Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Insect Science

# Direct sequencing of insect symbionts via nanopore adaptive sampling

Jonathan H Badger[1], Rosanna Giordano[2], Aleksey Zimin[3], Robert Wappel[4], Senem M Eskipehlivan[4], Stephanie Muller[4], Ravikiran Donthu[5], Felipe Soto-Adames[6], Paulo Vieira[7], Inga Zasada[8] and Sara Goodwin[4]

Insect symbionts can alter their host phenotype and their effects can range from beneficial to pathogenic. Moreover, many insects exhibit co-infections, making their study more challenging. Less than 1% of insect species have high-quality referenced genomes available and fewer still also have their symbionts sequenced. Two methods are commonly used to sequence symbionts: whole-genome sequencing to concomitantly capture the host and bacterial genomes, or isolation of the symbiont's genome before sequencing. These methods are limited when dealing with rare or poorly characterized symbionts. Long-read technology is an important tool to generate high-quality genomes as they can overcome high levels of heterozygosity, repeat content, and transposable elements that confound short-read methods. Oxford Nanopore (ONT) adaptive sampling allows a sequencing instrument to select or reject sequences in real time. We describe a method based on ONT adaptive sampling (subtractive) approach that readily permitted the sequencing of the complete genomes of mitochondria, *Buchnera* and its plasmids (pLeu, pTrp), and *Wolbachia* genomes in two aphid species, *Aphis glycines* and *Pentalonia nigronervosa*. Adaptive sampling is able to retrieve organelles such as mitochondria and symbionts that have high representation in their hosts such as *Buchnera* and *Wolbachia*, but is less successful at retrieving symbionts in low concentrations.

**Addresses**
[1] Genetics and Microbiome Core, Laboratory of Integrative Cancer Immunology, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA
[2] Institute of Environment, Florida International University, Miami, FL, USA
[3] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
[4] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[5] Centre for Life Sciences, Mahindra University, Bahadurpally, Hyderabad 500043, India
[6] Florida Department of Agriculture and Consumer Services, Department of Plant Industry, Gainesville, FL 32614, USA
[7] USDA-ARS Agricultural Research Center, Mycology & Nematology Genetic Diversity & Biology Laboratory, Beltsville, MD, USA
[8] USDA-ARS Horticultural Crops Research Laboratory, Corvallis, OR, USA

Corresponding author: Goodwin, Sara (sgoodwin@cshl.edu)

**Current Opinion in Insect Science** 2023, **61**:101135

This review comes from a themed issue on **Insect genomics**

Edited by **Robert DeSelle** and **Sara Oppenheim**

Available online 4 November 2023

https://doi.org/10.1016/j.cois.2023.101135



## Introduction

The study of arthropod genomics and metagenomics is critical to the study of ecology as arthropods play a significant role in many ecosystems. Arthropods are the most diverse group of animals on the planet, with over five million species thought to exist [1]. They play important roles in many ecological niches, from pollination to decomposition as well as to serve as food for other species. The study of arthropods can give researchers insights into their evolution, adaptation, interactions with other species, and mechanisms that regulate their behavior and ecological functions, including how bacterial and fungal symbionts can modulate host behavior and physiology.

Arthropods' success may be in part attributed to the symbiotic relationships they have with various microorganisms. Research on symbionts in arthropods has revealed their roles in nutrient acquisition, defense against pathogens, host reproduction, and more. Understanding the relationship between arthropod hosts and their symbionts has important implications for understanding the biology of arthropods in general, but also for the development of novel approaches to pest control and disease transmission and prevention.

One symbiont of particular interest is *Wolbachia*, a Rickettsia bacterium commonly found in arthropods and

nematodes and estimated to be one of the most common symbiotic bacteria [2]. *Wolbachia* is an obligate intracellular bacterium that has defied rearing outside a living cell environment. It perpetuates and favors its own reproduction via various methods such as cytoplasmic incompatibility, feminization, parthenogenesis, and male killing [3,4] and in rare cases can be pathogenic [5].

In recent years, there has been growing interest in using *Wolbachia* as a tool to control the spread of mosquito-borne diseases, such as dengue, Zika, and chikungunya [6–8]. This is because *Wolbachia* can be introduced into mosquito populations and reduces their ability to transmit these diseases to humans. This approach is known as *Wolbachia*-based vector control, and it has shown promising results in field trials in Indonesia and Brazil [9–11].

Despite the growing interest in arthropod/symbiont genomics, the field remains understudied, with less than 1% of insect species having high-quality reference genomes available and with fewer still having their symbionts sequenced [12,13].

In the last several years, a number of initiatives such as the i5k [14] and ag100Pest [15] have aimed to take advantage of long-read sequencing methods to increase the number and quality of publicly available genomes, transcriptomes, and symbionts for arthropods. Today, these initiatives are relying heavily on long-read methods to achieve their goals. The first publicly available sequence of *Drosophila melanogaster* [16,17] catapulted arthropod research and genome science in general, but these studies had significant limitations compared with genomes obtained currently with long-read technology.

Long-read sequencing is an essential element in arthropod research as high levels of heterozygosity, high repeat content, and high proportion of transposable elements pose significant challenges for traditional short-read sequencing technologies [18]. Long-read sequencing can overcome these challenges by generating reads that span repetitive regions, complex rearrangements, and gene fusions that may be missed by short-read sequencing.

These advances are important for understanding arthropod biology and evolution. While there have been hurdles to address, Oxford Nanopore (ONT) sequencing has been of particular interest for arthropod researchers due to its low cost and scalability relative to PacBio sequencing.

The quality of DNA extractions needed has been a limitation for the wider adoption of ONT for arthropod genomics. DNA extraction from insects can be a challenging process due to a number of factors. The size and structure of insects can make it difficult to obtain sufficient DNA from a single individual for analysis. Some insects have a tough exoskeleton that is resistant to chemical and mechanical disruption, making it difficult to break down the cells and release the DNA. In addition, many insects produce defensive chemicals, such as polyphenols and quinones, that can interfere with DNA extraction protocols. DNA extractions of insects also have high levels of DNA-degrading enzymes such as nucleases, which can break down genomic DNA if not properly preserved. In addition, insect samples obtained from the field can be accompanied by environmental contaminants such as bacteria and fungi. These challenges make it important to use appropriate sample preservation techniques and extraction methods to obtain high-quality DNA for downstream analysis.

PacBio established early protocols for dealing with low-input library preparations, such as the single-mosquito method for *de novo* assembly highlighted by PacBio in 2019 [19] and ultra-low-input methods [20]. While these methods have also been adapted for ONT sequencing, the higher sequencing coverage needed to address the higher error rate [21] has limited its adoption. Recently however, ONT sequencing has greatly improved its accuracy, with 114 chemistries, to >98% [22] and ultra-long methods are proving invaluable for creating highly contiguous or even gapless assemblies [23,24].

The sequencing of symbionts poses a particular challenge to sequencing and to long sequencing in general. Investigations studying symbionts have relied on PCR to amplify the target DNA from the host background [25–27]. While effective at increasing the abundance of the target sequence, the resulting product is usually fragmented and not representative of the entire genome. The PCR method is further confounded if the symbiont is poorly characterized, making primer design challenging and not possible in many cases. In some rare instances, symbionts can be isolated and cultured *in vitro* facilitating their augmentation for sequencing. Another approach has been to dissect infected ovaries from multiple individuals followed by whole-genome amplification [28]. However, the genomic amplification step can add artifacts to the data.

ONT has introduced a method unique to its platform, referred to as adaptive sequencing, that can be invaluable for dealing with issues of contamination and symbiont sequencing [29]. This method, given a reference, can selectively reject DNA or RNA fragments of interest. The method can be implemented in one of two ways. In an additive mode, a reference is provided and only the sequences that match the reference are allowed to move forward and sequence. Alternatively, in a subtractive mode, fragments that match the reference are rejected. This protocol can be used to selectively sequence the host genome, if this is known, thus reducing off-target contamination or to selectively sequence

symbionts or gut contents [30]. Herein, we describe results from the implementation of the ONT-subtractive approach with the goal of enriching for the endosymbiont component from genomic extractions of two aphid species, *Aphis glycines* (soybean aphid) and *Pentalonia nigronervosa* (banana aphid).

## Methods

### Origin of samples

*Aphis glycine* biotype-3 aphids were obtained from a culture maintained at the Soybean/Maize Germplasm, Pathology, and Genetics Research Laboratory, Urbana, IL. Aphids were reared in a Percival, TC-2 plant tissue culture chamber at 23ºC, 60% humidity and at 16-hours light regime, on whole soybean LD08-12435a (*Rag2*) and were obtained from the culture on March 22, 2022. *P. nigronervosa* adults and nymphs were collected from *Musa* sp. (Linnaeus) in Gainesville, FL, USA (N 29.62825 W 82.35839), and flash-frozen on dry ice on October 4, 2022. Voucher samples were deposited in the Florida State Collection of Arthropods and can be accessed under sample number 10052022-08884.

### DNA extraction and Oxford Nanopore library preparation

Approximately 10 aphids were placed in a standard 1.5-ml centrifuge tube and ground with a disposable pestle on dry ice (Fisher PN 12141364). The powder was then extracted using a Qiagen MagAttract kit following the manufacturer's instructions for a tissue sample. Samples were quantified with a Qubit Fluorometer, Nanodrop, and Femto Pulse for sizing. Oxford Nanopore DNA libraries were prepared with either SQK-LSK109 or SQK-LSK110 following the manufacturer's instructions. Individual libraries were sequenced on a GridION with R9.5 flow cells for 48 hours. The reference for adaptive sampling consisted of the host references or host reference plus the species-specific *Wolbachia* reference.

The reference for the genome of *P. nigronervosa* genome and its *Buchnera* and *Wolbachia* symbionts were obtained from (https://doi.org/10.5281/zenodo.3765644) [31]. Reads that mapped to the reference were rejected from the run. Real-time base calling with Guppy v 6.3.8, set to high accuracy, was used. Sequencing data are in process for deposit to NCBI and will be released pending publication of related manuscripts. Additionally, data may be accessed upon request.

### Analysis

Raw read data from each condition (control, host only, or host plus selected nonhost targets) were used as input for metaFlye [32] to assemble the mitochondria and bacterial genomes. A custom Kraken2 [33] database was used as a reference to classify the contigs for metagenome classification, the targets in the database are listed in Supplementary Table 1. The final assemblies were polished via Medaka [34] (Github) by comparing the raw reads to the metaFly reference. Reference comparison plots were generated by MUMmer2 [35]. The purging of *Wolbachia* sequences from *P. nigronervosa* was done by inputting the assembly into Kraken2, then removing the contigs classified as *Wolbachia* from the reference. The alignment yield from runs was calculated by aligning the read data with Minimap2 [36] to determine the number of reads that mapped to the hosts, symbionts, and the symbiont plasmids.

## Results

### Effectiveness of adaptive sampling

We performed several trials of adaptive sampling for two different aphid species, *A. glycines* and *P. nigronervosa*. These were selected as representative aphids grown in culture, *A. glycines*, and 'the wild' *P. nigronervosa*. Data illustrating the performance of each flow cell are found in Table 1. There was a notable difference in the yield of individual flow cells. This is likely the result of normal nanopore variations as well as reduction in performance due to the demands of adaptive sampling. Standard sequencing without adaptive sampling had the highest yield at 22Gb. The *P. nigronervosa* host plus *Wolbachia* run performed the worst with a maximum per flowcell yield of 1.4 Gb. The overall read length was also affected by adaptive sampling. During adaptive sampling, the sequence data from a pore are aligned to the provided references file in real time. In the depletion mode used here, if the sequence data align to the reference, the voltage in the pore is reversed and the DNA fragment is rejected. This occurs within the first few hundred bases. Ultimately, this means that the average length of the reads generated is shorter due to the rejection of undesirable reads. In this study, the mean length of the

**Table 1**

**Summary of performance of all runs executed to test adaptive sampling for *A. glycines* and *P. nigronervosa*.**

| Run | Run type | Reference used | Total yield (bp) | Mean read length (bp) | Total # of reads | Accepted yield | Accepted read length (bp) | Accepted Reads |
|---|---|---|---|---|---|---|---|---|
| 1 | *A. glycines*, No subtraction | None | 22,337,187,111 | 5,646 | 3,956,226 | NA | NA | NA |
| 2 | *A. glycines*, Host subtraction | *A. glycines* Bt3 host genome (in prep.) | 11,436,173,421 | 1,203 | 9,510,007 | 4,248,237,298 | 9,672 | 439,234 |
| 3 | *A. glycines*, Host + *Wolbachia* subtraction | *A. glycines* Bt3 host genome + *Wolbachia* (in prep.) | 2,617,486,584 | 1,137 | 2,301,501 | 202,073,729 | 12,369 | 16,337 |
| 4 | *P. nigronervosa*, Host + *Wolbachia* subtraction | https://zenodo.org/record/3765644#.ZFGDWnbMKMp | 1,853,530,986 | 2,068 | 896,211 | 307,183,015 | 7,088 | 43,340 |
| 5 | *P. nigronervosa*, Host subtraction | Genome at above zenodo link purged of *Wolbachia* data | 7,781,339,818 | 1,508 | 5,159,420 | 808,096,162 | 7,230 | 111,769 |

**Table 2**

**Yield mapping obtained for host, mitochondria, and symbiont genomes.**

| Run | Run type | Reference used | Yield mapping to host | Yield mapping to *Buchnera* | Yield mapping to *Wolbachia* | Yield mapping to mitocondri | Yield mapping to *Buchnera* pLeu plasmid | Yield mapping to *Buchnera* pTrp plasmid |
|---|---|---|---|---|---|---|---|---|
| 1 | *A. glycines*, No subtraction | None | 19,912,941,688 | 130,929,950 | 59,247,732 | 53,094,258 | 335,772 | 709,890 |
| 2 | *A. glycines*, Host subtraction | *A. glycines* Bt3 (in preperation) | 3,827,559,347 | 7,511,327,009 | 251,672,467 | 2,252,561 | 13,196,645 | 44,908,132 |
| 3 | *A. glycines*, Host + *Wolbachia* subtraction | *A. glycines* Bt3 unpublished + *Wolbachia* (In prep) | 1,865,119,318 | 507,309,360 | 652,673 | 850,113 | 741,787 | 7,629,611 |
| 4 | *P. nigronervosa*, Host + *Wolbachia* information | https://zenodo.org/record/3765644#.ZFGDWnbMKMp | 502,187,429 | 532,386,626 | 0 | 34,436,733 | 1,219,294 | 2,032,572 |
| 5 | *P. nigronervosa*, Host | Above zenodo genome purged of *Wolbachia* data | 5,848,225,326 | 1,458,706,195 | 0 | 239,452,418 | 4,765,696 | 10,914,019 |

accepted reads was 5–10 kb longer than the mean length of all reads (Table 1).

To test the feasibility of adaptive sampling to deplete the representation of the abundant host genome, we conducted control nonsubtractive and host-only subtractive sequencing for *A. glycines*. A comparison of the nonsubtractive control run (Tables 1 and 2, Run 1) with a run rejecting the host genome (Tables 1 and 2, Run 2) shows that for the former, the host was represented by >89% of the data, while in the latter, only 33% of the data mapped to the host. Likewise, the total bases mapping to nonhost genomes (mitochondria, *Buchnera*, and *Wolbachia*) in Run 1 were substantially less abundant than in Run 2 (Table 2). The *Buchnera* bases increased more than 50-fold, the bases mapping to *Wolbachia* increased fivefold, and the *Buchnera*-associated plasmids increased more than 1000-fold.

When *Wolbachia* was included as a subtraction target, the overall flowcell performance declined notably, down to 2.6 total Gbs (Table 1, Run 3). Despite this, the proportional abundance of *Buchnera* was much higher, accounting for nearly 20% of all bases and 300% more abundant than in the control run. As expected, the *Wolbachia* abundance was drastically reduced, accounting for just 652 kb of total sequence (Table 2, Run 3). Interestingly, the abundance of both the pLeu and pTrp plasmid was much lower than would be expected based on the reduced flowcell yield. While their abundance remained higher than in the control run, they were considerably less abundant than in the host-only subtractive run, implying that reads mapping to these plasmids were rejected more than expected. The *Buchnera* plasmids pLeu and pTrp are plasmids that code for the essential amino acids leukine and tryptophan and are necessary for the proper functioning of *Buchnera*.

We next evaluated host depletion in a sample of *P. nigronervosa* collected from the wild.

Based on published data from the sequencing of a genome of a banana aphid population from Kenya [37], we surmised that *P. nigronervosa* is infected with *Wolbachia*. As with *A. glycines*, the initial run of *P. nigronervosa* (Table 2, Run 4), using the public *P. nigronervosa* genome as reference [37], showed that the abundance of *Buchnera* represented >36% of the total data generated. This result is compatible with the *Buchnera* abundances in Runs 2 (65%) and 3 (20%). Given the expectation that *P. nigronervosa* is infected with *Wolbachia*, the absence of reads mapping to *Wolbachia* from Run 4 was unexpected. Our results concur with the conclusion of [37] that different geographic populations of *P. nigronervosa* differ with respect to their *Wolbachia* infection status and the association is not likely to be obligate.

We closely examined the original reference genome of *P. nigronervosa* and determined that it contained data mapping to *Wolbachia*. We purged these data and generated a new reference used for a subsequent *P. nigronervosa* experiment. The results of the new experiment were comparable to the first, resulting in ~19% of the data mapping to *Buchnera* and its plasmids, while 75% of the data mapped to the host (Table 2, Run 5).

## Examination of symbiont genomes and mitochondria obtained for each run

To generate the assemblies of constituent symbionts, for all runs, the reads were assembled with metaFlye. The length in base pairs of the metaFlye assemblies and coverage for the mitochondria, *Wolbachia*, and *Buchnera* and its plasmids (pLeu, pTrp) are found in Table 3.

With the exception of a few cases that were slightly shorter in length, most of the assembled genomes were complete (Table 3). The resulting coverage generated for the *A. glycines* host subtraction (Run 2) can be seen in Table 2. We used Medaka to polish the metaFlye assemblies of mitochondria, *Wolbachia* and *Buchnera* and its plasmids generated from this run. These results can be seen in Figure 1. Similarly, the host subtraction run conducted for *P. nigronervosa* can be seen in Run 4 and the Medaka-polished metaFlye assemblies in Figure 2.

Variations in coverage impacted the assembly in two cases. In the *A. glycines* nonsubtractive run (Run 1), *Wolbachia* was assembled into two contigs, due to lower overall coverage, while in the *P. nigronervosa* host-subtractive run (Run 5), *Buchnera* was assembled into eight contigs (Table 3) due to excessive coverage that negatively impacted the assembly.

**Table 3**

**Assessment of mitochondria and metaFlye symbiont assemblies for each run performed.**

| | | | | | Reference | | Generated Assembly | | |
|---|---|---|---|---|---|---|---|---|---|
| Run | Code | Species | Subtraction type | Genome | Genome Size bp | Source | # contigs | Length bp from MetaFlye | Coverage (X) |
| 1 | A1 | *Aphis glycines* Bt3 | None | *Buchnera* | 626,817 | unpublished | 1 | 626,510 | 208 |
| 1 | A2 | *Aphis glycines* Bt3 | None | Mitochondria | 18,196 | unpublished | 1 | 17,636 | 2,918 |
| 1 | A3 | *Aphis glycines* Bt3 | None | pLeu | 7,255 | unpublished | 1 | 7,710 | 46 |
| 1 | A4 | *Aphis glycines* Bt3 | None | pTrp | 3,048# | CP009255.1 | 5 | 11,580 | * |
| 1 | A5 | *Aphis glycines* Bt3 | None | *Wolbachia* | 1,538,935 | unpublished | 2 | 1540048 | 38 |
| 2 | B1 | *Aphis glycines* Bt3 | Host only | *Buchnera* | 626,817 | unpublished | 1 | 626,817 | 11,983 |
| 2 | B2 | *Aphis glycines* Bt3 | Host only | Mitochondria | 18,196 | unpublished | 1 | 17,845 | 124 |
| 2 | B3 | *Aphis glycines* Bt3 | Host only | pLeu | 7,255 | unpublished | 1 | 15,273 | 1,819 |
| 2 | B4 | *Aphis glycines* Bt3 | Host only | pTrp | 3,048# | CP009255.1 | 1 | 14,242 | * |
| 2 | B5 | *Aphis glycines* Bt3 | Host only | *Wolbachia* | 1,538,935 | unpublished | 1 | 1,542,977 | 164 |
| 3 | C1 | *Aphis glycines* Bt3 | Host and *Wolbachia* | *Buchnera* | 626,817 | unpublished | 1 | 626,519 | 809 |
| 3 | C2 | *Aphis glycines* Bt3 | Host and *Wolbachia* | Mitochondria | 18,196 | unpublished | 1 | 17,866 | 47 |
| 3 | C3 | *Aphis glycines* Bt3 | Host and *Wolbachia* | pLeu | 7,255 | unpublished | 1 | 7,642 | 102 |
| 3 | C4 | *Aphis glycines* Bt3 | Host and *Wolbachia* | pTrp | 3,048# | CP009255.1 | 1 | 21,669 | * |
| 3 | C5 | *Aphis glycines* Bt3 | Host and *Wolbachia* | *Wolbachia* | 1,538,935 | unpublished | 0 | 0 | less than 1 |
| 4 | D1 | *Pentalonia nigronervosa* | Host only | *Buchnera* | 617,483 | Zenodo | 1 | 615,849 | 862 |
| 4 | D2 | *Pentalonia nigronervosa* | Host only | Mitochondria | 15,642 | Zenodo | 1 | 17,866 | 2,002 |
| 4 | D3 | *Pentalonia nigronervosa* | Host only | pLeu | 7,706 | Zenodo | 1 | 15,388 | 158 |
| 4 | D4 | *Pentalonia nigronervosa* | Host only | pTrp | 2,172 | Zenodo | 1 | 1,816 | * |
| 5 | E1 | *Pentalonia nigronervosa* | Host and *Wolbachia* | *Buchnera* | 617,483 | Zenodo | 8 | 732,167 | 2,362 |
| 5 | E2 | *Pentalonia nigronervosa* | Host and *Wolbachia* | Mitochondria | 15,642 | Zenodo | 1 | 22,463 | 15,308 |
| 5 | E3 | *Pentalonia nigronervosa* | Host and *Wolbachia* | pLeu | 7,706 | Zenodo | 1 | 7689 | 618 |
| 5 | E4 | *Pentalonia nigronervosa* | Host and *Wolbachia* | pTrp | 2,172 | *Zenodo* | 3 | 3,269 | * |

# The accurate size of the pTrp plasmid is uncertain because the number of operons has not been confirmed.

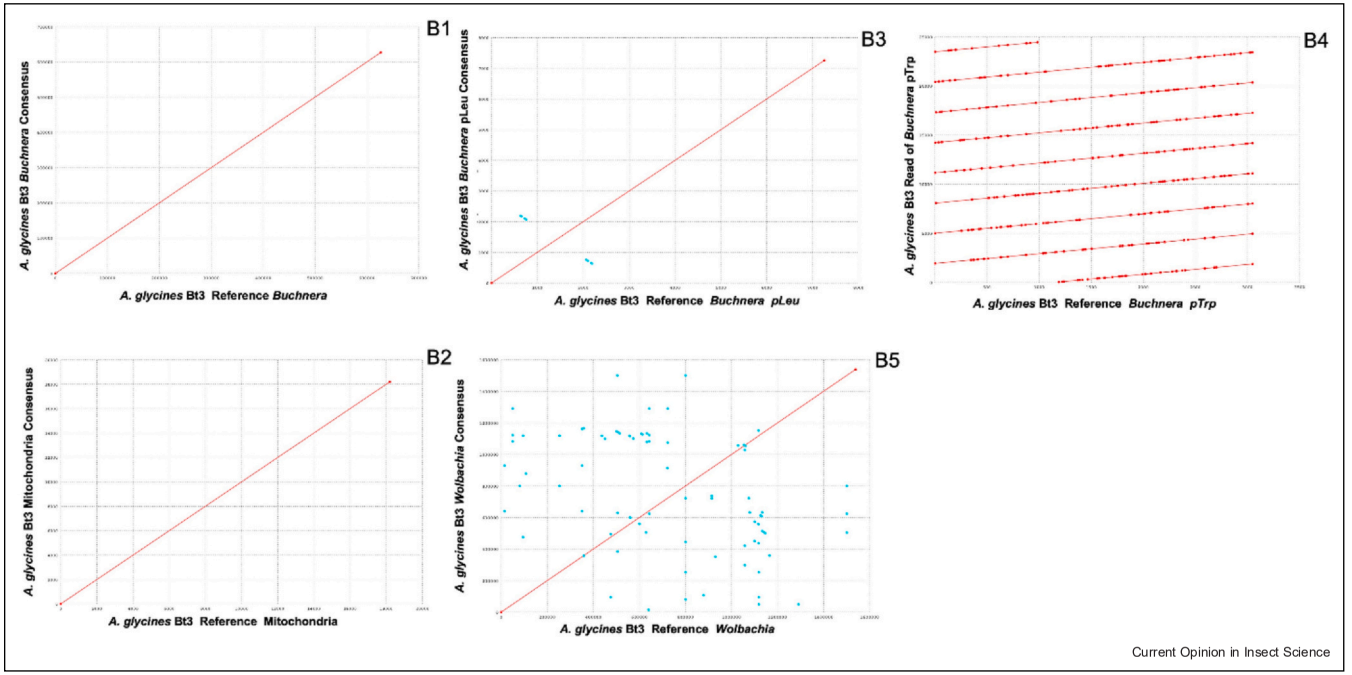
We successfully assembled the pLeu plasmid for the *Buchnera* of both *A. glycines* and *P. nigronervosa*, but it was not possible to assemble the complete pTrp plasmid due to its highly repetitive nature. When Medaka polishing was attempted to improve the assembly, it collapsed into a single operon matching the published length for this plasmid of ~3000 bp. To further explore this assembly, we extracted the longest read from each of the data sets and used NCBI BLAST to visualize what it aligned to (Sup. Figure 9). From the alignments obtained, we concluded that for both *A. glycines* and *P. nigronervosa*, the plasmid was composed of multiple copies of genes for tryptophan biosynthesis. In the case of *A. glycines*, the plasmid is composed of at least 17 identical copies of the pTrp operon, similar to what has been reported for other aphid species [38] (Figure 1, B4). While the pTrp plasmid in *P. nigronervosa* is composed of two complete operon copies and many additional incomplete copies (Figure 2, D4).

A visualization of the coverage for mitochondria, *Wolbachia*, and *Buchnera* and its plasmids for each run, for both *A. glycines* and *P. nigronervosa*, can be found in Supplementary Figures 1–5. Likewise, Medaka-polished plots of the mitochondria and symbiont assemblies for Run 1 (*A. glycines*, no subtraction), Run 3 (*A. glycines*, host and *Wolbachia* subtraction), and Run 5 (*P. nigronervosa*, host and *Wolbachia* subtraction) can be found in Supplementary Figures 6–8.

## Metagenome analysis

Subsequent to generating the metaFlye assemblies, Kraken2 was used to classify the contigs generated with metaFlye to assess the composition of genomes obtained from the subtractive runs for both *A. glycine* and *P. nigronervosa*. As expected, Kraken2 classified the host genomes and the most abundant symbionts (Table 4). In addition, the metagenome analysis also revealed the presence of environmental contaminants. Of these, *Alkalihalobacillus miscanthi*, a bacteria found in soil [39], was identified in the *A. glycines* runs. The combined length of the contigs obtained for *A. miscanthi* was equal to the size of its genome (Table 4). Other environmental contaminants such as *Microbacterium* sp. and *Escherichia coli* were also identified with the latter being present in both the *A. glycines* and *P. nigronervosa* preparations. However,

**Figure 1**




Run 2: MUMmer plots of Medaka-polished assemblies for *A. glycines* Bt3 with host subtraction: B1 — *Buchnera,* B2 — mitochondria, B3 — *Buchnera* pLeu plasmid, B4 — *Buchnera* pTrp plasmid, B5 — *Wolbachia.*

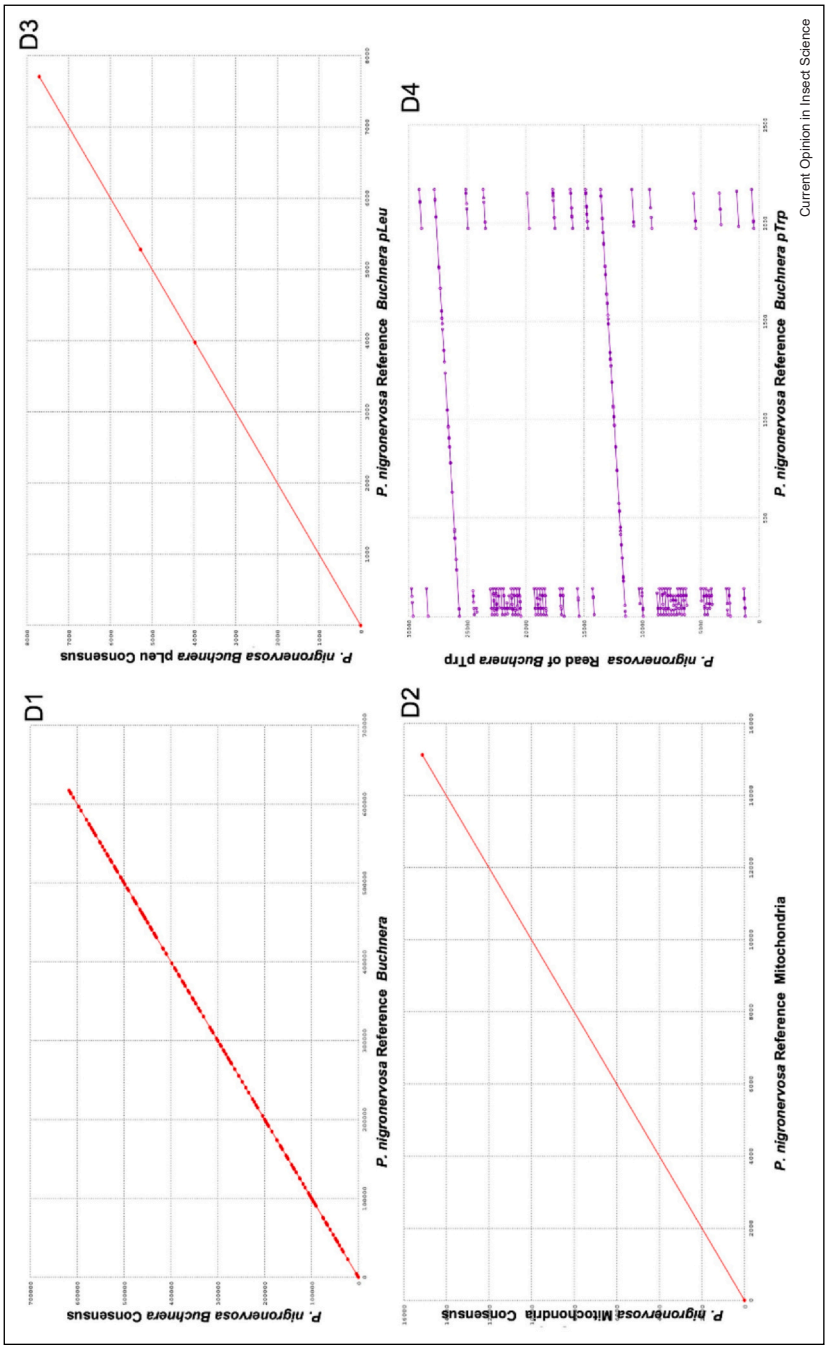



**Figure 2**

Run 4: MUMmer plots of Medaka-polished assemblies for *P. nigronervosa* with host subtraction: D1 — *Buchnera*, D2 — mitochondria, D3 — *Buchnera* pLeu plasmid, D4 — *Buchnera* pTrp plasmid.

**Table 4**

**Classification of MetaFlye contigs from *A. glycines* and *P. nigronervosa*.**

| ***Aphis glycines*** | | |
|---|---|---|
| **Organism** | **Length** | **Counts** |
| *Alkalihalobacillus miscanthi* (taxid 2598861) | 4,499,461 | 18 |
| *Aphis gossypii* (taxid 80765) | 3,034,180 | 148 |
| *Wolbachia* endosymbiont of *Aphis glycines* (taxid 1078297) | 1,543,368 | 1 |
| *Buchnera aphidicola* (*Aphis glycines*) (taxid 1265350) | 1,179,434 | 40 |
| *Aphis glycines* (taxid 307491) | 311,693 | 13 |
| *Microbacterium homini* s (taxid 162426) | 164,909 | 2 |
| *Microbacterium resistens* (taxid 156977) | 156,337 | 2 |
| *Microbacterium paludicola* (taxid 300019) | 126,278 | 1 |
| *Microbacterium terricola* (taxid 344163) | 90,222 | 2 |
| *Bacillus methanolicus* PB1 (taxid 997296) | 71,225 | 1 |
| *Escherichia coli* str+A2 K-12 substr. MG1655 (taxid 511145) | 68,712 | 2 |
| *Gigaspora margarita* (taxid 4874) | 40,186 | 1 |
| *Wolbachia* endosymbiont (group A) of *Icerya purchasi* (taxid 2954019) | 39,811 | 1 |
| *Microbacterium lemovicicum* (taxid 1072463) | 26,220 | 1 |
| *Bacillus yapensis* (taxid 2492960) | 13,141 | 1 |

| ***Pentalonia nigronervosa*** | | |
|---|---|---|
| **Organism** | **Length** | **Counts** |
| *Pentalonia nigronervosa* (taxid 693967) | 898,961 | 52 |
| *Buchnera aphidicola* (*Myzus persicae*) (taxid 98795) | 615,879 | 1 |
| *Buchnera aphidicola* (Therioaphis trifolii) (taxid 1241884) | 83,513 | 2 |
| *Buchnera aphidicola* (taxid 9) | 36,957 | 4 |
| *Acinetobacter soli* (taxid 487316) | 19,461 | 2 |
| *Buchnera aphidicola* (*Schizaphis graminum*) (taxid 98794) | 15,381 | 1 |
| *Escherichia coli* str. K-12 substr. MG1655 (taxid 511145) | 14,093 | 2 |

their contig lengths were less than their genome sizes (Table 4). It is important to note that the Kraken2 classification is limited to the database used and that the genomes not found in the database would not be classified by this method. A list of the taxa used in the custom database created for this analysis can be found in Supplementary Table 1.

## Discussion

Long-read sequencing of arthropods is often a daunting task confounded by the small size and variable biochemistry of insects. Full-length sequencing of their symbionts is more challenging in that they account for only a small proportion of the mass of a single insect and are rarely amenable to isolation and culturing. Significant strides have been made in adapting long-read sequencing to obtain the genomes of arthropods and their symbionts. Assemblers such as metaFlye are adept at grouping and assembling long reads from different species present in a complex sample, an inevitability when individuals are difficult or impossible to dissect. Furthermore, as the accuracy of long reads increases, the coverage needed to assemble a genome decreases, thus, less mass is needed to carry out sequencing. Despite these advances, working with small organisms to sequence their genome and their symbionts remains a challenge. Adaptive sampling strategies show promise for managing some of these difficulties. In this work, we tested adaptive sequencing using ONT technology. By selectively rejecting the host genome, we show that it is possible to generate contiguous assemblies of mitochondria, *Wolbachia*, and *Buchnera* and its plasmids without the need for other enrichment strategies. The subtractive approach used in

this work substantially reduces the effort and cost needed to acquire the genomes of insect symbionts and avoids many of the pitfalls that come with short-read methods or long-range PCR. While we were successful in enriching for mitochondria, *Wolbachia*, and *Buchnera* and its plasmids, we were also confounded by high levels of variability in flowcell performance, reductions in flowcell yield due to the burden of adaptive sampling, and environmental contaminants that accompanied the insect samples. For *A. glycines*, it was not possible to enrich for *Hamiltonella* and *Arsenophonus*, likely due to their low abundance in comparison to the levels of the obligate *Buchnera* symbiont and the opportunistic *Wolbachia*. While these low-abundance symbionts have been demonstrated to be present in other samples of the same species [40], reads mapping to them were not found. Future work will explore active enrichment that consists of positively selecting for low-abundance targets, as an alternative method to enrich for symbionts that exist in low concentration.

While further work is needed to optimize adaptive sampling for the symbionts of specific insects, we have demonstrated that it is effective for the ecologically important *Wolbachia* symbiont. As long reads improve in accuracy, drop in cost, and computational approaches such as adaptive sampling improve, the number of arthropods that will be sequenced with long reads will increase. Some initiatives such as the USDA pest initiative and the i5k consortium have already committed to generating 100% of their genomes using long-read methods. The higher accuracy and long reads possible with these technologies will result in more accurate, complete, and high-quality chromosome-level assemblies, opening a new era in arthropod genome research and their symbionts.

## Authors contribution

SG drafted the paper, organized sequencing, and performed analysis. RG drafted the paper and provided the samples. SME, RW, and SM carried out the laboratory procedures. JB, RD, and AZ carried out the analysis. PV provided nematode genome data. FSA collected banana aphid samples and IZ collected nematode samples. All authors reviewed the paper.

## Data Availability

Raw data are uploaded to SRE but are temporarily embargoed. Referenced unpublished assemblies will be available after publication of a secondary manuscript. Please contact the authors for more information.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

The authors thank Glen Hartman, Theresa Herman, Steve Clough, and Adam Mahan for providing the soybean aphid samples. We also thank Churamani Khanal and Hehe Wang for screening candidate nematode species for *Wolbachia*. Our thanks also to Susan Halbert for identifying a location where the banana aphid could be collected. This is publication #1658 from the Institute of Environment at Florida International University. This work is funded in part by National Institutes of Health, USA Grant 5R50CA243890 to SG and United States Department of Agriculture ARS, USA FIU#800015405 to RG and G. Hartman and National Institutes of Health Grants R01-HG006677 and R35-GM130151 to AZ (PI Steven Salzberg, JHU).

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.cois.2023.101135.

## References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Ødegaard F: **How many species of arthropods? Erwin's estimate revised**. *Biol J Linn Soc Lond* 2000, **71**:583-597.

2. Zug R, Hammerstein P: **Still a host of hosts for *Wolbachia*: analysis of recent data suggests that 40% of terrestrial arthropod species are infected**. *PLoS One* 2012, **7**:e38544.

3. Werren JH, Baldo L, Clark ME: ***Wolbachia*: master manipulators of invertebrate biology**. *Nat Rev Microbiol* 2008, **6**:741-751.

4. Stevens L, Giordano R, Fialho RF: **Male-killing, nematode infections, bacteriophage infection, and virulence of cytoplasmic bacteria in the genus *Wolbachia***. *Annu Rev Ecol Syst* 2001, **32**:519-545.

5. Woolfit M, Iturbe-Ormaetxe I, Brownlie JC, Walker T, Riegler M, Seleznev A, Popovici J, Rancès E, Wee BA, Pavlides J, *et al*.: **Genomic evolution of the pathogenic *Wolbachia* strain, wMelPop**. *Genome Biol Evol* 2013, **5**:2189-2204.

6. Hoffmann AA, Montgomery BL, Popovici J, Iturbe-Ormaetxe I, Johnson PH, Muzzi F, Greenfield M, Durkan M, Leong YS, Dong Y, *et al*.: **Successful establishment of *Wolbachia* in *Aedes* populations to suppress dengue transmission**. *Nature* 2011, **476**:454-457.

7. Aliota MT, Peinado SA, Velez ID, Osorio JE: **The wMel strain of *Wolbachia* reduces transmission of Zika virus by *Aedes aegypti***. *Sci Rep* 2016, **6**:28792.

8. Aliota MT, Walker EC, Uribe Yepes A, Velez ID, Christensen BM, Osorio JE: **The wMel strain of *Wolbachia* reduces transmission of chikungunya virus in *Aedes aegypti***. *PLoS Negl Trop Dis* 2016, **10**:e0004677.

9. • Ribeiro dos Santos G, Durovni B, Saraceni V, Souza Riback TI, Pinto SB, Anders KL, Moreira LA, Salje H: **Estimating the effect of the wMel release programme on the incidence of dengue and chikungunya in Rio de Janeiro, Brazil: a spatiotemporal modeling study**. *Lancet Infect Dis* 2022, **22**:1587-1595.
Introduction of Wolbachia bacteria *Aedes aegypti* mosquitos reduces the incidence of Dengue. Release program in Brazil aimed to determine the effect of these mosquitos on the incidence of dengue and chikungunya.

10. Utarini A, Indriani C, Ahmad RA, Tantowijoyo W, Arguni E, Ansari MR, Supriyati E, Wardana DS, Meitika Y, Ernesia I, *et al*.: **Efficacy of *Wolbachia*-infected mosquito deployments for the control of dengue**. *N Engl J Med* 2021, **384**:2177-2186.

11. Lenharo M: **Massive mosquito factory in Brazil aims to halt dengue**. *Nature* 2023, **616**:637-638.

12. Hotaling S, Kelley JL, Frandsen PB: **Toward a genome sequence for every animal: where are we now?** *Proc Natl Acad Sci USA* (52) 2021, **118**:e2109019118.

13. • Hotaling S, Sproul JS, Heckenhauer J, Powell A, Larracuente AM, Pauls SU, Kelley JL, Frandsen PB: **Long reads are revolutionizing 20 years of insect genome sequencing**. *Genome Biol Evol* (8) 2021, **13**:evab138.

An update on the current state of insect genomics and the impact of long read on the field.

14. i5K Consortium: **The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment**. *J Hered* 2013, **104**:595-600.

15. Childers AK, Geib SM, Sim SB, Poelchau MF, Coates BS, Simmonds TJ, Scully ED, Smith TPL, Childers CP, *et al*.: **The USDA-ARS Ag100Pest initiative: high-quality genome assemblies for agricultural pest arthropod research**. *Insects* (5461) 2021, **287**:2196-2204.

16. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, *et al*.: **A whole-genome assembly of *Drosophila***. *Science* 2000, **287**:2196-2204.

17. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al*.: **The genome sequence of *Drosophila melanogaster***. *Science* 2000, **287**:2185-2195.

18. Li F, Zhao X, Li M, He K, Huang C, Zhou Y, Li Z, Walters JR: **Insect genomes: progress and challenges**. *Insect Mol Biol* 2019, **28**:739-758.

19. Kingan SB, Heaton H, Cudini J, Lambert CC, Baybayan P, Galvin BD, Durbin R, Korlach J, Lawniczak MKN: **A high-quality de novo genome assembly from a single mosquito using PacBio sequencing**. *Genes* (1) 2019, **10**:62.

20. Raley, C., Munroe, D., Jones, K., Tsai, Y.C., Guo, Y., Tran, B., Gowda, S., Troyer, J.L., Soppet, D.R., Stewart, C. and Stephens, R., 2014. Preparation of next-generation DNA sequencing libraries from ultra-low amounts of input DNA: Application to single-molecule, real-time (SMRT) sequencing on the Pacific Biosciences RS II. *bioRxiv*, 2014, p.003566.

21. Delahaye C, Nicolas J: **Sequencing DNA with nanopores: troubles and biases**. *PLoS One* 2021, **16**:e0257521.

22. Cuber P, Chooneea D, Geeves C, Salatino S, Creedy TJ, Griffin C, Sivess L, Barnes I, Price B, Misra R: **Comparing the accuracy and efficiency of third generation sequencing technologies, oxford nanopore technologies, and Pacific Biosciences, for DNA barcode sequencing applications**. *Ecological Genetics and Genomics* 2023,100181.

23. •• Hotaling S, Wilcox ER, Heckenhauer J, Stewart RJ, Frandsen PB: **Highly accurate long reads are crucial for realizing the potential of biodiversity genomics**. *BMC Genom* 2023, **24**:117.

A comprehensive examination of the state of a diverse set of public genomes and the impact long reads on the quality of genome science.

24. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, *et al*.: **Nanopore sequencing and assembly of a human genome with ultra-long reads**. *Nat Biotechnol* 2018, **36**:338-345.

25. Mavingui P, Valiente Moro C, Tran-Van V, Wisniewski-Dyé F, Raquin V, Minard G, Tran F-H, Voronin D, Rouy Z, Bustos P, *et al*.: **Whole-genome sequence of *Wolbachia* strain wAlbB, an endosymbiont of tiger mosquito vector *Aedes albopictus***. *J Bacteriol* 2012, **194**:1840.

26. Leichty AR, Brisson D: **Selective whole genome amplification for resequencing target microbial species from complex natural samples**. *Genetics* 2014, **198**:473-481.

27. Hongoh Y, Toyoda A: **Whole-genome sequencing of unculturable bacterium using whole-genome amplification**. *Methods Mol Biol* 2011, **733**:25-33.

28. Martinez J, Ant TH, Murdochy SM, Tong L, da Silva Filipe A, Sinkins SP: **Genome sequencing and comparative analysis of *Wolbachia* strain wAlbA reveals *Wolbachia*-associated plasmids are common**. *PLoS Genet* 2022, **18**:e1010406.

29. •• Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M: **Readfish enables targeted nanopore sequencing of gigabase-sized genomes**. *Nat Biotechnol* 2021, **39**:442-450.

The seminal paper on the adaptive sampling approach described in this manuscript.

30. • Kipp EJ, Lindsey LL, Milstein MS, Blanco CM, Baker JP, Faulk C, Oliver JD, Larsen PA: **Nanopore adaptive sampling for targeted mitochondrial genome sequencing and bloodmeal identification in hematophagous insects**. *Parasit Vectors* 2023, **16**:68.

An examination of a similar method used to explore the mitochondrial genomes of mosquitos and their most recent bloodmeal.

31. Mathers TC, Mugford ST, Hogenhout SA, Tripathi L: **Genome sequence of the banana aphid, *Pentalonia nigronervosa* Coquerel (Hemiptera: Aphididae) and its symbionts**. *G3* 2020, **10**:4315-4321.

32. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL, *et al*.: **metaFlye: scalable long-read metagenome assembly using repeat graphs**. *Nat Methods* 2020, **17**:1103-1110.

33. Wood DE, Lu J, Langmead B: **Improved metagenomic analysis with Kraken 2**. *Genome Biol* 2019, **20**:257.

34. *medaka*: Sequence Correction Provided by ONT Research. Github; [date unknown].

35. Delcher AL, Salzberg SL, Phillippy AM: **Using MUMmer to identify similar regions in large sequence sets**. Current Protocols in Bioinformatics. Wiley; 2003.

36. Li H: **Minimap2: pairwise alignment for nucleotide sequences**. *Bioinformatics* 2018, **34**:3094-3100.

37. Manzano-Marín A: **No evidence for *Wolbachia* as a nutritional co-obligate endosymbiont in the aphid *Pentalonia nigronervosa***. *Microbiome* 2020, **8**:72.

38. Gil R, Sabater-Muñoz B, Perez-Brocal V, Silva FJ, Latorre A: **Plasmids in the aphid endosymbiont *Buchnera aphidicola* with the smallest genomes. A puzzling evolutionary story**. *Gene* 2006, **370**:17-25.

39. Lee YS, Park W: **Enhanced calcium carbonate-biofilm complex formation by alkali-generating *Lysinibacillus boronitolerans* YS11 and alkaliphilic *Bacillus* sp. AK13**. *AMB Express* 2019, **9**:49.

40. Wille BD, Hartman GL: **Two species of symbiotic bacteria present in the soybean aphid (Hemiptera: Aphididae)**. *Environ Entomol* 2009, **38**:110-115.