bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To B or not to B: *Arsenophonus* as a source of B-vitamins in whiteflies.

Diego Santos-Garcia[1,†,*], Ksenia Juravel[1], Shiri Freilich[2], Einat Zchori-Fein[3], Amparo Latorre[4,5], Andrés Moya[4,5], Shai Morin [1] and Francisco J. Silva[4,5,†,*]

[1] Department of Entomology, The Hebrew University of Jerusalem, Rehovot, Israel
[2] Institute of Plant Sciences, Newe-Ya'ar Research Center, ARO, Ramat-Yishai, Israel
[3] Department of Entomology, Newe-Ya'ar Research Center, ARO, Ramat-Yishai, Israel
[4] Institute for Integrative Systems Biology (I2SysBio), Universitat de València-CSIC, Spain
[5] Unidad Mixta de Investigación en Genómica y Salud (FISABIO/Universitat de València-I2SysBio), València, Spain

[†] Equal contribution.

[*] For correspondence: diego.santos@mail.huji.ac.il and francisco.silva@uv.es

1 Abstract

Insect lineages feeding on nutritionally restricted diets such as phloem, xylem, or blood, were able to diversify by acquiring bacterial species that complemented the missing nutrients. These bacteria, considered obligate/primary endosymbionts, share a long evolutionary history with their hosts. In some cases, however, these endosymbionts are not able to fulfill all the nutritional requirements of their host, driving the acquisition of additional symbiotic species. Whiteflies, which feed on phloem, established an obligate relationship with *Candidatus* Portiera aleyrodidarum, who provides essential amino acids and carotenoids to the host. As many Whiteflies species harbor additional endosymbionts, they could provide their hosts with missing nutrients. To test this hypothesis, genomes of several endosymbionts from the whiteflies *Aleurodicus dispersus*, *A. floccissimus* and *Trialeurodes vaporariorum* were sequenced and analyzed. All three species were found to harbor an endosymbiont from the genus *Arsenophonus*, and the two former also host *Wolbachia*. A comparative analysis of the three *Arsenophonus* genomes revealed that although all of them are capable of synthesizing B-vitamins and cofactors, such as pyridoxal, riboflavin, or folate, their genomes and phylogenetic relationship vary greatly. *Arsenophonus* of *A. floccissimus* and *T. vaporariorum* belong to the same clade and display characteristics of facultative endosymbionts, such as large genomes (3 Mb) with thousands of genes, many pseudogenes, intermediate GC content, and mobile genetic elements (MGEs). In contrast, *Arsenophonus* of *A. dispersus* belongs to a different lineage and displays the characteristics of a primary endosymbiont, such as a reduced genome (670 kb) with 400 genes, 32% GC content, and no MGEs. However, the presence of 274 pseudogenes suggests that this symbiotic association is more recent than other reported hemipteran's primary endosymbionts. *Arsenophonus* of *A. dispersus* gene repertoire is completely integrated in the symbiotic consortia, and the biosynthesis of most vitamins occurs in shared pathways with its host. In addition, *Wolbachia* have also retained the ability to produce riboflavin, FAD, and folate, and may have a nutritional contribution. Taken together, our results show that *Arsenophonus* have a pivotal place in whiteflies nutrition by their ability to produce B-vitamins, even if they diverge and/or go through a genome reduction process.

Keywords: whitefly, symbiosis, vitamins, genome reduction, *Arsenophonus*, *Wolbachia*, metabolic complementation, riboflavin

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2 Introduction

Diverse eukaryotes live together with bacterial symbionts. Mutualistic, commensal, or parasitic relationships have been described for these associations (Moya et al., 2008). The Class Insecta is an example of mutualistic associations and several lineages are known to live in intimate relationship with an obligate symbiont for millions of years. Such symbionts are usually harbored inside specialized cells, termed bacteriocytes, and are vertically transmitted from the mother to her offspring. These bacterial symbionts, living inside cells and transmitted for long evolutionary periods, has been denominated primary (or obligatory) endosymbionts. The evolution of these symbioses had several effects on the gene repertoires of both hosts and endosymbionts, but is most notedly characterized by a drastic number of gene reductions in the latter (Latorre and Manzano-Marín, 2016). Host diet complementation has frequently initiated the symbiosis, but in some cases, the host has nutritional requirements which the primary endosymbiont is unable to satisfy, and secondary (or facultative) endosymbiont are thus required for diet complementation. When these requirements became vital, it is possible that a secondary endosymbiont becomes primary, resulting in a consortium of primary endosymbionts.

Whiteflies and psyllids are two sister hemipteran lineages, which probably started their evolutionary success through the ancient association with ancestral bacterial species of the family Halomonadaceae (Santos-Garcia et al., 2014a). This bacterial lineage splits into *Candidatus* Portiera aleyrodidarum (hereafter, the term Candidatus will be used only at the first time a species is mentioned) in whiteflies and *Candidatus* Carsonella ruddii in psyllids. Based on gene repertoires, it was proposed that *Portiera* collaborates with its hosts in the synthesis of amino acids and carotenoids (Santos-Garcia et al., 2012, 2015; Sloan and Moran, 2012). Because whiteflies feed only on the phloem sieve-element, their diet is also deficient in vitamins and cofactors. Individual whiteflies may carry one or several additional endosymbionts belonging to seven genera (Zchori-Fein et al., 2014). It has been proposed that some of these additional endosymbionts are likely involved in vitamins and cofactors supply to their host, as for example, *Candidatus* Hamiltonella defensa (Rao et al., 2015). *Candidatus* Arsenophonus is a genus of secondary endosymbiont that has been observed in many whitefly species (Thao and Baumann, 2004; Zchori-Fein et al., 2014; Santos-Garcia et al., 2015). Similar to *Hamiltonella*, it is only found inside bacteriocytes, while other endosymbionts like *Candidatus* Wolbachia sp. or *Candiddatus* Rickettsia sp. present different tissue tropisms (Gottlieb et al., 2008; Skaljac et al., 2010, 2013; Marubayashi et al., 2014). *Arsenophonus* lineages have also evolved intimate associations with other insect taxa with different diets, such as blood-sucking insects, where the supplementation of B-vitamins and cofactors were their proposed role (Šochová et al., 2017).

Whiteflies (Sternorrhyncha:Aleyrodidae) are composed mainly by two subfamilies, the Aleurodicinae and the Aleyrodinae. These two subfamilies probably originated between the Jurassic (Drohojowska and Szwedo, 2015; Shcherbakov, 2000) and the Middle Cretaceous (Campbell et al., 1994). While the Aleyrodinae contains the largest number of species described so far (140 genera, 1440 species), the Aleurodicinae has a relatively low number of described species (17 genera, 120 species) (Ouvrard and Martin, 2018). Analysis of the genome content of *Portiera* from the whitefly species *Aleurodicus dispersus* and *A. floccissimus* (Aleurodicinae) and *Trialeurodes vaporariorum* (Aleyrodinae) suggested that they are unable to complement their host diet with essential cofactors and vitamins (Santos-Garcia et al., 2015). The work presented here was thus initiated in order to test the prediction that *Arsenophonus* and/or *Wolbachia*, the other symbionts present in these species, are capable of providing their hosts with the lacking cofactors and vitamins.

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3 Material and Methods

Samples preparation and sequencing

Metagenome-assembled genomes (MAGs) of bacteria other than *Portiera* were retrieved by shotgun sequencing of three whiteflies species: *A. dispersus*, *A. floccissimus*, and *T. vaporariorum* (Santos-Garcia et al., 2015). All the whiteflies were collected from field populations in Barcelona (*T. vaporariorum*) and Canary Islands (*A. dispersus* and *A. floccissimus*) (Spain). All three samples contained *Portiera* and *Arsenophonus*, while both *Aleurodicus* species also harbored *Wolbachia*. Genomic DNA was obtained from bacteriome enriched samples after a Whole Genome Amplification process (WGA; GenomiPhi V2, GE Healthcare). For each species, ten independent WGA reactions were performed and pooled. WGA DNAs were sequenced by Illumina HiSeq 2000 using a mate-paired library (100bp*2, 3kb insert size). A full description can be found at Santos-Garcia et al. (2015).

Metagenome-assembled genomes (MAGs)

For each species library, RAW reads were quality checked, and trimmed/clipped if necessary, with FastQC v0.11.3 (Andrews, 2010) and TrimmomaticPE v0.33 (Bolger et al., 2014). Kraken v0.10.6 (Wood and Salzberg, 2014) was used to classify the RAW Illumina reads using a custom genomic database including: *Portiera*, *Hamiltonella*, *Rickettsia*, *Wolbachia*, *Arsenophonus*, *Candidatus* Cardinium hertigii, several whiteflies' mitochondria, *Bemisia tabaci* MEAM1, and *Acyrthosiphon pisum* (Table S1). Reads classified as mitochondrial, insect or *Portiera* were discarded. The rest of reads were assembled with SPADES v3.11.0 (–meta –careful –mp) (Nurk et al., 2017). Obtained contigs were classified using Kraken and the custom database. Contigs belonging to known whiteflies endosymbionts were recovered and added to the custom database. Trimmed/clipped reads were re-classified and reads belonging to known whitefly endosymbionts were re-assembled alone with SPADES (–careful –mp) to produce MAGs contigs. Illumina reads were digitally normalized (khmer v1.1) before the re-assembly stage (Crusoe et al., 2015). MAGs contigs were scaffolded and gap-filled with SSPACE v3 (Boetzer et al., 2011) and Gapfiller v1.10 (Boetzer and Pirovano, 2012), respectively. Finally, a manual iterative mapping approach, using Illumina trimmed/clipped reads, was performed with Bowtie2 v2.2.6 (Langmead and Salzberg, 2012), MIRA v4.9.5 (mapping mode) (Chevreux, B., Wetter, T. and Suhai, 1999), and Gap4 (Staden et al., 2000) until no more contigs/reads were joined/recovered for each MAG. Finally, Bowtie2 and Pilon v.1.21 (–jumps) (Walker et al., 2014) with Illumina reads were used to correct MAGs contigs.

MAGs annotation

MAGs initial annotations were performed with prokka v1.12 (Seemann, 2014). Enzyme commission numbers were added with PRIAM March_2015 release (Claudel-Renard et al., 2003). Gene Ontology, PFAM, and InterPro terms were added with InterProScan v5.27-66 (Jones et al., 2014). Putative pseudogenes and their positions in the genome were detected with LAST using several bacterium relative proteomes as a query (Kielbasa et al., 2011). Insertion Sequences (IS) were predicted with ISsaga (Varani et al., 2011). *B. tabaci* MEAM1 genome annotations (GCA_001854935.1) were downloaded and refined with PRIAM and InterProScan. MAGs, their corresponding *Portiera* (Santos-Garcia et al., 2015) and *B. tabaci* metabolism comparisons were performed on PathwayTools v21.5 (Karp et al., 2002). Heatmap and clustering was performed on R (R Core Team, 2018) with ggplots2 (Wickham, 2009).

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Comparative genomics

OrthoMCL v2.0.9 was used to compute clusters of orthologous proteins (Li et al., 2003). Cluster of Orthologous Groups (COGs) terms were assigned using diamond v0.8.11.73 (July 2016 bacterial RefSeq database, Buchfink et al. (2015)) and MEGAN6 (Huson et al., 2016). Average Nucleotide Identity (ANI) and Average Amino Acid Identity (AAI) values were obtained with the Enveomics tools (Rodriguez-R and Konstantinidis, 2016). Synteny between MAGs was checked with Mummer v3 (Kurtz et al., 2004) and Mauve, using *Arsenophonus* ARAD as reference for contig re-ordering (Darling et al., 2010). GenoPlotR was used to plot Mauve results (Guy et al., 2010).

The presence of potential orthologs of *B. tabaci* genes in the other whiteflies species was assessed using the Illumina cleaned data and diamond, with a blastx search strategy (minimum alignment length of 25 bp and $1e^{-3}$ e-value) against the selected *B. tabaci* proteins (Table S2). Recovered reads were classified with Kraken, using the custom and the mini-kraken database, to discard bacterial reads. Finally, non-bacterial reads were used as query for a blastx search against the nr database (last access: March 2018) (Altschul et al., 1990) and their best hit classified with MEGAN6. Only reads assigned to the phylum "Arthropoda" were considered as genomic reads of the screened genes in the other whiteflies species.

Phylogenetics

General phylogenetic position of the newly sequenced *Arsenophonus* and *Wolbachia* was assessed using several 16S rRNA genes downloaded from the GenBank. Genes were aligned with ssu-align (default masking) (Nawrocki, 2009). IQ-TREE was used to select the best substitution model (TVM+F+R3 and TN+F+R2, respectively) and to infer the majority rule maximum likelihood (ML) tree, with its associated support values (Nguyen et al., 2015; Kalyaanamoorthy et al., 2017).

Phylogenomic trees were generated using 82 and 97 conserved proteins, selected using PhyloPhlAn v0.99 (Segata et al., 2013), from *Arsenophonus* and *Wolbachia* inferred proteomes, respectively. Protein files were sorted by species name, using fastasort from exonerate v2.2.0 (Slater and Birney, 2005) and aligned with MAFFT v7.215 (Katoh et al., 2002). Proteins alignments were concatenated by species index using Geneious version 11.1.2 (Kearse et al., 2012). IQ-TREE was used to infer majority rule ML tree, and associated support values, under the substitution models JTTDCMut+F+R3 for *Arsenophonus* and JTT+F+I+G4 for *Wolbachia*.

4 Results

Genomic features and comparative genomics

Five MAGs, two *Arsenophonus* and two *Wolbachia* from *A. dispersus* (ARAD, WBAD) and *A. floccissimus* (ARAF, WBAF) and one *Arsenophonus* from *T. vaporariorum* (ARTV) were recovered (Table 1). All MAGs were in a draft status but with different quality. Only ARAD, was assembled as a circular scaffold, supported by some mate-paired reads. However, a gap at the contig edges was not closed due to the presence of a *dnaX* duplication and a long AT low complexity region. The assembly of ARAF genome was also of high quality, with only 11 contigs and a N50 value higher than 500 kb. The rest of recovered MAGs were in different draft status, being the *Wolbachia* the most fragmented ones.

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1: Assembly statistics and genomic features of the five endosymbionts sequenced.

	ARAD	ARAF	ARTV	WBAD	WBAF
Species	*Arsenophonus* sp.	*Arsenophonus* sp.	*Arsenophonus* sp.	*Wolbachia* sp.	*Wolbachia* sp.
Host	*A. dispersus*	*A. floccissimus*	*T. vaporariorum*	*A. dispersus*	*A. floccissimus*
Genome size (bp)	663125	3001875	3080136	1244696	1239495
Contigs	1	11	93	237	422
N50 (bp)	-	510728	236293	9433	5991
Coverage	1698	2500	182	268	106
%GC	32	37	37	34	34
Genes	428	1880	2312	990	909
CDS	381	1781	2209	946	862
Noncoding RNA genes	47	99	103	44	47
Coding density†	54	41	51	70	58
Avg. CDS length	941	689	712	918	840
CDS %GC	36	39	39	34	34
1st+2nd %GC	42	42	42	38	38
3rd %GC	23	31	32	26	26
Pseudogenes	274	1213	946	346	638
Avg. pseudo length	461	1030	1011	586	527
Pseudo %GC	30	38	38	34	35
rRNA (16S|23S|5S)	1|1|3	1|1|4	1|0|9*	1|1|1	1|1|1
tRNA	34	43	43	34	32
tmRNA	1	1	1	1	1
other RNAs	7	49	49	6	11
Insertion Sequences (IS)	0	11	21*	5*	2*
Prophage regions	0	12	12	2	0
Prophage regions size (bp)	-	430067	336650	9948	-
Complete Prophage	-	7	7	0	-
Incomplete Prophage	-	5	5	2	-

*rRNA genes and ISs could be underestimated due to the draft status of the assembly

†Total number of nucleotides in functional CDS/total number of nucleotides in the genome

According to their genomic features, the *Arsenophonus* sequenced can be divided into two groups. While ARAD displays typical characteristics of primary endosymbionts, ARAF and ARTV display genomic features closer to facultative or secondary endosymbionts (Table 1). This pattern is emphasized by the different genome sizes (0.67 *vs.* 3 Mb), their GC content (32 *vs.* 37%), the number of genes (428 *vs.* around 2000), and the presence of insertion sequences (ISs) and prophages (0 *vs.* 12). However, they have in common the low coding density (range 41 to 54%) due to the presence of almost as many pseudogenes as genes. Most of the prophages in ARAF and ARTV were unrelated (based on their ANI values) except the pairs of prophages 2/8 and 3/7 from ARAF and the pair of prophage 9 from ARAF and prophage 7 from ARTV (Figure S1, Figure S2 and Table S3).

Regarding *Wolbachia*, both strains display genome sizes and genomic features similar to other sequenced strains from this genus (Ellegaard et al., 2013), with WBAF being more fragmented and lacking two tRNAs. Two incomplete prophages were detected in the genome of WBAD (Figure S1).

Phylogenetic placement

The phylogenetic position of each sequenced *Arsenophonus* within the genus was tested using the 16S rRNA phylogeny, phylogenomics and ANI/AAI analyses. Based on 16S rRNA phylogenetic analysis, ARAD belongs to a different lineage than ARAF and ARTV (Figure S3). While ARAD was placed in a not well supported cluster, including *A. nasoniae* and *Arsenophonus* from other insects (the aphid *Stomaphis* and the dipteran *Ornithomya*), ARAF and ARTV clustered with endosymbionts of several whitefly species and other insects. The use of a species

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

threshold of 95% ANI (Konstantinidis and Tiedje, 2005) suggests that *A. triatominarum*, *A. nilaparvata*, *A.* of *Entylia carinata*, ARAF and ARTV belong to the same species, whereas *A. nasoniae*, *A. lipopteni* and ARAD form three different species (Table S3, Figure S4). The AAI analysis showed the presence of two clusters, one for ARAD and *A. nasoniae* and the other for the remaining species, except the endosymbiont *A. liptoteni*, with a value smaller than 80% AAI. Phylogenomics results show a partial congruence with ANI/AAI, placing *A. triatominarum* as the closest relative of ARAD (Figure 1A). The two endosymbionts form a sister monophyletic clade to all the others *Arsenophonus*, which form a second monophyletic clade. *A. nasoniae* was placed at the basal position of both clades.

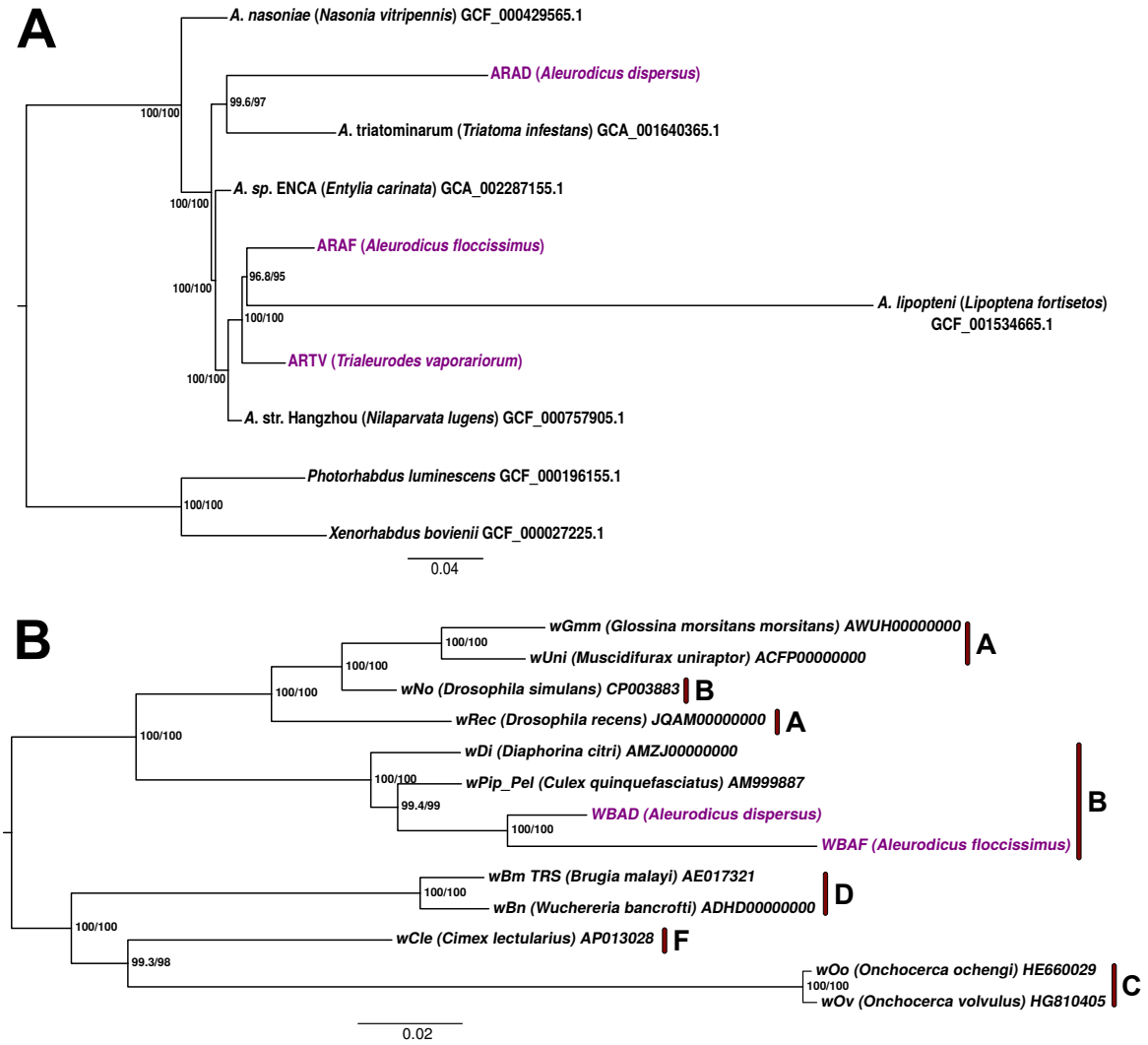

Figure 1: **Phylogenomic trees of Arsenophonus and Wolbachia.** (**A**) *Arsenophonus* (rooted) and (**B**) *Wolbachia* (unrooted) ML trees were inferred using 82 and 97 concatenated conserved protein alignments under a JTTDCMut+F+R3 and JTT+F+I+G4 substitution model respectively. Support values were obtained with 1000 ultrafast bootstraps (right node labels) and 5000 SH-aLRT (left node labels). Names for the eukaryotic hosts are shown after the strain names inside parentheses. Accession numbers are displayed for each of the proteomes used. *Arsenophonus* ARAD, ARAF, and ARTV and *Wolbachia* WBAD and WBAF are highlighted in purple. (B) Red bars and letters denote the different *Wolbachia* supergroups.

The phylogenetic positions of the two *Wolbachia* were tested with the same analyses. In the 16S rRNA phylogeny, WBAD (*A. dispersus*) and WBAF (*A. floccissimus*) form a well-supported

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

clade inside a major clade, which contains *Wolbachia* strains from other whiteflies and other insect taxa (Figure S5). In addition, based on phylogenomics, WBAD and WBAF can be placed inside the *Wolbachia* B super group, very close to *Wolbachia* from *Culex quinquefasciatus* (Figure 1B). Finally, the high level of nucleotide identity (97.6% ANI value) suggests that both *Wolbachia*, WBAD and WBAF, are strains from the same species.

Comparative genomics: synteny and functional categories

When the inferred proteomes from *Arsenophonus* ARAD, ARAF and ARTV were compared (Table S4), a core genome composed of only 289 clusters was obtained (Figure 2B). Interestingly, the core genome includes many clusters related to vitamins and cofactors biosynthetic pathways. The small number of shared protein clusters results from the reduced proteome of ARAD, which is mainly a subset of the larger ARTV and ARAF proteomes. Only 4 out of the 10 ARAD specific clusters were not hypothetical proteins: the pyruvate kinase II (pyruvate kinase I is present in the three proteomes), the 3-oxoacyl-[acyl-carrier-protein] synthase (*fabF*), which is involved in fatty acid and biotin biosynthesis, and proteins encoded by the genes *recA* and *nudE*. In case only ARAF and ARTV were compared, the core proteome would have contained more than 1,000 clusters. The number of species specific clusters of ARTV (997) and ARAF (609) were higher compared to ARAD, which is in accordance with their genome sizes. Synteny analysis highlights the strong genome reduction underwent in ARAD compared to ARAF and ARTV (Figure 2A). This includes the loosing of macrosynteny (general genome architecture), while maintaining microsynteny (e.g. operons). In contrast, ARTV and ARAD still show high level of macrosynteny, although some rearrangements were observed (Figure 2C).

Wolbachia ARAD and ARAF showed a core genome of 686 clusters, and 253 and 169 species specific clusters, respectively (Figure 3A, Table S5). It should be noted, however, that the highly fragmented status of these genomes could produce a large number of artificial specific clusters. Comparison of the two Wolbachia genomes revealed that some of the largest contigs show the same gene order and a high level of nucleotide identity (Figure 3B), suggesting that both macrosynteny and microsynteny are probably kept.

The proteomes of the five endosymbionts were functionally classified according to the Cluster of Orthologous Groups (COGs). Because of the strong reduction of the gene repertoire in *Arsenophonus* ARAD, this endosymbiont shows a smaller number of hits in each functional category, except J (translation and ribosomal structure and biogenesis), where the number of hits was similar to the other endosymbionts (Figure 4 left). In fact, in ARAD, this category has a relative frequency in the proteome higher than 0.20 4 right). In addition, category J and the other categories related to Information storage and processing (A, K, and L) showed higher relative frequencies in ARAD than in the other two *Arsenophonus* (Figure 4 right). Based on the number of hits and relative frequencies, the loss of the gene repertoire in categories G (carbohydrate metabolism), I (lipid metabolism), and O (post-translational modification and chaperones) was lower in ARAD. As expected, *Wolbachia* proteomes show the retention of the informational categories but also retain COGs related, among others, to environmental response.

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

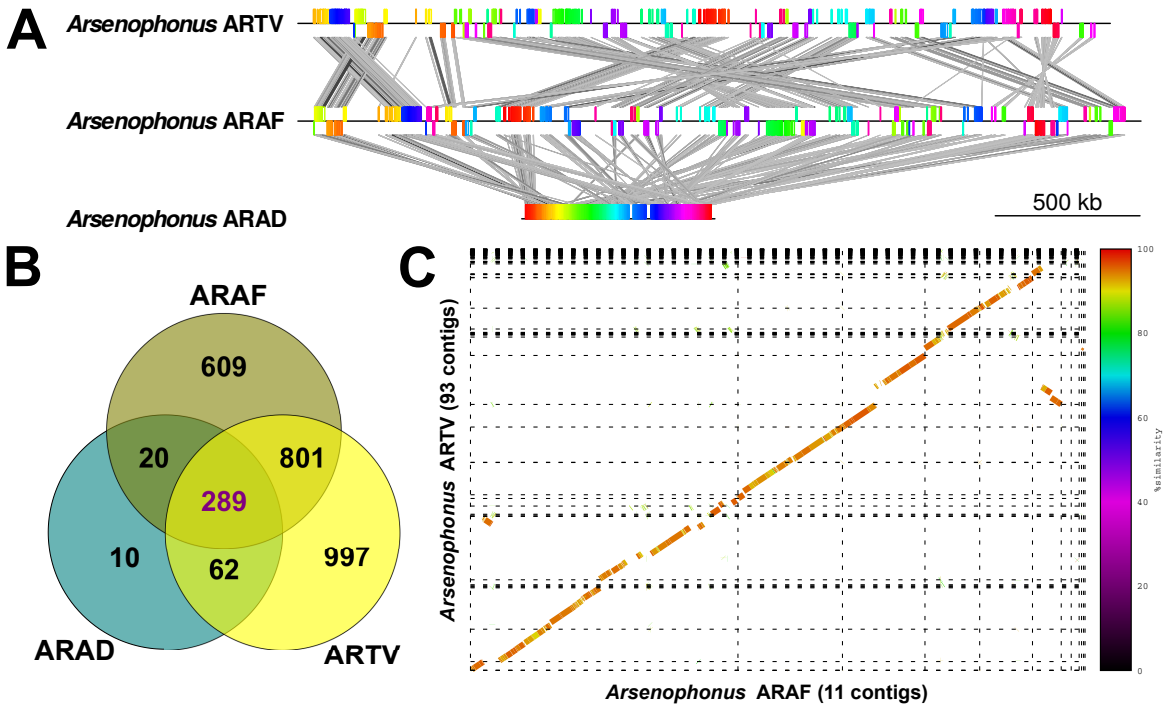

Figure 2: **Comparative analyses of the three *Arsenophonus* MAGs.** (**A**) Graphic linear representation of the genome rearrangements observed comparing ARTV with ARAF and ARAF with ARAD. Synteny blocks display the same color. ARAF and ARTV contigs were ordered according to ARAD before synteny blocks calculations with Mauve for a better visualization. (**B**) Venn diagram displaying the number of clusters found on each subspace of the *Arsenophonus* pangenome (Table S4). (**C**) Synteny between the draft genomes of ARTV and ARAF. Contig edges are displayed as dashed lines. ARTV contigs were ordered according to ARAF before synteny blocks calculations with Mauve for a better visualization.

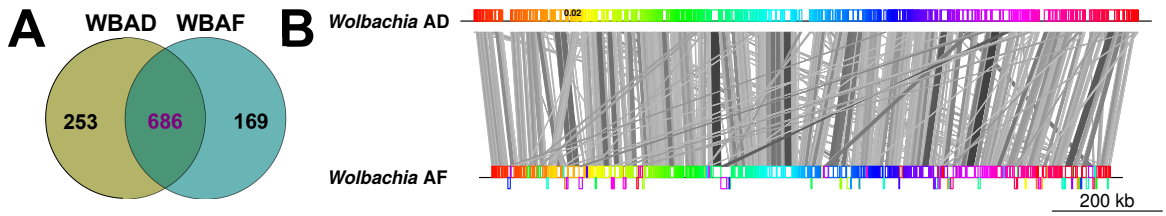

Figure 3: **Comparative analyses of the two *Wolbachia* MAGs.** (**A**) Venn diagram showing the shared and specific coding gene clusters between WBAD and WBAF (Table S5). (**B**) Synteny between the draft genomes (see Table 1) WBAD and WBAF. Contigs of WBAF were ordered according to WBAD before synteny blocks calculations with Mauve for a better visualization.

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

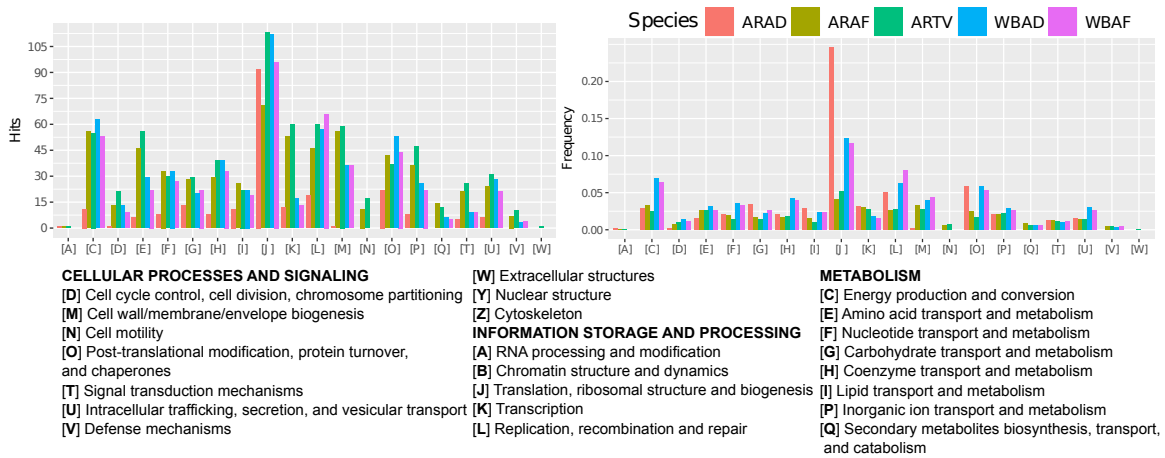

Figure 4: **Distribution of the endosymbiont proteomes in functional categories.** Bar plots showing the number of hits (right) and its relative frequency (left) in each functional category (COG) in the proteomes of the three analyzed *Arsenophonus* and the two analyzed *Wolbachia* MAGs.

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Biosynthetic metabolic potential

An integrated analysis of the metabolism of the whitefly bacteriocytes of the three species was performed in order to predict the potential biosynthetic capabilities of the harbored *Arsenophonus*, *Wolbachia* and *Portiera* (Santos-Garcia et al., 2015). That of the insect host was inferred using the *B. tabaci* genome (GCA_001854935.1). Amino acid biosynthesis in the three whiteflies is mainly conducted by the *Portiera* strains (PAAF, PAAD, and PATV), which maintain the ability to produce four (PATV)/five (PAAF and PAAD) out of ten essential amino acids, and require some complementation from the host for the synthesis of the others six (Figure 5A, cluster A). While some cofactors/vitamins and precursors are produced by almost all the endosymbionts and the host (cluster B), others are mainly produced by some *Arsenophonus* (cluster C), both *Arsenophonus* ARAF and ARTV, *Wolbachia* and the host (cluster D), or ,apparently, by none of them (cluster E). In general, both *Wolbachia* present a biosynthetic potential that is lower than that of the most reduced *Arsenophonus*, ARAD. In addition, while both *Wolbachia* still present all the electron transport chain, *Arsenophonus* ARAF and ARTV have lost the Ubiquinol oxidase and ARAD only maintains the Cytochrome-c oxidase and has lost the ATP synthase (not shown).

Although the essential amino acid biosynthetic potential of *Arsenophonus* and *Wolbachia* is limited (Figure 5A), ARAF and ARTV still retain some genes involved in the synthesis of amino acids or their precursors, such as lysine (its precursor, meso-diaminopimelate is also involved in the peptidoglycan biosynthesis) and chorismate (the precursor of phenylalanine). In ARTV, the complete pathway, encoded by seven genes, from D-erythrose 4-phosphate and D-ribulose-5-phosphate to chorismate is functional (*aroF*, *aroB*, *aroQ*, *aroE*, *aroL*, *aroA*, and *aroC*). This, in combination with the presence of *tyrA*, *tyrB*, and an additional monofunctional chorismate mutase gene, suggests the ability of ARTV to produce tyrosine and a potential collaboration with *Portiera*, or the host, for the production of phenylalanine. The ability to synthetize amino acid was completely lost in ARAD, suggesting this not to be the reason why the lineage of ARAD evolved a close association with *A. dispersus*.

The contribution of the different partners to the synthesis of vitamins and cofactors is heterogeneous. ARTV and ARAF, the two *Arsenophonus* with larger genomes, are those maintaining higher capabilities, with the loss of several genes required for some of the pathways. Despite the reductive evolutionary process in ARAD that has drastically reduced its potentialities, some capabilities are still retained, suggesting that they might be important in the symbiotic relation (Figure 5A and 5B). ARAD, but also ARAF and ARTV, encode the complete pathway to synthesize pyridoxal 5'-phosphate (vitamin B6) and riboflavin (vitamin B2). However, while ARAF and ARTV are able to produce FMN and FAD because they harbor a functional *ribF* gene, it is pseudogenized in ARAD. Since this gene is present in all whitefly hosts, both flavin cofactors may be produced in *A. dispersus* through a complementation between ARAD and its host (Figure 5B, Table S2). Also, being complemented again by the host, ARAD could synthesize thiamine (vitamin B1) and folate (vitamin B9, from imported chorismate) (Figure 5B). Although ARAD has lost *thiH* gene, some host encoded proteins can replace its function (Table S2). ARAF and ARTV also require complementation by their host in order to produce folate. ARTV lost most of the pathway to produce thiamine, and the missing steps are not complemented by its host. Additionally, *B. tabaci* seems to be able to produce biotin using some horizontally transferred genes (*bioA*, *bioB*, and *bioD*) (Luan et al., 2015), a protein capable of replacing the BioC function, and the insect's own fatty acid biosynthesis pathway. Genes with similar function to *bioABCDFH* genes were found to be present in *T. vaporariorum* genomic reads, suggesting that the whitefly is also capable of producing biotin by using only host genes. Because only genes similar to *bioBCFH* were identified in the genomic reads from both

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Aleurodicus species (Table S2), the presence of *bioA* and *bioD* in ARAD and ARAF is required for the cooperative synthesis of biotin. However, ARAD, unlike ARAF, has only retained the two missing genes in the genomes of both *Aleurodicus*.

In ARAF, the gene *yigB* which is part of the riboflavin pathway, is pseudogenized. This gene encodes for a Haloacid Dehalogenase (HAD) phosphatase, that belongs to a superfamily of enzymes. Therefore, the activity of the protein encoded by the gene, can be substituted by related genes in the endosymbiont or the host genomes (Manzano-Marín et al., 2015). Finally, all *Arsenophonus* conserve the glycolysis and the pentose phosphate pathways to produce the precursors required for riboflavin, thiamine, and pyridoxal. Indeed, in ARAD, the enzyme encoded by the *gatA* gene from the glycolysis pathway seems to have replaced the *epd* product in the pyridoxal pathway. The combination of *ilvC* encoded enzyme in *Portiera* (Price and Wilson, 2014) and the *panBC* genes that were horizontally transferred to *B. tabaci*, allows this whitefly to produce its own pantothenate (vitamin B5) and Coenzyme A. However, *panBC* genes are likely not present in the genomes of *A dispersus*, *A. floccissimus*, and *T. vaporariorum*, where only bacterial hits were recovered showing no similarity to the ones present in the *B. tabaci* genome. This suggests that these whiteflies are not able to produce pantothenate and they acquire it from the diet or other sources (Figure 5B and Table S2). However, failure on horizontal transferred genes detection could be an artifact from the methods used, since they are unable to discover new horizontally transferred genes in the whiteflies screened.

The third endosymbiont of *A. dispersus* and *A. floccissimus*, WBAD and WBAF, may be also contributing to the vitamin/cofactor synthesis of the system, as they have the complete pathway for riboflavin-FMN-FAD synthesis and potentially are able to produce folate from some intermediate precursors (Figure 5B).

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

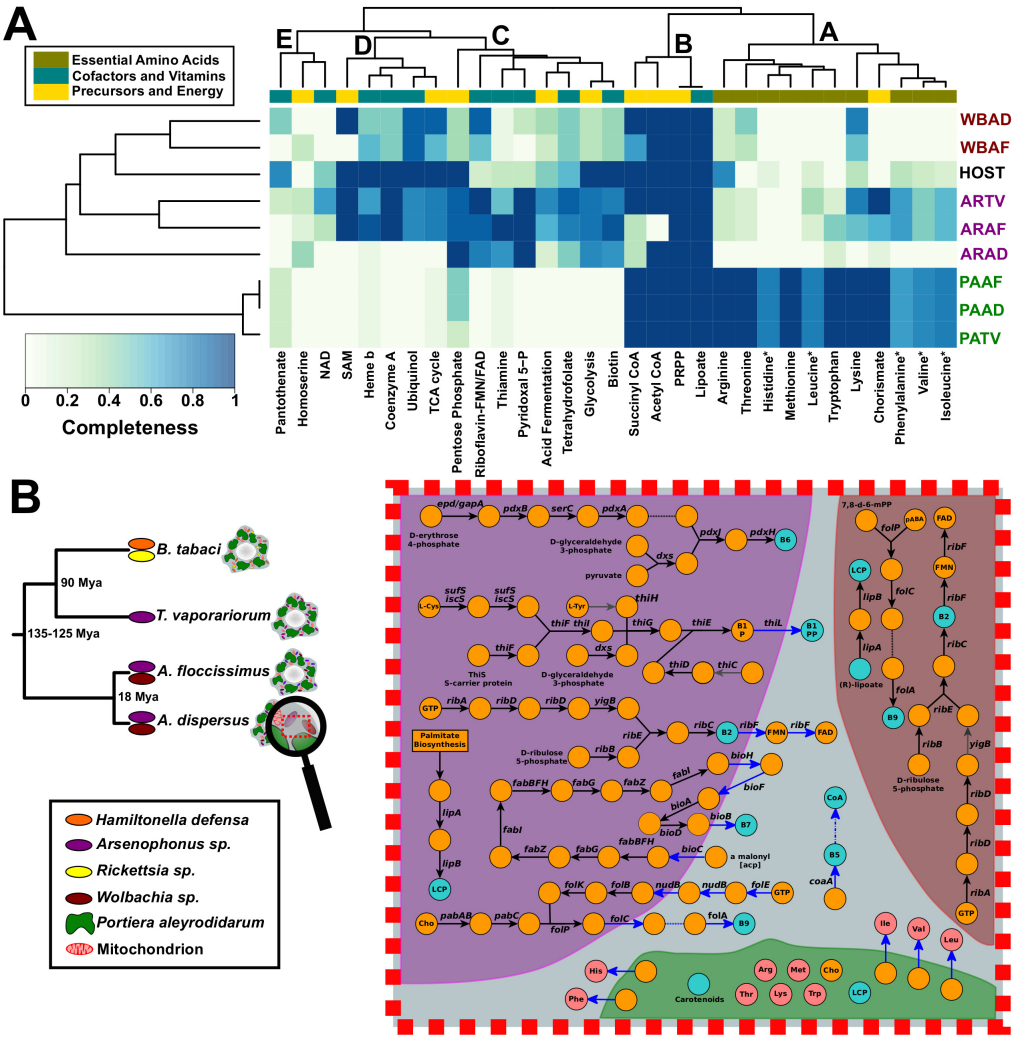

Figure 5: **Biosynthetic potential of *Arsenophonus*, *Wolbachia*, *Portiera* and white-flies.** (**A**) Heat map and hierarchical clustering showing in a color scale the degrees of completeness of the biosynthetic pathways of for several compounds, including essential amino acids, cofactors and vitamins biosynthetic pathways. Completeness was measured as number of enzymes present in the analyzed genome over the total number of enzymes in the pathway. *Arsenophonus* (AR), *Wolbachia* (WB) and *Portiera* (PA) from *A. dispersus* (AD), *A. floccissimus* (AF) and *T. vaporariorum* (TV) are highlighted in purple, red and green respectively. *B. tabaci* MEAM1 genome was used as a general host representative. (**B**) Phylogenetic relationship and endosymbionts content of the *Aleurodicus*, *T. vaporariorum* and *B. tabaci* whiteflies (left) (Santos-Garcia et al., 2015) and schematic representation of some vitamins and precursors metabolic pathways occurring at the bacteriocyte of *A. dispersus* (right). On the right, *Arsenophonus* ARAD is represented in purple, *Wolbachia* WBAD in redwood, *Portiera* in green and the host's cytosol in blue. Black arrows denote enzymes present in the corresponding endosymbiont, blue arrows enzymes present in the host, grey arrows denote pseudo/absent genes, dashed arrows denote simplified pathways and dashed lines spontaneous reactions. Abbreviations: lipoyl carrier protein (LCP), chorismate (Cho), acyl carrier protein (acp), 4-aminobenzoate (pABA), (7,8-dihydropterin-6-yl)methyl diphosphate (7,8-d-6-mPP), diphosphate (PP).

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

5 Discussion

Some bacterial lineages show, predominantly, facultative symbiotic lifestyles associated with insects. In such cases, they can inhabit the cytoplasm of bacteriocytes and coexist with the primary endosymbiont. These bacterial lineages usually have large gene repertoires and genomes, allowing them to be autonomous and capable of infecting individuals from very different taxa. Examples of such clades are the genera *Arsenophonus*, *Sodalis*, and *Serratia* (Nováková et al., 2009; Manzano-Marín et al., 2017; Santos-Garcia et al., 2017). Because their presence may bare fitness costs to the host, they do not become fixed in the insect lineage except if their nutritional contribution (or other kind of benefit) is important and is required for periods of many generations. Under such a scenario, a facultative endosymbiont may evolve towards a more intimate association, becoming a co-primary symbiont (Lamelas et al., 2011). Although there are several features associated with this new lifestyle, the most relevant is the decrease in the genome size. In the lineage *Arsenophonus-Riesia*, the sizes of several genomes have been reported, ranging from the 0.58 Mb of *Candidatus* Riesia pediculicola (Kirkness et al., 2010) and 0.84 Mb of *A. lipopteni* (Nováková et al., 2016) to 3.86 Mb of *A. triatominarum*, but with several genomes of intermediate sizes (Šochová et al., 2017). Similarly, *Sodalis* endosymbionts display also a large range of genome sizes, from the 0.35 Mb of *Candidatus* Mikella endobia (Husnik and McCutcheon, 2016) to the *circa* 4.5 Mb of *S. glossinidius* or *S. pierantonius*, and with several genome sizes in the range of 1-2 Mb (Santos-Garcia et al., 2017). This range has also been observed in *Serratia symbiotica*, with genome sizes from 0.65 to 3.86 Mb (Manzano-Marín et al., 2016).

Previous findings by Nováková et al. (2009) placed *Arsenophonus* from *A. dispersus* and *A. duguesii* (a close relative), together with an *Arsenophonus* from *B. tabaci*, in a clade with long branches, which is apart from other *Arsenophonus* of whiteflies. Our phylogenomics confirms these results, and have revealed that ARAD belongs to a different clade from the other two whiteflies' *Arsenophonus*, ARAF and ARTV. Two scenarios can explain the fact that ARAD and ARAF belong to different clades. In the first, the ancestor of ARAD infected the lineage of *A. dispersus* after its divergence from *A. floccissimus* and later started its shift to a primary symbiosis and its associated genome reduction process. In this case, the process could have taken place for as long as 18 My, the estimated divergence time between these two whitefly species (Santos-Garcia et al., 2015). An alternative scenario is that this association began prior to the divergence of the two *Aleurodicus*, but later, ARAD-type symbiont was replaced in the *A. floccissimus* lineage by another *Arsenophonus* lineage (ARAF). Although both scenarios are possible, the large number of pseudogenes in the ARAD genome suggests that the initiation of the process was relatively recent, as suggested by the first scenario. Still, the second scenario cannot be excluded as symbiont replacement has been documented in many insect lineages (Sudakaran et al., 2017).

As far as we know, *Arsenophonus* ARAD is among the smallest genomes in its genus, between *R. pediculicola* and *A. lipopteni* (Šochová et al., 2017). ARAD is still under a genome reduction process that could end up below 0.4 Mb if the current 54 % coding density is considered. ARAD gene repertoire is clearly a subset of other whiteflies' *Arsenophonus*. It is totally dependent on the host' environment, and putatively supplies/complements its host with cofactors/vitamins not produced by *Portiera* and the host. In conclusion, ARAD can be considered a co-primary endosymbiont. On the other hand, ARAF and ARTV are still at the beginning of the genome reductive process (e.g. high number of pseudogenes, inactivation of mobile genetic elements, and virulence/secretion factors). They are heritable facultative endosymbionts with the potential to establish a co-primary symbiotic relationship with *Portiera* and their host. However, as frequently observed, all *Arsenophonus*, independently of their symbiotic status, can be replaced

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

by other bacterium of the same, or different genera, as long as the new comers have similar biosynthetic capabilities (Thao and Baumann, 2004; Russell et al., 2017; Sudakaran et al., 2017).

We hypothesize that the ability to synthesize riboflavin and other vitamins is probably the main reason for the presence of *Arsenophonus* endosymbionts in whiteflies and the evolution of ARAD towards a co-primary symbiont in *A. dispersus*. As indicated above, nutritionally poor diets such as blood and phloem sap, cannot provide insects with their dietary requirements, including several vitamins. For example, the essentiality of dietary riboflavin was demonstrated in aposymbiotic aphids but not in aphids harboring *Buchnera aphidicola*, which is able to produce riboflavin (Nakabachi and Ishikawa, 1999). Moreover, *B. aphidicola* of aphids from the subfamily Lachninae already lost the genes encoding the pathway in the Lachninae ancestor. This led to the establishment in this group of a new association with an additional bacterium, capable of producing riboflavin. In most lineages this is the co-primary *S. symbiotica*, a facultative endosymbiont in other aphid lineages (Manzano-Marín and Latorre, 2014). The analyses of the co-symbiosis in species of Lachninae revealed a complex system with *S. symbiotica* endosymbionts harboring a very small genome (in *Tuberolachnus salignus*), intermediate genome sizes (in some species of the genus *Cinara*), or replaced by other endosymbionts belonging to different genera, such as, *Sodalis*, *Erwinia*, or *Hamiltonella*, among others (Manzano-Marín et al., 2016, 2017; Meseguer et al., 2017). The essential role riboflavin biosynthesis might play in establishing symbiotic relationships has also been demonstrated in several blood-feeding arthropods harboring different endosymbionts, such as *A. lipopteni* and its host the biting fly *Lipoptena cervi* (Nováková et al., 2016), *Wolbachia* associated with the bedbug *Cimex lectularius* Moriyama et al. (2015), and *Coxiella*-like bacteria found in several ticks (Gottlieb et al., 2015).

Although, thiamine, pantothenate, coenzyme A, niacin, pyridoxal and biotin can be found in the phloem (free or bounded to transporter proteins) (Pirson and Zimmermann, 1975), their concentrations might be not sufficient for many phloem-feeding species. This could explain why the capability to produce some of these compounds has been kept in bacterial endosymbionts while it has been lost in others. However, the production of one or several B vitamins (thiamine, riboflavin, pyridoxal, folate, and biotin) and *de novo* lipoate by whiteflies' *Arsenophonus* seems to be the main contribution of these endosymbionts to their hosts, allowing them to feed on the limited diet present in the phloem (Dale et al., 2006; Nováková et al., 2015, 2016; Mao et al., 2017). Regarding biotin biosynthesis, the bacterial origin of *bioAB* and *bioD* was reported by Luan et al. (2015) and Ankrah et al. (2017). We show here that not all whiteflies seem to have the horizontally transferred *bioAD* genes, although the rest of the *bio* operon homologs, including the *bioB*, are contained. This suggest that whiteflies acquired, by several horizontal gene transfer events, the ability to produce biotin. This is expected to produce a gradual loss of the biotin pathway in secondary endosymbionts of the different whiteflies lineages, at least from the Aleyrodinae subfamily, as this group harbors the full *bio* operon. Indeed, *Candidatus* Cardinium hertigii infecting *B. tabaci* has recently lost its ability to produce biotin, raising the possibility that biotin is provided by the co-present *Hamiltonella* endosymbiont or by the host metabolic potential (Santos-Garcia et al., 2014b; Rao et al., 2015). As an alternative, this can also lead to intricate complementation patterns as shown in the *Arsenophonus* from *Aleurodicus*. It is important to notice that supplementation of vitamins was the role proposed for *Hamiltonella* in *B. tabaci* (Rao et al., 2015). In fact, the results reported here support a similar function for *Arsenophonus* and interestingly, these two endosymbionts are never found together in the same individual (Gottlieb et al., 2008; Skaljac et al., 2010, 2013; Marubayashi et al., 2014). It could be possible that the two endosymbionts might compete for the same host resources to produce the same compounds, undermining the unique benefit to the host if harbored together due to the cost associated of maintaining high endosymbionts loads (Ferrari and Vavre, 2011).

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Finally, and in contrast to *Arsenophonus*, *Wolbachia* only produces riboflavin and, probably, folate. It has been proven that the bedbug *C. lectularius* requires the riboflavin produced by *Wolbachia* for its proper development. Indeed, the riboflavin pathway is conserved among *Wolbachia* species infecting different insects, suggesting that it may provide a fitness benefit to the infected host (Moriyama et al., 2015). However, it is yet to be determined if *Wolbachia* can be considered as a general mutualistic symbiont (at the metabolic level) in whiteflies, like *Hamiltonella* or *Arsenophonus*. First, while *Hamiltonella* and *Arsenophonus* are restricted to bacteriocytes, *Wolbachia* is found in different tissues (Gottlieb et al., 2008; Skaljac et al., 2010, 2013; Marubayashi et al., 2014). While the first pattern is more characteristic of hemipteran mutualistic endosymbiosis, the second can be found in endosymbionts with a wide range of symbiotic interactions, including parasitism. Second, in most reported cases of whiteflies harboring *Wolbachia*, the individual insects usually harbor in addition *Arsenophonus* or *Hamiltonella* (Skaljac et al., 2010, 2013; Kapantaidaki et al., 2014; Marubayashi et al., 2014; Zchori-Fein et al., 2014). The few cases in which *Wolbachia* was found to be the unique secondary endosymbiont should be handled with caution, as they may simply result from failure of the general PCR primers to amplify other symbionts species/strains (Augustinos et al., 2011).

In conclusion, the loss of the genes encoding the enzymes for the synthesis of vitamins in the ancestral *Portiera*, many million years ago, likely generated the requirement of a co-symbiont in whiteflies. *Arsenophonus* species seem to be the most common co-symbiont. In the lineage of *A. dispersus*, the harbored *Arsenophonus* lineage presents a highly reduced genome content. This *Arsenophonus* lineage is in the process of becoming a co-primary symbiont which putatively supplies/complements its host with cofactors/vitamins that are not produced by its co-partner *Portiera*.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

DSG and FJS conceived the study. DSG and FJS performed bioinformatics analysis. KJ performed phylogenetics analysis. All authors analyzed and discussed the data. DSG and FJS drafted the manuscript with inputs from AL, AM, EZF, SF and SM. All authors participated in the revision of the manuscript.

Funding

This work was supported by the grants PROMETEOII/2014/065 (Conselleria d'Educació, Generalitat Valenciana, Spain) to AM, BFU2015-64322-C2-1-R (co-financed by FEDER funds and Ministerio de Economía y Competitividad, Spain) to AL, and by the ISRAEL SCIENCE FOUNDATION (Israel) grants No. 1039/12 to SM and No. 484/17 to SF and EZF.

Acknowledgments

DSG is recipient of a Golda Meir Post Doctoral Fellowship at the Hebrew University of Jerusalem.

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplemental Data

The Supplementary Material for this article can be found online at: Supplementary Material

Data Availability Statement

The *Arsenophonus* (ERZ502704-6) and *Wolbachia* (ERZ502707-8) annotated genomes and the whole genome shotgun libraries (ERR2532344-45 and ERR2534068-71) analyzed for this study have been deposited at the European Nucleotide Archive (ENA) under the project number PRJEB26014. The generated Pathway Tools databases and the blastx results of the genomic reads similar to *Bemisia tabaci* key cofactors/vitamins metabolic genes can be found at 10.6084/m9.figshare.6142307.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* 215, 403–410. doi:10.1016/S0022-2836(05)80360-2

[Dataset] Andrews, S. (2010). Fastqc: A quality control tool for high throughput sequence data.

Ankrah, N. Y., Luan, J., and Douglas, A. E. (2017). Cooperative metabolism in a three-partner insect-bacterial symbiosis revealed by metabolic modeling. *J Bacteriol* 199, 1–15. doi:10.1128/JB.00872-16

Augustinos, A. A., Santos-Garcia, D., Dionyssopoulou, E., Moreira, M., Papapanagiotou, A., Scarvelakis, M., et al. (2011). Detection and characterization of *Wolbachia* infections in natural populations of aphids: is the hidden diversity fully unraveled? *PLoS One* 6, e28695. doi:10.1371/journal.pone.0028695

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27, 578–9. doi:10.1093/bioinformatics/btq683

Boetzer, M. and Pirovano, W. (2012). Toward almost closed genomes with GapFiller. *Genome Biol* 13, R56. doi:10.1186/gb-2012-13-6-r56

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12, 59–60. doi:10.1038/nmeth.3176

Campbell, B. C., Steffen-Campbell, J. D., and Gill, R. J. (1994). Evolutionary origin of whiteflies (Hemiptera: Sternorrhyncha: Aleyrodidae) inferred from 18S rDNA sequences. *Insect Mol Biol* 3, 73–88

Chevreux, B., Wetter, T. and Suhai, S. (1999). Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)* 99, 45–56

Claudel-Renard, C., Chevalet, C., Faraut, T., and Kahn, D. (2003). Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res* 31, 6633–6639. doi:10.1093/nar/gkg847

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Crusoe, M. R., Alameldin, H. F., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., et al. (2015). The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Res* 4, 900. doi:10.12688/f1000research.6924.1

Dale, C., Beeton, M., Harbison, C., Jones, T., Pontes, M., Al, D. E. T., et al. (2006). Isolation , Pure Culture , and Characterization of "*Candidatus* Arsenophonus arthropodicus", an Intracellular Secondary Endosymbiont from the Hippoboscid Louse Fly *Pseudolynchia canariensis*. *Society* 72, 2997–3004. doi:10.1128/AEM.72.4.2997

Darling, A. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5, e11147. doi:10.1371/journal.pone.0011147

Drohojowska, J. and Szwedo, J. (2015). Early Cretaceous Aleyrodidae (Hemiptera: Sternorrhyncha) from the Lebanese amber. *Cretac Res* 52, 368–389. doi:10.1016/j.cretres.2014.03.015

Ellegaard, K. M., Klasson, L., Näslund, K., Bourtzis, K., and Andersson, S. G. E. (2013). Comparative genomics of *Wolbachia* and the bacterial species concept. *PLoS Genet.* 9, e1003381. doi:10.1371/journal.pgen.1003381

Ferrari, J. and Vavre, F. (2011). Bacterial symbionts in insects or the story of communities affecting communities. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 366, 1389–1400. doi:10.1098/rstb.2010.0226

Gottlieb, Y., Ghanim, M., Gueguen, G., Kontsedalov, S., Vavre, F., Fleury, F., et al. (2008). Inherited intracellular ecosystem: symbiotic bacteria share bacteriocytes in whiteflies. *The FASEB Journal* 22, 2591–2599. doi:10.1096/fj.07-101162

Gottlieb, Y., Lalzar, I., and Klasson, L. (2015). Distinctive Genome Reduction Rates Revealed by Genomic Analyses of Two *Coxiella*-Like Endosymbionts in Ticks. *Genome Biol Evol* 7, 1779–1796. doi:10.1093/gbe/evv108

Guy, L., Kultima, J. R., and Andersson, S. G. E. (2010). genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* 26, 2334–5. doi:10.1093/bioinformatics/btq413

Husnik, F. and McCutcheon, J. P. (2016). Repeated replacement of an intrabacterial symbiont in the tripartite nested mealybug symbiosis. *Proc. Natl. Acad. Sci. U.S.A.* 113, E5416–E5424. doi:10.1073/pnas.1603910113

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., et al. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput Biol* 12, 1–12. doi:10.1371/journal.pcbi.1004957

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi:10.1093/bioinformatics/btu031

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14, 587–589. doi:10.1038/nmeth.4285

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Kapantaidaki, D. E., Ovčarenko, I., Fytrou, N., Knott, K. E., Bourtzis, K., and Tsagkarakou, A. (2014). Low Levels of Mitochondrial DNA and Symbiont Diversity in the Worldwide Agricultural Pest, the Greenhouse Whitefly *Trialeurodes vaporariorum* (Hemiptera: Aleyrodidae). *J Hered* , 1–13doi:10.1093/jhered/esu061

Karp, P. D., Paley, S., and Romero, P. (2002). The Pathway Tools software. *Bioinformatics* 18 Suppl 1, S225–32

Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30, 3059–66. doi:10.1093/nar/gkf436

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi:10.1093/bioinformatics/bts199

Kielbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Res* 21, 487–493. doi:10.1101/gr.113985.110

Kirkness, E. F., Haas, B. J., Sun, W., Braig, H. R., Perotti, M. A., Clark, J. M., et al. (2010). Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc. Natl. Acad. Sci. U.S.A.* 107, 12168–73. doi:10.1073/pnas.1003379107

Konstantinidis, K. T. and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 2567–72. doi:10.1073/pnas.0409727102

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biol* 5, R12. doi:10.1186/gb-2004-5-2-r12

Lamelas, A., Gosalbes, M. J., Manzano-Marín, A., Peretó, J., Moya, A., and Latorre, A. (2011). *Serratia symbiotica* from the aphid *Cinara cedri*: a missing link from facultative to obligate insect endosymbiont. *PLoS Genet* 7, e1002357. doi:10.1371/journal.pgen.1002357

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357–9. doi:10.1038/nmeth.1923

Latorre, A. and Manzano-Marín, A. (2016). Dissecting genome reduction and trait loss in insect endosymbionts. *Ann N Y Acad Sci* , 1–23doi:10.1111/nyas.13222

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13, 2178–89. doi:10.1101/gr.1224503

Luan, J.-B., Chen, W., Hasegawa, D. K., Simmons, A. M., Wintermantel, W. M., Ling, K.-S., et al. (2015). Metabolic Coevolution in the Bacterial Symbiosis of Whiteflies and Related Plant Sap-Feeding Insects. *Genome Biol Evol* 7, 2635–2647. doi:10.1093/gbe/evv170

Manzano-Marín, A. and Latorre, A. (2014). Settling Down: The Genome of *Serratia symbiotica* from the Aphid *Cinara tujafilina* Zooms in on the Process of Accommodation to a Cooperative Intracellular Life. *Genome Biol Evol* 6, 1683–98. doi:10.1093/gbe/evu133

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Manzano-Marín, A., Oceguera-Figueroa, A., Latorre, A., Jiménez-García, L. F., and Moya, A. (2015). Solving a Bloody Mess: B-Vitamin Independent Metabolic Convergence among Gammaproteobacterial Obligate Endosymbionts from Blood-Feeding Arthropods and the Leech *Haementeria officinalis*. *Genome Biol Evol* 7, 2871–2884. doi:10.1093/gbe/evv188

Manzano-Marín, A., Simon, J.-C., and Latorre, A. (2016). Reinventing the Wheel and Making It Round Again: Evolutionary Convergence in *Buchnera – Serratia* Symbiotic Consortia between the Distantly Related Lachninae Aphids *Tuberolachnus salignus* and *Cinara cedri*. *Genome Biol Evol* 8, 1440–1458. doi:10.1093/gbe/evw085

Manzano-Marín, A., Szabó, G., Simon, J.-C., Horn, M., and Latorre, A. (2017). Happens in the best of subfamilies: establishment and repeated replacements of co-obligate secondary endosymbionts within Lachninae aphids. *Environ Microbiol* 19, 393–408. doi:10.1111/1462-2920.13633

Mao, M., Yang, X., Poff, K., and Bennett, G. (2017). Comparative Genomics of the Dual-Obligate Symbionts from the Treehopper, *Entylia carinata* (Hemiptera: Membracidae), Provide Insight into the Origins and Evolution of an Ancient Symbiosis. *Genome Biol Evol* 9, 1803–1815. doi:10.1093/gbe/evx134

Marubayashi, J. M., Kliot, A., Yuki, V. A., Rezende, J. A. M., Krause-Sakate, R., Pavan, M. A., et al. (2014). Diversity and Localization of Bacterial Endosymbionts from Whitefly Species Collected in Brazil. *PLoS One* 9, e108363. doi:10.1371/journal.pone.0108363

Meseguer, A. S., Manzano-Marín, A., Coeur d'Acier, A., Clamens, A. L., Godefroid, M., and Jousselin, E. (2017). *Buchnera* has changed flatmate but the repeated replacement of co-obligate symbionts is not associated with the ecological expansions of their aphid hosts. *Mol Ecol* 26, 2363–2378. doi:10.1111/mec.13910

Moriyama, M., Nikoh, N., Hosokawa, T., and Fukatsu, T. (2015). Riboflavin Provisioning Underlies *Wolbachia*'s Fitness Contribution to Its Insect Host. *MBio* 6, e01732–15–e01732–15. doi:10.1128/mBio.01732-15

Moya, A., Peretó, J., Gil, R., and Latorre, A. (2008). Learning how to live together: genomic insights into prokaryote-animal symbioses. *Nat Rev Genet* 9, 218–29. doi:10.1038/nrg2319

Nakabachi, A. and Ishikawa, H. (1999). Provision of riboflavin to the host aphid, *Acyrthosiphon pisum*, by endosymbiotic bacteria, *Buchnera*. *J Insect Physiol* 45, 1–6. doi:10.1016/S0022-1910(98)00104-8

Nawrocki, E. P. (2009). *Structural RNA Homology Search and Alignment Using Covariance Models*. Electronic theses and dissertations. paper 256., Washington University School of Medicine

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* 32, 268–274. doi:10.1093/molbev/msu300

Nováková, E., Husník, F., Šochová, E., and Hypša, V. (2015). *Arsenophonus* and *Sodalis* symbionts in louse flies: An analogy to the *Wigglesworthia* and *Sodalis* system in tsetse flies. *Appl Environ Microbiol* 81, 6189–6199. doi:10.1128/AEM.01487-15

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Nováková, E., Hypsa, V., and Moran, N. a. (2009). *Arsenophonus*, an emerging clade of intracellular symbionts with a broad host distribution. *BMC Microbiol* 9, 143. doi:10.1186/1471-2180-9-143

Nováková, E., Hypša, V., Nguyen, P., Husník, F., and Darby, A. C. (2016). Genome sequence of *Candidatus* Arsenophonus lipopteni, the exclusive symbiont of a blood sucking fly *Lipoptena cervi* (Diptera: Hippoboscidae). *Stand Genomic Sci* 11, 72. doi:10.1186/s40793-016-0195-1

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 27, 824–834. doi:10.1101/gr.213959.116

[Dataset] Ouvrard, D. and Martin, J. H. (2018). The White-files - Taxonomic checklist of the world's whiteflies (Insecta: Hemiptera: Aleyrodidae).

Pirson, A. and Zimmermann, M. H. (1975). *Transport in Plants I*, vol. 1 (Berlin, Heidelberg: Springer Berlin Heidelberg). doi:10.1007/978-3-642-66161-7

Price, D. R. G. and Wilson, A. C. C. (2014). A substrate ambiguous enzyme facilitates genome reduction in an intracellular symbiont. *BMC Biol* 12, 110. doi:10.1186/s12915-014-0110-4

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria

Rao, Q., Rollat-Farnier, P.-A., Zhu, D.-T., Santos-Garcia, D., Silva, F. J., Moya, A., et al. (2015). Genome reduction and potential metabolic complementation of the dual endosymbionts in the whitefly *Bemisia tabaci*. *BMC Genomics* 16, 226. doi:10.1186/s12864-015-1379-6

Rodriguez-R, L. M. and Konstantinidis, K. T. (2016). The enveomics collection : a toolbox for specialized analyses of microbial genomes and metagenomes. *Peer J Preprints* 4, e1900v1. doi:10.7287/peerj.preprints.1900v1

Russell, J. A., Oliver, K. M., and Hansen, A. K. (2017). Band-aids for *Buchnera* and B vitamins for all. *Mol Ecol* 26, 2199–2203. doi:10.1111/mec.14047

Santos-Garcia, D., Farnier, P.-A., Beitia, F., Zchori-Fein, E., Vavre, F., Mouton, L., et al. (2012). Complete genome sequence of "*Candidatus* Portiera aleyrodidarum" BT-QVLC, an obligate symbiont that supplies amino acids and carotenoids to *Bemisia tabaci*. *J Bacteriol* 194, 6654–5. doi:10.1128/JB.01793-12

Santos-Garcia, D., Latorre, A., Moya, A., Gibbs, G., Hartung, V., Dettner, K., et al. (2014a). Small but Powerful, the Primary Endosymbiont of Moss Bugs, *Candidatus* Evansia muelleri, Holds a Reduced Genome with Large Biosynthetic Capabilities. *Genome Biol Evol* 6, 1875–1893. doi:10.1093/gbe/evu149

Santos-Garcia, D., Rollat-Farnier, P.-A., Beitia, F., Zchori-Fein, E., Vavre, F., Mouton, L., et al. (2014b). The Genome of *Cardinium* cBtQ1 Provides Insights into Genome Reduction, Symbiont Motility, and Its Settlement in *Bemisia tabaci*. *Genome Biol Evol* 6, 1013–1030. doi:10.1093/gbe/evu077

Santos-Garcia, D., Silva, F. J., Morin, S., Dettner, K., and Kuechler, S. M. (2017). The all-rounder *Sodalis* - a new bacteriome-associated endosymbiont of the lygaeoid bug *Henestaris halophilus* (Heteroptera: Henestarinae) and a critical examination of its evolution. *Genome Biol Evol* 9, 1–54. doi:10.1093/gbe/evx202

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Santos-Garcia, D., Vargas-Chavez, C., Moya, A., Latorre, A., and Silva, F. J. (2015). Genome Evolution in the Primary Endosymbiont of Whiteflies Sheds Light on Their Divergence. *Genome Biol Evol* 7, 873–888. doi:10.1093/gbe/evv038

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi:10.1093/bioinformatics/btu153

Segata, N., Börnigen, D., Morgan, X. C., and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* 4, 2304. doi:10.1038/ncomms3304

Shcherbakov, D. E. (2000). The most primitive whiteflies (Hemiptera; Aleyrodidae; Bernaeinae subfam. nov.) from the Mesozoic of Asia and Burmese amber, with an overview of Burmese amber hemipterans. *Bull. nat. Hist. Mus. Lond. (Geol.)* 56, 29–37

Skaljac, M., Zanic, K., Ban, S. G., Kontsedalov, S., and Ghanim, M. (2010). Co-infection and localization of secondary symbionts in two whitefly species. *BMC Microbiol* 10, 142. doi:10.1186/1471-2180-10-142

Skaljac, M., Zanić, K., Hrnčić, S., Radonjić, S., Perović, T., and Ghanim, M. (2013). Diversity and localization of bacterial symbionts in three whitefly species (Hemiptera: Aleyrodidae) from the east coast of the Adriatic Sea. *Bull Entomol Res* 103, 48–59. doi:10.1017/S0007485312000399

Slater, G. S. C. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6, 1–11. doi:10.1186/1471-2105-6-31

Sloan, D. B. and Moran, N. A. (2012). Endosymbiotic bacteria as a source of carotenoids in whiteflies. *Biol Lett* 8, 986–989. doi:10.1098/rsbl.2012.0664

Šochová, E., Husník, F., Nováková, E., Halajian, A., and Hypša, V. (2017). *Arsenophonus* and *Sodalis* replacements shape evolution of symbiosis in louse flies. *PeerJ* 5, e4099. doi:10.7717/peerj.4099

Staden, R., Beal, K. F., and Bonfield, J. K. (2000). The Staden package, 1998. *Methods Mol Biol* 132, 115–30

Sudakaran, S., Kost, C., and Kaltenpoth, M. (2017). Symbiont Acquisition and Replacement as a Source of Ecological Innovation. *Trends Microbiol* 24, 932–976. doi:10.1016/j.tim.2017.02.014

Thao, M. L. and Baumann, P. (2004). Evidence for multiple acquisition of *Arsenophonus* by whitefly species (Sternorrhyncha: Aleyrodidae). *Curr Microbiol* 48, 140–4. doi:10.1007/s00284-003-4157-7

Varani, A. M., Siguier, P., Gourbeyre, E., Charneau, V., and Chandler, M. (2011). ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biol.* 12, R30. doi:10.1186/gb-2011-12-3-r30

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9. doi:10.1371/journal.pone.0112963

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York)

bioRxiv preprint doi: https://doi.org/10.1101/311076; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Wood, D. E. and Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 15. doi:10.1186/gb-2014-15-3-r46

Zchori-Fein, E., Lahav, T., and Freilich, S. (2014). Variations in the identity and complexity of endosymbiont combinations in whitefly hosts. *Front Microbiol* 5, 1–8. doi:10.3389/fmicb.2014.00310