MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 628–640 doi: 10.1111/1755-0998.12478

Assessment of a 16S rRNA amplicon Illumina sequencing procedure for studying the microbiome of a symbiont-rich aphid genus

E. JOUSSELIN,* A.-L. CLAMENS,* M. GALAN,* M. BERNARD,† S. MAMAN,‡ B. GSCHLOESSL,* G. DUPORT,§ A. S. MESEGUER,* F. CALEVRO§ and A. COEUR D'ACIER*

**INRA – UMR 1062 CBGP (INRA, IRD, CIRAD, Montpellier SupAgro), 755 avenue du Campus Agropolis CS 30016, F-34 988 Montferrier-sur-Lez, France, †INRA – UMR 1313 GABI–SIGENAE, INRA de Jouy en Josas, Domaine de Vilvert, 78352 Jouy en Josas, France, ‡INRA, GenPhySE, Sigenae, Chemin de Borde rouge -CS 52627, 31326 Castanet Tolosan, France, §UMR 203 BF2I, Biologie Fonctionnelle Insectes et Interactions, INRA, INSA de Lyon, Université de Lyon, 20 Avenue Einstein, F-69621 Villeurbanne, France*

Abstract

The bacterial communities inhabiting arthropods are generally dominated by a few endosymbionts that play an important role in the ecology of their hosts. Rather than comparing bacterial species richness across samples, ecological studies on arthropod endosymbionts often seek to identify the main bacterial strains associated with each specimen studied. The filtering out of contaminants from the results and the accurate taxonomic assignment of sequences are therefore crucial in arthropod microbiome studies. We aimed here to validate an Illumina 16S rRNA gene sequencing protocol and analytical pipeline for investigating endosymbiotic bacteria associated with aphids. Using replicate DNA samples from 12 species (Aphididae: Lachninae, *Cinara*) and several controls, we removed individual sequences not meeting a minimum threshold number of reads in each sample and carried out taxonomic assignment for the remaining sequences. With this approach, we show that (i) contaminants accounted for a negligible proportion of the bacteria identified in our samples; (ii) the taxonomic composition of our samples and the relative abundance of reads assigned to a taxon were very similar across PCR and DNA replicates for each aphid sample; in particular, bacterial DNA concentration had no impact on the results. Furthermore, by analysing the distribution of unique sequences across samples rather than aggregating them into operational taxonomic units (OTUs), we gained insight into the specificity of endosymbionts for their hosts. Our results confirm that *Serratia symbiotica* is often present in *Cinara* species, in addition to the primary symbiont, *Buchnera aphidicola*. Furthermore, our findings reveal new symbiotic associations with *Erwinia*- and *Sodalis*-related bacteria. We conclude with suggestions for generating and analysing 16S rRNA gene sequences for arthropod-endosymbiont studies.

Keywords: endosymbionts, Illumina, metagenomics, phloem-feeders, *Serratia symbiotica*

Received 13 March 2015; revision received 2 October 2015; accepted 6 October 2015

Introduction

Endosymbiotic bacteria inhabiting arthropods are increasingly recognized as major players in the ecology and evolution of their hosts (Frago *et al.* 2012; Jaenike 2012; White *et al.* 2013; Oliver *et al.* 2014). Aphids (Hemiptera: Aphididae) are model systems for the study of bacteria-arthropod endosymbiosis, as an obligate mutualistic association with *Buchnera aphidicola* (a ϒ-3 proteobacteria) was established in these insects more than 180 Mya (Moran *et al.* 1993). These bacteria provide the aphids with essential amino acids, and vitamins they cannot synthesize or find in sufficient amounts in the plant sap, their sole source of nutrients (Douglas 1998). Aphids also harbour a diversity of facultative endosymbionts. Unlike obligate endosymbionts, facultative endosymbionts are not required for survival or reproduction. However, their presence can increase the fitness of their hosts in specific environmental conditions (Oliver *et al.* 2010).

Most studies on the endosymbiotic communities of aphids have involved PCR-based detection with specific primers. The traits conferred by the endosymbionts are then generally identified by investigating the correlations between the presence of the endosymbiont and host

Correspondence: Emmanuelle Jousselin, Fax: +33 (0) 4 99 62 33; E-mail: jousseli@supagro.inra.fr

© 2015 John Wiley & Sons Ltd

ecological traits and environmental variables (e.g. Smith *et al.* 2015). These techniques rely on strong prior assumptions concerning the bacterial strains generally present in the organisms studied. High-throughput sequencing technologies are now replacing PCR for the investigation of microbial communities. These methods are based on the deep sequencing of PCR-amplified bacterial 16S ribosomal RNA gene fragments, potentially making it possible to detect all the bacteria present in a sample (Degnan & Ochman 2012). They have been widely used to study environmental samples and are becoming a method of choice in aphid microbiome studies (Jones *et al.* 2011; Bansal *et al.* 2014; Jing *et al.* 2014; Gauthier *et al.* 2015). These methods are undoubtedly powerful as they do not rely on prior knowledge about the diversity of the communities investigated. Furthermore, the sequencing read abundances assigned to each bacterial strain are sometimes used as a proxy for the relative abundance of each type of bacteria in a sample (e.g. Bansal *et al.* 2014; Otani *et al.* 2014 for studies on insect microbiomes).

However, the repeatability and reliability of results obtained with high-throughput sequencing technologies are rarely evaluated before undertaking large-scale studies on multiple host populations. Furthermore, unlike microbial ecology studies, which aim to characterize bacterial diversity across different environments (Goodrich *et al.* 2014), ecological studies on arthropod endosymbionts generally aim to detect the presence of a few specific bacteria in individuals from natural populations, with a high level of confidence, rather than focusing on diversity indices. The accurate taxonomic identification of bacteria and the exclusion of contaminants from the results are therefore crucial steps in arthropod-endosymbiont studies. Massively parallel sequencing of 16S rRNA genes is highly susceptible to contamination. Bacteria present in extraction kits and PCR mixes have been shown to distort the results of microbiome analyses, particularly for samples with a low microbial biomass (Goodrich *et al.* 2014; Salter *et al.* 2014), and the impact of such contaminants on arthropod-endosymbiont studies is unclear. In addition, studies on arthropod-endosymbiont associations often evaluate the specificity of the interaction between the partners (Jousselin *et al.* 2009; Hendry & Dunlap 2014). This requires an evaluation of the intrastrain genetic diversity of the bacteria. The sequence data generated by high-throughput sequencing technologies are generally clustered into operational taxonomic units (OTUs) on the basis of sequence similarity. This procedure is designed to cluster sequence from the same 'bacterial species' together and also aggregate sequences that are probably derived from base incorporation errors to prevent the artificial inflation of diversity indices. It is followed by the assignment of representative sequences of the OTU to particular taxa. OTUs are valuable tools for the investigation of complex bacterial communities, but their use can introduce bias, and they are not necessarily appropriate for endosymbiont studies, for several reasons. First, OTU-picking algorithms and sequence similarity thresholds strongly influence the taxonomic composition of samples (Goodrich *et al.* 2014; Mahé *et al.* 2014). Second, clustering approaches often erase information about within-strain diversity and may render taxonomic assignment ambiguous, as sequences not assigned to the same bacterial lineage may be clustered into a single OTU (Mahé *et al.* 2014). In addition, arthropod-endosymbiont communities are generally dominated by a few known taxa (Aylward *et al.* 2014; Jing *et al.* 2014; Vanthournout & Hendrickx 2015) and clustering sequences into OTUs might not be necessary when assessing their composition.

In this study, we assessed the validity of a 16S rRNA amplicon-based Illumina sequencing and analytical procedure, by comparing the results (in terms of taxonomic identification and relative abundance of the reads assigned to a phylum) obtained with several DNA extracts from the same aphid colony, and for PCR replicates for each DNA extract. We used aphids from the genus *Cinara* (Aphididae: Lachninae), as studies using PCR with specific primers for detection have indicated that the incidence of bacterial endosymbiont infection is high in *Cinara* species (Lamelas *et al.* 2008; Burke *et al.* 2009). As our aim was to identify the principal bacteria associated with aphids and investigate their specificity for their hosts, we removed individual sequences that did not meet a minimum threshold number of reads in each sample rather than grouped them into OTUs and analysed the distribution of the remaining sequences across samples. Negative controls were sequenced at every step of the procedure, to evaluate the impact of laboratory contaminants on the results obtained. This study focused on *Cinara* species, but the approach used could probably be transferred to any microbiome associated with an aphid or arthropod species.

Methods

DNA samples

We examined 12 species of *Cinara*. Each species was comprised of multiple individuals collected from the same aphid colony. The specimens were kept in 70% ethanol at 6 °C immediately after collection (Table S1, Supporting information).

We obtained three DNA extracts from each sample (Fig. 1).

1 DNA samples enriched in bacteria were prepared with a slightly modified version of the protocol described

© 2015 John Wiley & Sons Ltd

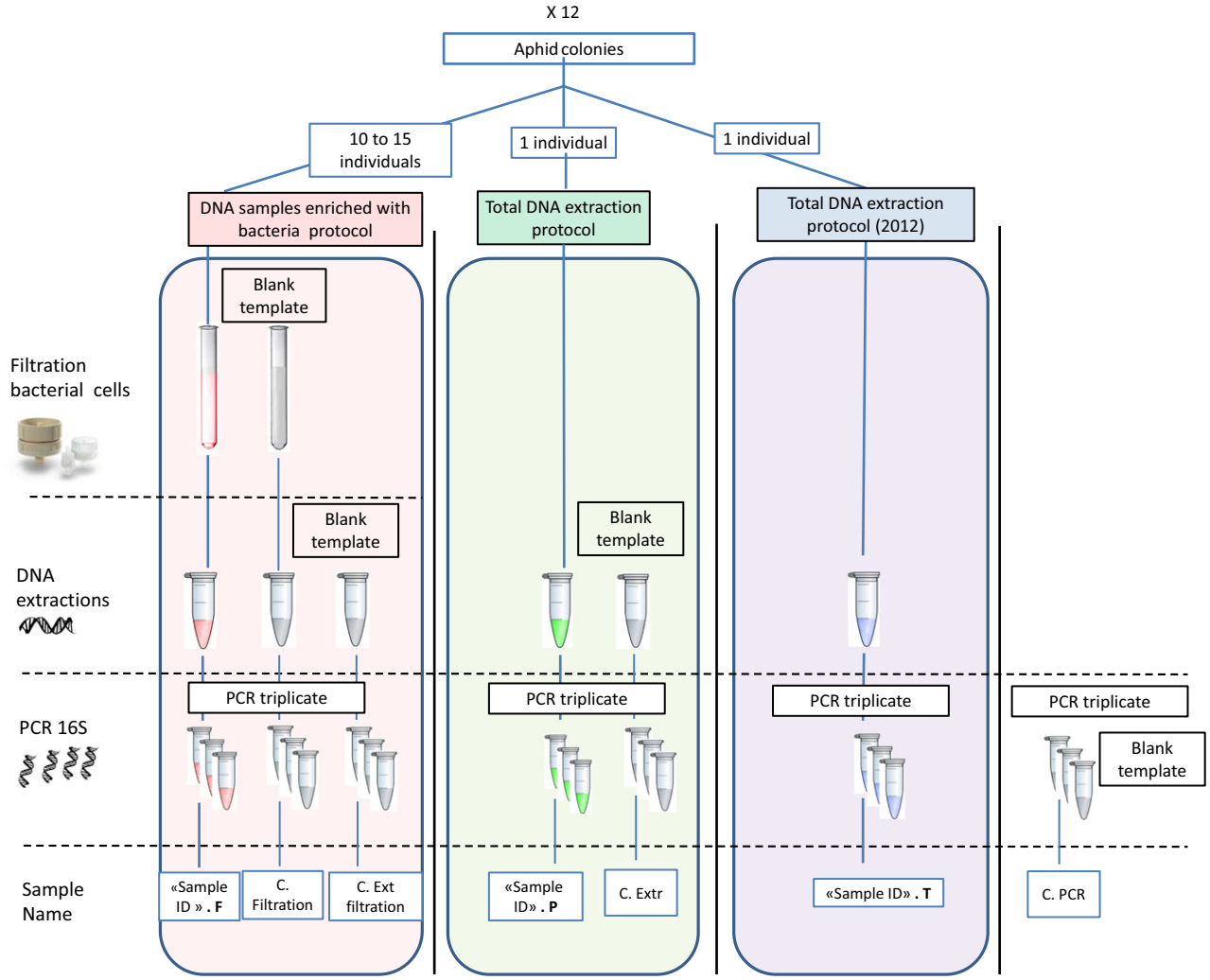


Fig. 1 Workflow of the laboratory procedure used to obtain *Cinara* DNA extracts, PCR products and related negative controls.

by Charles & Ishikawa (1999). For each *Cinara* species, 10–15 aphids per sample were first rinsed three times in ultrapure water (Qiagen, Germany) and then crushed in 1 mL of buffer A (35 mM Tris-HCl, 25 mM KCl, 10 mM $MgCl_2$ and 25 mM sucrose, pH 7.5) with a Teflon pestle and a glass mortar. The resulting homogenate was then successively filtered through two filters (the first with 100 μm pores and the second with 30 μm filter pores) made of plankton net fabric. It was then filtered through two Millipore® Isopore™ white polycarbonate membranes, the first with 10 μm pores and the second with 5 μm pores. The filtrates were centrifuged at 5000 × *g* for 10 min, and the supernatants were discarded. These successive filtrations eliminated eukaryotic cells, which are generally about 10 times larger than bacterial cells (bacterial cells are 0.5–5.0 μm long, on average). We extracted DNA from the residual pellets, with the DNeasy Blood & Tissue Kit (Qiagen, Germany), according to the manufacturer's recommendations. The DNA was eluted in 40 μL of elution buffer. These extracts concentrate bacterial DNA in comparison with extracts conducted on a whole aphid individual; this should decrease the influence of contaminants on the results (Salter *et al.* 2014).

2 A single individual from each colony was washed three times in ultrapure water. Total genomic DNA was extracted from whole individuals with the same extraction kit as used for the procedure described above.

3 In parallel, we used DNA extracts from our previous phylogenetic investigations of the genus *Cinara* (Jousselin *et al.* 2013). These extracts were obtained with the EZ-10 96-Well Plate Genomic DNA Extraction Kit, Animal Samples (Bio Basic Inc., ON, Canada), in 2012, from individuals from the same aphid colonies

© 2015 John Wiley & Sons Ltd

as used in the other two protocols. Samples had already been taken from these DNA extracts for several PCR amplifications, and the extracts were stored at −20 °C. The aim of this test was to determine whether DNA extracts that had been stored for up to 3 years gave results similar to those for fresh extracts. DNA degradation and contamination during previous laboratory procedures might distort the results of 16S rRNA gene sequencing. The feasibility of using stored samples is an issue of particular relevance for investigations of natural populations of arthropods that can be difficult and costly to sample, and, particularly, for small arthropods, such as aphids and mites, for which DNA extraction necessarily involves the destruction of the entire individual (DNA can therefore only be extracted once from any given individual).

We included several negative controls. We filtered 1 mL of buffer A, using the laboratory reagents and material used for the filtration procedure. This negative control template was then processed in the same way as the DNA extracts. During the extraction procedures conducted for this study, a 'blank template' of ultrapure water was processed with the same extraction kit (Fig. 1). There was no negative control for the extractions conducted in 2012.

As positive DNA controls, we used DNA extracts from four pure bacterial strains and DNA extracts from seven arthropod specimens with known bacterial endosymbionts (Table S1, Supporting information).

All DNA samples were stored at −20 °C.

16S amplification and sequencing

We used a modified version of the protocol of Kozich *et al.* (2013) for the targeted sequencing of an indexed bacterial 16S rRNA gene fragment on a MISEQ (Illumina) platform. We amplified a 251-bp portion of the V4 region that has been shown to be one of the most effective markers for assessing bacterial diversity (Mizrahi-Man *et al.* 2013). We used slightly modified forms of the universal primers described by Kozich *et al.* (2013) (16S-V4F: GTGCCAGCMGCCGCGGTAA and 16S-V4R: GGAC TACHVGGGTWTCTAATCC). The 5′ regions of these primers were modified by adding the appropriate Illumina adapter (P5 or P7), an 8-nt index sequence (i5 or i7), a 10-nt pad sequence and a 2-nt linker. This strategy is called dual-index sequencing (Kozich *et al.* 2013).

Each DNA sample and negative control was amplified in triplicate. Each of these replicates was conducted on a different 96-well microplate. A negative control for PCR was included on each plate (Fig. 1). The PCR mixture contained 5 μL Qiagen Multiplex PCR Master Mix (1× final concentration, including *Taq* polymerase, dNTPs and $MgCl_2$ at a final concentration of 3 mM), 2 μL each of the forward and reverse indexed primers (1 μM final concentration) and 2 μL genomic DNA, or ultrapure water for the PCR negative controls. The reaction mixture was heated at 95 °C for 15 min to denature the DNA and then subjected to 40 cycles of 95 °C for 20 s, 55 °C for 15 s and 72 °C for 5 min for amplification, followed by a final extension for 10 min at 72 °C. We obtained 150 PCR products in total. Each pair of indices (i5 and i7) was unique to a PCR well, aiming to re-cover the origin of each sequence to a sample.

Combinations of 24 i5 index primers and 36 i7 index primers were used for identification, and we were therefore able to multiplex up to 24 × 36 (=864) amplicons in one MISEQ FLOWCELL. Our samples were pooled with other libraries, including 711 amplicons from another project.

PCR products were pooled, and the resulting mixture was subjected to gel electrophoresis. The bands corresponding to the PCR products were excised from the gel, purified with a PCR clean-up and gel extraction kit (Macherev-Nagel) and quantified with the Kapa Library Quantification Kit (Kapa Biosystems). The pool of 854 libraries was submitted for paired-end sequencing in a MISEQ (Illumina) FLOWCELL equipped with a version 2 500-cycle reagent cartridge.

Data analyses

Sequence filtering. Sequence filtering criteria were applied through Illumina's quality control procedure. We then used a pipeline of MOTHUR (v1.3.3) software package (Schloss *et al.* 2009) functions implemented on a Galaxy workbench (Goecks *et al.* 2010) (http://galaxy-workbench.toulouse.inra.fr/). The overlapped paired-end reads were assembled with the *make.contigs* function of MOTHUR, and the contigs exceeding 280 bp in length and/or containing ambiguous base pairs were filtered out and excluded from further analyses. A FASTA file containing unique contigs and a file reporting the occurrences of these sequences in each sample were created. Unique sequences from the FASTA file were then aligned with the V4 portion of reference sequences from the SILVA 16S reference database (v119) (Quast *et al.* 2013). Sequences that did not align with the V4 fragment were excluded from further analyses. After this filtering step, a new file containing unique sequences was created. The number of reads resulting from sequencing errors was then reduced by merging rare unique sequences with frequent unique sequences with a mismatch of no more than 2 bp relative to the rare sequences (*pre.cluster* command in MOTHUR). We then used the UCHIME (Edgar *et al.* 2011) program implemented in MOTHUR to detect chimeric sequences (e.g. sequences resulting from the recombination of two

© 2015 John Wiley & Sons Ltd

sequences from two different taxa due to jumping PCR events). This procedure was performed on each sample, and the chimeric sequences identified were excluded from the data set. The number of reads for each sequence in each sample was determined and used to compile a contingency table.

Analyses of the diversity of bacteria, repeatability across PCR replicates and DNA extracts. For each sample, we transformed read numbers into frequencies (percentages). Many sequences were represented by very few reads (Fig. S1, Supporting information). Sequences accounting for <1% of all the reads for a given sample were excluded with an R script (Appendix S1, Supporting information). Before applying this threshold, we determined how the number of unique sequences per sample decreased with increasing threshold values (Fig. S2, Supporting information). The script used also generates a table summarizing, for each sample, the total number of reads, the total number of unique sequences and the number of unique sequences remaining after the removal of all the sequences accounting for <1% of all reads.

For all samples (including negative controls), we determined whether the total number of unique sequences and the number of unique sequences making up more than 1% of all reads differed between PCR replicates for the same sample, using a General Linear Model in R (the number of sequences was used as the response variable and was considered to follow a Poisson distribution; 'PCR replicate' and sample were set as factors). For *Cinara* samples, we used the model described above, but with the inclusion of 'DNA extract' as a factor.

We investigated differences in the bacterial community between PCR replicates and DNA extraction replicates for each aphid colony, by calculating the Shannon diversity index (H) of sequences making up more than 1% of all reads for each sample. We used ANOVA followed by Tukey post hoc tests to determine whether this index differed between PCR and DNA replicates. The dissimilarity between the bacterial communities of *Cinara* samples was also quantified by calculating Sørensen's index, a metric based on the presence/absence of taxa (here, 'taxa' were unique sequences). The Sørensen dissimilarity matrix was ordinated following a non-metric multidimensional scaling (nMDS). The result of nMDS ordination was plotted on a three-dimensional graph on which the position of each sample depended on its distance from all other samples. We then conducted an Adonis analysis with the Sørensen distance matrix as the response variable and *Cinara* species, PCR and DNA extract replicates as factors, to determine whether the mean similarity between samples from the same species (i.e. across PCR and DNA replicates) was greater than that between samples from different species. All community diversity analyses were conducted with the R package VEGAN V.2.3-0.

We investigated whether the frequencies of reads were similar across replicates for the same aphid colony. We transformed the frequencies of each unique sequence into ranks reflecting their relative abundance. For each *Cinara* species, we then performed a Friedman rank test, with bacterial types as 'blocks' and replicates (any replicate of the same aphid colony, that is DNA extractions and PCR) as 'treatments' in R.

The taxonomic affiliation of each unique sequence was determined with the naive Bayes classifier technique, as implemented in the *classify.seq* function of MOTHUR (Mizrahi-Man *et al.* 2013), with the SILVA reference database v119 (Quast *et al.* 2013) and a minimum confidence value of 0.80. This step corresponds to the 'sequence analysis' module of the Galaxy pipeline. Taxonomic identification was also achieved with the LEBIBI^{QBPP} program (http://pbil.univ-lyon1/bibi/lebibi.cgi) (Devulder *et al.* 2003). This online tool searches for *n* (this number was set to 75 for this analysis) sequences similar to the sequence submitted, by BLAST analyses of precompiled databases. We first used the 16S SSU-rRNA TS stringent database, which only includes sequences of validly denominated species. When taxonomic assignation of a sequence was 'undetermined' or reached only the family level, we used the 16S SSU-rRNA 'LAX' database, which includes all sequences corresponding to given criteria concerning sequence length and gene identification available in the NCBI database. Pairwise local alignments from the BLAST output are then extracted, and the *n* similar sequences are aligned with MAFFT (Katoh *et al.* 2005). A maximum-likelihood phylogenetic tree is generated with FASTTREE (Price *et al.* 2009) and a GTR-gamma model of evolution. Taxonomic assignment is based on the name of the sequence with the shortest patristic distance to the query sequence.

We used these assignations and the table of sequence frequencies per sample to determine the bacterial composition of each sample. We simplified the representation of the results by adding the frequencies of different unique sequences assigned to the same bacterial species (or genus). These taxonomic bins were used to visualize community compositions, whereas ungrouped unique sequences were used for all statistical analyses.

We investigated the intrastrain diversity and host specificity of the main bacterial taxa (i.e. *Buchnera aphidicola* and *Serratia* spp.) associated with our samples. We built a heatmap representing the distribution of unique sequences (i.e. haplotypes) and their relative abundance across *Cinara* samples. We used CLUSTALX 2.0.11 (Thompson *et al.* 1994) to obtain sequence alignments for *B. aphidicola* and *Serratia* spp. Sequences (Appendices S2 and S3: Supporting information); *p* distances between

© 2015 John Wiley & Sons Ltd

pairs of sequences were computed in MEGA v5.0 (Tamura *et al.* 2007), and neighbour-joining trees were reconstructed in PHYML-3.1 (Guindon *et al.* 2010). These phylograms were incorporated into the heatmaps to illustrate the relatedness of the sequences in each phylum.

Results

Data set description

Using Illumina's stringent quality control (QC), we obtained a total of 2 606 606 reads passing Illumina quality filters and assigned to a sample. The mean number of reads generated for each sample was 13 947 (SD = 3615) (excluding negative controls) (Table S2, Supporting information). The number of unique sequences after all filtering steps (i.e. the removal of sequences with ambiguous base pairs, long contigs, nonspecific sequences, chimeric sequences) was 37 509 represented by 2 018 186 reads. We removed 7142 chimeric unique sequences from the data set. Negative controls generated fewer reads than DNA samples (Fig. 2): from 364 on average for 'C. PCR' (PCR conducted on a blank template, Fig. 1) to 6185 on average for 'C. extractions' (the control for de novo extractions from whole individuals).

The total number of unique sequences per sample differed between PCR replicates ($\chi^2 = 56.8$, d.f. = 2, $P < 0.001$), but many of these sequences were represented by very few reads (18 314 of the 37 509 unique sequences were represented by a single read) (Fig. S1, Supporting information). Once the sequences making up <1% of the reads in a sample were discarded from that sample, our data set included only 191 unique sequences and the number of unique sequences per sample did not differ significantly between PCR replicates ($\chi^2 = 0.067$, d.f. = 2 $P = 0.95$). For *Cinara* colonies, the number of unique sequences per sample did not differ between DNA extracts ($\chi^2 = 5.76$, d.f. = 2, $P = 0.051$).

Bacterial diversity in negative and positive controls

Negative controls generally contained ubiquitous bacteria, such as *Anoxybacillus, Methylobacterium* and bacteria from the Comamonadaceae family (Figs 2 and S3, Supporting information). Overall, our negative controls contained very few bacteria that were also found in aphid samples: only 123 of the 37 509 unique sequences from all reads were common to the negative controls and *Cinara* species. Once the sequences accounting for <1% of reads had been discarded, only seven unique sequences were common to the negative controls and some of our *Cinara* DNA samples. These sequences were assigned to *Brevibacillus*, *Anoxybacillus*, *Geobacillus* and the Comamonadaceae family. The only negative control containing

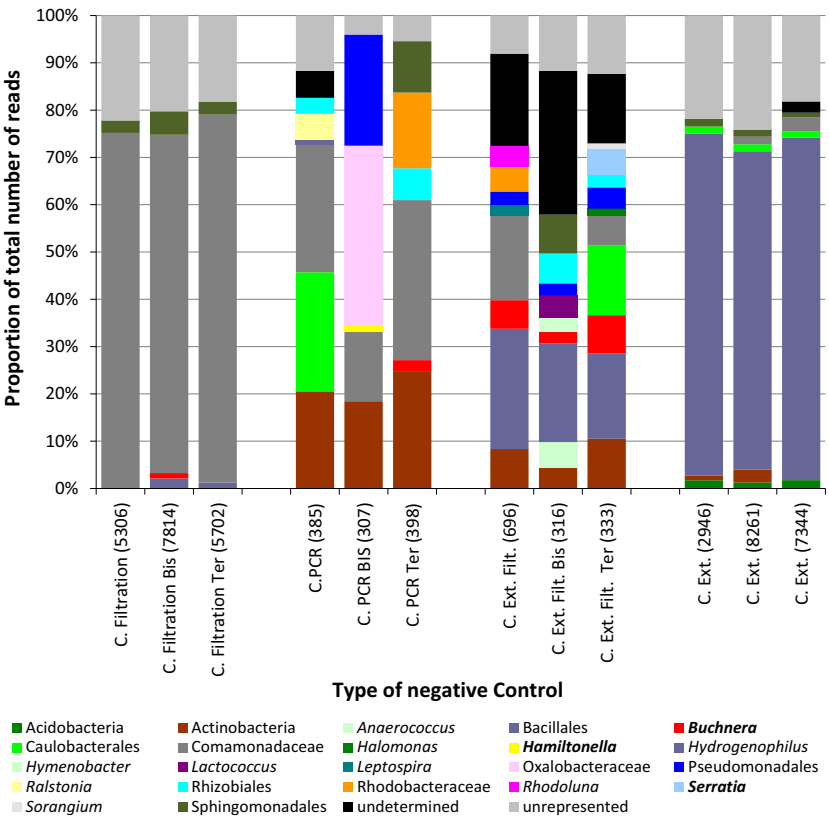


Fig. 2 Summary of 16S rRNA gene sequencing results and taxonomic assignment for negative controls. Each column shows the relative frequency of each type of bacterium obtained by sequencing a PCR product: 'Bis' and 'Ter' indicate PCR replicates for the same DNA extract (the numbers of reads obtained for each sample are indicated in brackets). The taxonomic assignments of sequences making up <1% of all the reads in a sample are not shown; these sequences are found in the 'unrepresented' part of each column. Taxonomic assignment was carried out to genus level whenever possible (Fig. S2). Here, we present a simplified version: except for bacteria known to be arthropod endosymbionts (indicated in bold) and genera that were unique representatives of a family, sequences were binned at the taxonomic level of family or order.

© 2015 John Wiley & Sons Ltd

bacteria identified as aphid endosymbionts in all PCR replicates was the control set up during the extraction of DNA from filtration products (C.ext.filtration) (Fig. 2). *B. aphidicola* accounted for up to 8% of the reads and *Serratia* for up to 5% of the reads in this control, but this corresponded to only a very small number of reads (26 and 12 reads, respectively, whereas *B. aphidicola* accounted for 8800 reads, on average, per aphid sample). Three different *Buchnera* haplotypes were detected in these samples. The main haplotype (corresponding to 13 reads) found in the C.ext.filtration control was also found in *C. brevispinosa* (3412); the other two haplotypes were not retrieved in any of our samples. The *Serratia* identified in this control also had the same haplotype as the *Serratia* detected in sample 3412. This suggests that this sample contaminated our control. A few *B. aphidicola* sequences were also found in one of the PCR replicates of the extraction control (C.filtration.Bis: 88 reads) and one of the PCR controls (C.PCR.Ter: eight reads).

In the positive controls, consisting of pure bacterial cultures, bacterial diversity was low and the cultured bacteria were successfully sequenced (Fig. S4, Supporting information). The *Ricketssia conorii* sequencing results included sequences from a bacterium identified as *Mycoplasma*, suggesting possible contamination of the culture or the DNA sample. Ubiquitous bacteria, such as *Acinetobacter*, were found at low frequency in the *Borrelia* culture. Sequencing of DNA from arthropods with known bacterial symbionts yielded the expected results in all cases. For example, the two aphid samples (*Acyrthosiphon pisum*) contained the expected secondary symbionts (*Regiella* and *Serratia*) in addition to *B. aphidicola*.

Diversity of bacterial symbionts associated with Cinara

The mean Shannon diversity index was low: 0.78 (SD = 0.29), suggesting that each sample was dominated by a few bacteria (Table S3, Supporting information). For the various *Cinara* colonies, this diversity index did not differ between PCR replicates ($\chi^2 = 0.002$, d.f. = 2, $P = 0.99$) or DNA extracts ($\chi^2 = 0.51$, d.f. = 2, $P = 0.72$). When percentages of reads were transformed into ranks, these ranks did not differ significantly between PCR and DNA extraction replicates, for 10 of 12 species (Table S3, Supporting information, Fig. 3). Thus, when a 16S rRNA gene haplotype was identified as the dominant feature in one of the PCR products from an aphid colony, it was also found to be the dominant feature in all the DNA and PCR replicates for the same aphid colony.

Buchnera aphidicola generally predominated in *Cinara* samples. *Cinara* also contained up to four other endosymbionts (Fig. 3). In eight of 12 *Cinara* species, *Serratia* strains were found to be associated with all DNA extracts, and in six of these eight species, *Serratia* was the second most abundant endosymbiont after *B. aphidicola*, even accounting for almost half the sequencing reads in five species (Fig. 3). Other common aphid endosymbionts, such as *Hamiltonella*, *Wolbachia* and *Regiella*, were found in our samples. Our results also revealed that *Erwinia*-related strains were associated with *Cinara pseudotaxifoliae* and *Cinara laricis*, and that *Sodalis* strains were found in *Cinara strobi*. Two of the DNA extracts obtained by filtration and one of the old DNA extracts contained ubiquitous bacteria, such as *Brevibacillus*, *Anoxybacillus*, *Geobacillus* and bacteria from the Comamonadaceae, Bradyrhizobiaceae and Chitinophagaceae families (Fig. 3). These bacteria were often represented by only a few reads (1% to 5% of the reads), and they were not systematically present in all PCR replicates for any given DNA extract (Fig. 3). None of the de novo extracts contained such bacteria in significant proportions (i.e. representing more than 1% of the reads).

The scatter plots for Sørensen's index revealed that all replicate samples from an aphid species contained very similar bacterial communities, whereas there was greater dissimilarity between the communities present in different aphid species (Fig. 4a). The R^2 statistic of the Adonis analysis was 0.99 for 'species' and highly significant ($P < 0.001$), whereas those for 'PCR' and 'DNA extracts' were 0.00036 ($P = 0.99$) and 0.004 ($P = 0.01$), respectively. The R^2 values suggest that there was greater dissimilarity between samples from different species than between samples from the same species. PCR replicates for the same DNA samples always yielded the same bacterial communities, except for the *C. schwartzii* 2012 extraction, for which a single PCR replicate revealed the occurrence of *Hamiltonella*, and newly extracted *C. laricis*, for which a single PCR replicate revealed the presence of *Erwinia* (Fig. 3). However, there were slight differences between DNA extracts from the same aphid colony. A single DNA extract from *Cinara confinis* contained *Serratia*, and *Serratia* was found in only two of the three DNA extracts from *C. strobi Hamiltonella* was found in a single DNA extract from *C. pseudotaxifoliae*, and a single extract from *C. schwartzi. Erwinia* was found in only two of the three DNA extracts from *C. laricis* (Fig. 3).

Each aphid species hosted *Buchnera* strains with distinctive 16S haplotypes (Fig. 4b), except for *C. fresai* and *C. confinis*, which shared *Buchnera* strains with the same haplotype. Seven *Cinara* species each contained a single *B. aphidicola* haplotype. By contrast, in five *Cinara* species, a second *B. aphidicola* 16S rRNA gene sequence was found at low abundance (about 1–6% of the reads) in some or all of the DNA extraction replicates. This second sequence was always very closely related to the most abundant haplotype (Fig. 4b). The mean *p* distance between the *B. aphidicola* sequences present in different *Cinara* species was 0.057 (SD = 0.035), whereas the mean

© 2015 John Wiley & Sons Ltd

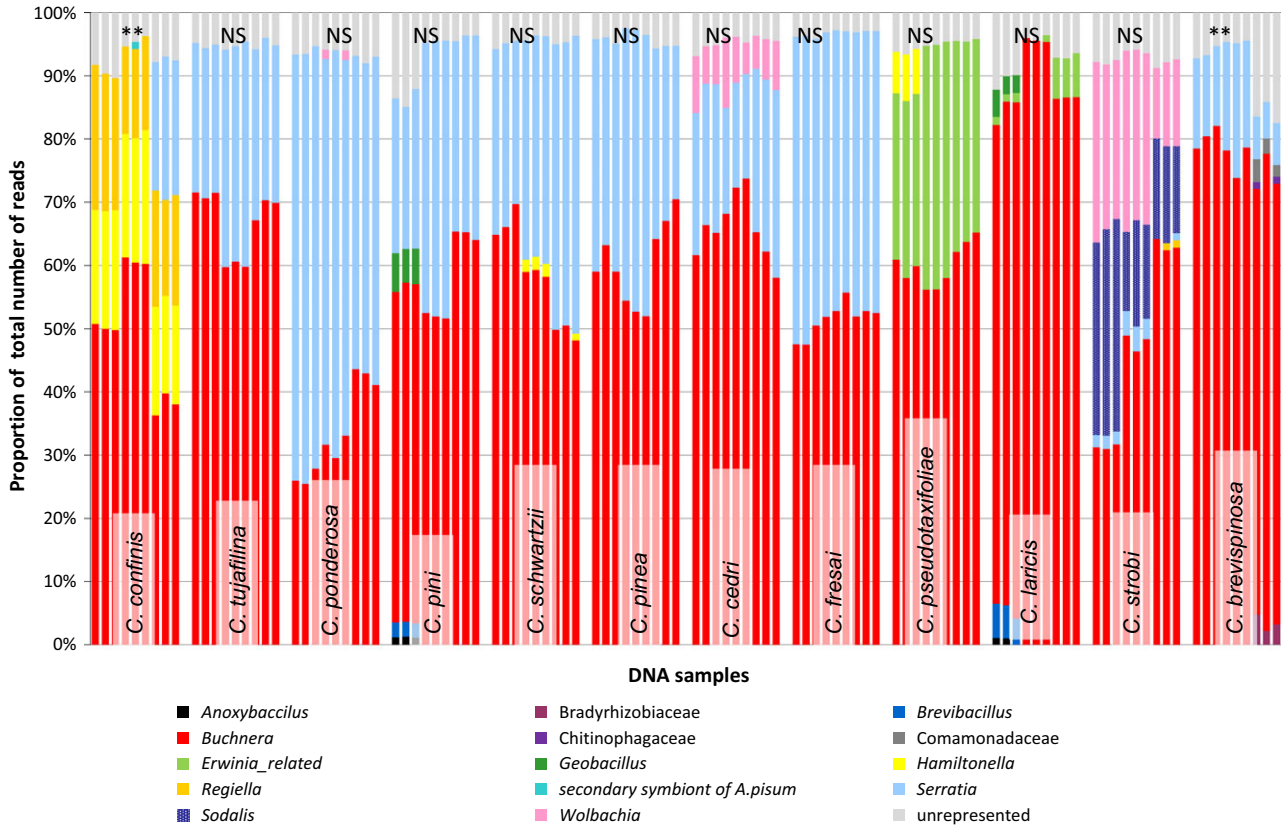


Fig. 3 Summary of 16S rRNA gene sequencing-based taxonomic assignment for *Cinara* samples. Each column represents the proportion of reads obtained from the sequencing of a PCR product. Unique sequences making up <1% of all the reads for a sample are found in the 'unrepresented' part of each column. For each *Cinara* colony, all PCR replicates for each type of DNA extraction are represented: the three-first columns correspond to extracts enriched in bacterial DNA, the next three columns correspond to *de novo* extractions, and the last three correspond to extractions from previous molecular studies. We indicate the results of Friedman rank tests (NS = non significant) and the name of each *Cinara* species above each sample.

p distance between sequences present in the same species was 0.012 (SD = 0.001).

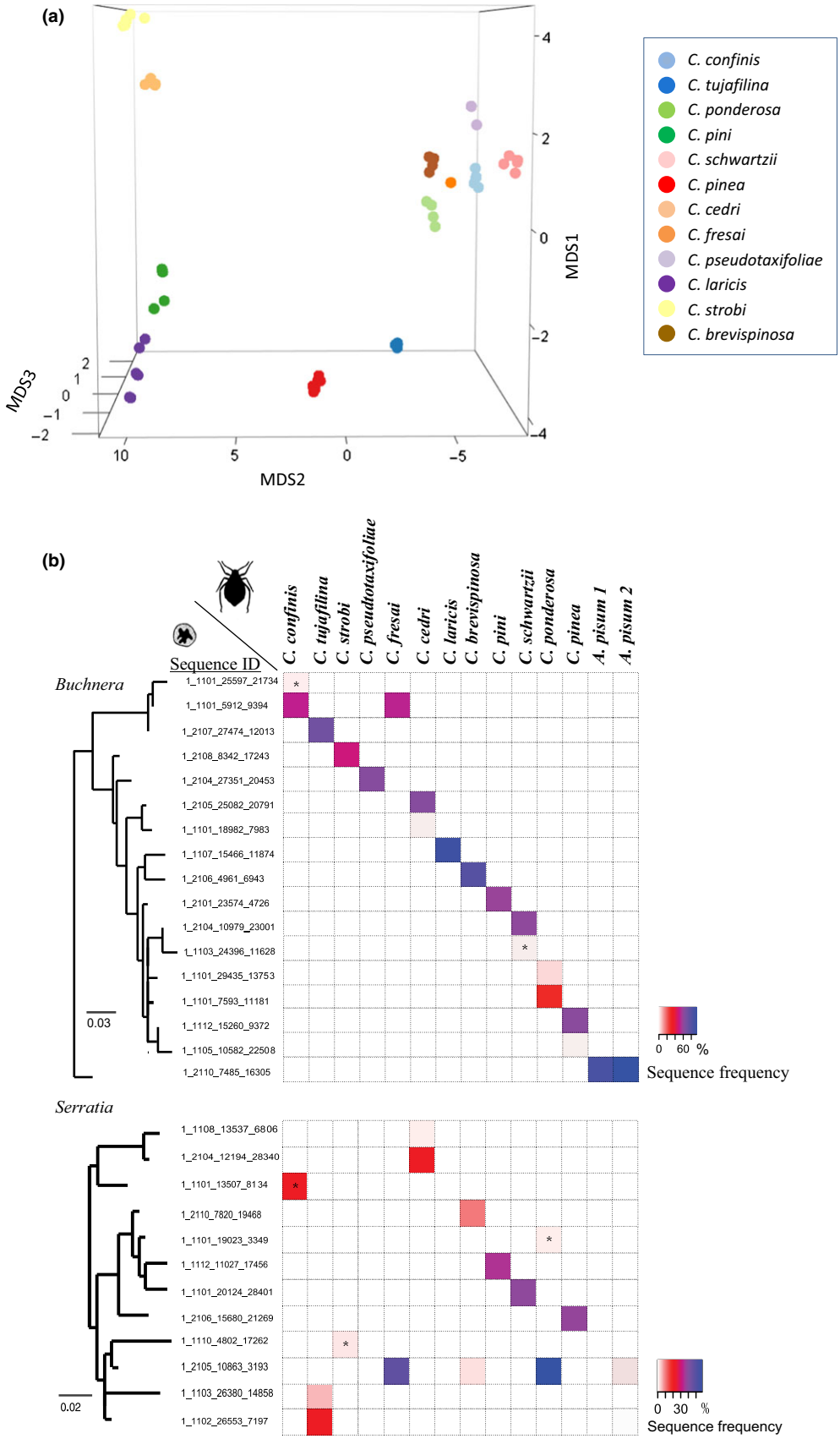
Cinara confinis, *C. pini*, *C. schwartzii*, *C. pinea*, *C. strobi* each contained a single, specific *Serratia* haplotype, whereas *C. cedri*, *C. ponderosa*, *C. brevispinosa* and *C. tujafilina* each contained two specific *Serratia* sequences, one of which occurred at low frequency (from 1% to 4% of the reads) (Fig. 4b). *C. ponderosa*, *C. fresai*, *C. brevispinosa* and *A. pisum* all contained a *Serratia* strain, of the same haplotype in each case. The mean *p* distance between the *Serratia* sequences present in different *Cinara* species was 0.038 (SD = 0.014), whereas the mean *p* distance between sequences present in the same species was 0.025 (SD = 0.011).

Discussion

Do contaminants affect aphid microbiome analyses?

The sequencing of negative controls often yielded sequences from many ubiquitous environmental bacteria, These communities were very similar to those described in Salter *et al.* (2014). However, aphid samples had few taxa in common with these controls and the predominant taxa present in aphids always corresponded to phyla known to be arthropod endosymbionts. Thus, the bacteria found in laboratory reagents had little impact on aphid endosymbiont analyses. Contrary to expectations, working with DNA preparations enriched in bacterial DNA had little impact on the taxonomic compositions determined for our samples, demonstrating that the concentration of bacterial DNA is not required to optimize investigations on aphid endosymbionts. These findings are consistent with those of Rubin *et al.* (2014) showing that insect DNA extraction procedures have no impact on the results obtained concerning endosymbiont community composition. However, DNA extraction by the cell filtration procedure and classical methods of DNA extraction should be compared for arthropods with more complex and less abundant microbiota than *Cinara*. Some of the negative controls contained a small percentage of aphid endosymbiont sequences. This may be due to

© 2015 John Wiley & Sons Ltd



© 2015 John Wiley & Sons Ltd

Fig. 4 (a) Nonmetric multidimensional scaling (NMDS) scores for the dissimilarity distance matrix between each sample. Each point represents a PCR product; each aphid colony is represented by a colour code. (b) Heatmaps showing the distribution and frequency (mean percentage) for *Buchnera* and *Serratia* 16S rRNA haplotypes across *Cinara* species. On the left-hand side of the figure, the NJ trees depict phylogenetic relationships between 16S rRNA haplotypes.*Indicates cases in which haplotypes were not present in all DNA extraction replicates.

cross-contamination of the samples, but it may also have resulted from tag-reading errors, which affect a small proportion of Illumina sequencing results in dual-index strategies (Kircher *et al.* 2012; Esling *et al.* 2015).

The removal of sequences accounting for <1% of the reads in aphid samples generally eliminated most of the sequences from bacteria common to negative controls and sequences not found in all the PCR replicates for a given sample. These low-depth sequences not present in all PCR replicates may originate from contaminants (e.g. airborne bacteria) present at low frequency and, therefore, not always successfully amplified, or spurious sequences. Their removal greatly increased the repeatability of the results. Nevertheless, sequences matching sequences from water- and soil-borne bacteria (*Geobacillus*, *Brevibacillus*, *Anoxybacillus* and some Bradyrhizobiaceae, Comamonadaceae and Chitinophagaceae) were found in some aphid samples: in *C. brevispinosa* extracts from 2012, and in DNA obtained after the filtration of *C. laricis* and *C. pini* samples to remove eukaryotic cells. They were never found in all the DNA extracts from a given aphid colony, and related bacteria were found in our negative controls. This suggests that these bacteria may be environmental contaminants. The filtration protocol might increase the risk of contamination, as more reagents are required than for extraction from a whole individual aphid. The extracts obtained in 2012 may also have been contaminated during previous laboratory manipulations.

We suggest that the detection of ubiquitous bacteria in aphid samples should lead to caution in interpretation concerning possible symbiotic associations. Bacteria from the genera *Acitenobacter*, *Brevundimonas* and *Brevibacillus*, which we found in several of our negative controls and a few of our samples and are listed among the contaminants found in laboratory reagents (Salter *et al.* 2014), have been detected in arthropod microbiome studies (in aphids, Bansal *et al.* 2014; and other arthropods, Rogers & Backus 2014) involving deep sequencing of the 16S rRNA gene. These bacteria have been considered to be possible symbiotic partners. In the absence of concurrent sequencing of negative controls in these studies, we believe that it would be risky to consider these bacteria to be present in the arthropod microbiota. They may be among the bacteria present in the gut or exoskeleton of these insects, but they may also simply be contaminants from the laboratory.

Reproducibility of taxonomic composition and bacteria relative abundance determinations

Once low-depth sequences had been removed, the taxonomic composition of our samples was highly similar across technical replicates for any given aphid colony. The few differences observed between DNA extracts from the same colony may be due to intracolony variation. Individuals from an aphid colony are generally the parthenogenetic progeny of a single female and, as aphid endosymbiotic bacteria are usually maternally transmitted (Michalik *et al.* 2014), it is therefore highly likely that all the individuals from a given colony have the same endosymbionts. However, it remains possible that some of the colonies we sampled were founded by several different females.

When transforming read frequencies into ranks reflecting the relative abundance of bacterial strains in each sample, the ranks obtained were found to be reproducible across the replicates for an aphid colony. Read frequencies are therefore not the result of a random process. Furthermore, our results concerning the abundance of *Serratia* in *Cinara* are consistent with those of previous studies: *Serratia symbiotica* bacteriocytes are abundant in *C. cedri* and *C. tujafilina* (Lamelas *et al.* 2008). However, as pointed out in many studies (e.g. Amend *et al.* 2010; Kembel *et al.* 2012; Bachy *et al.* 2013), the read frequencies of 16S rRNA genes cannot be used as direct estimates of bacterial cell abundance. Differences in read abundance may reflect biases in PCR success. Sequences from some bacterial species may be more likely to be amplified than those of other species, due to differences in primer specificity (Kurata *et al.* 2004). This hypothesis could easily be tested with different 16S rRNA gene primer sets. More specifically, for our biological model, read frequencies may reflect differences in 16S rRNA gene copy number rather than actual bacterial loads (Kembel *et al.* 2012). In the pea aphid, *B. aphidicola*, chromosome copy number per bacterial cell varies from 50 to 200, depending on the age of the aphid (Komaki & Ishikawa 2000). The ploidy levels of other symbionts are not known. It is therefore difficult to correct read abundances for gene copy numbers. Furthermore, the loads of *B. aphidicola* and secondary bacteria may vary during the nymphal development of aphids (Koga *et al.* 2003), and it is therefore risky to link quantitative estimates of endosymbiont loads to host-species life history traits

© 2015 John Wiley & Sons Ltd

(e.g. diet, reproductive mode) without assessing individual variations over time. In summary, the read frequencies obtained by Illumina 16S rRNA gene sequencing should be interpreted with care. They may indicate the bacteria abundant in an individual at a given point in time and can guide future studies investigating this association.

Can Illumina 16S rRNA amplicon sequencing provide insight into the specificity of aphid/endosymbiont associations?

The 16S rRNA gene fragment sequenced in the Illumina procedure is short (251 bp), but it can provide a first indication of intraspecific endosymbiont diversity and insight into the specificity of the host–endosymbiont interaction. The distribution of sequences across samples clearly showed that there was a specific haplotype of *B. aphidicola* associated with each species of *Cinara*, as expected from fine-scale studies of cospeciation between aphids and their primary symbiont (Jousselin *et al*. 2009). This analysis also suggests that slightly divergent *B. aphidicola* 16S rRNA gene copies (differing by 2–3 bp) can be found in a single colony of *Cinara*. These strains were always more closely related to each other than to the *B. aphidicola* present in other species, and one of the haplotypes was more abundant than the others. The less abundant haplotypes could be derived from base incorporation mistakes (occurring during the PCR or the sequencing procedure), but such errors should have been removed with the *pre.cluster* command in MOTHUR (see Methods). These haplotypes accounted for more than 400 reads in some samples and were present in all PCR replicates for the same DNA extract and, in some cases, in all DNA extracts from the same aphid colony, suggesting that they were not artefacts due to PCR errors (which are generally not reproducible) or sequencing errors (which have a frequency of 0.001–0.01 per base sequenced with the Illumina technology used here). This polymorphism may therefore result from *B. aphidicola* polyploidy: 16S rRNA gene copies may display slight variations.

Our results also revealed that three *Cinara* species and *A. pisum* harboured *Serratia* strains of the same haplotype. This suggests that these species have acquired closely related *Serratia* strains. The sequencing of multiple loci will be required to confirm this hypothesis. There were 11 *Serratia* haplotypes, each specifically associated with a single *Cinara* species. This species-specific pattern of association opens up the possibility of codiversification scenarios for *Serratia* and *Cinara*. We also observed some *Cinara* samples containing *Serratia* strains of two different haplotypes: these strains were sometimes as divergent as strains present in different species. This pattern of association may result from independent acquisitions of different strains or, as hypothesized for *B. aphidicola*, polyploidy resulting in variations in the number of copies of the 16S rRNA gene.

Conclusions and future prospects

The taxonomic composition of our samples was very similar across replicates for the same aphid sample, suggesting that our sequencing procedure and analytical approach are robust. De novo extracts from single individuals were less likely to contain environmental contaminants than DNA samples enriched in bacteria or extracts used in previous analyses. Nevertheless, when resampling is not possible, DNA extracts from previous studies can be used with no impact on the conclusions drawn concerning the presence/absence of the main endosymbionts. The sequencing of positive controls is useful to check amplification success, and it may also facilitate the detection of cross-contamination between samples. Our results for various arthropods also suggest that the primers used in our study successfully amplify many common endosymbionts. The removal of unique sequences corresponding to very small numbers of reads and taxonomic assignment of the remaining sequences excluded most of the contaminants from the results and provided insight into host–endosymbiont specificity. This is probably a sound approach for investigating arthropod endosymbiotic communities dominated by a few bacteria.

Finally, our results confirm that *Serratia* is a dominant feature of the microbiome of *Cinara* (Lamelas *et al.* 2011; Manzano-Marín & Latorre 2014). However, it was not present in all species, which suggests that it has been acquired several times or lost repeatedly in the evolutionary history of the genus. We also show that *Sodalis*-related bacteria dominate the *C. strobi* microbiome. This bacterial genus is relatively ubiquitous in insects (Snyder *et al*. 2011) and has previously been found in one *Cinara* species (Burke *et al.* 2009). We also revealed associations with *Erwinia*-like bacteria in two *Cinara* species: these bacteria are usually free-living phytobacteria, but they have also been found in aphid guts (Harada *et al*. 1997; Clark *et al.* 2012; Gauthier *et al*. 2015). Neither *Sodalis* nor *Erwinia* is common in aphids, and these genera may have been overlooked in PCR investigations based on the use of specific primers. Future work could focus on investigating a larger sample with several specimens per species to assess the prevalence of the predominant bacteria from the *Cinara* microbiome identified here.

Acknowledgements

This project was funded by an ANR (Phylospace), Projet Innovant INRA-EFPA 2013 to EJ and Projet Agropolis Fondation

© 2015 John Wiley & Sons Ltd

'*Cinara*s microbiome'. We wish to thank Hélène Vignes and the GPTR CIRAD-AGAP for the use of MISEQ platform, Jean-François Cosson for advice and the GenoToul bioinformatics facility Toulouse Midi-Pyrenees for computing resources. We would also like to thank Virginie Dupuy and Muriel Vayssier-Taussat for DNA extractions from bacterial cultures.

References

Amend AS, Seifert KA, Bruns TD (2010) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Molecular Ecology*, **19**, 5555–5565.

Aylward FO, Suen G, Biedermann PHW *et al*. (2014) Convergent bacterial microbiotas in the fungal agricultural systems of insects. *MBio*, **5**, e02077-14.

Bachy C, Dolan JR, Lopez-Garcia P, Deschamps P, Moreira D (2013) Accuracy of protist diversity assessments: morphology compared with cloning and direct pyrosequencing of 18S rRNA genes and ITS regions using the conspicuous tintinnid ciliates as a case study. *ISME Journal*, **7**, 244–255.

Bansal R, Mian MAR, Michel AP (2014) Microbiome diversity of *Aphis glycines* with extensive superinfection in native and invasive populations. *Environmental Microbiology Reports*, **6**, 57–69.

Burke GR, Normark BB, Favret C, Moran NA (2009) Evolution and diversity of facultative symbionts from the aphid subfamily Lachninae. *Applied and Environmental Microbiology*, **75**, 5328–5335.

Charles H, Ishikawa A (1999) Physical and genetic map of the genome of *Buchnera*, the primary endosymbiont of the pea aphid *Acyrthosiphon pisum*. *Journal of Molecular Evolution*, **48**, 142–150.

Clark EL, Daniell TJ, Wishart J, Hubbard SF, Karley AJ (2012) How conserved are the bacterial communities associated with aphids? A detailed assessment of the *Brevicoryne brassicae* (Hemiptera: Aphididae) using 16S rDNA. *Environmental Entomology*, **41**, 1386–1397.

Degnan PH, Ochman H (2012) Illumina-based analysis of microbial community diversity. *ISME Journal*, **6**, 183–194.

Devulder G, Perrière G, Baty F, Flandrois JP (2003) BIBI, a bioinformatics bacterial identification tool. *Journal of Clinical Microbiology*, **41**, 1785–1787.

Douglas AE (1998) Nutritional interactions in insect-microbial symbioses: Aphids and their symbiotic bacteria *Buchnera*. *Annual Review of Entomology*, **43**, 17–37.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Esling P, Lejzerowicz F, Pawlowski J (2015) Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*, **43**, 2513–2524.

Frago E, Dicke M, Godfray HCJ (2012) Insect symbionts as hidden players in insect-plant interactions. *Trends in Ecology & Evolution*, **27**, 705–711.

Gauthier J-P, Outreman Y, Mieuzet L, Simon J-C (2015) Bacterial communities associated with host-adapted populations of pea aphids revealed by deep sequencing of 16S ribosomal DNA. *PLoS ONE*, **10**, e0120664.

Goecks J, Nekrutenko A, Taylor J, Team TG (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, **11**, R86.

Goodrich JK, Di Rienzi SC, Poole AC *et al*. (2014) Conducting a microbiome study. *Cell*, **158**, 250–262.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, **59**, 307–321.

Harada H, Oyaizu H, Kosako Y, Ishikawa H (1997) *Erwinia aphidicola*, a new species isolated from pea aphid, *Acyrthosiphon pisum*. *Journal of General and Applied Microbiology*, **43**, 349–354.

Hendry TA, Dunlap PV (2014) Phylogenetic divergence between the obligate luminous symbionts of flashlight fishes demonstrates specificity of bacteria to host genera. *Environmental Microbiology Reports*, **6**, 331–338.

Jaenike J (2012) Population genetics of beneficial heritable symbionts. *Trends in Ecology & Evolution*, **27**, 226–232.

Jing X, Wong ACN, Chaston JM *et al*. (2014) The bacterial communities in plant phloem-sap-feeding insects. *Molecular Ecology*, **23**, 1433–1444.

Jones RT, Bressan A, Greenwell AM, Fierer N (2011) Bacterial communities of two parthenogenetic aphid species cocolonizing two host plants across the Hawaiian Islands. *Applied and Environmental Microbiology*, **77**, 8345–8349.

Jousselin E, Desdevises Y, Coeur d'acier A (2009) Fine-scale cospeciation between *Brachycaudus* and *Buchnera aphidicola*: bacterial genome helps define species and evolutionary relationships in aphids. *Proceedings of the Royal Society B: Biological Sciences*, **276**, 187–196.

Jousselin E, Cruaud A, Genson G *et al*. (2013) Is ecological speciation a major trend in aphids? Insights from a molecular phylogeny of the conifer-feeding genus *Cinara*. *Frontiers in Zoology*, **10**, 56.

Katoh K, Kuma K-I, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, **33**, 511–518.

Kembel SW, Wu M, Eisen JA, Green JL (2012) Incorporating 16S gene copy number information improves of microbial diversity and abundance. *PLoS Computational Biology*, **8**, e1002743.

Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, **40**, e3.

Koga R, Tsuchida T, Fukatsu T (2003) Changing partners in an obligate symbiosis: a facultative endosymbiont can compensate for loss of the essential endosymbiont *Buchnera* in an aphid. *Proceedings of the Royal Society of London Series B, Biological Sciences*, **270**, 2543–2550.

Komaki K, Ishikawa H (2000) Genomic copy number of intracellular bacterial symbionts of aphids varies in response to developmental stage and morph of their host. *Insect Biochemistry and Molecular Biology*, **30**, 253–258.

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and Environmental Microbiology*, **79**, 5112–5120.

Kurata S, Kanagawa T, Magariyama Y *et al*. (2004) Reevaluation and reduction of a PCR bias caused by reannealing of templates. *Applied and Environmental Microbiology*, **70**, 7545–7549.

Lamelas A, Perez-Brocal V, Gomez-Valero L *et al*. (2008) Evolution of the secondary symbiont "*Candidatus Serratia symbiotica*" in aphid species of the subfamily Lachninae. *Applied and Environmental Microbiology*, **74**, 4236–4240.

Lamelas A, Gosalbes MJ, Manzano-Marin A *et al*. (2011) Serratia symbiotica from the aphid *Cinara cedri*: a missing link from facultative to obligate insect endosymbiont. *Plos Genetics*, **7**, e1002357.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M (2014) Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, **2**, e593.

Manzano-Marín A, 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 Biology and Evolution*, **6**, 1683–1698.

Michalik A, Szklarzewicz T, Jankowska W, Wieczorek K (2014) Endosymbiotic microorganisms of aphids (Hemiptera: Sternorrhyncha: Aphidoidea): ultrastructure, distribution and transovarial transmission. *European Journal of Entomology*, **111**, 91–104.

Mizrahi-Man O, Davenport ER, Gilad Y (2013) Taxonomic classification of bacterial 16S rRNA Genes using short sequencing reads: evaluation of effective study designs. *PLoS ONE*, **8**, e53608.

Moran NA, Munson MA, Baumann P, Ishikawa H (1993) A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts.

© 2015 John Wiley & Sons Ltd

Proceedings of the Royal Society of London Series B, Biological Sciences, **253**, 167–171.
Oliver KM, Degnan PH, Burke GR, Moran NA (2010) Facultative symbionts in aphids and the horizontal transfer of ecologically important traits. *Annual Review of Entomology*, **55**, 247–266.
Oliver KM, Smith AH, Russell JA (2014) Defensive symbiosis in the real world -advancing ecological studies of heritable, protective bacteria in aphids and beyond. *Functional Ecology*, **28**, 341–355.
Otani S, Mikaelyan A, Nobre T *et al.* (2014) Identifying the core microbial community in the gut of fungus-growing termites. *Molecular Ecology*, **23**, 4631–4644.
Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, **26**, 1641–1650.
Quast C, Pruesse E, Yilmaz P *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, **41**, D590–D596.
Rogers EE, Backus EA (2014) Anterior foregut microbiota of the glassy-winged sharpshooter explored using deep 16S rRNA gene sequencing from individual insects. *PLoS ONE*, **9**, e106215.
Rubin BER, Sanders JG, Hampton-Marcell J *et al.* (2014) DNA extraction protocols cause differences in 16S rRNA amplicon sequencing efficiency but not in community profile composition or structure. *Microbiologyopen*, **3**, 910–921.
Salter SJ, Cox MJ, Turek EM *et al.* (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, **12**, 87.
Schloss PD, Westcott SL, Ryabin T *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**, 7537–7754.
Smith AH, Łukasik P, O'Connor MP *et al.* (2015) Patterns, causes and consequences of defensive microbiome dynamics across multiple scales. *Molecular Ecology*, **24**, 1135–1149.
Snyder AK, McMillen CM, Wallenhorst P, Rio RVM (2011) The phylogeny of *Sodalis*-like symbionts as reconstructed using surface-encoding loci. *FEMS Microbiology Letters*, **317**, 143–151.
Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**, 1596–1599.
Thompson JD, Higgins DG, Gibson TJ (1994) ClustalW: improving sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.
Vanthournout B, Hendrickx F (2015) Endosymbiont dominated bacterial communities in a dwarf spider. *PLoS ONE*, **10**, e0117297.
White JA, Giogini M, Strand MR, Pennacchio F (2013) Arthropod endosymbiosis and evolution. In: *Arthropod Biology and Evolution – Molecules, Development, Morphology*(eds Minelli A., Boxshall G., Fusco G.), pp. 441–477. Springer, Berlin Heidelberg, Germany.

E.J., A.L.C., M.G. and A.C.D.A. designed the study. A.C.D.A. collected and identified aphid specimens. A.L.C. performed the experiment. S.M. and M.B. implemented the Galaxy pipeline and helped analyse the data. B.G. and A.S.M. participated in data analyses. G.B. and F.C. conducted the filtration protocol. E.J., A.L.C., M.G., A.C.D.A. analysed the data and wrote the manuscript. All authors contributed to the final version of the M.S.

Data accessibility

Zip archives for the sequencing results for each PCR sample are available from the Dryad Digital Repository, with a table describing each sample. A table reporting sequence occurrences across samples and the final sequence alignment from the GALAXY pipeline (both files can be used as inputs for the R script) are also available (doi:10.5061/dryad.m8g90). The pipeline of MOTHUR functions is available at http://galaxy-workbench.toulouse.inra.fr/.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Positive controls and collection details for aphid samples (voucher numbers for aphids refer to the aphid colony).

Table S2 Summary of the number of reads and number of unique sequences per sample.

Table S3 Values of the Shannon and Simpson indices for each sample (calculated on unique sequences) and results of Friedman rank tests.

Appendix S1 R Script

Appendix S2 Alignment (fasta file) of *Buchnera* 16s rRNA haplotype

Appendix S3 Alignment (fasta file) of *Serratia* 16s rRNA haplotype

Fig. S1 Number of unique sequences (log-transformed) as a function of the number of sequencing reads (log-transformed).

Fig. S2 Each column represents the mean number of unique sequences obtained for a sample once sequences making up $<x$ % (x ranged from 0.01 to 10) of all the reads were removed.

Fig. S3 16S rRNA gene sequencing-based taxonomic assignments for negative controls.

Fig. S4 Summary of 16S rRNA gene sequencing-based taxonomic assignments for positive controls.

© 2015 John Wiley & Sons Ltd