Title

Blanking on blanks: few insect microbiota studies control for contaminants.

Permalink

https://escholarship.org/uc/item/3pm096tt

Authors

Williamson, Elisabeth M
Hammer, Tobin J
Hogendoorn, Katja
et al.

Publication Date

2025-02-25

DOI

10.1128/mbio.02658-24

Peer reviewed

Blanking on Blanks: Few Insect Microbiota Studies Control for Contaminants

Elisabeth Williamson[1], Tobin J. Hammer[2], Katja Hogendoorn[1], Raphael Eisenhofer[3]

[1] *The University of Adelaide, School of Agriculture, Food and Wine, Adelaide, 5005, Australia*

[2] *Department of Ecology and Evolutionary Biology, University of California, Irvine, USA*

[3] *Centre for Evolutionary Hologenomics, The University of Copenhagen, Globe institute, Denmark.*

5.0 Abstract

Research on insect-microbe relationships is booming, with DNA sequencing being the most commonly used method to describe insect microbiota. However, sequencing is vulnerable to contamination, especially when the sample has low microbial biomass. Such low-biomass samples are common across insect taxa, developmental stages, and tissue types. Identifying putative contaminants is essential to distinguish between true microbiota and introduced contaminant DNA. It is therefore important that studies control for contamination, but how often this is done is unknown. To investigate the status quo of contamination control, we undertook a systematic literature review to quantify the prevalence of negative control usage and contamination control across the literature on insect microbiota (specifically bacterial communities) over a 10-year period. Two-thirds of the 243 insect microbiota studies evaluated had not included blanks (negative controls), and only 13.6% of the studies sequenced these blanks and controlled for contamination in their samples. Our findings highlight a major lack of contamination control in the field of insect microbiota research. This result suggests that a large number of microbes reported in the literature may be contaminants as opposed to insect-associated microbiota, and that more rigorous contamination control is needed to improve research reliability, validity and reproducibility. Based on our findings, we recommend a modified version of the guidelines outlined in the RIDES checklist: Report methodology, Include negative controls, Determine the level of contamination, Explore contamination downstream, and State the amount of off-target amplification.

5.0.1 Importance

Our systematic review reveals a major lack of methodological rigor within the field of research on insect-associated microbiota. The small percentage of studies that control for contamination suggests that a considerable proportion of bacteria reported in the literature could be contaminants. The implication of this finding is that true microbiota may be masked or misrepresented, especially in insects with low microbial biomass.

5.1 Introduction

Research on insect-associated microbial communities is booming globally. Communities of microbial symbionts (microbiota) can have profound ecological and evolutionary impacts on insects, that in some cases form obligate, species-specific symbiotic relationships with their host [e.g. 1, 2, 3]. However, most insect microbiota are yet to be described, or are in early stages of investigation, and the extent of microbial reliance and the nature of these associations across the insect world is unknown. The most commonly used method for characterizing insect microbiota is DNA amplicon sequencing [4]. While DNA sequencing is a highly powerful, sensitive and accessible tool, interpretation of the data requires great care.

One of the biggest limitations for amplicon sequencing assessments is DNA contamination. Biological samples can become contaminated from exposure during collection, contact with research personnel and the laboratory environment, as well as from reagents, DNA extractions kits (called 'kitomes') and cross contamination between samples (called 'splashomes') [5, 6, 7, 8]. When uncontrolled, DNA contamination can result in erroneous community assessments by distorting taxonomic diversity, obscuring differences between samples, and misrepresenting true absences of microbiota [9]. For example, human placental tissue was once thought to harbour specific microbiota, but later studies found that almost all of the sequence data could be attributed to contamination [6, 10, 11].

Contamination is particularly problematic when biological samples have a low number of microbes (called 'low biomass samples'). Low biomass samples occur when microbes naturally exist in low abundance, or because the biomass of the sample itself is small. These can include (but are not limited to) certain types of insects [12], mammalian tissues [13, 14], glacial ice [15], rocks [16], air [17], and man-made environments [18]. Low biomass samples are more prone to DNA contamination because there are fewer "true" microbes to crowd out the contaminants [19, 5]. An empirical example of the relationship between low biomass and high contamination is given in Figure S1.

Universal primers designed to amplify sequences of bacterial 16S rRNA genes can also amplify chloroplast and mitochondrial genomes (referred to as 'off-target DNA') due to their common evolutionary history [20, 21]. Consequently, 16S rRNA sequencing assessments can experience amplification bias toward off-target DNA if samples are low biomass and contain animal or plant tissue. Chloroplasts and mitochondria are predominant in insect microbiota due to insect diets and/or tissues, however off-target DNA can also be an issue in other contexts, such as in plant, human microbiome, or food-web studies [22, 23, 24]. A high relative abundance of off-target DNA can distort and obscure community assessments by reducing the representation of low abundance taxa, and this can further exacerbate the impact of contamination. Furthermore, there are several cases where high proportions of off-target DNA in insect samples are predicted to be the result of low microbial biomass, such as in *Lepidoptera* [12, 25], *Hymenoptera* [26], *Thysanoptera* [26], and *Phasmatodea* [27].

There are established protocols to control for DNA contamination. The standard approach, and the method focused on in this systematic review, is to process no-template or negative controls (i.e., blank samples) alongside biological samples during DNA extraction, PCR, and sequencing. Amplicon sequence variants (ASVs) can then be removed from the biological sample data if they are detected and more prevalent in negative controls [8, 19, 5, Figure 1]. It is recommended to include a negative control per batch of DNA extractions, as the degree of contamination can vary between lots processed [28]. The accuracy of such techniques can be further improved by employing statistical packages, such as Decontam [9], that classify ASVs as contaminants using established and reproducible methods. Another benefit to using statistical packages is that the method can be tailored according to the dataset; for instance, Decontam provides the option to adjust the prevalence method used and modify the sensitivity of contamination identification [9].
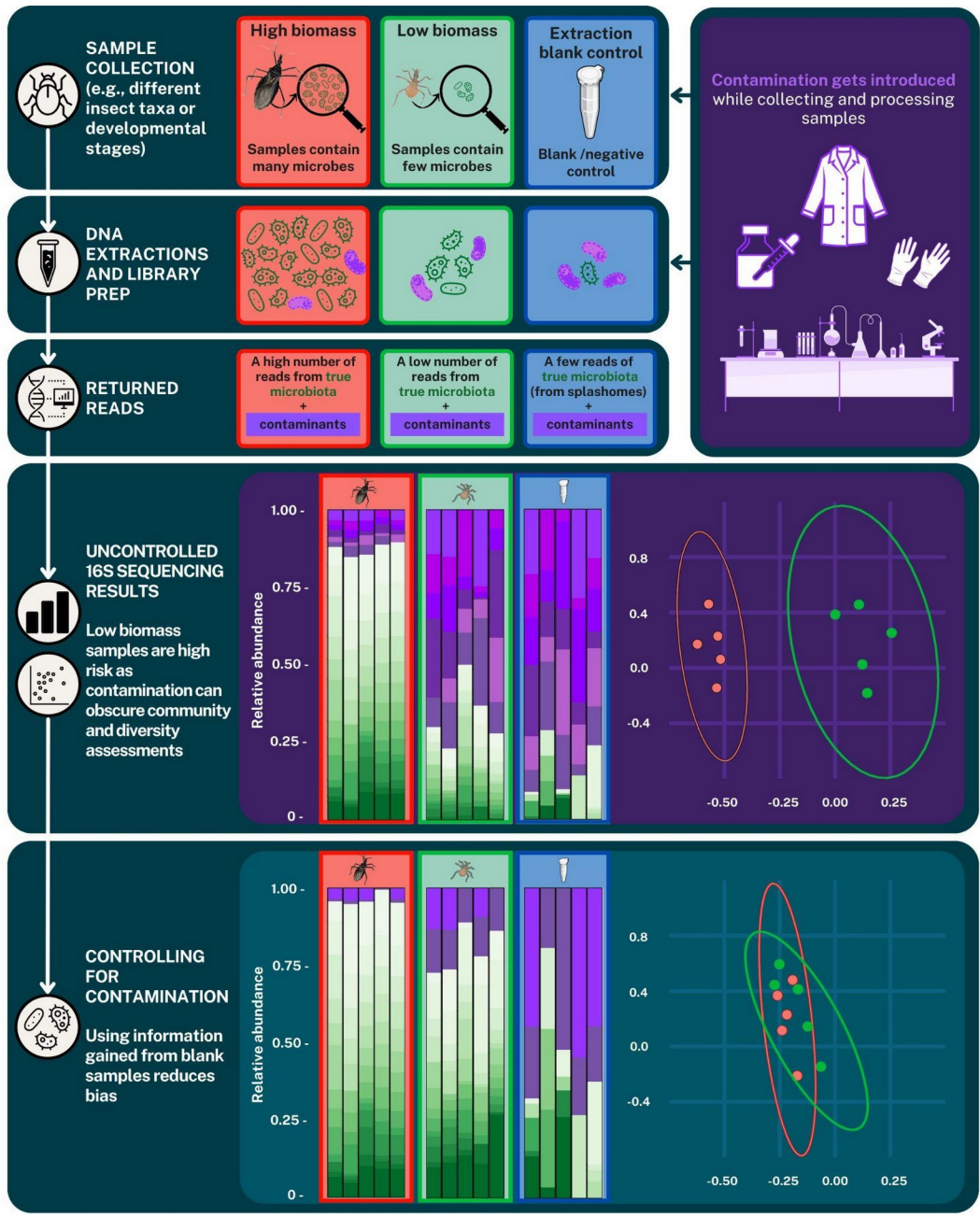
Figure 1: Conceptual schematic to demonstrate the different impact contamination (represented in purple) can have on high vs low biomass samples. Using relative abundance plots and principle coordinate analysis (PCoA) diversity assessments, uncontrolled contamination is shown to mask true taxa and similarities between samples. While not being perfect, due to differences in contaminant prevalence across samples, and/or splashomes, contamination control using sequence information from blanks reduces this bias.

Negative controls can also be used to measure the limit of detection (LoD). The LoD is a benchmark to determine the lowest amount of sample-derived DNA that can be reliably used to identify and quantify microbial taxa in a given dataset [29, 5]. The LoD can be measured using quantitative PCRs (qPCRs), where the absolute abundances in all samples and negative controls are measured. The average abundance in negative controls is used as the LoD, and anything above this is inferred to be amplified DNA from biological samples. If a biological sample falls below the LoD, it should be discarded as it does not meet the minimum threshold of 'true' DNA [5, 19, 28].

Controlling for contamination is essential, but how often it is actually done in the insect microbiota literature is unknown. Here, we systematically assess whether insect microbiota research over a 10-year period has appropriately controlled for DNA contamination. Specifically, we address the following questions: 1) What percentage of studies have used negative controls to control contamination? 2) Has there been an increase in the proportion of studies that control for contamination over the years? 3) How many studies have determined their experimental limit of detection? 4) What proportion of studies acknowledge off-target amplification from chloroplast or mitochondrial DNA? In light of our findings, we provide recommendations to improve the robustness of future insect microbiota research.

5.2 Materials and methods

5.2.1 Search strategy, KAPPA analysis, and Filtering

This review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses Statement (PRISMA) [30]. We systematically searched Scopus, Web of Science and Google Scholar for relevant articles and data papers. Due to the lack of standardization between electronic databases, we tailored our search strategy for each database to include certain terms and synonyms relating to our key concepts. Specifically, we searched for papers that described or investigated the microbiota of insects (searches provided in supplementary materials, Figure S2). Ordered by 'relevance', the first 200 search results from the three database searches were then imported into the Covidence platform [31] and duplicate papers were removed. Before filtering all imported studies using the Inclusion/Exclusion criteria (supplementary materials Table S3), 50 papers (sorted by 'Title') were assessed individually by two authors for comparison using a KAPPA analysis. Because the KAPPA analysis confirmed the screening was reproducible, we proceeded with the remaining articles to produce the final pool of papers for metadata collection. A summary of the information

gathered from our final pool is given in Figure 2, and the detailed data extraction method is given in the supplementary materials.

5.2.2 Metadata analysis

We first assessed the percentage of the most represented insect orders, to verify a lack of bias to certain taxa. We then examined trends in specimen sampling and processing by obtaining the percentages of the developmental stage(s) investigated (only adults, only juveniles, or multiple stages), the inclusion of a surface sterilization step, and the region of the 16S rRNA gene used. It was determined how many studies mentioned mitochondria and/or chloroplasts in their methods or results sections, and if mentioned, whether they disclosed the amount of off-target amplification. We calculated the proportion of studies that reported using a negative control, and that used qPCRs to measure the limit of detection. From the subset of studies that used a negative control, we further calculated the proportion of studies that sequenced their controls and used this information to control for contamination. After assessing the distribution of citation counts using a Shapiro-Wilk test for normality, we opted to use a Mann-Whitney-Wilcoxon (MWM) test to assess whether the number of citations differed between studies that did or did not control for contamination, as an indicator of their perceived usefulness.

5.3 Results and Discussion

5.3.1 Exploring the metadata

Our final pool consisted of 243 papers published between 2011-2022, which were used for metadata collection and subsequent analyses (supplementary materials Figure S4).

5.3.1.1 Representation of insect orders and sampling approaches

The top 5 insect orders represented in our final pool were *Diptera* (23%), *Hemiptera* (18.1%), *Lepidoptera* (16.9%), *Coleoptera* (14.8%), and *Hymenoptera* (14%). The remaining 13.2% studied *Blattodea*, *Odonata*, *Orthoptera*, *Psocodea*, *Siphonaptera*, or multiple different insect orders. The distribution of insect orders studied indicates there was no bias to one field in our metadata. For the insects studied, 77.5% of studies sampled one developmental stage, with 51.4% of papers focusing on the adult stage, 25.9% sampling only juveniles (e.g. eggs, larvae, nymphs, or pupae), and the remaining 22.6% investigated more than one stage of development.
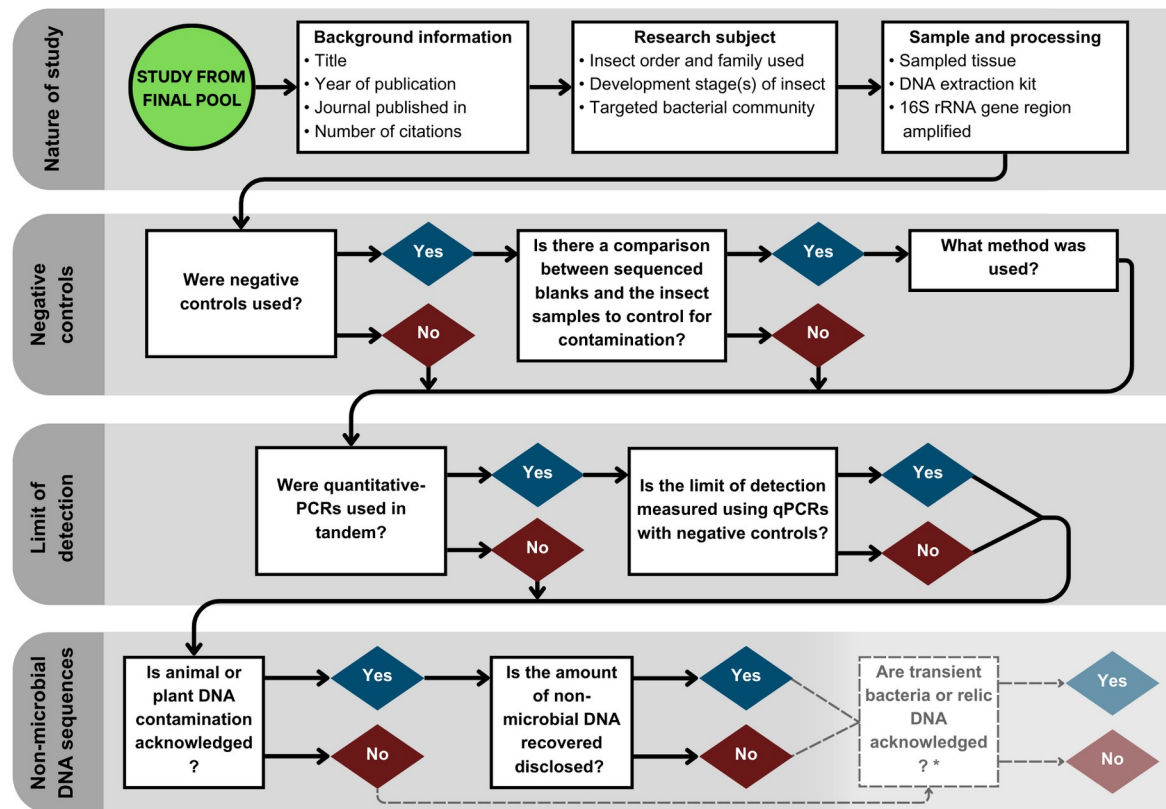
*Figure 2: The information mined from the final pool of papers that researched insect microbiota using 16S sequencing to create the metadata set. *Additional questions addressed in the supplementary material.*

A total of 56.8% of studies aimed to describe the microbial community of the gut specifically, 39.1% of studies described the community for the whole insect, and the remaining targeted specific organs or tissues.

5.3.1.2 Trends in specimen processing methods

A total of 66.7% of studies included a surface sterilization step. A range of substances were used to wash insects, with ethanol, bleach, sterile water, PBS, and detergent most frequently used. The variation in chemicals and whether or not insects were washed exemplifies the current lack in uniformity throughout insect-microbiota studies. In addition, many papers that included a surface sterilization step stated that the purpose of the procedure was to remove external microbes. Contaminating microbes on the surface of specimens can be introduced during the handling and storage, or, occur naturally as a consequence of environmental contact (i.e. transient bacteria). It is possible that in some cases, surface sterilization was presumed sufficient to control contamination. However, this fails to acknowledge downstream

contaminants introduced during molecular processing, and previous research has shown no detectable effect of surface sterilization on insect microbiota [32].

5.3.2 Prevalence of negative control usage and controlling contamination in insect microbiota studies

Only one third (80/243) of studies reported using at least one negative control. Negative controls were referred to as blanks, extraction blank controls (EBCs), no-template controls (NTCs), PCR blanks, sequencing controls, or surface washes (from insects or tool rinses). Fewer than half (33/80) of the studies that reported including a negative control subsequently sequenced these and then compared the taxa found in their blanks to the corresponding insect samples. Therefore, only 13.6% of the studies included in the systematic review assessed their sequencing results for contaminating taxa. While the number of papers investigating insect microbiota increased exponentially between 2011 and 2022, the proportion of studies that used a negative control or controlled contamination per year has not increased (Figure 3). Studies that did not control for contamination were not less frequently cited (p = 0.84), suggesting a lack of scrutiny when interpreting the conclusions derived from uncontrolled studies.

The low proportion of studies that report using negative controls (Figure 3) signifies that a large part of insect microbiota research has thus far not adhered to the latest guidelines that are recommended to control for contamination [5, 8, 33]. Although some studies may have used negative controls without mentioning them, this cannot be assumed.

The implication of this finding is that for 86.4% of studies included in our metadata, it can not be conclusively said whether the microbiota reported are truly insect-associated or the byproduct of contamination. One of the challenges in recognizing contamination in insect studies is that some taxa of commonly reported insect microbiota overlap with that of common contaminants. For example, *Pseudomonas* [34], *Acinetobacter* [35, 36], *Staphylococcus* [37], *Bacillus* [38] and *Burkholderia* [39] are all described insect-associated microbes yet are also common contaminant taxa [19, 28, 5, 14].
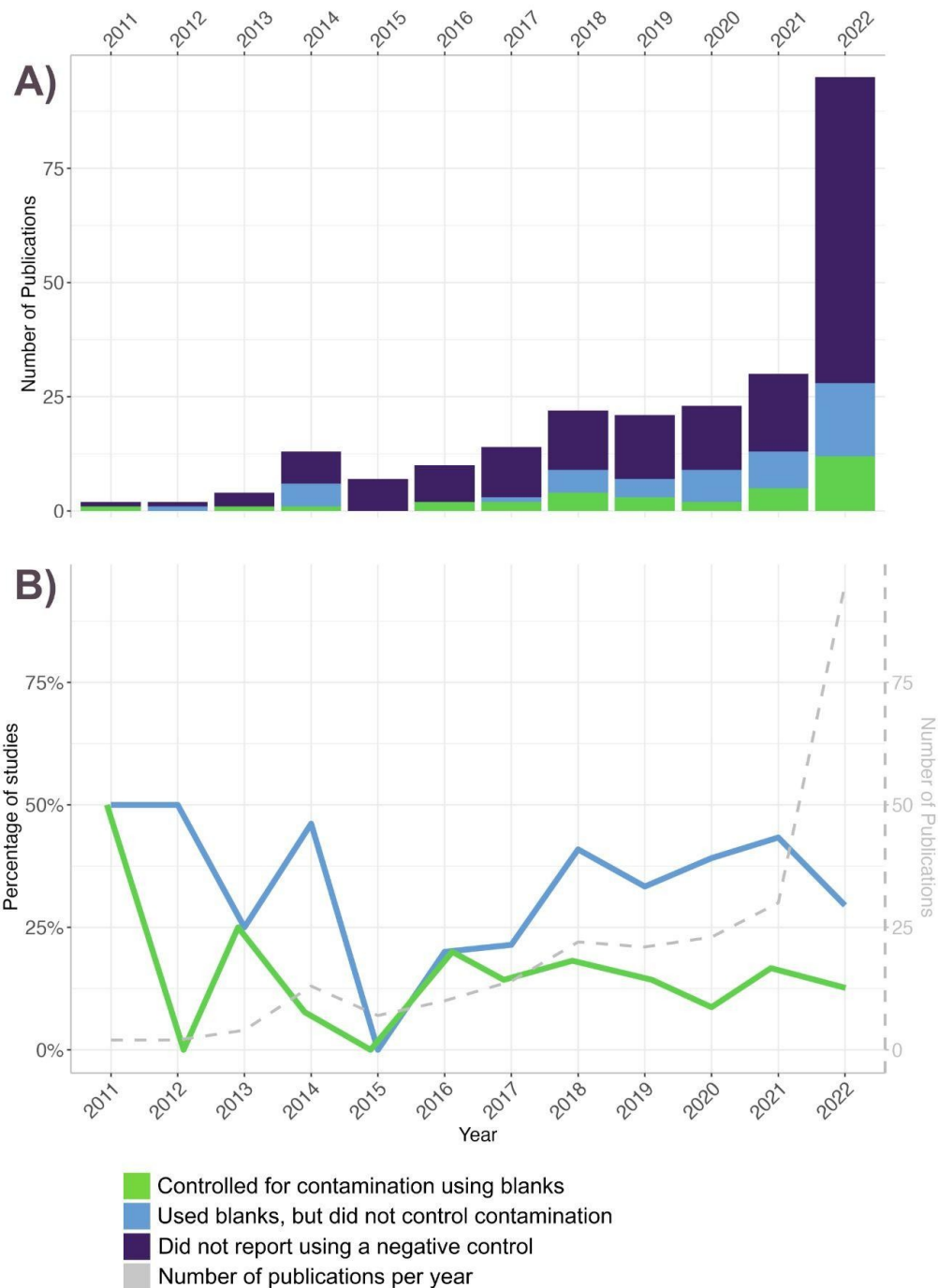
Figure 3: The trends in negative control use in our final pool of 243 papers from a 10-year period. A) The number of studies published per year that indicated including negative controls but did not control contamination (blue), included and sequenced negative controls to control for contamination (green), and those that did not use negative controls (purple). B) The same data replotted as percentages of studies. The total number of studies published per year is shown in grey.

Importantly, we also noted that there was considerable variation in the approaches used by the 33 studies that controlled for contamination. A total of 7/33 of studies used the statistical package Decontam [9] to call putative contaminant ASVs. The remaining studies used heuristic approaches, often involving a comparison between the communities identified, or manually filtering prevalent taxa found in negative controls from samples. Sometimes taxa were filtered based on the number of reads of putative contaminants from the controls, as a crude 'limit of detection' measure.

Without measuring the LoD, it is difficult to reliably distinguish true taxa from the contamination-induced 'background noise' that is inevitably introduced during sampling and sequencing processes. Only 1.6% (n = 4 studies) measured the LoD using qPCR. The absence of measurements for the LoD, coupled with the few papers which controlled for contamination, further support the finding that few authors accounted for potential contamination in their results. An alternative method to measure the absolute abundance and LoD is the inclusion of known amounts of artificial DNA or cells known as "spike-ins" [40]. While not assessed in this systematic literature review, as long as the spike-in's identity is carefully considered to ensure their absence in the target communities, they can offer some advantages over qPCR, including being cost-effective and providing quantification post-sequencing, accounting for the entire workflow including PCR amplification and sequencing biases.

5.3.3 Acknowledging off-target amplification from plant/animal DNA

Reporting the number of reads from off-target DNA is highly recommended to provide an accurate overview of the sequences recovered and to indicate the potential microbial biomass of samples and resulting sensitivity to introduced DNA contamination. We investigated what percentage of studies acknowledged the presence of chloroplast or mitochondrial DNA, the most predominant forms of off-target DNA in insect microbiota studies, and found 35.2% included such acknowledgment. Most were studies that included a filtering step to remove chloroplast and/or mitochondrial DNA from their sequencing results, but did not necessarily disclose the proportion of non-microbial DNA in the data. Only 5.8% (n = 14 studies) of total studies quantified the amount of off-target amplification. It is possible that studies did not provide this information because off-target DNA represented few reads or was absent. However, disclosing the number of off-target reads can still be useful for comparing 16S data between insect microbiota studies.

In addition to microbial biomass, the amount of off-target DNA amplification also depends on the diet of the specimen, gut content at time of sampling, and the primers used for PCR [20].

Therefore, there can be considerable variation in the amount of off-target DNA and amplification bias within insect species, depending on study design. The implication is that there can be differences in the bacterial abundances and diversity assessments, particularly in the representation of low abundance taxa, between studies. In these cases, reporting the proportion of off-target DNA can indicate the extent of the potential amplification bias and facilitate comparison between 16S datasets.

The vast majority of studies included in our systematic review did not follow best practice for controlling contamination. Consequently, a large proportion of published insect microbiota may be contaminated and this has the potential to mask and misrepresent true insect microbiota. While it is impossible to know the true number of studies affected, the lack of contamination control found in this systematic review suggests it could be high. This reveals a strong need for improved methodological rigor and adherence to standard guidelines across insect microbiota research.

5.4 Concluding remarks and recommendations

While this study focuses specifically on insect microbiota studies, the problem of contamination likely extends to other animal, plant, and environmental samples. Insects are not uniquely rife with contamination. All studies that use DNA sequencing are potentially susceptible to contamination, and low biomass samples can be common—though often unrecognized—in many environments. Further, we do not expect insect researchers to be uniquely likely (or not) to control for contamination. Hence, we speculate that a lack of contamination control is similarly widespread in other fields of research that use 16S sequencing.

To address contamination, we recommend the RIDE checklist, a minimum standards guide for low microbial biomass studies that can be easily integrated into study design [5]. In addition to the four original guidelines, based on our findings we propose including an additional guideline. The RIDES checklist stands for: 1) Report methodology, 2) Include controls in sequencing, 3) Determine the level of contamination to measure the limit of detection, 4) Explore the impacts of contamination in downstream analysis, and 5) State the amount of off-target DNA amplification to disclose potential amplification bias.

This checklist can and should be paired with biologically informed inspections of the data. For example, bacteria that are found at consistent levels across species, habitats, tissues or other

factors that often differentiate microbiomes should raise suspicion as potential contaminants. The dominant taxa should also be scrutinised. How plausible is it that they are true symbionts? Taxa characteristic of human skin, the ocean, or extreme environments should be treated with caution. Adopting these recommendations will improve the accuracy and standardization of future work, and reduce the uncertainty posed by undisclosed amplification bias and uncontrolled DNA contamination in the field of insect microbiota research.

5.5 Acknowledgements

5.6 Data availability statement

All data on our filtering and metadata set are available in the supplementary materials. Code used to analyse the data is available on Github at: https://github.com/hi-its-lisou/BlanksSLR.git

5.7 References

1. Braendle C, Miura T, Bickel R, Shingleton AW, Kambhampati S, Stern DL. 2003. Developmental origin and evolution of bacteriocytes in the aphid–*Buchnera* symbiosis. *PLOS Biology* 1 (1):e21.
2. Rio RVM, Symula RE, Wang J, Lohs C, Wu Yn, Snyder AK, Bjornson RD, Oshima K, Biehl BS, Perna NT, Hattori M, Aksoy S. 2012. Insight into the transmission biology and species-specific functional capabilities of Tsetse (Diptera: Glossinidae) obligate symbiont *Wigglesworthia*. *mBio* 3 (1):10.1128/mbio.00240–11.
3. Kwong WK, Medina LA, Koch H, Sing KW, Soh EJY, Ascher JS, Jaffé R, Moran NA. 2017. Dynamic microbiome evolution in social bees. *Science Advances* 3 (3):e1600513.
4. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal* 6 (8):1621–1624.

5. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. 2019. Contamination in low microbial biomass microbiome studies: issues and recommendations. *Trends in Microbiology* 27 (2):105–117.

6. Olomu IN, Pena-Cortes LC, Long RA, Vyas A, Krichevskiy O, Luellwitz R, Singh P, Mulks MH. 2020. Elimination of "kitome" and "splashome" contamination results in lack of detection of a unique placental microbiome. *BMC Microbiology* 20 (1):157.

7. Paniagua Voirol LR, Valsamakis G, Yu M, Johnston PR, Hilker M. 2021. How the 'kitome' influences the characterization of bacterial communities in Lepidopteran samples with low bacterial biomass. *Journal of Applied Microbiology* 130 (6):1780–1793.

8. Hornung BVH, Zwittink RD, Kuijper EJ. 2019. Issues and current standards of controls in microbiome research. *FEMS Microbiology Ecology* 95 (5):fiz045.

9. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. 2018. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 6 (1):226.

10. Lauder AP, Roche AM, Sherrill-Mix S, Bailey A, Laughlin AL, Bittinger K, Leite R, Elovitz MA, Parry S, Bushman FD. 2016. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome* 4 (1):29.

11. Leiby JS, McCormick K, Sherrill-Mix S, Clarke EL, Kessler LR, Taylor LJ, Hofstaedter CE, Roche AM, Mattei LM, Bittinger K, Elovitz MA, Leite R, Parry S, Bushman FD. 2018. Lack of detection of a human placenta microbiome in samples from preterm and term deliveries. *Microbiome* 6 (1):196.

12. Hammer TJ, Janzen DH, Hallwachs W, Jaffe SP, Fierer N. 2017. Caterpillars lack a resident gut microbiome. *Proceedings of the National Academy of Sciences* 114(36):9641–9646.

13. Lietaer L, Pascottini OB, Hernandez-Sanabria E, Kerckhof FM, Lacoere T, Boon N, Vlaminck L, Opsomer G, Van de Wiele T. 2021. Low microbial biomass within the reproductive tract of mid-lactation dairy cows: A study approach. *Journal of Dairy Science* 104 (5):6159–6174.

14. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. 2016. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut pathogens* 8:1–12.

15. Zhong ZP, Tian F, Roux S, Gazitúa MC, Solonenko NE, Li YF, Davis ME, Van Etten JL, Mosley-Thompson E, Rich VI, et al.. 2021. Glacier ice archives nearly 15,000-year-old microbes and phages. *Microbiome* 9:1–23.

16. Barton H, Taylor N, Lubbers B, Pemberton A. 2006. DNA extraction from low-biomass carbonate rock: an improved method with reduced contamination and the low-biomass contaminant database. *Journal of Microbiological Methods* 66 (1):21–31.

17. Luhung I, Uchida A, Lim SB, Gaultier NE, Kee C, Lau KJ, Gusareva ES, Heinle CE, Wong A, Premkrishnan BN, et al.. 2021. Experimental parameters defining ultra-low biomass bioaerosol analysis. *NPJ biofilms and microbiomes* 7 (1):37.

18. Bruckner J, Venkateswaran K. 2007. Overview of methodologies to sample and assess microbial burden in low biomass environments. *Japanese Journal of Food Microbiology* 24 (2):61–70.

19. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* 12 (1):87.

20. Hanshew AS, Mason CJ, Raffa KF, Currie CR. 2013. Minimization of chloroplast contamination in 16S rRNA gene pyrosequencing of insect herbivore bacterial communities. *J Microbiol Methods* 95 (2):149–55.

21. Dams E, Hendriks L, Van De Peer Y, Neefs JM, Smits G, Vandenbempt I, de Wachter R. 1988. Compilation of small ribosomal subunit RNA sequences. *Nucleic Acids Research* 16:r87 – r173.

22. Pompanon F, Deagle BE, Symondson WO, Brown DS, Jarman SN, Taberlet P. 2012. Who is eating what: diet assessment using next generation sequencing. *Molecular ecology* 21 (8):1931–1950.

23. Deissová T, Zapletalová M, Kunovsky L, Kroupa R, Grolich T, Kala Z,ˋ Borˇilová Linhartová P, Lochman J. 2023. 16S rRNA gene primer choice impacts off-target amplification in human gastrointestinal tract biopsies and microbiome profiling. *Scientific Reports* 13 (1):12577.

24. Song L, Xie K. 2020. Engineering CRISPR/Cas9 to mitigate abundant host contamination for 16S rRNA gene-based amplicon sequencing. *Microbiome* 8 (1):80.

25. Phalnikar K, Kunte K, Agashe D. 2019. Disrupting butterfly caterpillar microbiomes does not impact their survival and development. *Proceedings of the Royal Society B:Biological Sciences* 286 (1917):20192438.

26. Hammer TJ, De Clerck-Floate R, Tooker JF, Price PW, Miller DG, Connor EF. 2021. Are bacterial symbionts associated with gall induction in insects? *Arthropod-Plant Interactions* 15 (1):1–12.

27. Shelomi M, Lo WS, Kimsey LS, Kuo CH. 2013. Analysis of the gut microbiota of walking sticks (*Phasmatodea*). *BMC Research Notes* 6 (1):368.

28. Weyrich LS, Farrer AG, Eisenhofer R, Arriola LA, Young J, Selway CA, Handsley-Davis M, Adler CJ, Breen J, Cooper A. 2019. Laboratory contamination over time during low-biomass sample analysis. *Molecular ecology resources* 19(4):982–996.

29. Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L, Granger B, Green L, Howd T, Mason T, Vicente G, Dasilva M, Brodeur W, DeSmet T, Dodge S, Lennon NJ, Gabriel S. 2018. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* 19 (1):332.

30. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372:n71.

31. Innovation VH. Covidence systematic review software www.covidence.org.

32. Hammer TJ, Dickerson JC, Fierer N. 2015. Evidence-based recommendations on storing and handling specimens for analyses of insect microbiota. *PeerJ* 3:e1190.

33. Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C, Clarke E, Lauder A, Sherrill-Mix S, Chehoud C, Kelsen J, Conrad M, Collman RG, Baldassano R, Bushman FD, Bittinger K. 2017. Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* 5 (1):52.

34. Teoh MC, Furusawa G, Veera Singham G. 2021. Multifaceted interactions between the pseudomonads and insects: mechanisms and prospects. *Archives of Microbiology* 203 (5):1891–1915

35. Cohen H, McFrederick QS, Philpott SM. 2020. Environment shapes the microbiome of the blue orchard bee, *Osmia lignaria*: RRH: Environmental drivers of bee microbiome. *Microbial Ecology* 80 (4):897–907.

36. Holley JAC, Jackson MN, Pham AT, Hatcher SC, Moran NA. 2022. Carpenter bees (*Xylocopa*) harbor a distinctive gut microbiome related to that of honey bees and bumble bees. *Applied and Environmental Microbiology* 88 (13):e00203–22.

37. Feng W, Wang XQ, Zhou W, Liu GY, Wan YJ. 2011. Isolation and characterization of lipase-producing bacteria in the intestine of the silkworm, *Bombyx mori*, reared on different forage. *Journal of Insect Science* 11 (1):135.

38. Li G, Zheng X, Zhu Y, Long Y, Xia X. 2022. Bacillus symbiont drives alterations in intestinal microbiota and circulating metabolites of lepidopteran host. *Environmental Microbiology* 24 (9):4049–4064.

39. Kaltenpoth M, Flórez LV. 2020. Versatile and dynamic symbioses between insects and *Burkholderia* bacteria. *Annual review of entomology* 65 (1):145–170.

40. Tourlousse DM, Yoshiike S, Ohashi A, Matsukura S, Noda N, Sekiguchi Y. 2017. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. *Nucleic Acids Research* 45 (4):e23–e23.
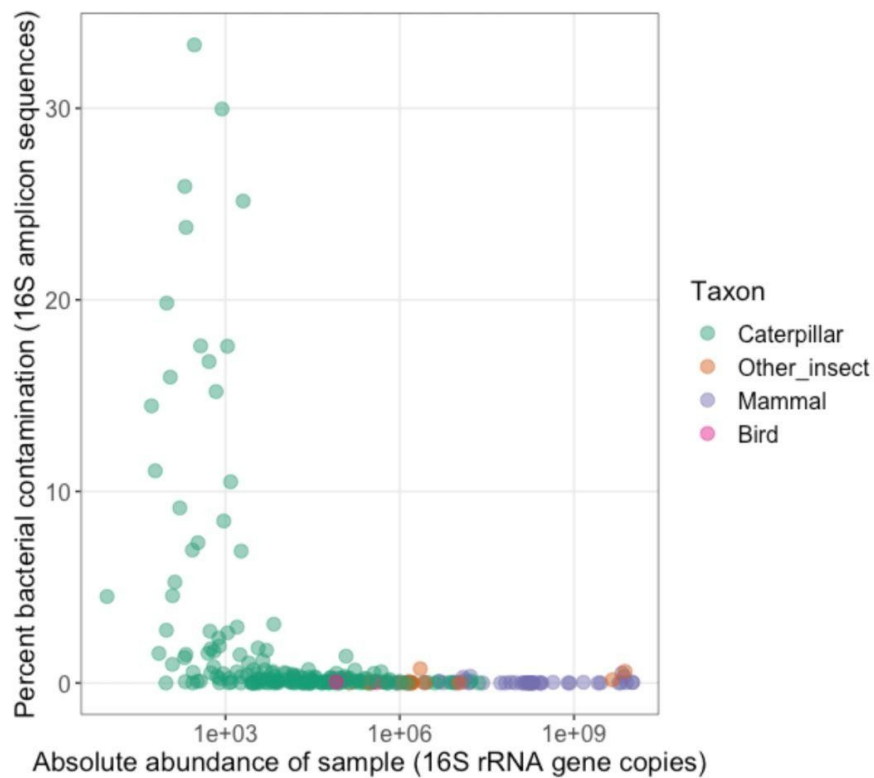
Supplementary materials

Figure S1: The empirical relationship between microbial biomass and contamination. Samples with varying levels of biomass are plotted against the percentage of bacterial contamination per sample. Low biomass samples are from caterpillars (green) moderate biomass samples are from other insects and birds (organs - pink), and high biomass samples are from mammals (purple).

Database advanced search inputs		
Scopus	**Web of science**	**Google Scholar**
ALL ("microbiome" OR "microbiota" OR "bacterial community" OR "bacterial communities" OR "microbial communities" OR "microbial community" OR "microbe" OR "bacteria") AND TITLE-ABS-KEY ("insect" OR "insects" AND NOT fish OR birds OR arthropods) AND ALL ("16S" OR "16S rRNA gene sequencing" OR "Next Generation DNA Sequencing" OR "High Throughput Nucleotide Sequencing" OR "Illumina Sequencing") AND PUBYEAR > 2010 AND PUBYEAR < 2023 AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "dp") OR LIMIT-TO (DOCTYPE , "English"))	TS=(insect OR insects OR entomology) AND ALL= (microbiome OR microbiota OR bacterial community OR bacterial communities OR microbial communities OR microbial community OR microbe OR bacteria) AND ALL=(16S OR 16S rRNA gene sequencing OR Next Generation DNA Sequencing OR High Throughput Nucleotide Sequencing OR Illumina Sequencing) AND DT=(Article OR Data Paper) AND LA=(English) AND PY=(2011-2022)	insect OR insects AND microbiome OR microbiota OR bacterial community OR microbial community AND 16S rRNA gene sequencing OR Next Generation DNA Sequencing
Number of hit results: 2937	**Number of hit results: 1215**	**Number of hit results: 17000**

Figure S2: The advanced search input used for Scopus, Web of Science, and Google Scholar. The search results were ordered by relevance, and the first 200 studies were extracted and uploaded to Covidence software to begin filtering.

Table S1: The inclusion and exclusion criteria used for advanced searches and manual screening. The final pool of papers strictly adheres to these requirements.

Criterium	Include	Exclude
Language:	English	Other
Years:	2011 - 2022	Other
Article type(s):	Peer-reviewed journal articles and data papers	Review, opinion, survey, thesis, or book
Subject type(s):	Insects (i.e., Class - Insecta)	Exclusively or mainly other animals, insect food, nest or environmental microbiota
Target microbiome(s):	Gut, whole insects, body segments, exoskeleton, glands using universal 16S primers	Exclusively fungal internal transcribed spacer (ITs), viruses, or specific 16s primer
Data output:	16S community analysis, primary dataset	For studies that use the same or overlapping datasets, the multiples are excluded to only include one
Sequencing method:	Next-generation/High throughput/Illumina sequencing, and 454 pyrosequencing	Culture dependent, Cloning, PCR-DGGE, TRFLP, bTEFAP, PCR-TTGE, or sanger sequencing

Exclusion Annotations

- Non-insect (microbiome descriptions for spiders, ticks, plants, soil, chickens, pigs, human, cadavers, genomic analyses, wine, annelids, or frass)
- No 16s (no original 16s results, no 16s in methodology, reviews, current opinions)
- Cultures (isolates, culture dependent, strain specific)
- Specific primers
- Cloning
- PCR DGGE
- Insect food (not including the insect itself)
- bTEFAP
- TRFLP
- Not English

Figure S3: The list of justifications given to a study when it was filtered/excluded from our final pool of studies for metadata extraction.

Table S2: The data extracted from our final pool of studies, and details on how/where the information was obtained from.

Metadata Extraction Information Gathered and Search Methods			
Metadata		**Information gathered**	**Search method**
Background info	**Title**	[Title of the study]	Exported from databases, imported into Covidence
	Year	[Year of publication]	
	Authors	[Authors of the study]	
	Journal published in	[Journal where the study was published]	
Research subject	**Order of insect**	[Order of insect used in the study]	From methods section
	Family of insect	[Family of insect used in the study]	
	Development stage(s) used in study	• 'Adults only' • 'Juveniles only' • '>1 development stage'	
	Target community of study	• 'Gut' • 'Body' (whole insect) • 'Other' (specific glands/organs, cuticular, or both internal and external communities targeted)	
	16S rRNA gene target region	[Region covered based on the primers used] (cross-referenced with Abellan-Schneyder et al. 2021).	
Sampling and extraction method	**Sampled**	• 'Gut' • 'Whole' • 'Other' (sample both dissected guts and whole insects, washes, or specific glands)	From methods section
	DNA extraction kit used	[Manufacturer of DNA extraction kit used, or 'Protocol' if no kit used]	
	Surface sterilisation	• 'No' • 'Yes'	

| Metadata		Information gathered	Search method

Use of negative controls	Did they use a negative control?	• 'No'
		• 'Yes'
		↳ [What kind?]
		↳ [How many?]
		→ Do they sequence them?
		  • 'No'
		  • 'Yes'
		   → Did they control contamination using them?
		    • 'No'
		    • 'Yes'
		     ↳ [Method used]
			Keywords search - "negative", "control", "blank", "extract", "contam"

Quantification	Do they use qPCRs in tandem?	• 'No'
		• 'Yes'
		→ Do they determine the limit of detection?
		  • 'No'
		  • 'Yes'
			Keywords search - "quanti-", "qPCR"

Alternative sources of DNA	Animal/plant DNA	• 'No' • 'Yes' (must be acknowledged in methods, results, or discussion)	Keywords search - "mito-", "chloro-", "plant", "animal"
	Transiency	• 'No' • 'Yes'	Keywords search - "transi-", "environm-"
	Relic DNA	• 'No' • 'Yes'	Keywords search - "relic", "dead", "remnant", "residual"
Data availability	Do they provide a project or accession number for sequence data?	• 'No' • 'Yes'	Keywords search - "proj'", "acc-"
Citations	How many citations do they have?	[Number of citations]	Crossref API via the rcrossref package in R
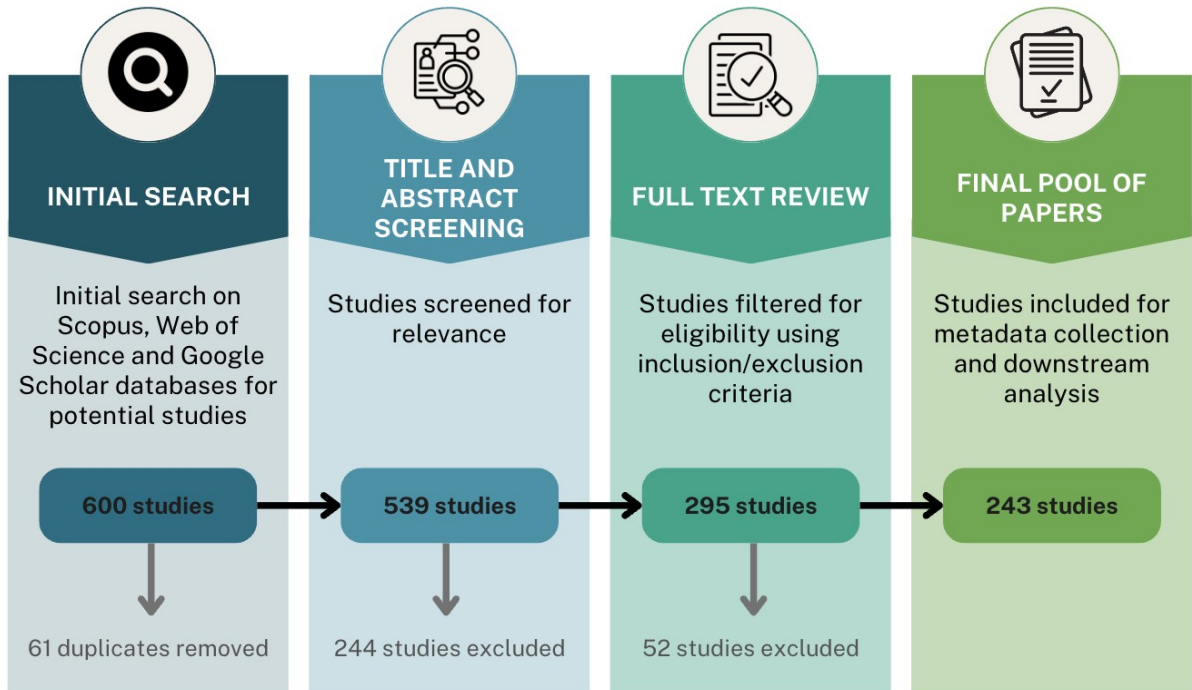
Figure S4: The number of papers that passed each filtering step, with the final pool of papers (n=243) being used for metadata collection.

Supplementary analysis: Additional considerations when assessing the nature of 16S sequences - Transient bacteria and Relic DNA

There are many important factors to consider when using 16S sequencing to characterize insect microbiota, with the focus of this systematic review being to investigate how contamination is dealt with in the literature. However, DNA contamination isn't the only factor that can influence assessments on taxonomic diversity and ecological function. Transient bacteria and relic DNA are also important considerations when interpreting sequencing results and assessing microbial assemblages.

While contamination, mitochondria and chloroplasts can be filtered based on taxonomy, transient microbes and relic DNA are much harder to distinguish, yet just as ubiquitous in the environment. Transient microbes do not replicate within a host, are often temporarily passing through, having positive, negative or negligible effects on insect hosts [1]. Relic DNA is the genetic material left behind from dead cells that can still be amplified and sequenced [2]. Both transients and relic DNA have the potential to influence ecological assessments of microbial communities as they interact differently with hosts compared to beneficial symbionts or pathogens [2, 3].

Despite this, only 22.5% of papers acknowledged transient bacteria and 3.6% acknowledged relic DNA as potential sources of DNA in their sequencing results. This finding suggests that some studies may be assuming symbionts, and overestimating their importance to the host. Without acknowledging alternative explanations, such as relic or transiency, we may be limiting our understanding of these complex ecological systems.

References:

1. Hammer TJ, Sanders JG, Fierer N. Not all animals need a microbiome. FEMS Microbiology Letters 2019;366(10)
2. Lennon JT, Muscarella ME, Placella SA, Lehmkuhl BK. How, When, and Where Relic DNA Affects Microbial Diversity. mBio 2018;9(3):10.1128/mbio.00637–18.
3. Snell Taylor SJ, Evans BS, White EP, Hurlbert AH. The prevalence and impact of transient species in ecological communities. Ecology 2018;99(8):1825–183