.

I believe that the authors should also include a table in the supplementary data highlighting the MAG quality information (completeness, contamination) including what sample and what project these MAGs were identified in. For example, authors make mention of a specific instance of PRJNA385349 not containing any high-quality MAGs and I would be interested to see what the quality of MAGs were (*i.e.*, were they nowhere close to being high-quality or were they on the

threshold?).

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Metagenomics, bioinformatics, symbiotic interactions

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 1

Reviewer Report 17 April 2023

https://doi.org/10.21956/wellcomeopenres.21234.r56050

?

Taylor Reiter
[1] Arcadia Sciences, Berkeley, California, USA
[2] Arcadia Sciences, Berkeley, California, USA

In, “Recovery of metagenomic data from the *Aedes aegypti* microbiome using a reproducible snakemake pipeline: MINUUR,” Foo *et al.* describe a snakemake workflow to leverage contamination in whole genome sequencing data from mosquitos to discover microbiome

constituents. Contamination is rampant in WGS projects, so this workflow cleverly leverages that information to inform new biological insights.

The steps the authors chose for their workflow make sense. I have alternative preferences for some steps — using fastp instead of FastQC → cutadapt → FastQC and using sourmash gather for taxonomic annotation of reads instead of kraken2. However, none of these choices are terribly consequentially important to the downstream outcomes. I think the tool could do a better job of assessing residual host contamination in unmapped reads and in resultant MAGs. This is particularly concerning as most high- and medium-quality MAGs were larger than expected. Some strategies I can think of to do this would be:

1. Mapping against a pangenome instead of a reference genome in the initial step to remove more reads.

2. Using bbduk to remove contamination with a low k-mer size (k=21 or k=17).

3. Assessing for eukaryote or host-specific genes in the unmapped reads (e.g. using DIAMOND or something similar) or in the assemblies (using BUSCO or something similar).

I also find it unfortunate that that authors chose not to include GTDBtk as part of the workflow, as most people would probably like to run a similar step.

Another concern I have with this approach is detecting true host contamination (e.g. symbionts) vs. process-oriented contamination (kit contamination, index hopping, etc). It would be interesting to compare the kraken2 or GTDB-tk labels against lists of organisms that are frequently contaminants. If the authors could also back track and determine the concentration of DNA used in sequencing and compare this to the number of organisms detected, this might highlight whether process-oriented contamination drives results in any samples. The authors may also look into GUNC as another tool for assessing contamination separate from checkM.

I have concerns around the implementation of the MINUUR workflow. As a data analysis project, the process the authors went through to run this analysis is phenomenally documented and I found it easy to follow. However, the authors present the workflow as a tool that others could use for other WGS projects. I think in this context, the workflow lacks some adherence to software engineering principles that could greatly facilitate downstream use:

1. Some of the software tools that are installed via conda by the workflow do not have recorded versions (humann, samtools, bam2fastq to name a few).

2. Some of the R packages that are used are not installed via conda, which will make it difficult for new users to use the pipeline.

3. The workflow lacks tests. I don't think third party tools need unit tests, but the workflow could be greatly improved by including unit tests for the new R and python scripts that are introduced there in.

4. It appears that users must download databases themselves, which creates cognitive burden associated with getting the workflow up and running.

5. It would be super useful to have a small toy dataset that will quickly run through the whole pipeline so that users could validate that the workflow is up and running on their system.

Some smaller things I highlighted while reading include:

1. In the first paragraph of the results section, it would be helpful if you could report the percent of unmapped reads instead of absolute numbers.

2. It would be interesting to compare the kraken2 results against the GTDB tk results.

3. Some of the font in the figures is very small, making it difficult to read.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Metagenomics, automation with workflows, contamination detection, bioinformatics of sequencing data

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 18 May 2023

Eva Heinz

Reviewer Responses

I think the tool could do a better job of assessing residual host contamination in unmapped reads and in resultant MAGs. This is particularly concerning as most high- and medium-quality MAGs were larger than expected. Some strategies I can think of to do this would be:

1. **Mapping against a pangenome instead of a reference genome in the initial step to remove more reads**

2. **Using bbduk to remove contamination with a low k-mer size (k=21 or k=17).**
3. **Assessing for eukaryote or host-specific genes in the unmapped reads (e.g. using DIAMOND or something similar) or in the assemblies (using BUSCO or something similar).**

We thank the reviewer for these suggestions. We have added BUSCO to the pipeline as a step to assess the final assemblies. The step can either be used to look for Eukaryotic (--auto-lineage-euk) or prokaryotic marker genes (--auto-lineage-prok). To address the reviewers query, we used BUSCO to search for eukaryotic specific genes within high and medium-quality MAGs. The manuscript has been updated with the following sections:

- Within the methods section we added the following sentence to account for the inclusion of BUSCO: *"In addition, BUSCO (v5.4.7) can be included to search for eukaryotic or prokaryotic specific genes in the final MAGs as an additional quality assurance metric"*

- And in the results, we include the following section: *"We further applied BUSCO to assess eukaryotic contamination in the final assemblies. Eukaryotic contamination was 3.24% (0 – 4.47%) and 2.26% (0.4 = 4.7%) on average in high and medium quality MAGs respectively (Supplementary Figure 2, Extended data)"*

This graph is now included in Supplementary Figure 2 (under the new doi https://doi.org/10.5281/zenodo.7941078): **Supplementary Figure 2: BUSCO (v5.4.7) Assessment of Eukaryotic Contamination**. BUSCO percentage of complete eukaryotic marker genes from the eukaryota_odb10 database. Y-axis represents the percentage of genes from the eukaryota_osb10 database present in the assembly, X-axis represents the sample and bin id (sample.bin).

I also find it unfortunate that that authors chose not to include GTDBtk as part of the workflow, as most people would probably like to run a similar step.

This is a fair comment, however, we decided to exclude GTDB-Tk from the workflow since another large database dependency (The Genome Taxonomy Database) would be problematic. Also, since MAGs are not always acquired using this pipeline, setting up GTDB-Tk and downloading the required reference databases as part of the workflow would be computationally and time expensive, with no potential return. We decided it would be better for individuals who wanted to repeat this analysis to run GTDB separately if they were able to acquire MAGs from their host of interest using the workflow.

Another concern I have with this approach is detecting true host contamination (e.g. symbionts) vs. process-oriented contamination (kit contamination, index hopping, etc). It would be interesting to compare the kraken2 or GTDB-tk labels against lists of organisms that are frequently contaminants. If the authors could also back track and determine the concentration of DNA used in sequencing and compare this to the number of organisms detected, this might highlight whether process-oriented contamination drives results in any samples. The authors may also look into GUNC as another tool for assessing contamination separate from checkM.

We thank the reviewer for highlighting this. We agree and have addressed this limitation in text in the final paragraph of the discussion, with citations of key papers discussing the issue. Within the discussion, we have added the following paragraph:

"A limitation of this workflow is knowing whether this data is indicative of symbionts associated to the *Aedes aegypti* microbiome, or sequencing contamination (78–80). We believe the majority of our results support the identification of true microbial symbionts to the mosquito microbiome; our taxonomic classifications from MAGs and KRAKEN2 read assignments are congruent with previous studies (4,10,13,18,19,25,72,74). However, we also identify likely contaminants such as *Brevidensovirus* (Figure 3), which have previously been identified as ZIKV stock contamination (81). Discerning between symbiont or contaminant from results generated through this workflow requires further cross reference to the host-microbiome literature in question. Further analysis can also help answer questions of contamination or symbiont results, such as measuring species genetic similarity with other sequenced host-symbionts."

I have concerns around the implementation of the MINUUR workflow. As a data analysis project, the process the authors went through to run this analysis is phenomenally documented and I found it easy to follow. However, the authors present the workflow as a tool that others could use for other WGS projects. I think in this context, the workflow lacks some adherence to software engineering principles that could greatly facilitate downstream use

1. **Some of the software tools that are installed via conda by the workflow do not have recorded versions (humann, samtools, bam2fastq to name a few).**

   We have updated all of these packages accordingly – we thank the reviewer for pointing this out. (https://github.com/aidanfoo96/MINUUR/tree/main/workflow/envs)

2. **Some of the R packages that are used are not installed via conda, which will make it difficult for new users to use the pipeline.**

   The R packages have been added with specified versions (https://github.com/aidanfoo96/MINUUR/tree/main/workflow/envs)

3. **The workflow lacks tests. I don't think third party tools need unit tests, but the workflow could be greatly improved by including unit tests for the new R and python scripts that are introduced there in.**

   We have added a unit test for the pipeline which tests all third party and R packages are can be properly configured within their conda environments specified in the pipeline. Github actions re-runs this with each push request. We have also added the GitHub actions badge to the Readme to make users aware of the workflow's status.

4. **It appears that users must download databases themselves, which creates cognitive burden associated with getting the workflow up and running.**

   We certainly agree with this point. However, we decided for a long-term perspective that users should download the respective databases themselves. The first reasoning for this was that the teams who host these databases do a fantastic job maintaining the databases in dedicated repositories – this is something we are unable to do.

Second, these databases are updated often to reflect the growing number of reference data available; this is something we do not have the resources to do. Instructions for installing and configuring these databases is well documented, and repositories are hosted (such as Ben Langmead's repository cited in the paper) to provide precompiled databases that a user can download and implement.

5. **It would be super useful to have a small toy dataset that will quickly run through the whole pipeline so that users could validate that the workflow is up and running on their system**.

   This is a great suggestion by the reviewer – we have added a toy dataset into the data repository of the pipeline (https://github.com/aidanfoo96/MINUUR/tree/main/workflow/data) and configured the sample_list and sample_table to automatically point to this dummy data. We have also added instructions to the JupyterBooks page so that when a user initializes the pipeline with the required databases they can run through the workflow with the dummy data.

6. **In the first paragraph of the results section, it would be helpful if you could report the percent of unmapped reads instead of absolute numbers**

   We have changed the following section.

7. **Extraction of Non-Reference Reads Post-Alignment to *Aedes aegypti*,** now reads: *"91.9% of reads aligned to the AaegL5.3 reference genome, while 4.6% of reads did not map to the AaegL5.3 reference genome"*.

8. **It would be interesting to compare the KRAKEN2 results against the GTDB-Tk results**

   Thank you to the reviewer for the suggestion – we included in the results section of the original manuscript a summary of the KRAKEN2 classifications and GTDB-Tk classifications. Within the section **"Relationship between KRAKEN2 Taxonomic Classifications and MAG Recovery"** we included the following sentence: *"taxonomic classifications with a high number of KRAKEN2 assigned reads (>100,000) are akin to the taxonomic classifications of medium and high-quality MAGs (Supplementary Figure 3)."*

   We have expanded on this in the updated version of the manuscript with the following: "*For example, across multiple samples, Serratia, Elizabethkingia and Wolbachia were assigned >100,000 reads and resulted in six, 27 and 13 medium and high-quality MAGs respectively."*

9. **Some of the font in the figures is very small, making it difficult to read.**
   We have examined our figures and made the bar charts in Figure 4 larger to better distinguish the x-axis labels.

Competing Interests: No competing interests were disclosed.