MINUUR: Microbial INsight Using Unmapped Reads

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

***Correspondence**
Eva Heinz; eva.heinz@lstmed.ac.uk

1 Abstract

The microbiome is a collection of microbes that exist in symbiosis with a host. Whole genome sequencing produces off-target, non-specific reads, to the host in question, which can be used for metagenomic inference of a microbiome. This data is advantageous over barcoding methods since higher taxonomic resolution and functional predictions of microbes are possible. With the growing number of genomic sequencing data publicly available, comes opportunity to elucidate reads pertaining to the microbiome. However, characterization of these reads can be complex, with many steps

required to perform a robust analysis. To address this, we developed MINUUR (**M**icrobial **IN**sights **U**sing **U**nmapped **R**eads); a snakemake pipeline to characterize non-host reads from existing genomic data. We apply this pipeline to ten, publicly available, high coverage *Aedes aegypti* (*Ae. aegypti*) genomic samples. Using MINUUR, we describe species level microbial classifications; predict microbe associated genes and pathways and find bacterial metagenome assembled genomes (MAGs) associated to the *Ae. aegypti* microbiome. Of these MAGS, 19 are high-quality representatives with over 90% completeness and under 5% contamination. In summary, we present an in-depth analysis of non-host reads from *Ae. aegypti* whole genome sequencing data within a reproducible and open-access pipeline.

2 Introduction

A microbiome refers to the collection of microbes and their genomic content that exist in symbiosis with a host (1). To identify taxa within a microbiome, culture independent approaches are commonly used (2,3) such as amplicon-based sequencing with taxonomic barcodes (3) or metagenomic shotgun sequencing (4). The later approach is advantageous because higher taxonomic resolution and functional predictions are possible, either from the

52 sequenced reads directly or using contiguous assemblies. Sequencing the
53 DNA of a whole organism to obtain its genomic information yields data from
54 the primary organism of interest (referred throughout the manuscript as
55 "host" for simplicity), but potentially also reads corresponding to
56 endo/ectosymbionts, pathogens or environmental contamination that is not
57 readily removed from the host during sample preparation. Indeed, studies in
58 *Drosophila*, bumble bees, killer whales, moths and nematodes have shown
59 existing whole genome sequencing (WGS) data is a rich source to
60 characterize their associated symbionts (5–10). These studies employ
61 approaches including specific enrichment of non-host with bait sequences
62 targeting a specific taxon of interest (5,8); or steps following the sequencing
63 experiment without prior enrichment, such as *de novo* metagenome
64 assemblies (9–11); prediction of microbial genes and pathways (5) or
65 classification-based methods using predefined taxonomic libraries (6).

66 Mosquitoes are important vectors for human pathogens. A prominent
67 example of this is *Aedes aegypti* (*Ae. aegypti*) which transmits pathogens
68 including dengue virus, yellow fever virus, chikungunya virus and Zika virus.
69 Dengue cases alone are estimated to cause 10,000 deaths and 100 million
70 infections per year, contributing to significant burden of human morbidity and
71 mortality worldwide (12). Studies show the mosquito microbiome influences

vectorial capacity (13), blood feeding propensity (14) and life history traits (15–17). Mosquito microbiomes are understood to be highly variable, dependent on a suite of deterministic processes such as the environment (18–21), host factors (22,23), microbial interactions (14,24,25) and mosquito-microbe interactions (26,27). These important findings have been aided by amplicon based 16S rRNA sequencing approaches to characterize the microbiome. Complementary to this, we believe a metagenomic approach would add further insight of the mosquito microbiome by adding the genomic context of key symbionts. Whole genome shotgun sequencing is commonly used to study mosquito genomics (28,29), population genomics (30) and insecticide resistance (31); meaning non-mosquito sequence data (we refer to these as unmapped reads for the remainder of the manuscript) 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 (32) and, at time of writing, currently 100,514 *Ae. 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 mosquito-microbiomes from their unmapped sequences.

To make use of this large resource of already-available data we developed MINUUR, a user configurable 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 a robust analysis of unmapped reads associated to a host in question. We used MINNUR on an existing *Ae. aegypti* study (30) and describe the associated microbes based on taxonomic read classifications; predicted genes and metabolic pathways; and reconstruct quality checked metagenome assembled genomes (MAGs) pertaining to mosquito-associated bacteria using *de novo* metagenome assemblies. The application of MINUUR can provide additional insights of existing WGS data to investigate microbes associated with their host of interest.

3 Materials and Methods

3.1 Specifications

The MINUUR pipeline (Figure 1) is implemented in Snakemake (33) and available from github at https://github.com/aidanfoo96/MINUUR. Details of the pipeline are discussed in the following section.

3.2 Database Setup

MINUUR requires several databases. This includes a high quality bowtie2-indexed reference genome (34) to separate host and non-host reads; a KRAKEN2 (35), BRACKEN (36) and MetaPhlAn3 (37) database for taxonomic read classification; and ChocoPhlAn (38) and UniRef (39) databases for functional read profiling with HUMAnN3 (38). All databases are available in their respective GitHub repositories. The databases used in this study include the default MetaPhlAn3 marker gene database, default ChocoPhlAn and UniRef90 databases, and KRAKEN2 (35) and BRACKEN (36) indexes from the Ben Langmead repository located here: https://benlangmead.github.io/aws-indexes/k2. For our study, we downloaded and compiled these default databases.

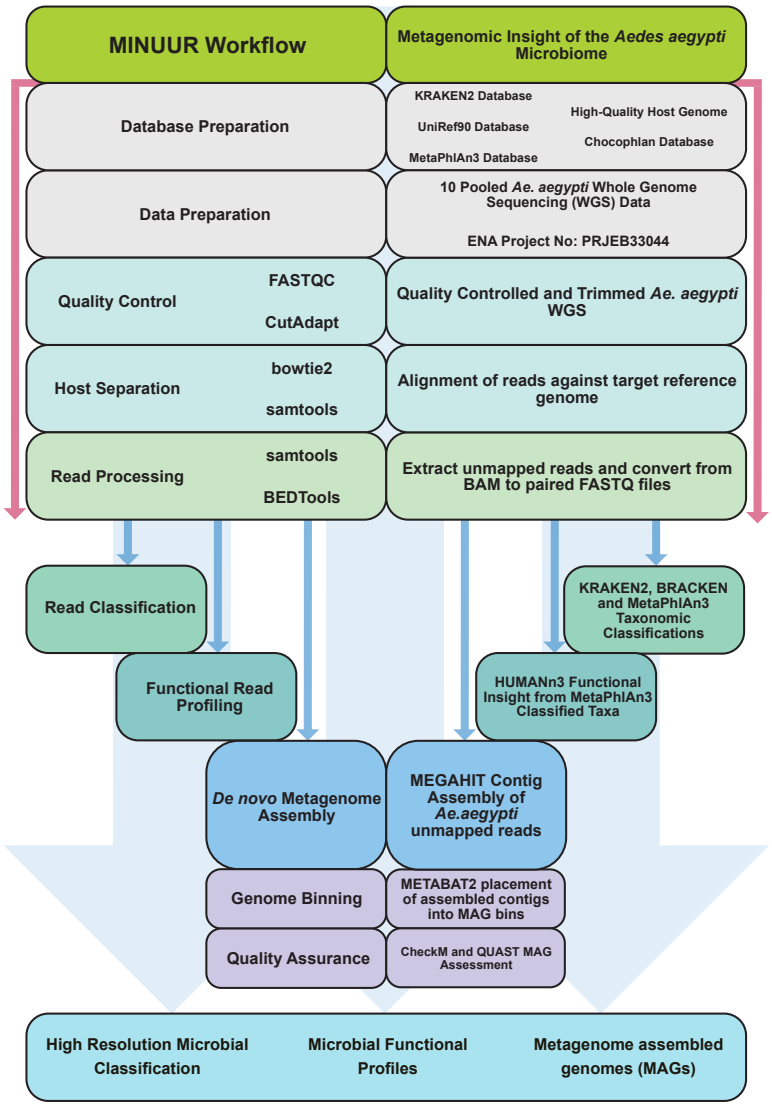


Figure 1: Microbial Insight Using Unmapped Reads (MINUUR). Workflow describing the pipeline's main steps. Top to bottom describes the workflow of the pipeline. Left panels describe MINUUR's key steps and tools, right panels describe our application of MINUUR on *Ae. aegypti* samples used in this study. Initial steps highlighted by the red arrows indicate pre-processing steps, including database preparation, quality control, read trimming, host-alignment and host-separation. Blue arrows indicate the characterization of unmapped reads. The three main outputs of MINUUR, indicated in the bottom panel, are microbial classifications, functional profiles and assembly of metagenome assembled genomes.

3.3 Data Preparation

MINUUR accepts either BAM or paired FASTQ inputs. For FASTQ inputs, MINUUR performs quality control (QC) using FASTQC (v0.11.9) (40), 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) (41) 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, reads are aligned using Bowtie2 (v2.4.4) (34) against a user defined indexed reference genome (the relevant host genome). Alignment sensitivity and type (global or local) 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 (42). As a result, unmapped reads used in subsequent steps are likely to contain a substantial number of host data. In this instance, we suggest users extract KRAKEN2 classified reads pertaining to known microbes to improve functional read profiling and metagenome assembly

(see later section). Unmapped reads within the coordinate sorted binary alignment (BAM) file are extracted using Samtools (v.1.14) (43) (”samtools -view -f 4”) and converted to FASTQ format using bedtools (v2.3.0) (44) (”–bamToFastq”). Since a large number of existing data is available in BAM format, the user may also define a BAM input, from which the pipeline will begin at the BAM separation stage.

MINUUR performs best when the initial number of reads from the host is high and library preparation stages minimize loss of prokaryotic DNA. For our study, we used an example dataset (30) which described genetic variation in *Ae. aegypti* infected with *Wolbachia* from high and low dengue virus blocking populations (30). The published sequencing data was retrieved from the European Nucleotide Archive (ENA) under the project accession number PRJEB33044 (30). We retrieved ten FASTQ files representing 90 pooled mosquitoes, sequenced with an Illumina HiSeq 3000 with 150bp paired-end reads to high coverage (>400,000,000 reads per pair) in the original study (30).

3.4 Read Classification

MINUUR uses two read classification approaches to infer taxonomy. KRAKEN2 (v2.1.2) (35) uses a k-mer based approach to map read

171 fragments of k-length against a taxonomic genome library of k-mer
172 sequences, whereas MetaPhlAn3 (v3.0.13) (37) aligns reads against a
173 library of marker genes using Bowtie2 (34). Both strategies are employed to
174 provide a wide classification range to the user, with the results subsequently
175 used in downstream analysis steps. Specifically, MINUUR provides the
176 option to use KRAKEN2 classified reads, parsed from KrakenTools (v1.2),
177 to select a specific set of reads (for example bacterial) for metagenome
178 assembly. MetaPhlAn3 taxonomic classifications are used in conjunction
179 with the ChocoPhlAn database to identify microbe associated genes and
180 pathways. The output of MetaPhlAn3 is the relative abundance of microbes
181 within a sample, whereas KRAKEN2 reports the number of reads associated
182 to a specific taxonomic ID. To estimate the relative taxonomic abundance
183 from KRAKEN2 classifications, MINUUR will parse KRAKEN2 read
184 classifications to BRACKEN (v2.6.2) (36) which uses a Bayesian probability
185 approach to redistribute reads assigned at higher taxonomic levels to lower
186 (species) taxonomic levels.

187 MINUUR outputs classified and unclassified reads to paired FASTQ files and
188 generates BRACKEN estimated taxonomic abundance profiles for further
189 analysis. Furthermore, the user can specify KrakenTools to extract a specific
190 taxon or group of taxa from KRAKEN2 - these can be used in later stages of

191 the pipeline to reduce non-specific reads for metagenome assemblies or
192 further statistical analysis at the user's discretion.

193 **3.5 Functional Read Profiling**

194 Functional profiling aims to infer microbial function directly from read
195 sequences without metagenome assembly. MINUUR implements
196 HUMAnN3 (v3.0.0) (the HMP Unified Metabolic Analysis Network) (38) to
197 functionally classify read sequences. Taxonomic classifications from
198 MetaPhlAn3 (37) are identified using the ChocoPhlAn pan-genome database
199 (37) annotated with UniRef90 (39) cluster annotations. In addition, non-
200 classified reads are searched against UniRef90 clusters to identity
201 unclassified taxonomic genes. HUMAnN3 produces taxonomic
202 classifications using MetaPhlAn3 and associated gene family abundance in
203 RPK (reads per kilobase) with UniRef90 annotations and metabolic pathway
204 abundances (RPK) and coverage.

205 **3.6 Metagenome Assembly, Binning and Quality Assurance (QA)**

206 MINUUR will perform *de novo* metagenome assembly to produce contiguous
207 sequences (contigs) from either all unmapped reads or KRAKEN2 classified
208 reads. MEGAHIT (v1.2.9) (45), a rapid and memory efficient metagenome

assembler, is used for *de novo* metagenome assembly. Assembled contigs are quality checked using QUAST (v5.0.2) (46) to assess contig N50 and L50 scores. The resultant contigs, which are ultimately 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) (47), and the original unmapped or KRAKEN2 classified 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) (48) to produce a depth file of contig coverage. The depth file and assembled contigs are input to the metagenome binner MetaBAT2 (v2.12.1) (48), to group contigs in defined genomic bins. Each bin is a predicted metagenome assembled genome (MAG). CheckM (v1.1.3) (49) is 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.

228 **3.7 Pipeline Configuration**

229 Ten paired Illumina HiSeq 3000 raw FASTQ reads were used as input in the
230 'data' directory of MINUUR, with names of each sample listed in the
231 'samples.tsv' file within the configuration directory. To implement the
232 pipeline, the configuration file was set to the following parameters: FASTQ =
233 True, QC = True, CutadaptParams = "–minimum-length 50 -q 30",
234 RemoveHostFromFastqGz = True, AlignmentSensitivity = "–sensitive-local",
235 ProcessBam = True, From- Fastq = True, KrakenClassification = True,
236 ConfidenceScore = 0, KrakenSummaries = True, GenusReadThreshold =
237 1000, SpeciesReadThreshold = 30000, ExtractKrakenTaxa = True, taxon
238 choice = "2" (bacteria), BrackenReestimation = True, ClassificationLvl = 'S'
239 (species) and 'G' (genus), DistributionThresh = 10, MetaphlanClassification
240 = True, HumannAnalysis = True, GetBiologicalProcess = True, Process =
241 'siderophore', MetagenomeAssm = True, MetagenomeBinning = True
242 (UseKrakenExtracted was set to True and False in separate pipeline runs),
243 MinimumContigLength = 1500, CheckmBinQA = True. All databases were
244 installed from their respective repositories from Github into the `resources`
245 directory of MINUUR. The pipeline was run on an Ubuntu Linux system with
246 660gb of available memory and 128 CPUs. For our analysis, with the above
247 settings and 10 cores available, MINUUR took 72 hours to complete; 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 4.1Tb (terabytes) across all 10 samples used in this study.

3.8 Taxonomic Classification of MAGs with GTDB-Tk

Separate from MINUUR, all bins produced from MetaBAT2 were taxonomically classified with GTDB-Tk (50) (v1.5.0) using ”–classify-wf” against the Genome Taxonomy Database (GTDB) (release 06-RS202, 27/04/21). GTDB-Tk assigns genes to MAGs using Prodigal (v2.6.3) (51); ranks the taxonomic domain of each MAG using 120 bacteria and 122 archaea marker genes with HMMER (52) using a published database (53). With this information, MAGs are placed into domain specific reference trees with pplacer (v1.1) (54). Taxonomic classification with GTDB-Tk is based on placement within the GTDB reference tree, relative evolutionary divergence and average nucleotide identity (ANI) scores. The relative evolutionary divergence score is used to refine ambiguous taxonomic rank assignments and ANI scores used to define species classifications. Using this approach,

strain variants are defined when average nucleotide identity is greater than 95% - below this threshold a MAG is classified as a novel species.

4 Results

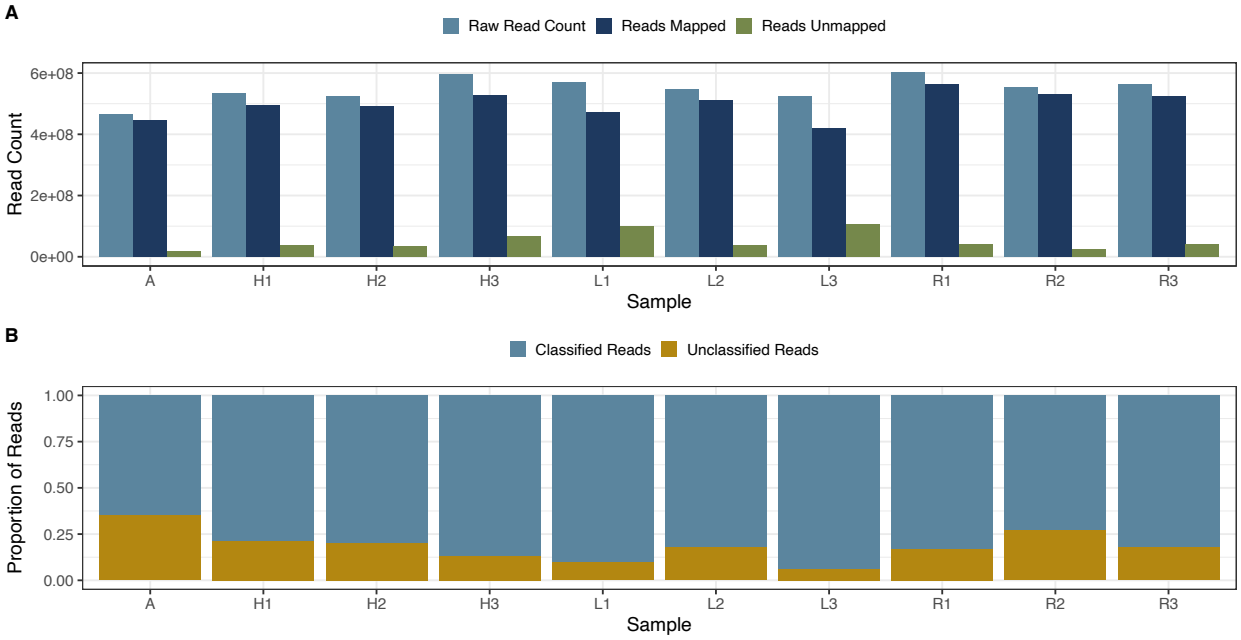
4.1 MINUUR Application

MINUUR is a Snakemake pipeline that separates and characterizes non-host, unmapped reads from WGS data using a series of metagenomic tools. The pipeline is broadly split into three paths; i) read classification with k-mers (KRAKEN2) or marker genes (MetaPhlAn3); ii) functional read profiling with HUMAnN3 and iii) *de novo* metagenome assembly with MEGAHIT followed by binning (MetaBAT2) and MAG quality assurance (QUAST and CheckM) (Figure 1). Taxonomic classifications and functional profiles are produced as 'tidy data' formats to easily parse for further analysis. MINUUR is open source and available on Github: https://github.com/aidanfoo96/MINUUR, with an accompanying WIKI page available here: https://github.com/aidanfoo96/MINUUR/wiki.

4.2 MINUUR Extracts Unmapped Reads from Host-aligned WGS Data

The initial number of reads were counted per sample. The mean number of paired reads per sample was 548,373,996, ranging between 466,236,232 to

603,970,014 reads (Figure 2A). After alignment to the *Ae. aegypti* reference genome (AaegL5.3 GCA_002204515.1) with bowtie2, the proportion of mapped and unmapped reads was calculated. On average, 497,688,151 reads (range: 418,353,293 to 563,050,257) aligned to the AaegL5.3 reference genome, averaging 90.8% read alignment (Figure 2A). To estimate the number of reads associated to the microbiome, we calculated the overall number of KRAKEN2 classifications from all unmapped reads (Figure 2B). On average, MINUUR classified 81.3% of reads that did not map to the *Ae. aegypti* genome (Figure 2B). The mean number of classified reads using KRAKEN2 was 19,910,928, ranging between 4,887,498 to 48,298,960 reads, and the number of unclassified reads was 3,331,246 on average, ranging between 2,702,596 and 4,614,580 reads (Figure 2C).



C

ENA Run No.	Experimental Treatment	Study Sample	Unclassified	Classified	Classification Percentage
ERR3376856	Random Population	R1	3116655	15293703	83.07119
ERR3376857	Random Population	R2	2975347	7831741	72.46856
ERR3376858	Random Population	R3	3356756	15089648	81.80265
ERR3376859	Low Dengue Resistant Population	L1	4614580	42918457	90.29185
ERR3376860	Low Dengue Resistant Population	L2	2857744	12814919	81.76606
ERR3376861	Low Dengue Resistant Population	L3	3067684	48298960	94.02787
ERR3376862	High Dengue Resistant Population	H1	3478887	12934491	78.80456
ERR3376863	High Dengue Resistant Population	H2	3065712	12245507	79.97735
ERR3376864	High Dengue Resistant Population	H3	4076494	26794255	86.79496
ERR3376865	Ancestral Population	A	2702596	4887598	64.39358

Figure 2: Read Alignment and Classification Statistics. **A** Grouped bar graphs depicting total reads (light blue), aligned reads (dark blue) and unaligned reads (green) from 10 pooled *Ae. aegypti* samples after alignment to the AaegL5.3 reference genome using Bowtie2 (v2.4.4) (33) within MINUUR. **B** Stacked bar graphs showing KRAKEN2 classified read proportions from unmapped read sequences. Classified reads = light blue, unclassified reads = gold. **C** Table showing (left to right) the original sequencing run, experimental treatment, sample number, unclassified read count, classified read count and percentage of classified reads.

317 **4.3 MINUUR produces Genus and Species Classifications from**
318 **Unmapped Reads**

319 41 different genera were classified with KRAKEN2 and relative abundance
320 estimated with BRACKEN (Figure 3). Genera present across all samples
321 include *Wolbachia*, *Staphylococcus, Salmonella, Pseudomonas*,
322 *Phytobacter, Klebsiella, Escherichia, Enterobacter*, *Elizabethkingia*,
323 *Clostridium*, *Citrobacter, Chryseobacterium, Bacillus* and *Acinetobacter*
324 (Figure 3A). Several genera, summed across all samples in this study,
325 contained high read numbers including *Wolbachia* (73,746,796 reads),
326 *Elizabethkingia* (92,471,561), *Pseudomonas* (24,455,843), *Acinetobacter*
327 (1,524,766), *Stenotrophomonas* (1,823,242), *Delftia* (346,944),
328 *Chryseobacterium* (665,791) and *Klebsiella* (345,966 reads) (Figure 3C).

329 Each sample represents a pool of *Ae. aegypti* mosquitoes with different
330 dengue blocking phenotypes (high = H, low = L, random = R) as described
331 in the original publication (30). We were interested to see if certain bacteria
332 were uniquely present in a given experimental group. Within the high dengue
333 blocking populations (H1, H2, H3), *Bacteroides, Lactobacillus, Lactococcus*
334 and *Pedobacter* were uniquely present (Figure 3D). Conversely,
335 *Achromobacter, Acidovorax, Aeromonas, Bradyrhizobium, Comamonas,*

336 *Cronobacter, Delftia, Kosakonia, Paraburkholderia, Rhizobium* and *Vibrio*
337 were only present in low dengue blocking (L1, L2, L3) populations (Figure
338 3D).

339 Of reads that were classified with KRAKEN2, 81 different species were
340 classified above a 30,000 read threshold (Figure 4A). Species present in all
341 samples include *Wolbachia pipientis*, *W.* endosymbionts of *Aedes aegypti*,
342 *Drosophila simulans*, *D. melanogaster*, *D. ananassae, D. ceratosolen* and
343 *Elizabethkingia anophelis* (Figure 4A). *Pseudomonas* was a highly abundant
344 genus classification (Figure 4A). Here, 43 different *Pseudomonas* species
345 were identified, with *P. protegens*, *P. frederiksbergensis* and *P. koreensis*
346 the most abundant *Pseudomonas* species identified across samples (Figure
347 4A). We identify species with high read associations totaled across samples,
348 notably *E. anophelis* (58,372,887 reads), *P. protegens* (4,742,920 reads), *P.*
349 *flourescens* (5,384,368 reads), *P. koreensis* (1,968,181 reads), *W. pipientis*
350 (1,774,201 reads), *Acinetobacter seifertii* (1,062,679 reads) and
351 *Stentrophomonas maltophilia* (1,459,193 reads) (Figure 4C).

352 For this analysis, we set a read cutoff-threshold of 30,000 reads (filtering for
353 high abundant classifications). However, species of low abundance are
354 important to consider when interrogating a microbiome. MINUUR produces

355 trellis plots (facets) of each genus with the distribution of species relative
356 abundance (Supplementary Fig 1), estimated with BRACKEN. To exemplify,
357 we describe the classification of the commonly identified symbiont *Serratia*
358 in *Ae. aegypti* (50; 26; 56). We find *Serratia* contains 14 classified species,
359 with *S. marcescens* present at the highest abundance compared to other
360 species and across all samples (Supplementary Fig 1). Other notable
361 classifications include *S. fonticola* and *S. symbiotica*.

362 High sequence similarity among microbes of the same genus and species is
363 common. With KRAKEN2, reads with a classification that overlap with two or
364 more taxa will be assigned to the highest taxonomic level where a delineation
365 is detected. To this end, genus or species level taxonomic classifications,
366 interpreted as relative abundance, could lead to underestimation since reads
367 may be assigned at higher taxonomic levels. MINUUR implements
368 BRACKEN to infer relative abundance from KRAKEN2 classified reads at
369 lower taxonomic levels (genus or species). Of the original KRAKEN2
370 classified reads (197,240,903), 1,821,131 reads were redistributed to genus
371 level (Figure 3B). On average, 182,113 reads were added per sample (range
372 = 80,503 to 188,441 reads). Genera with the most added reads include
373 *Elizabethkingia* (368,564 reads), *Salmonella* (340,228 reads), *Escherichia*
374 (217,985 reads), *Enterobacter* (184,844) and *Pseudomonas* (159,493)

375 (Figure 3B). From the original KRAKEN2 classified reads at species level,
376 109,895,493 reads were redistributed with BRACKEN. On average,
377 10,989,659 reads were added per sample (range = 3,316,449 to 21,706,942
378 reads) (Figure 4B). Species with the most added reads include the
379 *Wolbachia* endosymbiont of *Ae. aegypti* (59,708,135 reads), *E. anophelis*
380 (32,166,299 reads), *P. fluorescens* (1,364,494 reads) and *Pseudomonas sp.*
381 *S150* (1,017,521 reads) (Figure 4B).

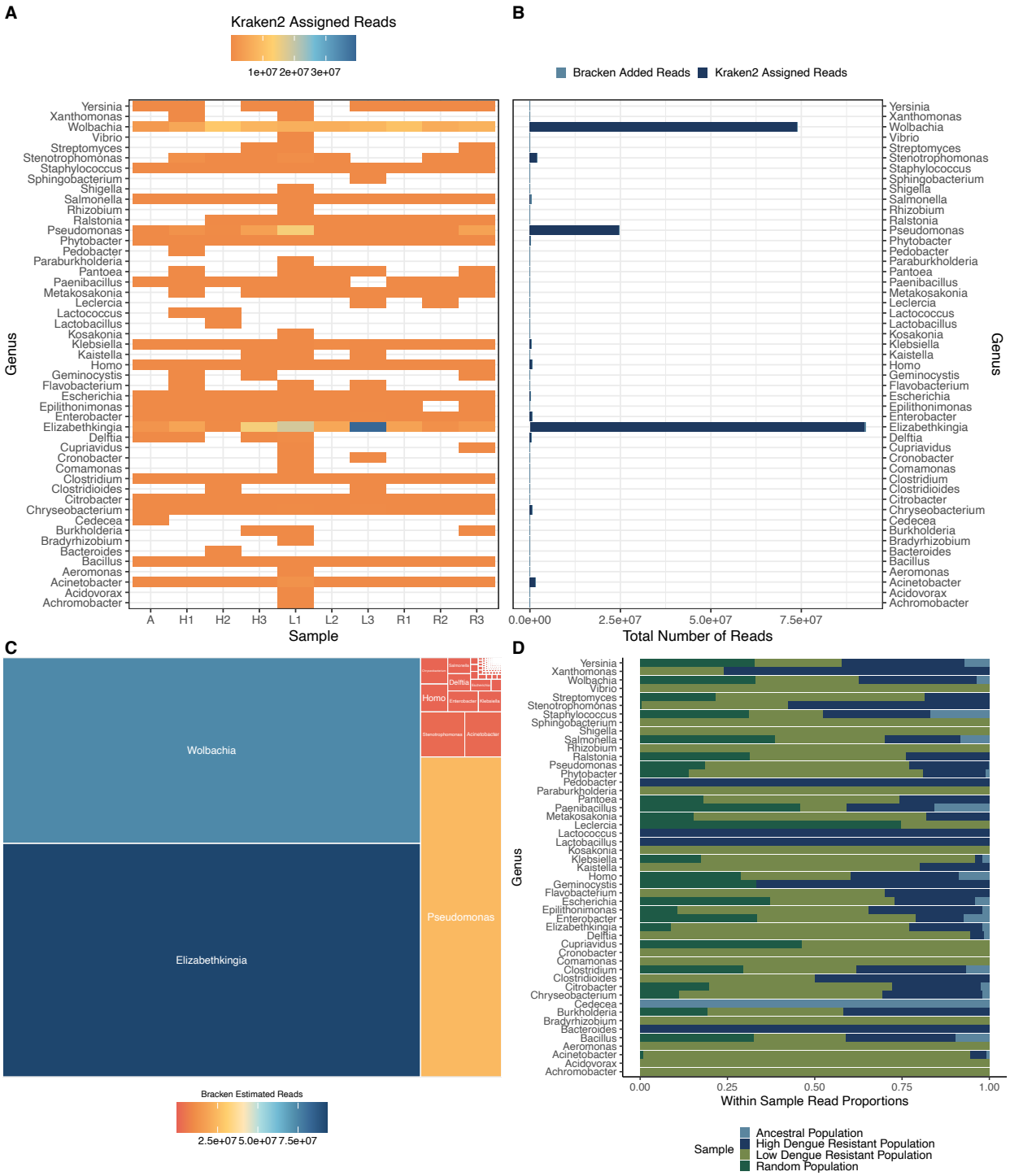


Figure 3: KRAKEN2 (v1.2) Genus Classifications and BRACKEN (v2.6.2) Abundance Estimation. **A**. Heatmap depicting genus level classifications and read abundance per sample. Genera shown are those with >1,000 assigned reads. Orange = low relative taxonomic abundance. Blue = high relative taxonomic abundance. **B**. Bar chart depicting total number of reads associated to each genus. Light blue = BRACKEN added reads, dark blue = KRAKEN2 classified reads. **C**.

Spatial chart depicting BRACKEN estimated read abundance within KRAKEN2 classified genera. Each block size is proportional to the total reads classified to each genus. **D.** Proportional taxonomic assignments within each experimental group. Each colour denotes the experimental group each taxon originated from.

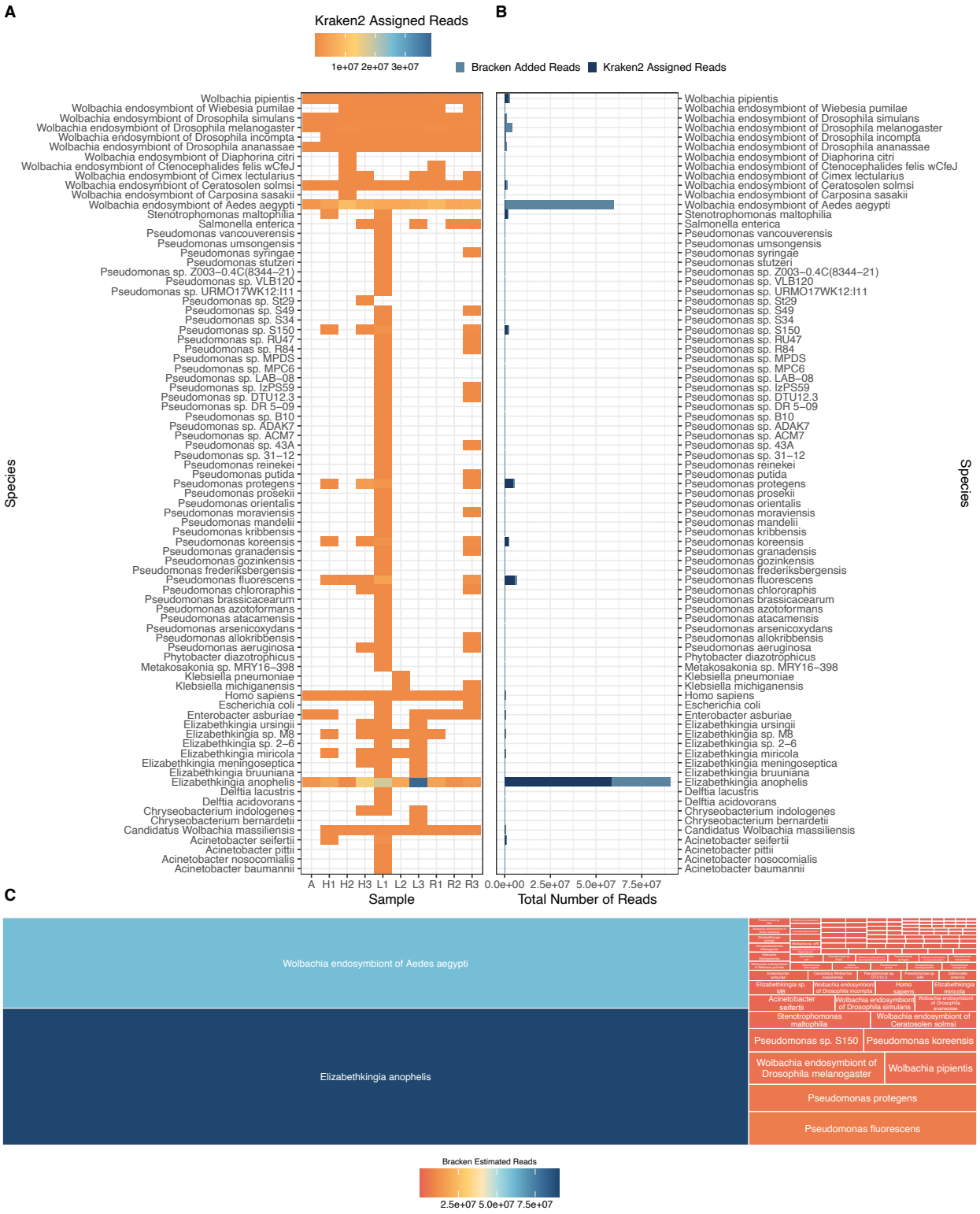


Figure 4: KRAKEN2 (v1.2) species classification and BRACKEN (v2.6.2) abundance estimation. **A.** Heatmap depicting species level classifications and total assigned read number . Species shown are those with >30,000 assigned reads. Orange = low taxonomic abundance. Blue = high

taxonomic abundance. **B.** Bar chart depicting total number of reads associated to each species. Light blue = BRACKEN added reads, dark blue = KRAKEN2 classified reads **C**. Spatial chart depicting BRACKEN estimated read number within KRAKEN2 classified species. Each block size is proportional to the total reads classified to each species.

4.4 MINUUR Predicts Microbial Function of Mosquito Associated Bacteria Using Unmapped Read Sequences

MINUUR uses HUMAaN3 to infer taxonomic functional profiles from read sequences directly. HUMAaN3 intakes classified taxa (which have been classified using MetaPhlAn3 against a library of clade-specific marker genes) and identifies gene profiles and metabolic pathways using UniRef90 annotations. In total, 107,196 genes with an ANI score of 90.5% (range = 72.5% to 100%) were identified across ten taxa classified with MetaPhlAn3 (Supplementary Figure 2A). On average, 10,720 genes (range = 4069 to 32, 359) and 214 pathways (range = 84 to 719) were identified per sample (Supplementary Figure 2B). The ten identified taxa with associated genes and metabolic pathways consisted of *E. anophelis* (40,243 genes, 850 metabolic profiles), *Stenotrophomonas maltophilia* (11,099 genes, 182 metabolic profiles), *Chryseobacterium sp ISE14* (10,948 genes, 840 metabolic profiles), *P. moraviensis* (10,045 genes, 246 metabolic profiles), *Klebsiella oxytoca* (8105 genes, 149 metabolic profiles), *K. michiganensis*

417 (7706 genes, 175 metabolic profiles), *Kluyvera intestini* (6133 genes, 206
418 metabolic pathways), *Acinetobacter seifertii* (6112 genes, 129 metabolic
419 profiles), *Delftia acidovorans* (5803 genes, 83 metabolic profiles) and
420 *Wolbachia* endosymbiont of *Brugia malayi* (1002 genes, 27 metabolic
421 pathways) (Supplementary Figure 2B).

422 Users can search specific genes and metabolic profiles of interest from
423 HUMAaN3's output within MINUUR. For gene profiles, we show an example
424 using the search term "siderophore" which are of interest given their previous
425 functional characterization in *Anopheles gambiae* associated bacteria (55).
426 Here, 17 siderophore related genes associated to nine MetaPhlAn3
427 classified taxa were identified (Figure 5B). The TonB dependent siderophore
428 receptor is present in seven bacteria. While both *K. michiganensis* and *K.*
429 *oxytoca* contain the catecholate siderophore receptor *fiu* and the OMR family
430 siderophore receptor (Figure 5B), suggesting an alternative mechanism for
431 siderophore acquisition. Furthermore, we chose to examine metabolic
432 pathways relating to 'heme'. We identify four pathways present in 5/8
433 associated taxa, with these identified bacteria containing the pathways for
434 heme b biosynthesis II. However, all identified pathways are incomplete with
435 respect to the genes used to reconstruct the pathway (Figure 5A).

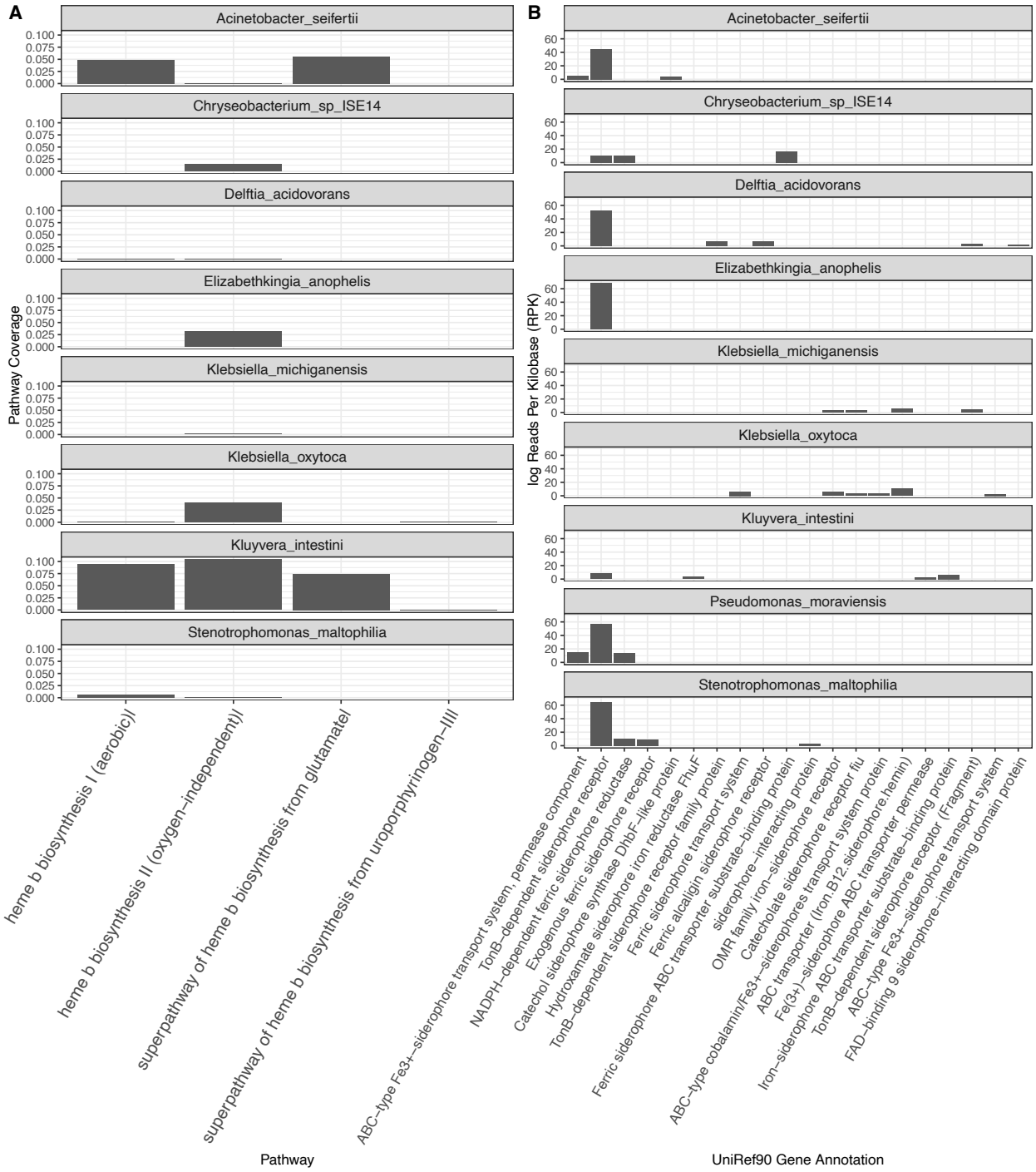


Figure 5: HUMAaN3 (v3.0.0) functional profile of MetaPhlAn3 (v3.0.13) classified taxa. **A.** Faceted plot showing 'heme' related pathway coverages stratified across species. X-Axis shows pathway coverage, denoted by the number of genes present that have reconstructed the pathway (0 = no coverage, 1 = complete pathway coverage), Y axis shows pathway. **B**. Faceted plot of log reads per kilobase (RPK) on the X axis, identified relating to 'Siderophores' on the Y axis. Gene number is stratified across MetaPhlAn3 classified bacterial species.

443 **4.5 Bacterial Metagenome-assembled Genomes from *Ae. aegypti***

444 *De novo* metagenome assembly aims to reconstruct contiguous sequences
445 or MAGs for further analysis. Here, we used two different approaches for
446 metagenome assembly; i) using KRAKEN2 classified reads and ii) using all
447 unmapped *Ae. aegypti* reads (composed of both taxonomically classified and
448 unclassified reads). We used CheckM to assess MAG completeness and
449 contamination based on the presence and copy number of single copy core
450 genes. Assembly with KRAKEN2 classified reads produced MAGs with less
451 contamination (Figure 6A, Figure 6C), while assembly with all unmapped
452 reads produced a higher number of MAGs, but with higher contamination
453 and lower completeness (<90% completeness) (Figure 6B, Figure 6D).

454 In total, we report the assembly of 43 *Ae. aegypti* associated MAGs using
455 KRAKEN2 classified reads and 57 *Ae. aegypti* associated bacterial MAGs
456 from all unmapped (classified and non-classified) reads (Figure 6A, Figure
457 6B). Community accepted standards of MAG quality are defined by the
458 genome standards consortium (GSC) (56). The GSC define high-quality draft
459 MAGs as >90% complete and <5% contamination - from our study, 19 MAGs
460 assembled using KRAKEN2 classified reads and 18 MAGs from all
461 unmapped reads met the GSC defined high-quality threshold (Figure 6A,

462 Figure 6B) (56). Medium quality draft MAGs are defined by the GSC as >50%
463 complete and <10% contamination; in which 30 MAGs from classified reads
464 and 31 MAGs from all unmapped reads were identified in our study. Finally,
465 low quality draft MAGs are defined by the GSC as <50% complete and <10%
466 contamination, in which 12 MAGs from classified reads and 26 MAGs from
467 all unmapped reads were identified.

468 Of MAGs with completeness >90%, the average genome size obtained from
469 KRAKEN2 classified reads was 2.94Mb (megabases), ranging between
470 1.06Mb and 5.70Mb (Figure 6F). The average genome size of MAGs
471 obtained from all unmapped reads was 3.03Mb, ranging between 1.13Mb
472 and 6.14Mb (Figure 6H). The mean N50 (the minimum contig length of an
473 assembled contig that covers 50% of the genome) of MAGs from KRAKEN2
474 classified reads was 97.5Kb (kilobases), ranging between 15.8Kb to 338Kb
475 (Figure 6G). The mean N50 of MAGs from all unmapped reads was 126Kb,
476 ranging between 5.78Kb to 591Kb (Figure 6I).

477 Outside of MINUUR, we used the taxonomic classifier GTDB-Tk to classify
478 MAGs against the Genome Taxonomy Database (GTDB). 41 MAGs were
479 classified with a mean FastANI score of 98.5%, ranging between 95.6% to
480 100% (Figure 7C). No MAGs were identified with FastANI scores <95%,

481 meaning no novel *Ae. aegypti* associated bacterial species were found,
482 however, MAGs with FastANI scores <99% are strain or subspecies variants.
483 Species classified from all assembled MAGs include *E. anophelis*,
484 *Wolbachia pipientis*, *Pseudomonas. E koreensis B*, *P. E protegens*,
485 *Stenotrophomonas sp002192255*, *Acinetobacter seifertii*, *Comamonas*
486 *acidovorans*, *Enterobacter cloacae M* and *Klebsiella. A michiganensis*
487 (Figure 7C). We also compared genome sizes of each MAG to its closest
488 reference genome (Figure 7A). 16 MAGs were smaller to their reference
489 genome by mean = 284kb, and two MAGs were larger by mean = 82.8kb
490 (Figure 7A). Congruent with the pairwise size differences between MAG and
491 reference genome, we found the overall distribution of MAG vs reference
492 genome size to be similar (Figure 7B). Two genomes skew this distribution
493 (references pertaining to MAGs ERR3376859.4 and ERR3376862.4), which
494 is consistent with the pairwise comparisons (Figure 7A).

495

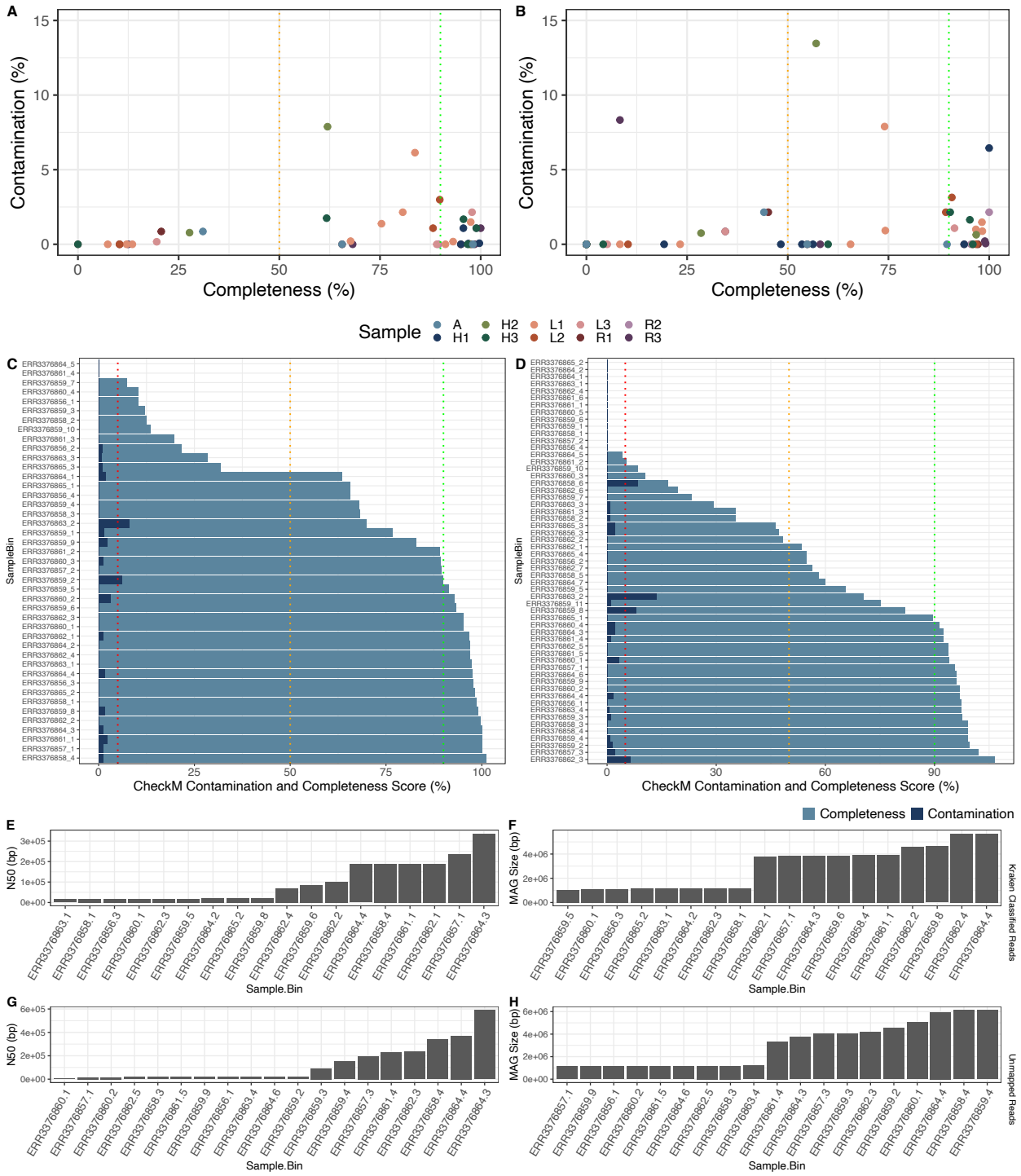


Figure 6: *Ae. aegypti* associated bacterial MAG statistics. **(A).** MAGs assembled from KRAKEN2 classified reads or **(B).** unmapped reads using MEGAHIT (v1.2.9) and binned with MetaBAT2 (v2.12.1). Colours denote each sample shown in the legend. 50% and 90% MAG completeness is specified by the orange and green dotted line. x-axis = CheckM completeness, y-axis = CheckM contamination. **(C, D)** Bar graph depicting completeness and contamination scores of each MAG,

light blue = completeness, dark blue = contamination, left = MAGs assembled using KRAKEN2 associated reads, right = MAGs assembled using unmapped reads. Red dotted line indicates 5% contamination threshold **(E, F)** N50 and MAG size (base pairs) of MAGs assembled using KRAKEN2 classified reads, with completeness over 90%. **(G, H)** N50 and MAG size (base pairs) of MAGs assembled using unmapped reads, with completeness over 90%.

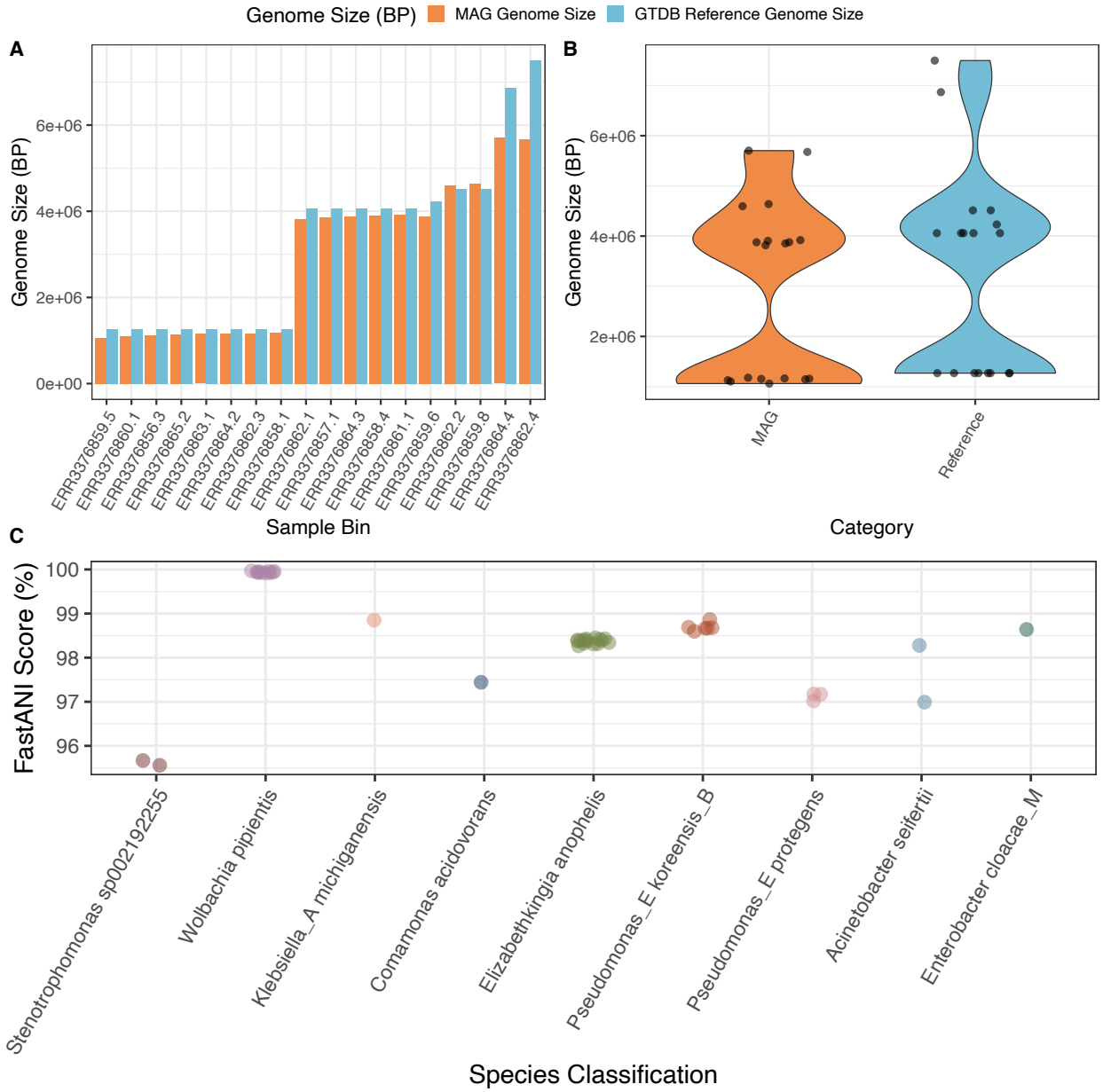


Figure 7: *Ae. aegypti* associated bacterial MAG GTDB-Tk (v1.5.0) classifications **A.** Genome size (base pairs) of *Ae aegyti* associated MAGs (orange) with CheckM completeness >90% and <5% contamination compared to genomes size of its closest GTDB-Tk reference genome (blue). Each bar denotes the sample originated and the bin number (sample.bin) **B.** Violin plot showing

the distribution of *Ae. aegypti* MAG genome size (orange) compared to their GTDB-Tk classified reference genome. Each point denotes a MAG with completeness >90% and <5% contamination. **C**. Taxonomic classifications from the Genome Taxonomy Database (GTDB) of MAGs obtained from unmapped *Ae. aegypti* reads. X-Axis = GTDB-Tk species classification, Y-Axis = FastANI (%) score to the closest related reference genome in the GTDB.

5 Discussion

The analysis of non-host reads from existing WGS data has previously been applied in other organisms (5,8,9,11). We developed MINUUR to facilitate a reproducible and robust metagenomic analysis of non-host sequences from whole genome sequencing data, which can be applied to other hosts at the user's discretion. For this study, we used MINUUR to characterize the microbiome of an existing WGS dataset of *Ae. aegypti* mosquitoes (30).

Read classifications reveal the high abundance of *Elizabethkingia*, *Pseudomonas*, *Acinetobacter*, *Stenotrophomonas* and *Wolbachia*. The presence of *Wolbachia* in these *Ae. aegypti* samples at high titers is expected as this line was transinfected with this bacterium (30). Our species level classifications support previous amplicon sequencing studies that show the adult *Ae. aegypti* microbiome is dominated by phyla from Proteobacteria (*Pseudomonas*, *Acinetobacter*, *Stentophomonas*, *Enterobacter*, *Klebsiella*) and Bacteroides (*Elizabethkingia*, *Chryseobacterium*) (18,19,21,25,57). *E.*

533 *anophelis* is an abundant bacterial species identified in this study. This
534 bacterium has previously been implicated in response to iron fluxes in *An.*
535 *gambiae* (58), and blood meals in *Ae. albopictus* (59) and *Ae. aegypti* (60).
536 The high abundance of this symbiont across these samples could be that
537 DNA was extracted from mosquitoes shortly after blood-feeding, which
538 would support the previous studies above. Furthermore, studies of mosquito
539 bacterial interactions show a strain of *E. anophelis*, *E. anophelis Ag1*,
540 interacts with *Pseudomonas Ag1* by up-regulating expression of the *hemS*
541 gene in *E. anophelis*, promoting heme breakdown into biliverdin catabolites
542 (24). Interestingly, *Pseudomonas sp Ag1* is closely related to *P. flourescens*
543 (61), identified as an abundant species in our study. Further work should
544 elucidate if a similar bipartite interaction is present in *Ae. aegypti*.

545 The samples used here originate from a study looking at genetic variation of
546 artificially selected *Ae. aegypti* for *Wolbachia*-mediated dengue blocking. As
547 such, we were interested in patterns between microbiome members and low
548 / high dengue blocking mosquito samples (Figure 3D). We found several
549 bacteria uniquely identified in high dengue resistant populations. Most
550 notably, our results support a previous study showing *Pedobacter*
551 significantly associated with a dengue virus refractory *Ae. aegypti* strain,
552 MAYO-R (62). *Pedobacter* identified in our study is uniquely present in high

553 dengue blocking mosquito populations, suggesting an association with low
554 dengue virus titers in *Ae. aegypti.* The mechanism of this association is
555 unknown but could be due to a specific interaction with dengue virus or
556 immune priming of *Ae. aegypti* to elicit an anti-viral response (15).

557 HUMANn3 was used to annotate MetaPhlAn3 classified bacteria using the
558 pan-genome ChocoPhlAn database and UniRef90 annotations. We show
559 the example of searching for "siderophore" related genes, which resulted in
560 the identification of 16 genes across 9 bacterial species. The TonB
561 dependent siderophore receptor was identified across seven / nine bacteria,
562 suggesting involvement of siderophore mediated iron uptake in *Ae. aegypti*
563 associated bacteria. However, while the TonB-dependent receptor has high
564 affinity for siderophores, it is also specific to other substrates including
565 vitamin B12s, carbohydrates and nickel chelates (63). The "siderophore"
566 gene profiles reported in this study also suggest different siderophore
567 acquisition mechanisms across *Ae. aegypti* bacteria (64). For example,
568 *Klebsiella* identified in this study does not contain the TonB dependent
569 siderophore receptor, but instead contains a catecholate siderophore
570 specific receptor *fiu*. A similar observation is noted for *Acinetobacter* which
571 uniquely contains the Catechol synthase *DhbF*.

572 We assembled 19 high quality MAGs with CheckM completeness scores
573 >90% and <5% contamination, which were subsequently classified, outside
574 of MINUUR, against the Genome Taxonomy Database. High-quality MAG
575 reconstructions are applied in large scale metagenomic studies from
576 chickens (65), humans (66) to cows (67–69), with these studies yielding
577 between 400 to 92,000 MAGs per study. We apply a similar approach with
578 unmapped *Ae. aegypti* sequencing reads to reconstruct high-quality
579 community accepted standard MAGs. Our study expands the genomic
580 representation of known mosquito-associated bacterial symbionts,
581 specifically to *Ae. aegypti*, adding these newly assembled MAGs to the 33-
582 mosquito associated bacterial genomes currently stored on the NCBI.
583 Overall, these provide a valuable resource for researchers in the field and
584 can be used in further work such as facilitating biosynthetic gene cluster
585 discovery (69) or to identify genetic targets for symbiont pathogen blocking
586 approaches (13).

587 In summary, we developed a pipeline to facilitate analysis of unmapped
588 reads from host-associated WGS data, with application in the pathogen
589 vector *Ae. aegypti*. Future considerations and prospects of mosquito
590 microbiome research were recently established by the Mosquito Microbiome
591 Consortium (70). A key point highlighted in this statement is the need for

592 (meta)genomics approaches with solid reproducibility for data analysis within
593 the field. Our pipeline provides a robust set of analyses to assess non-host
594 reads from existing genome sequence data. Within *Ae. aegypti,* we show the
595 reads that do not map to its reference genome can be taxonomically
596 classified to its microbiome members at genus and species level; associated
597 microbial genes and pathways predicted and high-quality mosquito-
598 associated MAGs reconstructed. We hope this pipeline and approach will
599 facilitate further analysis of existing WGS data within *Ae. aegypti* and other
600 organisms.

601 Data Availability Statement

602 MAGs are available in GenBank under the project accession number:
603 PRJNA866910.

604 **Author contributions**

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

Acknowledgments

AF was supported by a DTP scholarship (Medical Research Council MR/N013514/1). EH acknowledges funding from Wellcome (217303/Z/19/Z) and the BBSRC (BB/V011278/1).

Bibliography

1. Theis KR, Dheilly NM, Klassen JL, Brucker RM, Baines JF, Bosch TCG, et al. Getting the Hologenome Concept Right: an Eco-Evolutionary Framework for Hosts and Their Microbiomes. mSystems. 2016 Apr 26;1(2). Available from: https://journals.asm.org/doi/10.1128/mSystems.00028-16

2. 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. 6(8). Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000409

3. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018 Jul;16(7):410–22.

4. Lapidus AL, Korobeynikov AI. Metagenomic Data Assembly – The Way of Decoding Unknown Microorganisms. Front Microbiol. 2021 Mar 23;12:613791.

5. 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.

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

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

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

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

648 10. Fierst JL, Murdock DA, Thanthiriwatte C, Willis JH, Phillips PC.
649 Metagenome-Assembled Draft Genome Sequence of a Novel Microbial
650 *Stenotrophomonas maltoph*ilia Strain Isolated from *Caenorhabditis*
651 *remanei* Tissue. Genome Announc. 2017 Feb 16;5(7). Available from:
652 https://journals.asm.org/doi/10.1128/genomeA.01646-16

653 11. Gerth M, Hurst GDD. Short reads from honey bee (*Apis* sp.)
654 sequencing projects reflect microbial associate diversity. PeerJ. 2017
655 Jul 12;5:e3529.

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

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

662 14. Kozlova EV, Hegde S, Roundy CM, Golovko G, Saldaña MA, Hart CE,
663 et al. Microbial interactions in the mosquito gut determine Serratia
664 colonization and blood-feeding propensity. ISME J. 2021 Jan;15(1):93–
665 108.

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

668 16. Coon KL, Valzania L, McKinney DA, Vogel KJ, Brown MR, Strand MR.
669 Bacteria-mediated hypoxia functions as a signal for mosquito
670 development. Proc Natl Acad Sci. 2017 Jul 3;114(27):E5362–9.

671 17. Valzania L, Coon KL, Vogel KJ, Brown MR, Strand MR. Hypoxia-
672 induced transcription factor signaling is essential for larval growth of the
673 mosquito *Aedes aegypti*. Proc Natl Acad Sci. 2018 Jan 16;115(3):457–
674 65.

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

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

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

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

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

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

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

25. 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.

26. 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.

27. 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.

28. Chen C, Compton A, Nikolouli K, Wang A, Aryan A, Sharma A, et al. Marker-assisted mapping enables effective forward genetic analysis in the arboviral vector *Aedes aegypti*, a species with vast recombination deserts. Genetics; 2021 Apr.

29. Crava C, Varghese FS, Pischedda E, Halbach R, Palatini U, Marconcini M, et al. Immunity to infections in arboviral vectors by integrated viral sequences: an evolutionary perspective . Evolutionary Biology; 2020 Apr

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

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

32. 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 Oct;30(10):1533–46.

33. Köster M. Sustainable data analysis with Snakemake. F1000Research. 2022 Apr 21;10(33). Available from:

https://f1000researchdata.s3.amazonaws.com/manuscripts/56004/06ebda8f-ff09-4b68-a0e1-c12c245aca3b_29032_-_johannes_koster_v2.pdf?doi=10.12688/f1000research.29032.2&numberOfBrowsableCollections=55&numberOfBrowsableInstitutionalCollections=4&numberOfBrowsableGateways=40

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

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

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

37. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods. 12, 902-903. 2015 Sep 29. Available from: https://www.nature.com/articles/nmeth.3589.

38. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife. 2021 May 4;10:e65088.

39. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, the UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015 Mar 15;31(6):926–32.

40. Andrew, S A S. FASTQC: A Quality Control Tool for High Throughput Sequence Data. Available from: www.bioinformatics.babraham.ac.uk/projects/fastqc/

41. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet Journal. 2011 Aug 02; 17, 10-12.

42. Valiente-Mullor C, Beamud B, Ansari I, Francés-Cuesta C, García-González N, Mejía L, 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 Jan 27;17(1):e1008678.

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

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

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

774 46. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality
775 assessment tool for genome assemblies. Bioinformatics. 2013 Apr
776 15;29(8):1072–5.

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

779 48. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT2: an
780 adaptive binning algorithm for robust and efficient genome
781 reconstruction from metagenome assemblies. PeerJ. 2019 Jul
782 26;7:e7359.

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

787 50. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit
788 to classify genomes with the Genome Taxonomy Database.
789 Bioinformatics. 2019 Nov 15;26(6):1925-1927.

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

793 52. Finn RD, Clements J, Eddy SR. HMMER web server: interactive
794 sequence similarity searching. Nucleic Acids Res. 2011 Jul 1;39:W29–
795 37.

796 53. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A,
797 Chaumeil PA, et al. A standardized bacterial taxonomy based on

genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018 Nov;36(10):996–1004.

54. 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 Dec;11(1):538.

55. Ganley JG, Pandey A, Sylvester K, Lu KY, Toro-Moreno M, Rütschlin S, et al. A Systematic Analysis of Mosquito-Microbiome Biosynthetic Gene Clusters Reveals Antimalarial Siderophores that Reduce Mosquito Reproduction Capacity. Cell Chem Biol. 2020 Jul;27(7):817-826.e5.

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

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

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

59. 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.

60. 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. 2018 Oct 24; Available from: https://academic.oup.com/femsec/advance-article/doi/10.1093/femsec/fiy213/5144212

61. Alvarez C, Kukutla P, Jiang J, Yu W, Xu J. Draft Genome Sequence of *Pseudomonas sp*. Strain Ag1, Isolated from the Midgut of the Malaria

Mosquito *Anopheles gambiae*. J Bacteriol. 2012 Oct;194(19):5449–5449.

62. Charan SS, Pawar KD, Severson DW, Patole MS, Shouche YS. Comparative analysis of midgut bacterial communities of *Aedes aegypti* mosquito strains varying in vector competence to dengue virus. Parasitol Res. 2013 Jul;112(7):2627–37.

63. Fujita M, Mori K, Hara H, Hishiyama S, Kamimura N, Masai E. A TonB-dependent receptor constitutes the outer membrane transport system for a lignin-derived aromatic compound. Commun Biol. 2019 Dec;2(1):432.

64. Kramer J, Özkaya Ö, Kümmerli R. Bacterial siderophores in community and host interactions. Nat Rev Microbiol. 2020 Mar;18(3):152–63.

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

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

67. Wilkinson T, Korir D, Ogugo M, Stewart RD, Watson M, Paxton E, 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 Dec;21(1):229.

68. Watson M. New insights from 33,813 publicly available metagenome-assembled-genomes (MAGs) assembled from the rumen microbiome. Microbiology; 2021 Apr.

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

70. Dada N, Jupatanakul N, Minard G, Short SM, Akorli J, Villegas LM. Considerations for mosquito microbiome research from the Mosquito Microbiome Consortium. Open Science Framework; 2020 Jun. Available from: https://osf.io/2s8he