Identification and Removal of Potential Contaminants in 16S rRNA Gene Sequence Datasets from Low Microbial Biomass Samples: An Example from the Mosquito Gut

Sebastián Díaz

Universidad de Antioquia

Juan Sebastián Escobar

Vidarium Nutrition, Health and Wellness Research Center

Frank William Avila (✉ grupotandem.mosquito@udea.edu.co)

Universidad de Antioquia <https://orcid.org/0000-0002-7278-0646>

Methodology

Keywords: Mosquito, Gut microbiota, Contamination, Metabarcoding analysis

Posted Date: July 22nd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-45329/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background: The bacterial gut microbiota of the female mosquito influences numerous physiological processes, including vector competence. As a low-microbial-biomass ecosystem, mosquito gut tissue is prone to contamination from the laboratory environment and from reagents commonly used to dissect and/or isolate DNA from gut tissue. In this report, we analyze five 16S rRNA datasets, including new data obtained by us, to gain insight into the impact of potential contaminating sequences on the composition, diversity, and structure of the mosquito gut microbial community.

Results: We present a clustering-free approach that, based on the relative abundance of amplicon sequence variants (ASVs) in gut and negative control samples, allowed for the identification of candidate contaminating sequences. Some of these sequences belong to bacterial taxa previously identified as common contaminants in metagenomic studies; they have also been identified as part of the mosquito core gut microbiota, with putative physiological relevance for the host. By using different relative abundance cutoffs, we show that contaminating sequences have a significant impact on gut microbiota diversity and structure.

Conclusions: The approach presented here allows the identification and removal of purported contaminating sequences in datasets obtained from low-microbial biomass samples. While it was exemplified with the analysis of gut microbiota from mosquitoes, it can easily extend to other datasets dealing with similar technical artifacts.

4. Background

Mosquitoes have a resident bacterial gut microbiota that is fundamental for several physiological processes such as larval growth, blood digestion, and immune function [1–3], making gut bacteria a prospective avenue for reducing vector competence [4]. For this reason, a multitude of studies have described microbial gut communities of mosquitoes from relevant medical genera (e.g. *Aedes*, *Anopheles*, and *Culex*), in different settings (lab-reared or field-collected), and in different physiological contexts (e.g. blood-fed, virus- or parasite-infected) (reviewed in [1, 2, 5]). These studies have demonstrated the taxonomic diversity of bacteria present in mosquito gut tissues, even at the intraspecific level. Most of the identified bacteria belong to the Proteobacteria phylum (e.g., *Acinetobacter*, *Aeromonas*, *Asaia*, *Comamonas*, *Enterobacter*, *Klebsiella*, *Pantoea*, *Pseudomonas*, *Serratia*, among others) [5].

Parallel to advances of the past decade in microbiota identification techniques that have allowed the description of mosquito gut microbiota in greater detail, interest has grown in understanding and describing the impact of contaminants introduced by molecular reagents, such as microbiota from DNA extraction kits (referred to as the “kitome” [6–8]), PCR master mix [7, 9], laboratory facilities [8, 10] and technical issues such as well-to-well contamination [11] or index switching in sequencing platforms [12, 13]. Interestingly, many contaminating sequences belong to bacteria commonly associated with mosquito gut tissues, including *Acinetobacter*, *Chryseobacterium*, *Enterobacter*, or *Pseudomonas* [6, 8, 14]. This raises a question about the identification of these bacteria in the mosquito gut: is it an actual presence or the signal from undesired contamination? To counter the effect of contamination, especially when studying low-biomass samples, where contaminants can become dominant in the sampling [14], researchers have proposed precautionary measures, such as randomizing sample types and treatment groups, decontaminating working areas, and sequencing of negative controls (e.g. sampling blanks and DNA extraction reagents) and positive controls (e.g. mock communities) [6, 8, 14–16].

Interest in standardizing the methodologies has grown for mosquito microbiota research [17, 18], recognizing that mosquito tissues are low-biomass samples, expected to be prone to sequencing artifacts [17]. However, no consensus approach has been developed for the identification and removal of possible contaminating sequences from mosquito tissues to date. Published studies in mosquito gut microbiota that have reported the sequencing of negative control samples (Additional Table 1), only a few have addressed the quantification and reduction of putative contaminating sequences. One reported the complete removal of sequences detected in controls [19] and two others reported the removal of shared OTUs with relative abundances at least 10 times greater in control samples compared to tissue samples [20, 21]. Unfortunately, the taxonomic identity of the removed OTUs in these studies was not reported.

In this report, we examine mosquito gut microbiota datasets from some previously published studies and add a newly generated dataset obtained by us that included several negative controls, in order to identify and remove purported contaminating sequences. We next quantified the impact of contaminating sequences on gut microbiota diversity and structure. Finally, we develop a simple strategy for the removal of contaminating sequences from mosquito gut tissue samples.

5. Results

We analyzed five datasets (Table 1) using the same pipeline to remove low-quality sequences (i.e. sequences that did not align with the analyzed region, chimeric and/or non-bacterial sequences), subsequently defining amplicon sequence variants (ASVs; sequences clustered at 100% identity, see Methods). For negative control samples, we determined the total number of ASVs and those with relative abundance thresholds of $\geq 1\%$, 5% , and 10% (Table 2). We also quantified the proportion of ASVs shared between negative control and tissue samples. In four datasets, these shared ASVs represented more than 10% of sequences obtained in the tissue samples (66.84% in *Aedes*, 56.76% in *Albopictus*, 56.68% in *Anopheles2* and 11.79% in *Anopheles1*), while in the *Aegypti* dataset, shared ASVs represented less than 1% . As expected, increasing the relative abundance threshold

reduced the number of shared ASVs, as well as their presence in tissue samples. For the 10% threshold, the contamination abundance in tissue samples was 38.94% in *Aedes*, 18.68% in *Anopheles2*, 4.38% in *Anopheles1*, and 2.76% in *Albopictus*.

We identified 46 ASVs in control samples with a relative abundance threshold of $\geq 1\%$, representing 20 bacterial genera (Table 3). Aside from *Chryseobacterium* and *Cloacibacterium* (Bacteroidetes), all belonged to the phylum Proteobacteria, with *Acinetobacter* and *Serratia* being the most common (three of the five datasets). Most ASVs found in control samples had a low relative abundance ($< 5\%$). Seven ASVs had a relative abundance $\geq 5\%$, and all were detected in the *Aedes* and *Anopheles2* datasets: two *Enterobacter* (25.89% and 13.05%) and one *Serratia* (10.16%) in the *Aedes* dataset; *Aeromonas* (9.93%), *Pantoea* (8.75%), *Acinetobacter* (5.31%) and *Serratia* (5.04%) in the *Anopheles2* dataset.

Next, we examined how removing of sequences found in control samples affected the composition and structure of tissue samples using the following treatments: complete removal of ASVs found in negative controls, and removal of ASVs with abundance thresholds $\geq 1\%$, 5% , or 10% . Given the differences in targeted 16S rRNA fragments, DNA extraction methods, and the variety of negative controls utilized in each study (Table 1), alpha and beta diversities were not directly comparable across the datasets. Therefore, we compared changes in microbial diversity before and after ASV removal treatments in each dataset separately. For alpha diversity, we observed significant reductions in the number of OTUs for all datasets after total removal of the sequences found in control samples—the most drastic treatment, but also with the other removal treatments, except for the *Aegypti* dataset (Fig. 1A). The Shannon diversity index also showed statistically significant changes with the total removal treatment (except for *Albopictus* and *Anopheles1* datasets) and with the threshold removal treatments (except for *Aegypti*) (Fig. 1B). For the evenness (Pielou's index), we observed changes in three datasets—*Aedes*, *Anopheles1*, and *Anopheles2*—the first two with significant differences for all removal treatments (Fig. 1C). For beta diversity, we used the Jaccard and Bray-Curtis indices. Like the alpha diversity analysis, statistically significant differences against the non-removal treatment were observed for all treatments (Fig. 2).

6. Discussion

With our survey of available mosquito gut datasets and a new one reported here, we highlight the impact of potential contaminants on the composition, structure, and diversity of low-microbial biomass samples. We used a clustering-free approach to make precise identification of potentially contaminating sequences. This strategy works well for this purpose because contaminants are likely very specific, as has been previously demonstrated [22, 23]. Our removal strategy recognizes their uniqueness and puts forward a way to trim them that is not dependent on reference taxonomic databases. Therefore, it can be implemented in different datasets dealing with DNA sequences and extended to metagenomic pipelines.

The use of a clustering-free approach allows separating potential contamination and actual sequences with the same taxonomy but having a different biological origin. In our study, we found that abundant ASVs detected in negative controls were classified within *Acinetobacter*, *Chryseobacterium*, *Enterobacter*, or *Pseudomonas*. These genera have previously been described as common contaminants [6–8, 14]. However, other ASVs classified in these same genera were found in tissue samples and were not detected in negative controls. This illustrates the advantage of our clustering-free approach, as these bacterial groups have been reported as part of the core microbiota in *Aedes* and *Anopheles* mosquitoes [1, 2, 5], having putative functional roles in their host. *Enterobacter* has hemolytic activity associated with blood digestion and egg production [24]; *Enterobacter* and *Pseudomonas* reduce vector competence for *Plasmodium* infection [25, 26] and LaCrosse virus [27]; *Acinetobacter* and *Chryseobacterium* may contribute to larval development [28].

Not only have we identified purported contaminating sequences, but we also evaluated trimming strategies to reduce their effects on data analysis. Almost all removal strategies affected microbial inference, a major result that highlights the impact of ignoring contamination and the crucial role of negative controls to remove potential sources of noise. The first strategy used, removing all the sequences found in negative controls, is considered a very conservative method, where it is preferable to pay the cost of eliminating the true positives than to keep contaminants in the final dataset. This loss of biological data due to cross-contamination between experimental and the negative controls samples (e.g. well-to-well contamination [11] or index switching [12]) is a reason some authors discourage using this method, suggesting that removal of sequences present in negative controls should only be performed when it can be ensured that they correspond to actual contaminants [16] and propose the use of alternative methods [14] (see below).

Distinguishing laboratory or reagent contamination from cross-contamination with experimental samples is very challenging in low-microbial biomass samples, where the presence of ASVs with low abundance are ubiquitous. Our analysis showed that most ASVs found in negative controls had low abundance ($\leq 1\%$), and some were identified as symbionts, suggesting potential cross-contamination from tissue samples to negative controls. For instance, the endosymbiont *Wolbachia* was found in control samples in the *Aedes*, *Albopictus* and *Anopheles2* datasets in low abundance ($< 1\%$). *Thorsellia*, another bacteria reported as a natural mosquito symbiont [29–31], was present in the control sample of the *Anopheles2* dataset with an abundance of 1.54% (Table 3). As de Gouffau et al. [15] pointed out, ecological data should be considered when evaluating if these unexpected results make sense; in this instance, they do not.

Another approach tested here was the use of abundance thresholds for the removal of contaminating sequences. We observed that similar to the complete removal treatment, there were significant changes in alpha and beta diversities after removal of sequences with relative abundances $\geq 1\%$, 5% , or 10% in control samples. Previous studies have employed this approach based on two assumptions: (i) that contaminating sequences have frequencies that correlate inversely with the DNA concentration of the samples; (ii) that contaminating sequences have a higher prevalence in control

samples than in experimental samples [32]. However, these assumptions are not valid in the analysis of samples with low-microbial biomass, where contaminating sequences can dominate the entire library. Some authors have employed this approach in the study of mosquito-associated microbiota. For instance, Minard et al. [20] removed all shared OTUs with relative abundances at least 10 times greater in control samples than in tissue samples. Instead, we showed that microbial inference is severely affected by the removal of contaminants with varying abundance thresholds. However, we do not consider the total removal of sequences found in negative controls or any predefined abundance threshold as universal to determine contamination. Each study needs to define its proper criteria according to the data obtained from sequencing and the quality of the controls used.

Complementary to the strategy proposed here, it is necessary to establish additional measures to identify and reduce contamination in the analysis of low-microbial biomass samples [6, 14, 16]. These procedures include: (1) maximizing the starting sample biomass by choice of sample type, filtration or enrichment, (2) randomization of samples and treatments to avoid batch/day effects, (3) recording batch numbers of reagents, (4) sequencing of many negative controls that cover all sample processing steps (i.e. dissection, DNA extraction, library preparation), (5) sequencing of positive controls (e.g. mock community, high-biomass samples with known composition) that can help to detect cross-contamination and (6) reporting negative control sequences in genomic repositories, along with tissue sample sequences.

7. Conclusion

Our analysis of mosquito gut microbiota datasets revealed the common presence of contaminant sequences that significantly affected the composition, diversity, and structure of the inferred microbial community. To minimize this impact, we proposed a clustering-free approach to precisely identify potential contaminants and evaluated different abundance thresholds to gauge the impact of high and low abundance ASVs on the inferred microbial community. This strategy should be complemented with laboratory protocols to identify and reduce sample contamination including as many controls as possible.

8. Methods

Data acquisition

We analyzed microbial DNA sequences from mosquito guts and control samples using two data sources: a newly developed dataset which we report here, and four datasets retrieved from previous studies. The new dataset of *Aedes aegypti* and *Ae. albopictus* gut samples (hereafter referred to as *Aedes*) was obtained from adults from lab colonies raised at the Max Planck Tandem Group in Mosquito Reproductive Biology at the Universidad de Antioquia (Medellin, Colombia). Gut tissue was dissected in 1X PBS under sterile conditions to obtain 23 samples (21 females and 2 males) made of pools of 20 tissues per sample stored in STE buffer. Samples were lysed by adding 6 μ L of lysozyme (20 μ g/ μ L) for two hours at 37 °C, followed by an overnight incubation adding 24 μ L of proteinase K (20 μ g/ μ L) at 56 °C. DNA was extracted using a phenol-chloroform protocol and eluted in 50 μ L of AE buffer (Qiagen, Valencia, CA, USA). Experimental samples and a PBS sterile blank control were randomly seeded in five different extraction rounds, each with a DNA extraction control. Positive samples for a diagnostic 16S rRNA gene PCR using primers P338F and 1492R and negative controls were sent to Macrogen (Seoul, South Korea) for sequencing on the Miseq Illumina platform using the primers Bakt_341F (5'-CCT ACG GGN GGC WGC AG-3') and Bakt_805R (5'-GAC TAC HVG GGT ATC TAA TCC-3') [33], which amplified the V3-V4 hypervariable regions of 16S rRNA gene with an average sequencing depth of 100K reads per sample.

To identify published reports of mosquito gut microbiota, we used PubMed (<https://pubmed.ncbi.nlm.nih.gov/>) to search for articles with title words [Mosquito] OR [*Aedes*] OR [*Anopheles*] OR [*Culex*] AND [Gut] AND [Microbiota], complementing this effort with a more extensive manual search. We pursued studies that used high-throughput gene sequencing with data made available at the NCBI's BioProject or Sequence Read Archive (SRA). We found 17 articles from 2011–2019 (Additional Table 1) that matched our criteria. From this group, we selected five datasets that included sequences from both gut samples and negative controls (blank and/or DNA extraction sample(s)). One dataset was discarded given their limited number of sequences in control samples [34]. The four remaining datasets used in our analysis were from *Ae. aegypti* (referred to as *Aegypti*) [21], *Ae. albopictus* (*Albopictus*) [20], *An. gambiae*/*An. coluzzii* (*Anopheles1*) [31] and *An. darlingi*/*An. nuneztovari* (*Anopheles2*) [35] (Table 1).

Microbiota analyses

For the five datasets analyzed here, raw reads were processed following the SOP for MiSeq sequences of Mothur v. 1.43.0 [36]. Low-quality sequences were filtered out according to established parameters: (a) presence of ambiguous nucleotides, (b) sequences with more than 8 homopolymers, (c) sequence length lower than the 2.5% percentile, (d) sequence length higher than the 97.5% percentile. Remaining sequences were pre-clustered to reduce sequencing errors (allowing one difference every 100 bp), chimeras removed with VSEARCH [37], as well as non-bacterial sequences, based on a preliminary classification using the SILVA v132 database [38]. Singletons were removed from the final dataset. For the *Aedes* and *Albopictus* datasets, data were normalized to 25,000 sequences per sample because of their high sequencing depth.

To evaluate the effect of removing contaminating sequences on microbial composition and diversity, we made the following analyses. First, we used a clustering-free approach to identify independent bacterial subpopulations shared by gut samples and negative controls. Then, each unique

sequence (i.e. 100% nucleotide identity) was defined as an amplicon sequence variant (ASV). Afterward, we used different relative abundance criteria for removing ASVs shared between negative controls and tissue samples. We obtained five different subsets: (a) original dataset with no removal of ASVs found in negative controls, (b) removal of all ASVs present in control samples from tissue samples, (c) removal of ASVs with an overall relative abundance in control samples $\geq 1\%$ (d) removal of ASVs with an overall relative abundance in control samples $\geq 5\%$, and (e) removal of ASVs with an overall relative abundance in control samples $\geq 10\%$. Finally, for each of the above subsets, we clustered sequences at 97% sequence identity to obtain standard OTUs using the OptiClust algorithm implemented in Mothur [39].

To evaluate the changes in microbiota composition, diversity, and structure in the subsets described above, we calculated alpha and beta diversity indices. Specifically, we calculated the number of OTUs as a measure of microbial richness, Shannon diversity index, and Pielou's evenness index. Also, we calculated the Jaccard and Bray-Curtis indices as measures of beta diversity. After evaluating the normality of the indices for each subset using a Shapiro-Wilk test, we used paired t-tests and Wilcoxon signed-rank tests to determine whether there was a statistically significant difference between alpha and beta diversity indices in the subsets where putative contaminants were removed compared to the original, unaltered datasets.

9. List Of Abbreviations

STE: Sodium Chloride-Tris-EDTA

OTU: Operational taxonomic unit

ASV: Amplicon sequence variant

10. Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The datasets generated and analyzed during the current study are available in the NCBI Sequence Read Archive (SRA) repository, under the Bioproject with accession code PRJNA644640 (<http://www.ncbi.nlm.nih.gov/bioproject/644640>).

Competing interests

J.S.E. is employed by a food company. S.D. and F.W.A. had no competing interests.

Funding

This work was supported by the COLCIENCIAS, Universidad de Antioquia and Max Planck Society cooperation grant 566-1 (2014) to FWA.

Authors' contributions

SD, JSE, and FWA designed the research; SD performed the research and analyzed the data; SD, JSE, and FWA wrote the paper.

Acknowledgments

We are thankful to members of the Avila laboratory for their support in the rearing of mosquito colonies, the Centro de Computación Científica Apolo at Universidad EAFIT (Medellin, Colombia) for hosting supercomputing resources, and Ruta N for laboratory support.

References

1. Guégan M, Zouache K, Démichel C, Minard G, Potier P, Mavingui P, et al. The mosquito holobiont: fresh insight into mosquito-microbiota interactions. *Microbiome*.2018;6:49.

2. Scolari F, Casiraghi M, Bonizzoni M. *Aedes* spp. and their microbiota: a review. *Front Microbiol.* 2019;10:2036.
3. Strand MR. Composition and functional roles of the gut microbiota in mosquitoes. *Curr Opin Insect Sci.* 2018; 28:59–65.
4. Shaw WR, Catteruccia F. Vector biology meets disease control: using basic research to fight vector-borne diseases. *Nat Microbiol.* 2018; 4: 20–34.
5. Minard G, Mavingui P, Moro CV. Diversity and function of bacterial microbiota in the mosquito holobiont. *Parasit Vectors.* 2013;6:146.
6. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 2014;12:87.
7. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog.* 2016;8:24.
8. Weyrich LS, Farrer AG, Eisenhofer R, Arriola LA, Young J, Selway CA, et al. Laboratory contamination over time during low-biomass sample analysis. *Mol Ecol Resour.* 2019;19:982–96.
9. Tilburg JJHC, Nabuurs-Franssen MH, van Hannen EJ, Horrevorts AM, Melchers WJG, Klaassen CHW. Contamination of commercial PCR master mix with DNA from *Coxiella burnetii*. *J Clin Microbiol.* 2010;48:4634–5.
10. Laurence M, Hatzis C, Brash DE. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS One.* 2014;9.
11. Minich JJ, Sanders JG, Amir A, Humphrey G, Gilbert JA, Knight R. Quantifying and understanding well-to-well contamination in microbiome research. *MSystems.* 2019;4:e00186-19.
12. Costello M, Fleharty M, Abreu J, Farjoun Y, Ferreira S, Holmes L, et al. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics.* 2018;19:332.
13. MacConaill LE, Burns RT, Nag A, Coleman HA, Slevin MK, Giorda K, et al. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics.* 2018;19:30.
14. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in low microbial biomass microbiome studies: issues and recommendations. *Trends Microbiol.* 2019;27:105–17.
15. de Goffau MC, Lager S, Salter SJ, Wagner J, Kronbichler A, Charnock-Jones DS, et al. Recognizing the reagent microbiome. *Nat Microbiol.* 2018;3:851–3.
16. Hornung BVH, Zwittink RD, Kuijper EJ. Issues and current standards of controls in microbiome research. *FEMS Microbiol Ecol.* 2019;95:fiz045.
17. Dada N, Jupatanakul N, Minard G, Short SM, Akorli J, Villegas LM. Considerations for mosquito microbiome research from the Mosquito Microbiome Consortium. *OSF Preprints*; 2020.
18. Rodríguez-Ruano SM, Juhaňáková E, Vávra J, Nováková E. Methodological insight into mosquito microbiome studies. *Front Cell Infect Microbiol.* 2020;10:86.
19. Dickson LB, Ghazlane A, Volant S, Bouchier C, Ma L, Vega-Rua A, et al. Diverse laboratory colonies of *Aedes aegypti* harbor the same adult midgut bacterial microbiome. *Parasit Vectors.* 2018;11:1–8.
20. Minard G, TRAN F-H, Goubert C, Bellet C, Lambert G, Khanh HKL, et al. French invasive Asian tiger mosquito populations harbor reduced bacterial microbiota and genetic diversity compared to Vietnamese autochthonous relatives. *Front Microbiol.* 2015;6:970.
21. Dickson LB, Jiolle D, Minard G, Moltini-Conclois I, Volant S, Ghazlane A, et al. Carryover effects of larval exposure to different environmental bacteria drive adult trait variation in a mosquito vector. *Sci Adv.* 2017;3:e1700585.
22. Caruso V, Song X, Asquith M, Karstens L. Performance of microbiome sequence inference methods in environments with varying biomass. *MSystems.* 2019;4:e00163-18.
23. Karstens L, Asquith M, Davin S, Fair D, Gregory WT, Wolfe AJ, et al. Controlling for contaminants in low-biomass 16S rRNA gene sequencing experiments. *MSystems.* 2019;4:e00290-19.
24. de O Gaio A, Gusmão DS, Santos A V, Berbert-Molina MA, Pimenta PFP, Lemos FJA. Contribution of midgut bacteria to blood digestion and egg production in *Aedes aegypti* (Diptera: Culicidae)(L.). *Parasit Vectors.* 2011;4:105.
25. Cirimotich CM, Dong Y, Clayton AM, Sandiford SL, Souza-Neto JA, Mulenga M. Natural microbe-mediated refractoriness to *Plasmodium* infection in *Anopheles gambiae*. *Science.* 2011;332:855-8.
26. Tchioffo MT, Boissiere A, Churcher TS, Abate L, Gimonneau G, Nsango SE, et al. Modulation of malaria infection in *Anopheles gambiae* mosquitoes exposed to natural midgut bacteria. *PLoS One.* 2013;8: e81663.
27. Joyce JD, Nogueira JR, Bales AA, Pittman KE, Anderson JR. Interactions between La Crosse virus and bacteria isolated from the digestive tract of *Aedes albopictus* (Diptera: Culicidae). *J Med Entomol.* 2011;48:389–94.
28. Coon KL, Vogel KJ, Brown MR, Strand MR. Mosquitoes rely on their gut microbiota for development. *Mol Ecol.* 2014;23:2727–39.
29. Kämpfer P, Lindh JM, Terenius O, Haghdoust S, Falsen E, Busse H-J, et al. *Thorsellia anophelis* gen. nov., sp. nov., a new member of the Gammaproteobacteria. *Int J Syst Evol Microbiol.* 2006;56:335–8.

30. Kämpfer P, Glaeser SP, Nilsson LKJ, Eberhard T, Håkansson S, Guy L, et al. Proposal of *Thorsellia kenyensis* sp. nov. and *Thorsellia kandunguensis* sp. nov., isolated from larvae of *Anopheles arabiensis*, as members of the family Thorselliaceae fam. nov. *Int J Syst Evol Microbiol.* 2015;65:444–51.
31. Segata N, Baldini F, Pompon J, Garrett WS, Truong DT, Dabiré RK, et al. The reproductive tracts of two malaria vectors are populated by a core microbiome and by gender-and swarm-enriched microbial biomarkers. *Sci Rep.* 2016;6:1–10.
32. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome.* 2018;6:226.
33. Herlemann DP, Labrenz M, Jürgens K, Bertilsson S, Waniek JJ, Andersson AF. Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J.* 2011;5:1571–9.
34. Guégan M, Minard G, Tran F-H, Tran Van V, Dubost A, Valiente Moro C. Short-term impacts of anthropogenic stressors on *Aedes albopictus* mosquito vector microbiota. *FEMS Microbiol Ecol.* 2018;94:fiy188.
35. Bascuñán P, Niño-García JP, Galeano-Castañeda Y, Serre D, Correa MM. Factors shaping the gut bacterial community assembly in two main Colombian malaria vectors. *Microbiome.* 2018;6:1–12.
36. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol.* 2009;75:7537–41.
37. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ.* 2016;4:e2584.
38. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 2012;gks1219.
39. Westcott SL, Schloss PD. OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. *MSphere.* 2017;2.

Tables

Table 1.

Information on the examined datasets. Name of each dataset, mosquito host species, number of samples and controls, and other relevant metadata.

Dataset Name	Bioproject accession number	Host species	Sample source	DNA extraction method	16S rRNA region	No. Gut samples	No. Controls	Normalization	Reference
Aedes	To be published once the paper is accepted	<i>Ae. aegypti</i> / <i>Ae. albopictus</i>	Lab	Phenol-Chloroform	V3-V4	23 (Pools)	6	Yes	This study
Aegypti	PRJEB16334	<i>Ae. aegypti</i>	Field	DNeasy Kit (Qiagen)	V5-V6	26 (Individuals)	5	No	Dickson et al. (2017)
Albopictus	PRJEB6896	<i>Ae. albopictus</i>	Field	DNeasy Kit (Qiagen)	V5-V6	33 (Individuals)	1	Yes	Minard et al. (2015)
Anopheles1	PRJNA172065	<i>An. gambiae</i> / <i>An. coluzzii</i>	Field	DNeasy Kit (Qiagen)	V4	50 (Individuals)	1	No	Segata et al. (2016)
Anopheles2	PRJNA415615	<i>An. nuneztovari</i> / <i>An. darlingi</i>	Field	Salt precipitation	V2	62 (Individuals)	2	No	Bascuñan et al. (2018)

Table 2.

ASV distribution in the negative control samples. The total number of ASVs and ASVs with an abundance $\geq 1\%$, $\geq 5\%$, $\geq 10\%$ in control samples. The number of these ASVs shared with tissue samples and their relative abundance in the tissue samples.

Dataset	ASVs - Complete			ASVs $\geq 1\%$			ASVs $\geq 5\%$			ASVs $\geq 10\%$		
	# Total	# Shared	% in tissues samples	# Total	# Shared	% in tissues samples	# Total	# Shared	% in tissues samples	# Total	# Shared	% in tissues samples
Aedes	1992	138	66.84%	4	4	49.10%	3	3	49.10%	2	2	38.94%
Aegypti	307	5	0.08%	16	0	0%	2	0	0%	NA	NA	NA
Albopictus	1030	307	56.76%	7	7	7.99%	1	1	2.76%	1	1	2.76%
Anopheles1	26	16	11.79%	7	6	7.65%	4	3	6.48%	3	2	4.38%
Anopheles2	299	267	56.58%	12	12	42.70%	4	4	22.78%	2	3	18.68%

Table 3.
List of ASVs with an overall relative abundance $\geq 1\%$ in negative control samples.
Overall relative abundance for negative control samples and tissue samples for each dataset. ASVs sorted by abundance in control samples.

Dataset	ASV	Abundance Controls	Abundance Samples
Aedes	<i>Enterobacter</i>	40.06	25.89
	<i>Enterobacter</i>	19.64	13.05
	<i>Serratia</i>	5.69	10.16
	<i>Cutibacterium</i>	1.11	0
Aegypti	<i>Serratia</i>	6.34	0
	<i>Halomonas</i>	5.22	0
	<i>Halomonas</i>	4.54	0
	<i>Serratia</i>	3.79	0
	<i>Serratia</i>	2.18	0
	<i>Serratia</i>	2.16	0
	<i>Halomonas</i>	2.14	0
	<i>Halomonas</i>	2.06	0
	<i>Halomonas</i>	1.78	0
	<i>Halomonas</i>	1.67	0
	<i>Vibrio</i>	1.24	0
	<i>Marinimicrobium</i>	1.16	0
	<i>Serratia</i>	1.15	0
	<i>Serratia</i>	1.08	0
	<i>Halomonas</i>	1.06	0
	<i>Halomonas</i>	1.06	0
Albopictus	<i>Pseudomonas</i>	60.64	2.76
	<i>Chryseobacterium</i>	4.82	2.66
	<i>Janthinobacterium</i>	3.24	1.72
	<i>Pseudomonas</i>	2.89	0.25
	<i>Pseudomonas</i>	2.07	0.07
	<i>Acinetobacter</i>	1.53	0.03
	<i>Janthinobacterium</i>	1.08	0.5
Anopheles1	<i>Sphingomonas</i>	35.85	0.01
	<i>Acinetobacter</i>	21.94	4.37
	<i>Caulobacter</i>	19.83	0
	<i>Escherichia-Shigella</i>	8.26	2.1
	<i>Cloacibacterium</i>	4.46	0.86
	<i>Acinetobacter</i>	1.67	0.14
	<i>Diaphorobacter</i>	1.08	0.17
Aniopheles2	<i>Aeromonas</i>	17.74	9.93
	<i>Pantoea</i>	12.38	8.75
	<i>Chryseobacterium</i>	9.14	1.74
	<i>Acinetobacter</i>	6.36	2.36

<i>Acinetobacter</i>	3.55	5.31
<i>Serratia</i>	3.53	5.04
<i>Stenotrophomonas</i>	3.26	1.61
<i>Enhydrobacter</i>	1.82	3.16
<i>Thorsellia</i>	1.54	3.06
<i>Enhydrobacter</i>	1.38	0.23
<i>Pseudomonas</i>	1.36	1.4
<i>Acinetobacter</i>	1.17	0.11

Figures

Figure 1

Alpha diversity. (A) Number of OTUs, (B) Shannon diversity index, and (C) Pielou's evenness index. Asterisks indicate statically significant differences between the subset with purported contaminants removed vs. the original dataset. (*) p-value 0.05-0.005, (**) p-value 0.0049-0.0005, (***) p-value <0.00049.

Figure 2

Beta diversity. (A) Jaccard index and (B) Bray-Curtis index. Asterisks indicate statically significant differences between the subset with purported contaminants removed vs. the original dataset. (*) p-value 0.05-0.005, (**) p-value 0.0049-0.0005, (***) p-value <0.00049.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTable1.xlsx](#)