The draft genome of strain *c*Cpun from biting midges confirms insect *Cardinium* are not a monophyletic group and reveals a novel gene family expansion in a symbiont

Stefanos Siozios[1], Jack Pilgrim[2], Alistair C. Darby[1], Matthew Baylis[2,3] and Gregory D.D. Hurst[1]

[1] Institute of Integrative Biology, Faculty of Health and Life Sciences, University of Liverpool, Liverpool, UK
[2] Institute of Infection and Global Health, Faculty of Health and Life Sciences, University of Liverpool, Liverpool, UK
[3] NIHR Health Protection Research Unit in Emerging and Zoonotic Infections (HPRU-EZI), University of Liverpool, Liverpool, UK

ABSTRACT

Background: It is estimated that 13% of arthropod species carry the heritable symbiont *Cardinium hertigii*. 16S rRNA and gyrB sequence divides this species into at least four groups (A–D), with the A group infecting a range of arthropods, the B group infecting nematode worms, the C group infecting *Culicoides* biting midges, and the D group associated with the marine copepod *Nitocra spinipes*. To date, genome sequence has only been available for strains from groups A and B, impeding general understanding of the evolutionary history of the radiation. We present a draft genome sequence for a C group *Cardinium*, motivated both by the paucity of genomic information outside of the A and B group, and the importance of *Culicoides* biting midge hosts as arbovirus vectors.

Methods: We reconstructed the genome of *c*Cpun, a *Cardinium* strain from group C that naturally infects *Culicoides punctatus*, through Illumina sequencing of infected host specimens.

Results: The draft genome presented has high completeness, with BUSCO scores comparable to closed group A *Cardinium* genomes. Phylogenomic analysis based on concatenated single copy core proteins do not support *Cardinium* from arthropod hosts as a monophyletic group, with nematode *Cardinium* strains nested within the two groups infecting arthropod hosts. Analysis of the genome of *c*Cpun revealed expansion of a variety of gene families classically considered important in symbiosis (e.g., ankyrin domain containing genes), and one set—characterized by DUF1703 domains—not previously associated with symbiotic lifestyle. This protein group encodes putative secreted nucleases, and the *c*Cpun genome carried at least 25 widely divergent paralogs, 24 of which shared a common ancestor in the C group. The genome revealed no evidence in support of B vitamin provisioning to its haematophagous host, and indeed suggests *Cardinium* may be a net importer of biotin.

Discussion: These data indicate strains of *Cardinium* within nematodes cluster within *Cardinium* strains found in insects. The draft genome of *c*Cpun further produces new hypotheses as to the interaction of the symbiont with the midge host,

Submitted 19 September 2018
Accepted 15 January 2019
Published 21 February 2019

Corresponding author
Stefanos Siozios,
siozioss@liverpool.ac.uk

Academic editor
Joseph Gillespie

Additional Information and Declarations can be found on page 20

DOI **10.7717/peerj.6448**

How to cite this article Siozios S, Pilgrim J, Darby AC, Baylis M, Hurst GDD. 2019. The draft genome of strain *c*Cpun from biting midges confirms insect *Cardinium* are not a monophyletic group and reveals a novel gene family expansion in a symbiont. **PeerJ 7:e6448 DOI 10.7717/peerj.6448**

in particular the biological role of DUF1703 nuclease proteins that are predicted as being secreted by *c*Cpun. In contrast, the coding content of this genome provides no support for a role for the symbiont in provisioning the host with B vitamins.

Subjects Evolutionary Studies, Genomics, Microbiology
Keywords *Cardinium hertigii*, *Culicoides* biting midges, Genome sequence, Phylogenomic analysis, Gene family expansion, Heritable symbionts

INTRODUCTION

Invertebrates form a diverse range of symbiotic associations with heritable bacteria, microbes that pass from a female to her progeny. Ranging from less-intimate to highly specialized, these associations can confer novel phenotypic traits on their individual host, and thus may represent major drivers of both ecological and evolutionary dynamics (*McLean et al., 2016*; *Sudakaran, Kost & Kaltenpoth, 2017*; *Ferrari & Vavre, 2011*). Heritable bacteria can supplement the nutritionally imbalanced diet of hematophagous or sap feeding species with vitamins or essential amino acids, thus expanding the niche of the species (*Rio, Attardo & Weiss, 2016*; *Hansen & Moran, 2014*). Other symbionts exert protective effects against biotic or abiotic stress, including natural enemies (predators, parasitoids, fungi, bacteria, and viruses) (*Brownlie & Johnson, 2009*; *Hansen, Vorburger & Moran, 2012*) and heat stress (*Dunbar et al., 2007*). Notably, some heritable bacteria are parasitic and have evolved to manipulate host reproduction to increase the frequency of infected females and facilitate their own transmission (*Hurst & Frost, 2015*). These effects have further prompted their application in vector and pest management (*Iturbe-Ormaetxe, Walker & O' Neill, 2011*).

Cardinium (Bacteroidetes) is a bacterial genus found in a wide range of arthropod species that has a wide variety of impacts on host individuals, including feminization (*Weeks, Marec & Breeuwer, 2001*; *Groot & Breeuwer, 2006*), parthenogenesis induction (*Zchori-Fein et al., 2001*), and cytoplasmic incompatibility (CI) (*Hunter, Perlman & Kelly, 2003*; *Gotoh, Noda & Ito, 2006*; *Perlman, Kelly & Hunter, 2008*; *Ros & Breeuwer, 2009*), alongside the capacity to improve host fitness (*Weeks & Stouthamer, 2004*). First discovered in 1996 (*Kurtti et al., 1996*), it is now estimated that c. 13% of arthropod species carry the symbiont (*Weinert et al., 2015*). *Cardinium* infections are found in a diverse set of arthropods, but its incidence is heterogeneous, with pronounced "hotspots" in arachnids (including spiders, mites, and harvestmen), diaspidid scale insects, parasitoid wasps, planthoppers, whiteflies, and biting midges (*Duron et al., 2008*; *Zchori-Fein & Perlman, 2004*; *Gruwell, Wu & Normark, 2009*; *Nakamura et al., 2009*; *Chang et al., 2010*; *Morag et al., 2012*; *Lewis et al., 2014*; *Mee et al., 2015*). Further symbioses are observed with plant parasitic nematodes (*Noel & Atibalentja, 2006*; *Denver et al., 2016*), copepods (*Edlund et al., 2012*), non-marine ostracods (*Schön et al., 2018*), and oribatid mites (*Konecka & Olszanowski, 2018*) suggesting that the true diversity of the genus is yet to be appreciated. This wider clade

Cardinium represents the sister group to the amoeba symbiont *Amoebophilus asiaticus* (*Nakamura et al., 2009*; *Schmitz-Esser et al., 2010*; *Santos-Garcia et al., 2014*).

Phylogenetic analyses of *Cardinium* based on two gene sequences (*16S* rRNA and *gyr*B) inferred the existence of at least four monophyletic groups designated as A, B, C, and D (*Nakamura et al., 2009*; *Edlund et al., 2012*), resembling *Wolbachia* super-groups in terms of host-affinities (*Lo et al., 2002*). Group A is the largest and the most studied of the three groups and has been found in various arthropod species. Group B has been found in plant parasitic nematodes (*Noel & Atibalentja, 2006*; *Denver et al., 2016*) and is represented by *Cardinium* strains *c*HgTN10, an endosymbiont of the soybean cyst nematode *Heterodera glycines* (*Noel & Atibalentja, 2006*) and *c*Ppe, an endosymbiont of the plant parasitic nematode *Pratylenchus penetrans* (*Brown et al., 2018*). Group C consists of a phylogenetically distinct clade of *Cardinium* strains known only from species of *Culicoides* biting midges, an important group of hematophagous pests and vectors of arboviruses and parasites (*Nakamura et al., 2009*; *Morag et al., 2012*; *Lewis et al., 2014*; *Mee et al., 2015*). Finally, group D have been found as a constituent of the bacterial communities associated with the marine copepod *Nitocra spinipes* (*Edlund et al., 2012*).

To date, genomic characterization has been restricted to A and B group *Cardinium* strains. Three insect-associated A-group *Cardinium* strains have been sequenced. These include the CI-inducing *Cardinium* endosymbiont (*c*Eper1) of the parasitic wasp *Encarsia pergandiella* (*Penz et al., 2012*), the *Cardinium* endosymbiont (*c*BtQ1) of the whitefly *Bemisia tabaci* (*Santos-Garcia et al., 2014*) and the *Cardinium* endosymbiont (*c*Sfur) of the planthopper *Sogatella furcifera* (*Zeng et al., 2018*). These genome sequences have indicated that convergent phenotypes, like CI, have a divergent genetic basis in *Cardinium* from *Wolbachia*. Moreover, the *c*Eper1 *Cardinium* genome suggests the symbiont may supplement B-vitamin provision (*Penz et al., 2012*), a phenotype that would be important in bloodsucking vectors. More recently, the genome sequences for two B group *Cardinium* strains from nematodes have been completed. These are the genomes of the *Cardinium* endosymbiont (*c*HgTN10) from *H. glycines* (*Showmaker et al., 2018*) and the *Cardinium* endosymbiont *c*Ppe from *P. penetrans* (*Brown et al., 2018*). However, there is no available genome for the C clade *Cardinium,* which is particularly notable in the light of the pest and vector status of the host species.

In this paper, we present an annotated draft genome sequence for a *Cardinium* endosymbiont from clade C, carried by the biting midge *Culicoides punctatus*, hereafter *c*Cpun, and use this genome data to estimate the relationship between C clade *Cardinium* and those of A and B groups; improving our understanding of strain relatedness that currently rest on the sequence of two loci. We further use the genome sequence to infer potential aspects of the symbiosis between this microbe and *Culicoides* biting midges. The study of midge symbionts is important, as the symbiosis may potentially have an impact on the physiology of a bloodsucking host, and (by parallel with *Wolbachia*) its vector competence for arboviruses and other pathogens. The difficulty of growing midges in insectary culture has presented a challenge to determining the effect of the symbiont on the host experimentally. Analysis of the *c*Cpun genome and comparison to the previously sequenced *Cardinium* genomes as well as their sister species *A. asiaticus*

(*Schmitz-Esser et al., 2010*) was therefore undertaken to provide insight into the evolution and life style of clade C *Cardinium*.

MATERIALS AND METHODS

Genome sequencing, assembly, and annotation

Culicoides punctatus female midges were collected from Leahurst Campus, University of Liverpool, UK using UV light traps and identified from wing morphology and by cytochrome c oxidase subunit 1 barcoding as in *Pilgrim et al. (2017)*. DNA was extracted from single individuals using the QIAGEN DNAeasy^{TM} Blood & Tissue Kit following the protocol for purification of total DNA from Insect. All samples were tested for *Cardinium* infection using a PCR assay based on 16S rRNA *Cardinium* specific primers Car-sp-F 5′-CGGCTTATTAAGTCAGTTGTGAAATCCTAG-3′; Car-sp-R 5′-TCCTTCCTCCCGCTTACACG-3′ (*Nakamura et al., 2009*). Whole-genome sequencing was carried out by the Centre for Genomic Research, University of Liverpool using the Illumina TruSeq Nano library preparation protocol. Two short-insert (~550 bp insert size) paired-end libraries were constructed from two pooled DNA samples of three individuals each. The libraries were multiplexed and sequenced using 2/3 of a lane on an Illumina HiSeq 2500 platform, yielding 2 × 125 bp paired reads. Adapter removal and quality trimming of the raw Illumina reads were performed with Cutadapt version1.2.1 (*Martin, 2011*) and Sickle version 1.2 (*Joshi & Fass, 2011*).

Identification and filtering of symbiont reads were performed using a similar approach to that used previously (*Pilgrim et al., 2017*). Briefly, a preliminary assembly of the quality trimmed dataset was performed using SPAdes version 3.7.0 (*Nurk et al., 2013*) using the following parameters (-k 21,33,55,77, –careful, –cov-cutoff 5). The initial contigs were visualized using taxon-annotated GC-coverage plots (Fig. S1) with Blobtools (*Kumar et al., 2013*; *Laetsch, 2016*). Additional tblastx searches (*Altschul et al., 1997*; *Camacho et al., 2009*) were conducted against a local genomic database consisting of *Cardinium* genomes—*c*BtQ1 and *c*Eper1 endosymbionts of the whitefly *B. tabaci* and the parasitic wasp *E. pergandiella*, respectively (*Santos-Garcia et al., 2014*; *Penz et al., 2012*), that of *Cardinium* strain *c*HgTN10 from *H. glycines* (*Showmaker et al., 2018*) and the more distantly related *Acanthamoeba* endosymbiont *A. asiaticus* (*Schmitz-Esser et al., 2010*)—with an *e*-value cut-off of $1e^{-6}$. *Cardinium* contigs were extracted and checked for contamination by blastx searches against the non-redundant (nr) protein database. *Cardinium*-specific reads were subsequently retrieved using Bowtie2 (*Langmead & Salzberg, 2012*) and samtools (*Li et al., 2009*) and re-assembled de novo using SPAdes as described above. All contigs larger than 500 bp were checked for potential host or other bacteria contamination using blastx searches against nr database and all contaminant contigs were removed from the final assembly. Subsequently, we evaluated the quality of the assembled contigs using the reference-free assembly validation tool REAPR (*Hunt et al., 2013*). REAPR uses read pairs mapping information to identify potential assembly errors and assign quality scores on each base of the assembly. The error calls were then used to break the pre-assembled contigs at every potential

miss-assembly position using the aggressive option "-a." Finally, the broken assembly was scaffolded using SSPACE (*Boetzer et al., 2011*) using the default parameters.

The *c*Cpun draft genome was annotated using Prokka version 1.12 (*Seemann, 2014*) and the completeness was assessed using BUSCO v3 based on the presence of 148 universal bacterial marker genes (*Simão et al., 2015*). Clusters of Orthologous Groups (COG) functional categories were assigned using the eggNOG database (*Huerta-Cepas et al., 2016*) while additional domains were assigned by searches against the Pfam protein database (*Finn et al., 2016*). The k-mer fraction of the filtered reads were computed with Jellyfish v2.2.3 (*Marçais & Kingsford, 2011*) and used to determine the repeat fraction of *c*Cpun genome using GenomeScope (*Vurture et al., 2017*). Finally, comparison of the repeat density (repeats $\geq$ 200 bp and at least 95% identity) between the Amoebophilaceae genomes was performed using MUMmer-plots (*Kurtz et al., 2004*).

Ortholog identification, comparative, and phylogenetic analyses

The genome sequences of the three available arthropod-associated *Cardinium* strains *Cardinium hertigii* *c*Eper1 (*Penz et al., 2012*), *Cardinium hertigii* *c*BtQ1 (*Santos-Garcia et al., 2014*) and *Cardinium* *c*Sfur (*Zeng et al., 2018*), the two nematode-associated endosymbionts *c*HgTN10 and *c*Ppe (*Showmaker et al., 2018*; *Brown et al., 2018*) and the *Acanthamoeba* endosymbiont *A. asiaticus* (*Schmitz-Esser et al., 2010*) were obtained from GenBank and used for comparative analyses (accession numbers GCF_000304455.1, GCF_000689375.1, GCA_003351905.1, GCA_003176915.1, and GCF_000020565.1, respectively). The genomes of *Cyclobacterium marinum* DSM 745 (GCF_000222485.1) and *Marivirga tractuosa* DSM 4126 (GCF_000183425.1), two free living *Bacteroides* species, were used as outgroup for the phylogenetic analyses (based on *Santos-Garcia et al., 2014*). All GenBank retrieved genomes were re-annotated using Prokka software as described above in order to mitigate the effect of inconsistencies due to alternative annotation practices. Orthologous groups of proteins were identified between *c*Cpun, *c*Eper1, *c*BtQ1, *c*Sfur, *c*HgTN10, *c*Ppe, and *A. asiaticus* using an all-vs-all BLAST search and Markov Cluster (MCL) clustering approach as implemented in OrthoFinder method (*Emms & Kelly, 2015*). Core, accessory and strain-specific orthogroups between the five genomes were visualized with an UpSet plot using the UpSetR package (*Conway et al., 2017*).

Phylogenetic reconstruction was performed on a set of 278 single copy core protein sequences shared between the six *Cardinium* genomes, the genome of *A. asiaticus* and two free living *Bacteroides* species (*Cyclobacterium marinum* and *M. tractuosa*) that were used as outgroup. To this end, a super-matrix was generated by concatenating the protein alignments of the 278 core proteins and trimmed with trimAl version 1.4 (*Capella-Gutiérrez, Silla-Martínez & Gabaldón, 2009*) using the "automated" option. The best substitution model (LG+F+R4) was selected using ModelFinder (*Kalyaanamoorthy et al., 2017*) and phylogenetic inference was performed using the maximum likelihood (ML) criterion as implemented in IQ-TREE v1.6.6 (*Nguyen et al., 2015*). The robustness of the inferred tree was finally assessed with the ultrafast bootstrap approximation method as implemented in IQ-TREE using 1,000 replicates (*Hoang et al., 2018*). Alternative phylogenetic hypotheses were tested by constrained tree

searches using the approximately unbiased (AU) test (*Shimodaira & Goldman, 2002*) as implemented in IQ-TREE v1.6.6. Additionally, the distribution of the phylogenetic signal across the concatenated super-matrix was calculated as described in (*Shen, Hittinger & Rokas, 2017*). Briefly, for each of the 278 core protein alignments the log-likelihood score for the best ML tree topology under concatenation and an alternative conflicting topology was calculated under the same substitution model (LG+F+R4). The difference in the gene-wise log-likelihood scores (ΔGLS) between the two alternative topologies was used as a measure of the phylogenetic signal and to visualize the proportion of core genes supporting each conflicting phylogeny. Finally, an independent phylogenetic analysis was performed on a subset of 46 core ribosomal proteins in IQ-TREE v1.6.6 as described above in order to further test the robustness of our phylogenetic inference. Phylogenetic trees were drawn and annotated online using the EvolView tool (*He et al., 2016*).

Analyses of the DUF1703 gene family expansion

Genome analysis revealed an expansion of the DUF1703 gene family. To analyze this expansion further, a protein sequence alignment of the DUF1703 gene family from *Cardinium* together with selected Open Reading Frames (ORFs) with sequence similarity retrieved as best BLAST hits form NCBI's nr database was performed using MAFFT v7 and default parameters (*Katoh & Standley, 2013*). Ambiguously aligned positions were subsequently removed using trimAl version 1.4 and the "automated" option. A ML phylogenetic analyses was performed with IQ-TREE version 1.6.6 and the phylogenetic tree were constructed and annotated as described above. Additionally, a neighbor-net phylogenetic network was inferred from the translated nucleotide alignment of the *c*Cpun DUF1703 paralogs using SplitsTree version 4.12.6 (*Huson & Bryant, 2006*; *Bryant & Moulton, 2004*) and default parameters. A pairwise identity and similarity matrix of the *c*Cpun DUF1703 amino acid sequence paralogs were constructed using the Needleman–Wunsch global alignment method and the BLOSUM62 substitution matrix as implemented in EMBOSS package (*Rice, Longden & Bleasby, 2000*). Putative signal peptides were predicted on the SignalP 4.1 Server (*Petersen et al., 2011*) using the sensitive D-cutoff settings. Detection of putative recombination events was performed using the RDP4 software package (*Martin et al., 2015*). Recombination Detection Program (RDP) implements several methods for detecting recombination signals including MaxChi (*Smith, 1992*), GENECONV (*Padidam, Sawyer & Fauquet, 1999*), BottScan (*Salminen et al., 1995*), Chimera (*Posada & Crandall, 2001*), and RDP (*Martin & Rybicki, 2000*). Global parameters were as follow: *P*-value cutoff was set to 0.001 using a Bonferroni correction and significance was evaluated from a permutation test based on 1,000 permutations. Detected signals were considered significant only when they were confirmed by multiple methods. Inference of recombination signals can be particularly misleading when diverse sequences are analyzed. To avoid such misalignment artefacts, the 25 complete DUF1703 paralogs were grouped into three groups on the bases of nucleotide sequences similarity (>65%) and the analyses was repeated for each group separately. Finally, the results were also confirmed with PhiPack implementing the pairwise homoplasy index (PHI) algorithm (*Bruen, Philippe &*

Table 1 Genome Features of *c*Cpun draft genome and its closest relatives.

	*c*Cpun	*c*Eper1**	*c*BtQ1**	*c*Sfur	*c*HgTN10	*c*Ppe	*A. asiaticus*
Number of scaffolds	57*	1	11	1	1	27	1
Plasmids	0	1	1	0	0	0	0
Total size in kb	1,137	887 (58)	1,013 (52)	1,103	1,193	1,358	1,884
GC content (%)	33.7	36.6 (31.5)	35 (32)	39.2	38.2	35.8	35
CDS	917	841 (65)	709 (30)	795	974	1,131	1,557
Avg. CDS length (bp)	993	911 (733)	1,033 (1,389)	1,052	997	941	990
Coding density (%)	80	85.5 (82.1)	79.7 (80.1)	75.7	81.4	78.3	81.8
rRNAs	3	3	3	3	3	3	3
tRNAs	37	37	35	35	37	34	35
Ankyrin repeat proteins	46	18-19	26	29	27	32	54
Reference	this study	a	b	c	d	e	f

Notes:
[a] *Penz et al. (2012)*.
[b] *Santos-Garcia et al. (2014)*.
[c] *Zeng et al. (2018)*.
[d] *Showmaker et al. (2018)*.
[e] *Brown et al. (2018)*.
[f] *Schmitz-Esser et al. (2010)*.
* contigs > 500 bp.
** chromosome (plasmid).

Bryant, 2006). Residue composition and conservation within the core nuclease PD-(D/E)XK site of the DUF1703 homologs were illustrated with sequence logos using the Skylign tool (*Wheeler, Clements & Finn, 2014*).

Nucleotide sequence accession numbers

The raw reads and the *c*Cpun draft genome assembly have been submitted to the DDBJ/EMBL/GenBank database under the BioProject accession number PRJNA487198 (WGS project QWJI00000000).

RESULTS AND DISCUSSION

General features of *c*Cpun draft genomes

The final assembly of the *c*Cpun draft genome consists of 57 scaffolds larger than 500 bp (N50 = 41.6 kb, largest scaffold = 116 kb) comprising a total size of 1,137,634 bp (52 scaffolds $\geq$ 1,000 bp) with an average GC content of ~33% and an average depth of coverage 90× (Table 1; Fig. S2). Overall, the *c*Cpun genome shares many characteristics with those of the previously sequenced *Cardinium* strains *c*Eper1, *c*BtQ1, *c*Sfur, cHgTN10, and *c*Ppe including similar genome size of around one Mb and comparable GC content (33.7–38%) (Table 1). No plasmids were inferred based on the presence of scaffolds with atypically higher read coverage compared with the average coverage of the complete assembly, presenting a contrast to the previously sequenced arthropod-associated *Cardinium* (*c*Eper1 and *c*BtQ1) (Table 1; Fig. S2). Nevertheless, we were able to detect several regions with sequence similarity to elements of the two plasmids found in *c*Eper1 and *c*BtQ1. Matching regions were mainly transposases, suggesting that these might be remnants of ancestral plasmid invasion/s.

Although absence of plasmids has also been reported previously for *A. asiaticus,* the sister species of *Cardinium* clade (*Schmitz-Esser et al., 2010*), the presence of low-copy-number plasmids in *c*Cpun cannot be ruled out.

A total of 917 protein coding genes were identified with an average length of 993 bp corresponding to a coding density of around 80% (Table 1; Table S1). *c*Cpun harbors a single set of rRNA genes with the 16S separated from 5S and 23S and encode a complete set of 37 tRNA genes. The identification of 117 out of the 148 BUSCO marker genes (BUSCO score = C: 79% (S: 79%, D: 0%), F: 2.7%, M: 18.2%, *n*: 148) (Fig. S3) was comparable to that observed for the previously sequenced and complete *c*Eper1 *c*Sfur and cHgTN10 genomes, which suggests that *c*Cpun is a near complete genome. Overall, the redundancy in *c*Cpun as assessed through MUMmer-plots is lower than both *A. asiaticus* and *c*BtQ1 previously described as highly repetitive (*Santos-Garcia et al., 2014*) (Fig. S4). K-mer frequency analysis of the Illumina reads estimated the repetitive fraction of *c*Cpun genome to be circa 13%.

Phylogenomic analyses place *c*Cpun as an outgroup of both other insect and nematode *Cardinium* strains

Recently, a new family named Amoebophilaceae was proposed to include the *Cardinium* clades as well as the amoeba-associated *A. asiaticus* (*Santos-Garcia et al., 2014*). Currently, at least four major phylogenetic clades of *Cardinium* related bacteria have been described (*Nakamura et al., 2009*; *Edlund et al., 2012*) with possible evidence for additional clades (*Chang et al., 2010*). However, the phylogenetic (evolutionary) relationships between these clades are not clear. Previous phylogenetic studies based on partial 16S rRNA and gyrB sequences failed to provide a consistent phylogenetic placement for the arthropod and the nematode *Cardinium* clades (*Morag et al., 2012*; *Nakamura et al., 2009*).

We established the relationship of this group across a concatenated set of 278 single copy core protein coding genes as well as a subset of 46 ribosomal protein genes shared between the seven Amoebophilaceae genomes. The results of both analyses clearly support the position of the midge *Cardinium* clade (C) as a sister group to both the other arthropod and nematode *Cardinium* clades (clades A and B) and confirm that the arthropod-associated *Cardinium* do not form a monophyletic group (Fig. 1A). Constrained tree tests for two alternative topologies (a) nematode *Cardinium* as sister group of all other arthropod *Cardinium* and (b) *c*Cpun and nematode *Cardinium* as a monophyletic group resulted in significantly worse trees (AU test, $p < 0.01$). This inference was further supported by analysis of single protein phylogenies (Figs. 1B and 1C). A total of 157 out of the 278 single copy core genes (56%) support the monophyletic grouping of the B group *Cardinium* strains (*c*HgTN10, *c*Ppe) with the A group (*c*Eper1, *c*BtQ1 and *c*Sfur) in exclusion of *c*Cpun ($p < 0.001$, Fisher's exact test). In contrast, only 68 genes (24%) support the monophyletic grouping of *c*Cpun with the A group strains while a small subset of genes ($n = 52$; 19%) supports the monophyletic grouping of *c*Cpun with *c*HgTN10 and *c*Ppe.

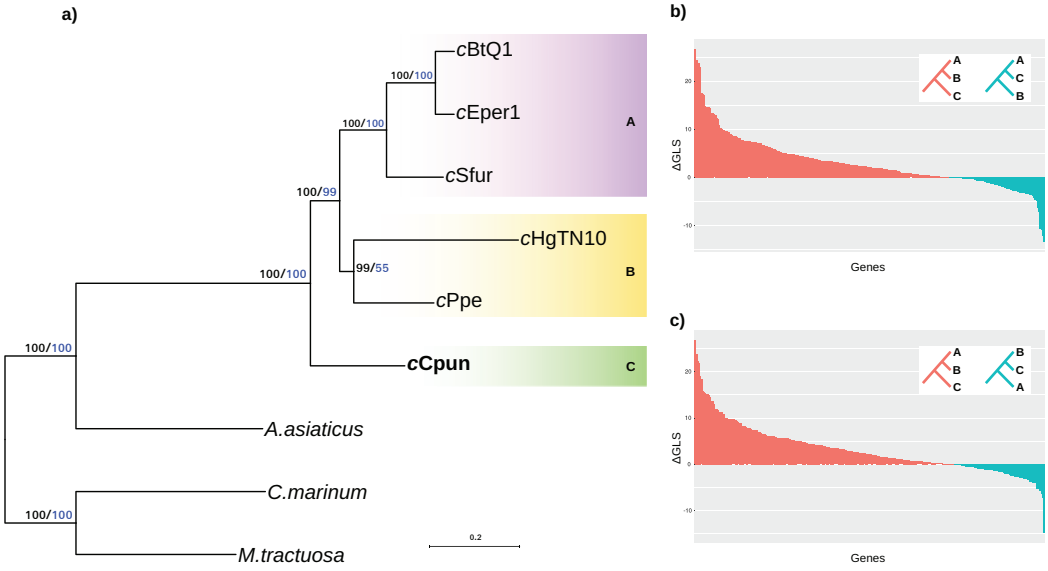

Figure 1 Phylogenetic relationships of *Cardinium* strains. (A) The phylogenetic tree was inferred from the concatenated analysis of 278 single copy core proteins and separately from a subset of 46 core ribosomal proteins using the Maximum Likelihood method as implemented in IQTRE v1.6.6 (model: LG+F+R4). Both datasets retrieved the same tree topology and here we present only the first one. The numbers on the branches represent support values based on 1,000 bootstrap replicates (black bold values: complete matrix; blue values: ribosomal dataset). The three major *Cardinium* groups A, B, and C are denoted with different color shading. *Cyclobacterium marinum* and *Marivirga tractuosa*, two free living members of Bacteroidetes were used as outgroups. (B, C) Distribution of the phylogenetic signal in *Cardinium* concatenated ML phylogeny. The gene-wise differences in log-likelihood scores (ΔGLS) between the concatenated Maximum likelihood tree in (A) versus two alternative topologies: A,C-groups monophyletic relative to B-group (B) and B,C-groups monophyletic relative to A-group (C) were calculated as described in (*Shen, Hittinger & Rokas, 2017*) and plotted in descending order. The red bars represent the genes supporting the Maximum likelihood tree while the blue bars represent the genes supporting each of the alternative topologies. Full-size DOI: 10.7717/peerj.6448/fig-1

Genome content comparisons estimate both a core *Cardinium* genome, genes associated with an insect-symbiont lifestyle, and *c*Cpun specific genes and gene families

The OrthoFinder clustering algorithm identified a total of 2,015 ortholog protein clusters across the seven Amoebophilaceae genomes (*A. asiaticus*, *c*HgTN10, *c*Ppe, *c*Cpun, *c*Eper1, *c*Sfur, and *c*BtQ1). The seven genomes share a core of 415 ortholog clusters of which 278 consist of single-copy genes (Fig. 2). The *c*Cpun genome codes for a substantial number of unique proteins (Fig. 2; Table S2). Specifically, among the 812 ortholog clusters predicted for *c*Cpun, 190 clusters—including 204 protein coding genes—were assigned as strain-specific (Fig. 2). Of these genes, 40 were predicted to code for proteins of less than 70 amino acids and likely represent either annotation artefacts or pseudogenised gene fragments.

The majority of *c*Cpun specific proteins, 138 (~67%), had neither significant matches (blastp, *e*-value $\leq 10^{-10}$) in the NCBI-nr database, nor predicted functional domain. These were assigned as hypothetical proteins. Amongst the remaining 66 predicted *c*Cpun-specific protein clusters, those with ankyrin-repeat domains were particularly well

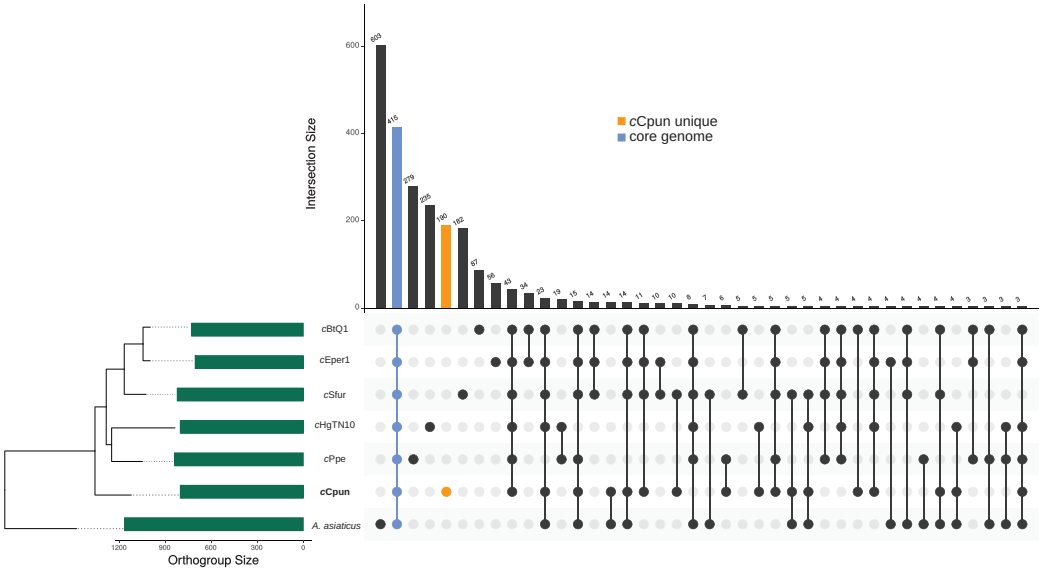

Figure 2 Genome content comparison across the seven *Amoebophilaceae* genomes. UpSet plot showing unique and overlapping protein ortholog clusters across the seven Amoebophilaceae genomes *c*Cpun, *c*Eper1, *c*BtQ1, *c*Sfur, *c*HgTN10, *c*Ppe, and *Amoebophilus asiaticus*. The intersection matrix is sorted in descending order. Green bars represent the orthogroup size for each genome ordered by their phylogenetic relationships. Connected dots represent intersections of overlapping orthogroups while vertical bars shows the size of each intersection. The core orthogroup and the *c*Cpun unique orthogroup cluster are shown with the blue and the orange bars respectively. The plot was generated using UpSetR package in R (*Conway et al., 2017*). Full-size DOI: 10.7717/peerj.6448/fig-2

represented in the strain specific set (Table S2). The abundance, diversity and presumably eukaryotic origin ANK repeat containing proteins has long led them to be considered likely to be involved in symbiotic interactions, and this has been demonstrated in a few cases (*Siozios et al., 2013*; *Nguyen, Liu & Thomas, 2014*; *Voth, 2011*; *Pan et al., 2008*). A total of 46 ANK repeat proteins were present in the *c*Cpun genome, which represents the largest expansion of this gene family in *Cardinium*, comparable to the expansion of this family in *A. asiaticus* (54 ANK proteins) (*Schmitz-Esser et al., 2010*). In total, 20 out of the 46 ankyrin repeat-containing proteins identified in *c*Cpun were not found in the other *Cardinium* strains, suggesting potential host-specific functions. Among the remaining strain-specific protein clusters, 13 were assigned as putative mobile elements (transposases), three putative transporters, a DNA repair protein RecN, two putative GNAT-family acetyltransferases and a homologue of the hemolysin transporter protein ShlB (Table S2). Finally, a folylpolyglutamate synthase (FolC) homologue involved in the tetrahydrofolylpolyglutamate biosynthesis pathway and a putative riboflavin biosynthesis protein RibBA were also detected. Absence of the complete pathway for the de novo biosynthesis of folate in *c*Cpun suggest that FolC probably participates in the folate salvage pathway (folate to polyglutamate) as suggested also by the presence of a dihydrofolate reductase homologue (*De Crécy-Lagard et al., 2007*).

Candidate proteins related to the adaptation of *Cardinium* to arthropod hosts (as opposed to Amoeba and nematode) were identified as being in the four

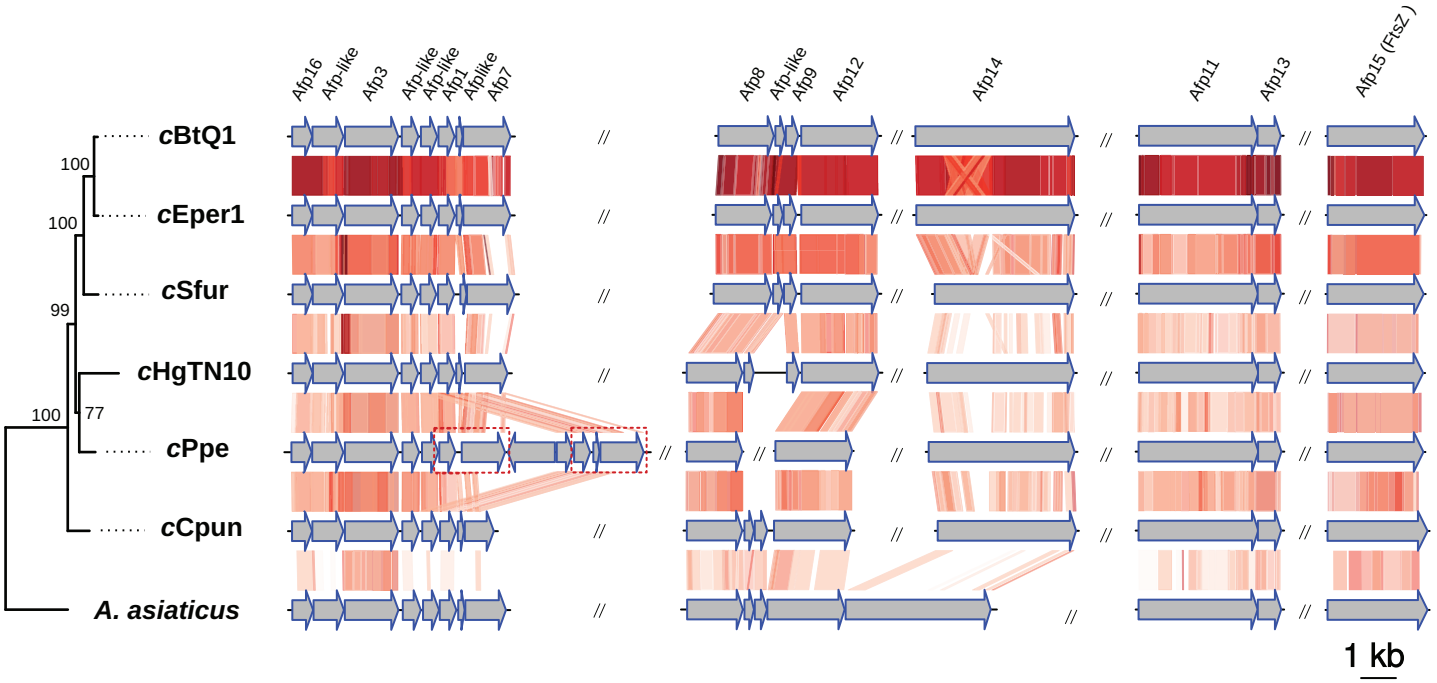

Figure 3 Organization and comparison of the antifeeding prophage (Afp-like) genes clusters in the seven Amoebophilaceae genomes. The phylogeny of the Afp-like secretion system was inferred with Maximum Likelihood based on the concatenated alignment of the 15 constituent protein sequences using IQTREE v1.6.6. Conserved regions are connected with a gradient of red shadings based on tblastx identities. The dash-line rectangles denote a duplicated region in *c*Ppe strain described in *Brown et al. (2018)*. The synteny and the phylogenetic tree of the Afp-like gene clusters were visualized using the genoPlotR package (*Guy, Roat Kultima & Andersson, 2010*). Full-size DOI: 10.7717/peerj.6448/fig-3

arthropod-associated *Cardinium* strains (*c*Cpun, *c*Sfur, *c*Eper1, and *c*BtQ1), and not *Amoebophilus* and the nematode-associated *Cardinium* strains (*c*HgTN10 and *c*Ppe). The four strains from whitefly, wasp, planthopper and midge uniquely share 11 ortholog protein clusters (Fig. 2). Among them we observed the virulence-associated E family protein previously detected in the plasmids harbored by *c*Eper1 and *c*BtQ1 (*Penz et al., 2012*; *Santos-Garcia et al., 2014*) and a nicotinamide mononucleotide transporter.

*c*Cpun possesses both afp-like and type IX secretion systems

Intracellular microbes utilize a variety of specialized protein secretion systems in order to invade and interact with their eukaryote host (*Tseng, Tyler & Setubal, 2009*; *Dale & Moran, 2006*). A common characteristic of the Amoebophilaceae genomes is that all encode for a putative afp-like protein secretion system presumably involved in host-microbe interactions (*Penz, Horn & Schmitz-Esser, 2010*; *Penz et al., 2012*; *Hurst et al., 2007*). This system was also observed in the *c*Cpun genome (Fig. 3) (*Penz, Horn & Schmitz-Esser, 2010*; *Penz et al., 2012*; *Santos-Garcia et al., 2014*). The organization of the AFP-like genes clusters is conserved between the four Amoebophilaceae genomes and suggests operon-like structures (Fig. 3).

We additionally identified seven components of the type IX secretion system (T9SS) in *c*Cpun, a system related to gliding motility and pathogenicity in several members of

the phylum *Bacteroidetes* (*McBride & Zhu, 2013*; *McBride & Nakane, 2015*). *c*Cpun is the third *Cardinium* strain reported to retain components of the T9SS system (*Santos-Garcia et al., 2014*; *Zeng et al., 2018*). Four of these protein clusters with homology to the core components of the T9SS (GldK, GldL, GldM, GldN) are shared between *c*Cpun, *A. asiaticus*, *c*BtQ1, and *c*Sfur while an additional three proteins with homology to the lipoproteins GldD, GldJ, and GldH are uniquely shared between *c*Cpun and *A. asiaticus* with exception the GldJ which was also found in the *c*Sfur genome in two identical copies (Table S3). More recently, core components of the T9SS secretion system were found on the plasmid of *Cardinium* *c*BtQ1 (*Santos-Garcia et al., 2014*).

Originally described in *Flavobacterium johnsoniae*, the T9SS is unique among the phylum *Bacteroidetes* having important role in secretion of proteins involved both in gliding motility and pathogenicity (*McBride & Nakane, 2015*; *Sato et al., 2010*). The presence of the Gld homologs in *c*Cpun as well as *A. asiaticus* supports an ancestral origin of the T9SS machinery which was subsequently lost from *c*Eper1 and the nematode clade (*c*HgTN10 and *c*Ppe). The functional role of the T9SS components in *Cardinium* is unknown. The gene set identified as present in the clade is small compared to that known for active T9SSs (which may have more than 18 components). The low number of genes identified may either reflect co-option of other (unidentified) genes into the secretion process, or a function outside of secretion. However, it is tempting to speculate that the T9SS machinery in Amoebophilaceae has progressively been replaced by the AFP-like protein secretion system. This hypothesis is supported by the complete absence of Gld homologs in both *c*Eper1 and the nematode strains, which suggests that the T9SS is dispensable and likely undergoing gradual loss due to genome reduction processes (*Toft & Andersson, 2010*).

The *c*Cpun genome contains an expansion of the DUF1703 gene family

Expansion and contraction of gene families in microbial genomes constitute a major source of both genetic and functional novelty, contributing to their adaptation to changing environments (*Bratlie et al., 2010*). Despite a tendency for evolution to eliminate redundancy and streamline genomes, endosymbiotic bacteria and intracellular pathogens often contain multi-gene families. Interestingly, the majority of the expanded gene families in these host-associated microbes encode putative effector proteins enriched in eukaryotic domains including ANK, LRR, and TPR repeats, F-box and U-box domains (*Domman et al., 2014*; *Wu et al., 2004*; *Siozios et al., 2013*; *Schmitz-Esser et al., 2010*).

Inspection of the *c*Cpun genome revealed the presence of an expansion of hypothetical proteins related to the DUF1703 protein family (*Knizewski et al., 2007*) not previously observed in other *Cardinium* genomes, or other heritable microbes. A total of 25 gene paralogs coding for hypothetical proteins of this family were identified (Fig. 4). The DUF1703 family contains a group of modular proteins consisting of an N-terminal AAA-ATPase like domain (Pfam ID: PF09820) and a C-terminal PDDEXK_9 nuclease domain (Pfam ID: PF08011). In addition to the 25 paralogs, six genes were found to contain only the AAA-ATPase like domain whilst two genes contained only the nuclease domain (Fig. 4B). All partial genes were detected near the borders of the

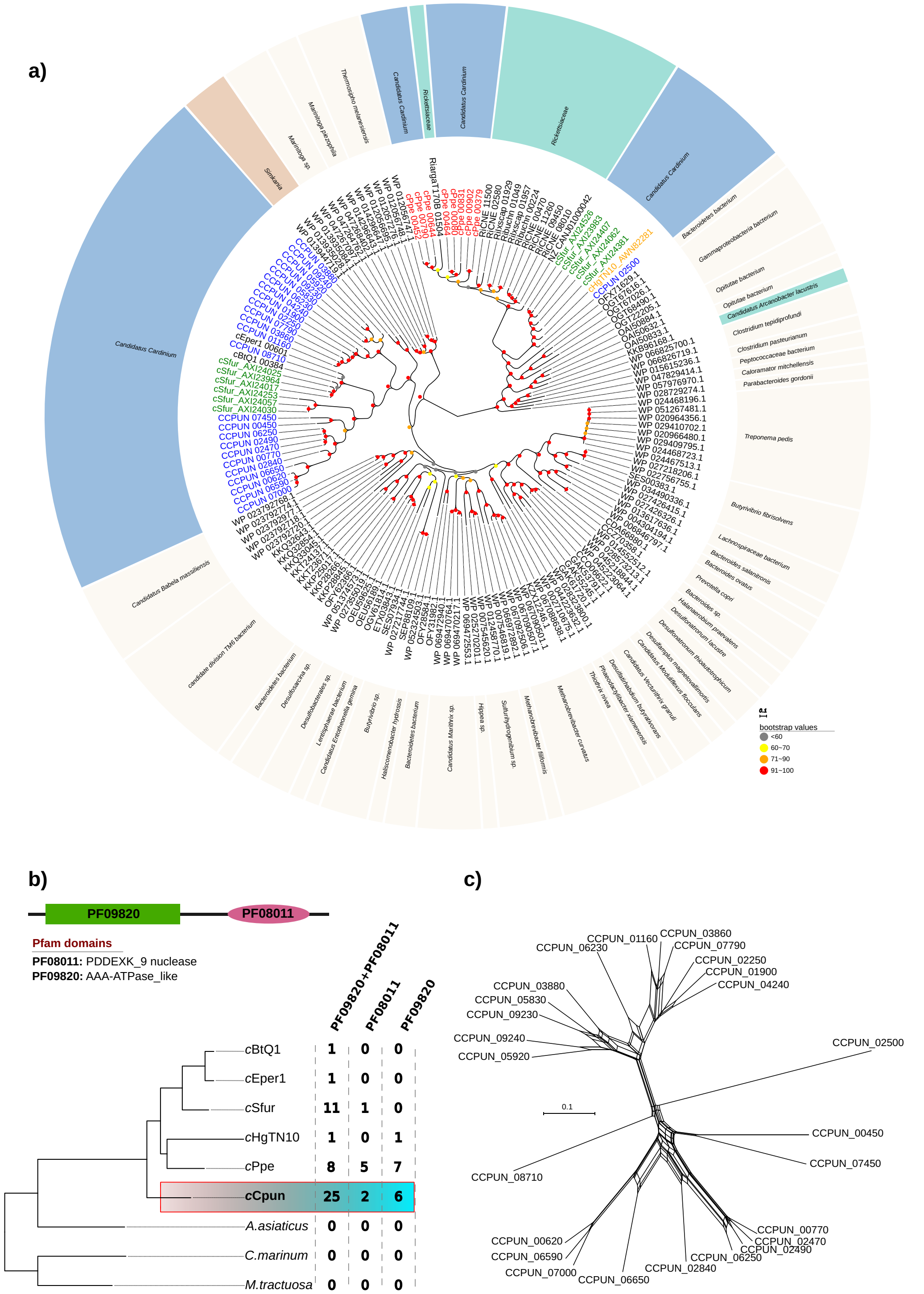

Figure 4 DUF1703 expansion in *c*Cpun genome. (A) Phylogenetic analysis of the *c*Cpun DUF1703 gene family. The unrooted phylogeny was inferred using maximum likelihood from the amino acid sequences of 156 DUF1703 homologs using IQ-TREE v1.6.6 (method: automated best model selection). *Cardinium*, *Simkania*, and *Rickettsia* homologs are shaded in blue, red, and green respectively. (B) The unique expansion of *c*Cpun DUF1703 gene family within the Amoebophilaceae. (C) Phylogenetic network showing the reticulated evolution of the *c*Cpun DUF1703 paralogs.

Full-size DOI: 10.7717/peerj.6448/fig-4

*c*Cpun scaffolds and may be artefactually truncated. Our estimate of gene family size is thus conservative.

The members of the DUF1703 gene family display in *c*Cpun are diverse, as attested by an average amino acid identity of just 39% amongst members (Fig. S5). This extensive divergence of paralogs suggests that the expansion of this gene family is not recent. Moreover, the pairwise comparison suggest at least three main expansion waves (Fig. S5). Phylogenetic analysis indicates that all but one of the *Cardinium* *c*Cpun DUF1703 carrying protein sequences form a single cluster closely related to those found in *Simkania*, an intracellular bacterium member of Chlamidiales known to be associated with protozoa (Fig. 4A). Notably, two of the simkania's paralogs are encoded on its pSn plasmid, suggesting possible roots for horizontal dissemination of the DUF1703 genes. The exception is the gene CCPUN_02500, which forms a distinct group with homologs identified in *Cardinium* strain *c*Sfur and the only intact DUF1703 carrying homolog in cHgTN10, and which is closely related to homologs found in *Rickettsia* and metagenomically-recovered sequences belonging to uncultured members of the Bacteroidetes and Gammaproteobacteria (*Anantharaman et al., 2016*).

Although larger, the expansion of the DUF1703 gene family is not unique to the *c*Cpun genome. Amongst the most recently sequenced *Cardinium* genomes (*c*Sfur and *c*Ppe) we identified smaller expansions of the DUF1703 family (Figs. 4A and 4B). In contrast, the genomes of *c*Eper1, *c*BtQ1, and *c*HgTN10 contain only a single gene homolog whilst no homologs were detected in *A. asiaticus* or free-living relatives (Fig. 4B). Reconstruction of the phylogenetic relationships between the homologs clearly show that members from the same organism group together suggesting that independent expansions took place after divergence from the common ancestor. Surprisingly, the eight paralogs identified in *c*Ppe genome are more closely related to their *Rickettsia* counterparts than the rest of the *Cardinium* homologs, indicating possible independent acquisition. Our results suggest that the DUF1703 genes have probably originated in *Cardinium* after they diverged from *A. asiaticus*, presumably by horizontal gene transfer (HGT) with later expansion in the lineage leading to *c*Cpun, *c*Sfur, and *c*Ppe.

Phylogenetic network analyses revealed several reticulation events within the DUF1703 gene family in *c*Cpun indicating frequent recombination among gene family members (Fig. 4C). We further investigated the extent of recombination using different methods implemented in RDP4 software (*Martin et al., 2015*). Due to the limited sequence similarity between the members of the DUF1703 family we restricted our analyses to group of sequences sharing at least 65–70% nucleotide similarities since misalignment artefacts can confound the identification of true recombination signals. We detected evidence of intragenic recombination in all examined groups with multiple methods (Table S4) suggesting that DUF1703 paralogs in *c*Cpun readily recombine. Despite the extensive recombination, no apparent homogenization between the members of this gene family is observed as suggested by the limited sequence similarity and the absence of monophyletic clustering of *c*Cpun paralogs. Overall, our results point to a HGT scenario for the origin of *Cardinium* DUF1703 gene family with subsequent expansion in the *c*Cpun genome, and variation produced both by mutation and recombination.

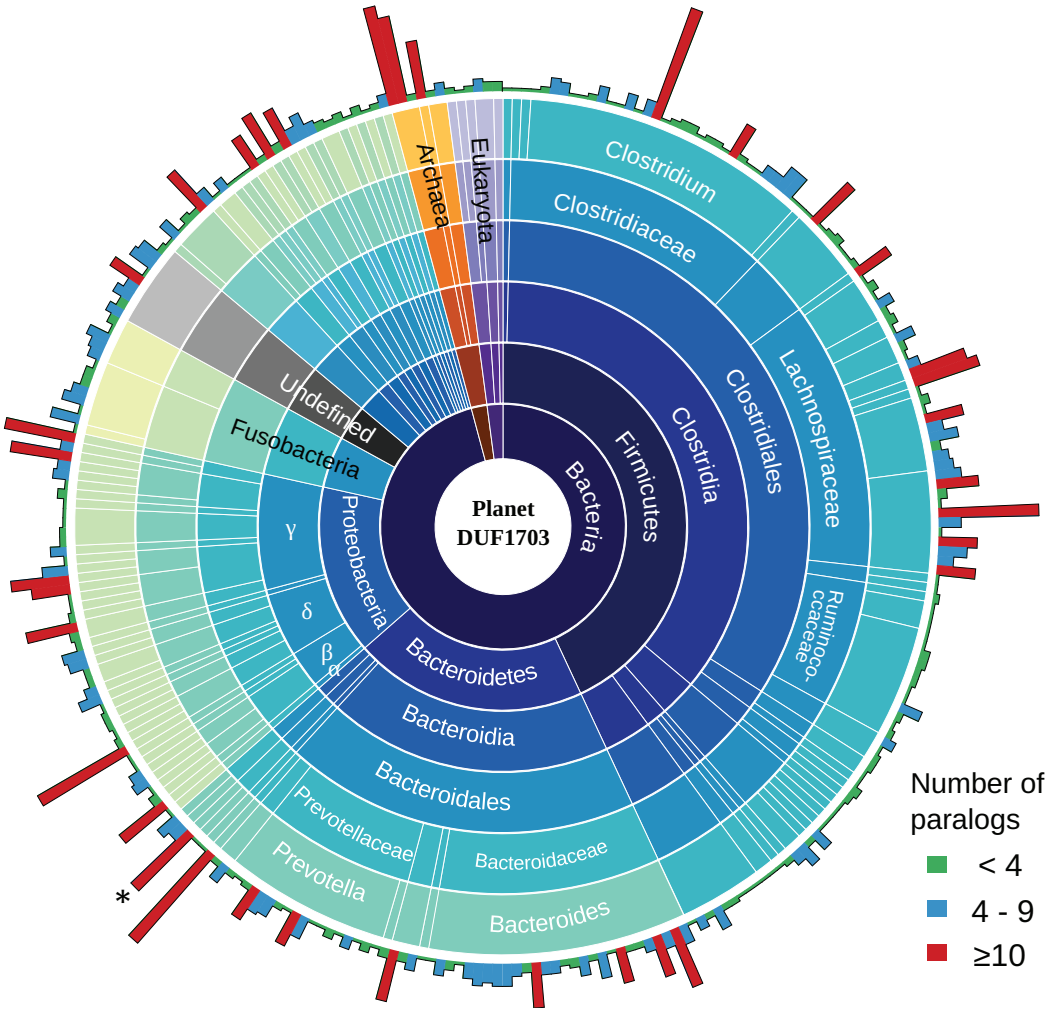

Figure 5 Planet DUF1703. Abundance and taxonomic distribution of DUF1703 proteins in PFAM database. *: cCpun genome. The graph was constructed using Circos v0.69 (*Krzywinski et al., 2009*).
Full-size DOI: 10.7717/peerj.6448/fig-5

To gain a better insight into the role of DUF1703 proteins we sought to investigate the distribution and abundance of proteins containing the AAA-ATPase and PDDEXK_9 domains in other prokaryotes and eukaryotes. We searched the Pfam database for protein sequences containing the two domains and exhibited similar architecture with *Cardinium* homologs. In most cases, DUF1703 containing genes occurred in low copy number per genome. Most species carried fewer than four copies whilst only 9.8% of the species contained 10 copies or more (Fig. 5), ranking *c*Cpun among the species with the largest number of DUF1703 paralogs. Species with higher abundance of DUF1703 paralogs are scattered across the prokaryotic taxonomy suggesting that DUF1703 protein expansion has occurred on multiple occasions within bacteria.

The reason for the expansion of the DUF1703 gene family in *c*Cpun and its putative functional role is yet unknown. It is notable that DUF1703 genes have been also identified in the *Rickettsia* endosymbiont infecting biting midges (*Pilgrim et al., 2017*).

Mirroring the pattern for midge *Cardinium*, the midge *Rickettsia* genome also contains multiple DUF1703 paralogs compared to other *Rickettsia* species with evidence of intragenic recombination ($p < 0.001$, PHI test, 1,000 permutations). However, *Cardinium* *c*Cpun and *Rickettsia* DUF1703 carrying genes are phylogenetically unrelated (Fig. 4A) suggesting independent evolutionary histories, and independent expansion of this gene family in the two groups of midge symbionts. These data suggest this gene family may have a particular function in symbiosis with midges.

The biological role of the DUF1703 is still unclear. A recent transcriptomic study of the *Cardinium* strain *c*Eper1 in its host *E. suzannae* showed that its only DUF1703 gene homolog is moderately transcribed in both sexes (*Mann et al., 2017*). Notably, a putative signal peptide cleavage site was predicted for 10 out of 25 DUF1703 paralogs in *c*Cpun (Table S5) suggesting that they are potentially secreted, acting against DNA/RNA outside of the symbiont. Surprisingly, no signal peptides were detected in any of the paralogs identified in *c*Sfur and *c*Ppe (data not shown). It is noteworthy that an intact DUF1703 homolog of bacterial origin has been previously reported as component of the Maternal-Effect Dominant Embryonic Arrest ("MEDEA") factor, a selfish genetic element reported in *Tribolium castaneum* (*Lorenzen et al., 2008*). PD-(D/E)XK nucleases constitute a large and functionally diverse superfamily of proteins which includes among others restriction endonucleases, Holliday junction resolvases, transposases, and DNA repair enzymes (*Steczkiewicz et al., 2012*). Recently, dual PD-(D/E)XK nuclease domains have been identified in a wide range of toxins from diverse intracellular bacteria (*Gillespie et al., 2018*; *Lindsey et al., 2018*). More interestingly, some of these domains have been directly linked with the induction of reproductive parasitism in the form of CI in *Wolbachia* (*Beckmann, Ronau & Hochstrasser, 2017*). Structural comparison of the PD-(D/E)XK core nuclease site from *Cardinium* and *Rickettsia* DUF1703 homologs and that of the CI-like toxins show considerable differences, especially in the sequence between the catalytic residues (Asp, Glu, and Lys) (Fig. S6). In addition, the AAA-ATPase domain associated with the DUF1703 nuclease is not found in the CI-like toxins of *Wolbachia* and related proteins (*Gillespie et al., 2018*) which might suggest these proteins have different functions. The biological role of *Cardinium* DUF1703 proteins remains to be determined.

Putative horizontal gene transfers as a source of genes in the *c*Cpun genome

Horizontal gene transfer has been previously reported as the source of several genes in *A. asiaticus*, *c*Eper1, and *c*BtQ1 (*Penz et al., 2012*; *Santos-Garcia et al., 2014*; *Schmitz-Esser et al., 2010*). Many of the HGT genes were found to be shared with members of the Alphaproteobacteria that have an intracellular lifestyle, especially species within the *Rickettsiales* order, consistent with HGT within the shared environment of the cell.

In accordance with previous observations of symbiont genomes, our results indicate that HGT has likely shaped the accessory genomes of *c*Cpun (Table 2). The majority of the accessory genes of *c*Cpun for which homologs could be assigned in the database are more similar to corresponding genes of bacterial species outside *Bacteroidetes,* with a bias

Table 2 Example of *c*Cpun genes likely originated from HGTs.

Gene id	Length (AA)	Annotation	Taxonomy of the Best BLAST hit, (GenBank accession)	*E*-value	AA identity (%)
CCPUN_00040	308	Hypothetical protein, putative transposase	*Rickettsia* endosymbiont of *Culicoides newsteadi*, (WP_094649760)	2E-128	64
CCPUN_00530	328	Hypothetical protein, putative transposase	*Rickettsia* endosymbiont of *Culicoides newsteadi*, (WP_094649760)	3E-124	62
CCPUN_01090	346	Hypothetical protein, putative transposase	Rickettsiales bacterium, (PCJ29205)	6E-133	58
CCPUN_02050	379	Hypothetical protein, putative transposase	Rickettsiales bacterium, (PCJ24349)	5E-55	44
CCPUN_04150	328	Hypothetical protein, putative transposase	*Rickettsia* endosymbiont of *Culicoides newsteadi*, (WP_094649760)	9E-125	59
CCPUN_04430	297	Hypothetical protein, putative transposase	Rickettsiales bacterium, (PCJ25778)	9E-136	65
CCPUN_01120	218	Carbonic anhydrase	*Lysobacter sp.* Root494, (WP_056131435)	2E-95	59
CCPUN_03570	551	DNA repair protein RecN	Rickettsiales bacterium, (PCJ29272)	2E-175	48
CCPUN_03900	258	Hypothetical protein, putative transposase	*Candidatus* Paracaedibacter acanthamoebae, (WP_038464592)	3E-114	67
CCPUN_06490	469	Arginine/agmatine antiporter	Gammaproteobacteria bacterium 39-13, (OJV90723)	4E-112	43
CCPUN_07910	266	Chromosome-partitioning protein Spo0J	*Candidatus* Phycorickettsia trachydisci, (WP_106874767)	9E-101	57
CCPUN_07920	327	Sporulation initiation inhibitor protein Soj	*Candidatus* Phycorickettsia trachydisci, (WP_106874768)	5E-135	62
CCPUN_08840	436	Folylpolyglutamate synthase	Wolbachia pipientis, (WP_010963010)	0E+00	76
CCPUN_08910	340	Hypothetical protein	*Rickettsia felis*, (WP_039595314)	2E-155	73
CCPUN_03830	426	Hypothetical protein	*Aedes aegypti*, (XP_001656120)	2E-60	39
CCPUN_08280	1,360	Hypothetical protein	Aedes albopictus, (KXJ68548)	5E-72	27

to genes within the Proteobacteria having closest sequence similarity (Table 2; Fig. S7). For *c*Cpun-specific genes, closest sequence matches lay within bacterial species known to be associated with other arthropods including *Rickettsia* and *Wolbachia*, as well as the amoeba-associated bacteria *Candidatus* Paracaedibacter acanthamoebae and *Candidatus* Jidaibacter acanthamoeba (Table 2). Four of these genes clustered with gene sequences from torix group *Rickettsia*, which are also found in midges. Three of these genes encode putative transposases, and one is a hypothetical protein that in other *Rickettsia* is located on a plasmid hypothesized to be important in determining the host-symbiont interaction (*Gillespie et al., 2015*).

Among the putatively horizontally exchanged gene set were ORFs encoding a carbonic anydrase (CA), an amino acid permease, and a putative chromosome-partitioning protein.

Finally, two *c*Cpun-specific genes encoding hypothetical proteins had their closest homologs within *Aedes* mosquitoes (Table 2; Fig. S7). Notably, the two proteins also have partial similarities with a large ankyrin repeat containing protein (Aasi_1610) previously identified in *A. asiaticus* (*Schmitz-Esser et al., 2010*). Although both *c*Cpun proteins had their ten top hits assigned to *Aedes* sequences, the partial similarities to *A. asiaticus* suggest that they might be fragments of an Aasi_1610 distant homolog. Note, the number of these genes derived from HGT may be even higher since the majority of the accessory genes did not have any significant matches on the GenBank database, and many of these likely represent HGT events from as yet uncharacterized genomes.

The presence of CAs gene is interesting. Among the Amoebophilaceae, CA homologs were detected only in *c*Cpun, *c*Sfur, and *c*HgTN10 and not in other *Cardinium* strains nor *A. asiaticus*, Notably, the three *Cardinium* homologs do not form a monophyletic group, with *c*HgTN10 and *c*Sfur homologs being clustered together and more closely associated with a putative CA previously identified in the *Rickettsia* endosymbiont previously found in biting midges (*Pilgrim et al., 2017*) (Fig. S8). Our results suggest that the *Cardinium* CA homologs have independent evolutionary histories and probably originated from independent horizontal transfer events into the three genomes.

The function of these CAs is not clear. CAs are ancient and ubiquitous multi-class zinc-containing metalloenzymes that catalyze the interconversion of CO_2 to bicarbonate (*Smith & Ferry, 2000*; *Smith et al., 1999*) and are involved in a variety of biochemical processes including respiration and pH homoeostasis (*Gai et al., 2014*). Studies have shown that CAs are essential for microbial growth in free living bacteria under ambient air with low levels of CO_2 (*Mitsuhashi et al., 2003*; *Merlin et al., 2003*; *Kusian, Sültemeyer & Bowien, 2002*). However, whilst CAs are common in many bacterial groups, they are less commonly observed in the genomes of obligate intracellular bacteria (*Ueda, Nishida & Beppu, 2012*). Studies suggest that intracellular pathogens may rely on CAs for virulence and survival within the host cell (*Valdivia & Falkow, 1997*), possibly through regulating the phagosome pH during the infection (*Nishimori et al., 2014*). An intriguing hypothesis is whether CAs might actually play a role in the survival of *Cardinium* outside of the host in comparable way to the role of CAs in free living bacteria, and thus facilitating its horizontal transmission. Interestingly, the plant-mediated horizontal transmission of *Cardinium* bacteria between phloem sap-feeding insects has been previously reported, supporting such a scenario (*Gonella et al., 2015*).

*c*Cpun lacks a biotin or other B-vitamin biosynthetic pathways, indicating it is unlikely to act as a source of these vitamins to its haematophagous host. Indeed, putative homologs of the complete biotin transport system (BioY: CCPUN_01590, BioM: CCPUN_08370, and BioN: CCPUN_08380) were detected, suggesting that *c*Cpun may depend on external provision of biotin from the host. The presence of a complete biotin transporter gene set contrasts with other *Cardinium* genomes, which lack these transporters, but may carry complete operons for the synthesis of biotin, lipoeta and pyridoxal 5'-phosphate (vitamin B6) (*Penz et al., 2012*). Exception is the recently

sequenced strain *c*Sfur which encode for both a biotin transport system and a complete operon for biotin synthesis (*Zeng et al., 2018*).

CONCLUSIONS

In the present study, we expanded the current genomic information from *Cardinium* lineages by presenting a new *Cardinium* draft genome belonging to the divergent and poorly studied group C. Phylogenomic comparison clearly nests the B group nematode-associated *Cardinium* symbionts within the clade A and C symbionts derived from insect strains, indicating that inference previously made on the basis of two gene sequences can now be regarded as supported robustly. The lack of monophyly of strains of *Cardinium* symbiotic with arthropods resembles the pattern for *Wolbachia*, where nematode *Wolbachia* strains are nested within a diverse set of arthropod *Wolbachia* strains (*Gerth et al., 2014*). Heritable microbes occasionally switching between distant host phyla may be more common than previously considered, with the pattern observed in *Wolbachia* (nematode and arthropod infections), torix *Rickettsia* (leech and arthropod lineages) and here in *Cardinium*.

Comparative genomics also provides some insight into whether the three *Cardinium* clades consist different species. The assignment of systematic names in symbiotic bacteria has been a controversial field, owing to the intimate association with their hosts and their ability to exchange genetic material. *Nakamura et al. (2009)* had previously proposed the use of the single species name "*Candidatus Cardinium hertigii*" to describe the three *Cardinium* clades (A, B, C) based on morphological similarities and comparable substitutions in the 16S rRNA gene with other symbiotic bacteria. The paucity of *Cardinium* genomic data and the complete absence of phenotypic information on all but clade-A suggest that is still early to apply an accurate systematic framework. However, the extensive genomic diversity between *Cardinium* clades suggest that *Cardinium* clades may be best described as separate species. Future genomic and phenotypic data will allow us to revise the taxonomy within *Cardinium* lineage.

The presence of *Rickettsia* alongside *Cardinium* in midges presents an opportunity to examine whether the genomes show any convergent properties and if HGT has occurred. Comparison of the gene content of the *c*Cpun *Cardinium* strain with the RiCNE *Rickettsia* symbiont of *Culicoides newsteadi* revealed some similarities. Expansion of the DUF1703 gene family and presence of a carbonic anhydrase gene were notable. However, neither case reflects HGT in the intracellular environment of midges, with the same pattern being independently derived. This separate derivation indicates the possession of these genes may be convergent properties, biologically related to symbiotic life in biting midge hosts, rather than HGT within a shared environment.

Finally, our data indicate that the *Cardinium* symbiont in biting midges is unlikely to serve as a source of B vitamins to its haematophagous host. Contrary to the *c*Eper1 genome, a biotin synthesis system was not observed in the *c*Cpun genome, and indeed the presence of a biotin transporter system indicates the symbiont may in fact be an importer of biotin, and thus a B vitamin sink rather than source. This result perhaps reflects the mixed trophic relationship of biting midges, where larval phases are aquatic

and detritivores, and the adult phase either haematophagous (female) or reliant only on sugar sources (males). It is likely that B vitamins are acquired heterotrophically in the larval phase in sufficient quantities such that selection for symbiont-mediated supplementation is low. Given that a major vector species, including *Culicoides imicola,* harbours *Cardinium* (*Morag et al., 2012*), future work should likely focus on their effects on vectorial capacity alongside the putative facilitation of *Cardinium*-midge interactions from the DUF1703 gene family and carbonic anhydrases.

ACKNOWLEDGEMENTS

The sequencing was carried out at the Centre for Genomic Research, University of Liverpool, United Kingdom. We would like to thank Kenneth Sherlock and Dr Georgette Kluiters for their support with the collection of midge samples.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by a Marie Curie Individual Fellowship (H2020-MSCA-IF-2014) grant 657135 "MIDGESYM" to Stefanos Siozios and a BBSRC DTP studentship to Jack Pilgrim. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Marie Curie Individual Fellowship (H2020-MSCA-IF-2014) grant 657135 "MIDGESYM" to Stefanos Siozios and a BBSRC DTP studentship to Jack Pilgrim.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Stefanos Siozios conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Jack Pilgrim performed the experiments, authored or reviewed drafts of the paper, approved the final draft.
- Alistair C. Darby analyzed the data, authored or reviewed drafts of the paper, approved the final draft.
- Matthew Baylis conceived and designed the experiments, authored or reviewed drafts of the paper, approved the final draft.
- Gregory D.D. Hurst conceived and designed the experiments, contributed reagents/materials/analysis tools, authored or reviewed drafts of the paper, approved the final draft.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

The raw reads and the *c*Cpun draft genome assembly have been submitted to the DDBJ/EMBL/GenBank database under the BioProject accession number PRJNA487198 (WGS project QWJI00000000).

Data Availability

The following information was supplied regarding data availability:

The supermatrix file used for the phylogenomic analysis (Fig. 1) and the alignment files used for the phylogenetic analyses of the DUF1703 and the Carbonic Anhydrase gene families (Fig. 3; Fig. S6, respectively) are provided as supplementary files. Both trimmed and untrimmed versions of the alignment files are provided.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.6448#supplemental-information.

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25(17)**:3389–3402 DOI 10.1093/nar/25.17.3389.

Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications* **7(1)**:13219 DOI 10.1038/ncomms13219.

Beckmann JF, Ronau JA, Hochstrasser M. 2017. A *Wolbachia* deubiquitylating enzyme induces cytoplasmic incompatibility. *Nature Microbiology* **2(5)**:17007 DOI 10.1038/nmicrobiol.2017.7.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27(4)**:578–579 DOI 10.1093/bioinformatics/btq683.

Bratlie MS, Johansen J, Sherman BT, Huang DW, Lempicki RA, Drabløs F. 2010. Gene duplications in prokaryotes can be associated with environmental adaptation. *BMC Genomics* **11(1)**:588 DOI 10.1186/1471-2164-11-588.

Brown AMV, Wasala SK, Howe DK, Peetz AB, Zasada IA, Denver DR. 2018. Comparative genomics of *Wolbachia–Cardinium* dual Endosymbiosis in a Plant-Parasitic Nematode. *Frontiers in Microbiology* **9**:2482 DOI 10.3389/fmicb.2018.02482.

Brownlie JC, Johnson KN. 2009. Symbiont-mediated protection in insect hosts. *Trends in Microbiology* **17(8)**:348–354 DOI 10.1016/j.tim.2009.05.005.

Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172(4)**:2665–2681 DOI 10.1534/genetics.105.048975.

Bryant D, Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* **21(2)**:255–265 DOI 10.1093/molbev/msh018.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10(1)**:421 DOI 10.1186/1471-2105-10-421.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25(15)**:1972–1973 DOI 10.1093/bioinformatics/btp348.

Chang J, Masters A, Avery A, Werren JH. 2010. A divergent Cardinium found in daddy long-legs (Arachnida: Opiliones). *Journal of Invertebrate Pathology* **105(3)**:220–227 DOI 10.1016/j.jip.2010.05.017.

Conway JR, Lex A, Gehlenborg N, Hancock J. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33(18)**:2938–2940 DOI 10.1093/bioinformatics/btx364.

Dale C, Moran NA. 2006. Molecular interactions between bacterial symbionts and their hosts. *Cell* **126(3)**:453–465 DOI 10.1016/j.cell.2006.07.014.

De Crécy-Lagard V, El Yacoubi B, De la Garza RD, Noiriel A, Hanson AD. 2007. Comparative genomics of bacterial and plant folate synthesis and salvage: predictions and validations. *BMC Genomics* **8(1)**:245 DOI 10.1186/1471-2164-8-245.

Denver DR, Brown AMV, Howe DK, Peetz AB, Zasada IA. 2016. Genome skimming: a rapid approach to gaining diverse biological insights into multicellular pathogens. *PLOS Pathogens* **12(8)**:e1005713 DOI 10.1371/journal.ppat.1005713.

Domman D, Collingro A, Lagkouvardos I, Gehre L, Weinmaier T, Rattei T, Subtil A, Horn M. 2014. Massive expansion of ubiquitination-related gene families within the *chlamydiae*. *Molecular Biology and Evolution* **31(11)**:2890–2904 DOI 10.1093/molbev/msu227.

Dunbar HE, Wilson ACC, Ferguson NR, Moran NA. 2007. Aphid thermal tolerance is governed by a point mutation in bacterial symbionts. *PLOS Biology* **5(5)**:e96 DOI 10.1371/journal.pbio.0050096.

Duron O, Hurst GDD, Hornett EA, Josling JA, Engelstädter J. 2008. High incidence of the maternally inherited bacterium Cardinium in spiders. *Molecular Ecology* **17(6)**:1427–1437 DOI 10.1111/j.1365-294X.2008.03689.x.

Edlund A, Ek K, Breitholtz M, Gorokhova E. 2012. Antibiotic-induced change of bacterial communities associated with the copepod *Nitocra spinipes*. *PLOS ONE* **7(3)**:e33107 DOI 10.1371/journal.pone.0033107.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16(1)**:157 DOI 10.1186/s13059-015-0721-2.

Ferrari J, Vavre F. 2011. Bacterial symbionts in insects or the story of communities affecting communities. *Philosophical Transactions of the Royal Society B: Biological Sciences* **366(1569)**:1389–1400 DOI 10.1098/rstb.2010.0226.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* **44(D1)**:D279–D285 DOI 10.1093/nar/gkv1344.

Gai CS, Lu J, Brigham CJ, Bernardi AC, Sinskey AJ. 2014. Insights into bacterial CO2 metabolism revealed by the characterization of four carbonic anhydrases in *Ralstonia eutropha* H16. *AMB Express* **4(1)**:2 DOI 10.1186/2191-0855-4-2.

Gerth M, Gansauge M-T, Weigert A, Bleidorn C. 2014. Phylogenomic analyses uncover origin and spread of the *Wolbachia* pandemic. *Nature Communications* **5(1)**:5117 DOI 10.1038/ncomms6117.

Gillespie JJ, Driscoll TP, Verhoeve VI, Rahman MS, Macaluso KR, Azad AF. 2018. A tangled web: origins of reproductive parasitism. *Genome Biology and Evolution* **10(9)**:2292–2309 DOI 10.1093/gbe/evy159.

Gillespie JJ, Driscoll TP, Verhoeve VI, Utsuki T, Husseneder C, Chouljenko VN, Azad AF, Macaluso KR. 2015. Genomic diversification in strains of *rickettsia felis* isolated from different arthropods. *Genome Biology and Evolution* **7(1)**:35–56 DOI 10.1093/gbe/evu262.

Gonella E, Pajoro M, Marzorati M, Crotti E, Mandrioli M, Pontini M, Bulgari D, Negri I, Sacchi L, Chouaia B, Daffonchio D, Alma A. 2015. Plant-mediated interspecific horizontal transmission of an intracellular symbiont in insects. *Scientific Reports* **5(1)**:15811 DOI 10.1038/srep15811.

Gotoh T, Noda H, Ito S. 2006. Cardinium symbionts cause cytoplasmic incompatibility in spider mites. *Heredity* **98(1)**:13–20 DOI 10.1038/sj.hdy.6800881.

Groot TVM, Breeuwer JAJ. 2006. Cardinium symbionts induce haploid thelytoky in most clones of three closely related *Brevipalpus* species. *Experimental and Applied Acarology* **39(3–4)**:257–271 DOI 10.1007/s10493-006-9019-0.

Gruwell ME, Wu J, Normark BB. 2009. Diversity and phylogeny of *cardinium* (Bacteroidetes) in armored scale insects (Hemiptera: Diaspididae). *Annals of the Entomological Society of America* **102(6)**:1050–1061 DOI 10.1603/008.102.0613.

Guy L, Roat Kultima J, Andersson SGE. 2010. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26(8)**:2334–2335 DOI 10.1093/bioinformatics/btq413.

Hansen AK, Moran NA. 2014. The impact of microbial symbionts on host plant utilization by herbivorous insects. *Molecular Ecology* **23(6)**:1473–1496 DOI 10.1111/mec.12421.

Hansen AK, Vorburger C, Moran NA. 2012. Genomic basis of endosymbiont-conferred protection against an insect parasitoid. *Genome Research* **22(1)**:106–114 DOI 10.1101/gr.125351.111.

He Z, Zhang H, Gao S, Lercher MJ, Chen W-H, Hu S. 2016. Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic Acids Research* **44(W1)**:W236–W241 DOI 10.1093/nar/gkw370.

Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* **35(2)**:518–522 DOI 10.1093/molbev/msx281.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, Von Mering C, Bork P. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research* **44(D1)**:D286–D293 DOI 10.1093/nar/gkv1248.

Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. 2013. REAPR: a universal tool for genome assembly evaluation. *Genome Biology* **14(5)**:R47 DOI 10.1186/gb-2013-14-5-r47.

Hunter MS, Perlman SJ, Kelly SE. 2003. A bacterial symbiont in the *Bacteroidetes* induces cytoplasmic incompatibility in the parasitoid wasp *Encarsia pergandiella*. *Proceedings of the Royal Society of London B: Biological Sciences* **270(1529)**:2185–2190 DOI 10.1098/rspb.2003.2475.

Hurst MRH, Beard SS, Jackson TA, Jones SM. 2007. Isolation and characterization of the *Serratia entomophila* antifeeding prophage. *FEMS Microbiology Letters* **270(1)**:42–48 DOI 10.1111/j.1574-6968.2007.00645.x.

Hurst GDD, Frost CL. 2015. Reproductive parasitism: maternally inherited symbionts in a biparental world. *Cold Spring Harbor Perspectives in Biology* **7(5)**:a017699 DOI 10.1101/cshperspect.a017699.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23(2)**:254–267 DOI 10.1093/molbev/msj030.

Iturbe-Ormaetxe I, Walker T, O' Neill SL. 2011. Wolbachia and the biological control of mosquito-borne disease. *EMBO Reports* **12(6)**:508–518 DOI 10.1038/embor.2011.84.

Joshi NA, Fass JN. 2011. *Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files*. Version 1.33. *Available at https://github.com/najoshi/sickle*.

Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14(6)**:587–589 DOI 10.1038/nmeth.4285.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software Version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30(4)**:772–780 DOI 10.1093/molbev/mst010.

Knizewski L, Kinch LN, Grishin NV, Rychlewski L, Ginalski K. 2007. Realm of PD-(D/E)XK nuclease superfamily revisited: detection of novel families with modified transitive meta profile searches. *BMC Structural Biology* **7(1)**:40 DOI 10.1186/1472-6807-7-40.

Konecka E, Olszanowski Z. 2018. A new *Cardinium* group of bacteria found in *Achipteria coleoptrata* (Acari: Oribatida). *Molecular Phylogenetics and Evolution* **131**:64–71 DOI 10.1016/j.ympev.2018.10.043.

Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Research* **19(9)**:1639–1645 DOI 10.1101/gr.092759.109.

Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M. 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Frontiers in Genetics* **4**:237 DOI 10.3389/fgene.2013.00237.

Kurtti TJ, Munderloh UG, Andreadis TG, Magnarelli LA, Mather TN. 1996. Tick cell culture isolation of an intracellular Prokaryote from the *TickIxodes scapularis*. *Journal of Invertebrate Pathology* **67(3)**:318–321 DOI 10.1006/jipa.1996.0050.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biology* **5(2)**:R12 DOI 10.1186/gb-2004-5-2-r12.

Kusian B, Sültemeyer D, Bowien B. 2002. Carbonic anhydrase is essential for growth of *Ralstonia eutropha* at Ambient CO2 concentrations. *Journal of Bacteriology* **184(18)**:5018–5026 DOI 10.1128/JB.184.18.5018-5026.2002.

Laetsch DR. 2016. *blobtools:blobtools*. v0.9.19.4. *Available at http://doi.org/10.5281/zenodo.61799*.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9(4)**:357–359 DOI 10.1038/nmeth.1923.

Lewis SE, Rice A, Hurst GDD, Baylis M. 2014. First detection of endosymbiotic bacteria in biting midges *Culicoides pulicaris* and *Culicoides punctatus*, important Palaearctic vectors of bluetongue virus. *Medical and Veterinary Entomology* **28(4)**:453–456 DOI 10.1111/mve.12055.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup 1000 Genome Project Data Processing. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25(16)**:2078–2079 DOI 10.1093/bioinformatics/btp352.

Lindsey ARI, Rice DW, Bordenstein SR, Brooks AW, Bordenstein SR, Newton ILG. 2018. Evolutionary genetics of cytoplasmic incompatibility genes cifA and cifB in Prophage WO of Wolbachia. *Genome Biology and Evolution* **10(2)**:434–451 DOI 10.1093/gbe/evy012.

Lo N, Casiraghi M, Salati E, Bazzocchi C, Bandi C. 2002. How many *Wolbachia* Supergroups exist? *Molecular Biology and Evolution* **19(3)**:341–346 DOI 10.1093/oxfordjournals.molbev.a004087.

Lorenzen MD, Gnirke A, Margolis J, Garnes J, Campbell M, Stuart JJ, Aggarwal R, Richards S, Park Y, Beeman RW. 2008. The maternal-effect, selfish genetic element Medea is associated with a composite Tc1 transposon. *Proceedings of the National Academy of Sciences of the United State of America* **105(29)**:10085–10089 DOI 10.1073/pnas.0800444105.

Mann E, Stouthamer CM, Kelly SE, Dzieciol M, Hunter MS, Schmitz-Esser S. 2017. Transcriptome sequencing reveals novel candidate genes for *Cardinium hertigii*-caused cytoplasmic incompatibility and host-cell interaction. *mSystems* **2(6)**:e0014117 DOI 10.1128/mSystems.00141-17.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27(6)**:764–770 DOI 10.1093/bioinformatics/btr011.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17(1)**:10–12 DOI 10.14806/ej.17.1.200.

Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evolution* **1(1)**:vev003 DOI 10.1093/ve/vev003.

Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16(6)**:562–563 DOI 10.1093/bioinformatics/16.6.562.

McBride MJ, Nakane D. 2015. *Flavobacterium* gliding motility and the type IX secretion system. *Current Opinion in Microbiology* **28**:72–77 DOI 10.1016/j.mib.2015.07.016.

McBride MJ, Zhu Y. 2013. Gliding motility and por secretion system genes are widespread among members of the phylum *bacteroidetes*. *Journal of Bacteriology* **195(2)**:270–278 DOI 10.1128/JB.01962-12.

McLean AHC, Parker BJ, Hrček J, Henry LM, Godfray HCJ. 2016. Insect symbionts in food webs. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371(1702)**:20150325 DOI 10.1098/rstb.2015.0325.

Mee PT, Weeks AR, Walker PJ, Hoffmann AA, Duchemin J-B. 2015. Detection of low-level *Cardinium* and *Wolbachia* infections in *Culicoides*. *Applied and Environmental Microbiology* **81(18)**:6177–6188 DOI 10.1128/AEM.01239-15.

Merlin C, Masters M, McAteer S, Coulson A. 2003. Why is carbonic Anhydrase essential to *Escherichia coli*? *Journal of Bacteriology* **185(21)**:6415–6424 DOI 10.1128/JB.185.21.6415-6424.2003.

Mitsuhashi S, Ohnishi J, Hayashi M, Ikeda M. 2003. A gene homologous to β-type carbonic anhydrase is essential for the growth of *Corynebacterium glutamicum* under atmospheric conditions. *Applied Microbiology and Biotechnology* **63(5)**:592–601 DOI 10.1007/s00253-003-1402-8.

Morag N, Klement E, Saroya Y, Lensky I, Gottlieb Y. 2012. Prevalence of the symbiont Cardinium in *Culicoides* (Diptera: Ceratopogonidae) vector species is associated with land surface temperature. *FASEB Journal* **26(10)**:4025–4034 DOI 10.1096/fj.12-210419.

Nakamura Y, Kawai S, Yukuhiro F, Ito S, Gotoh T, Kisimoto R, Yanase T, Matsumoto Y, Kageyama D, Noda H. 2009. Prevalence of *Cardinium* bacteria in Planthoppers and spider mites and taxonomic revision of "Candidatus *Cardinium hertigii*" based on detection of a new Cardinium group from biting midges. *Applied and Environmental Microbiology* **75(21)**:6757–6763 DOI 10.1128/AEM.01583-09.

Nguyen MTHD, Liu M, Thomas T. 2014. Ankyrin-repeat proteins from sponge symbionts modulate amoebal phagocytosis. *Molecular Ecology* **23(6)**:1635–1645 DOI 10.1111/mec.12384.

Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32(1)**:268–274 DOI 10.1093/molbev/msu300.

Nishimori I, Vullo D, Minakuchi T, Scozzafava A, Osman SM, AlOthman Z, Capasso C, Supuran CT. 2014. Anion inhibition studies of two new β-carbonic anhydrases from the bacterial pathogen *Legionella pneumophila*. *Bioorganic & Medicinal Chemistry Letters* **24(4)**:1127–1132 DOI 10.1016/j.bmcl.2013.12.124.

Noel GR, Atibalentja N. 2006. "*Candidatus* Paenicardinium endonii," an endosymbiont of the plant-parasitic nematode *Heterodera glycines* (Nemata: Tylenchida), affiliated to the phylum *Bacteroidetes*. *International Journal of Systematic and Evolutionary Microbiology* **56(7)**:1697–1702 DOI 10.1099/ijs.0.64234-0.

Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, Prjibelsky A, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, McLean J, Lasken R, Clingenpeel SR, Woyke T, Tesler G, Alekseyev MA, Pevzner PA. 2013. Assembling genomes and mini-metagenomes from highly chimeric reads. In: Deng M, Jiang R, Sun F, Zhang X, eds. *Research in Computational Molecular Biology*. Springer Berlin Heidelberg: Lecture Notes in Computer Science, 158–170.

Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new geminiviruses by frequent recombination. *Virology* **265(2)**:218–225 DOI 10.1006/viro.1999.0056.

Pan X, Lührmann A, Satoh A, Laskowski-Arce MA, Roy CR. 2008. Ankyrin repeat proteins comprise a diverse family of bacterial type IV effectors. *Science* **320(5883)**:1651–1654 DOI 10.1126/science.1158160.

Penz T, Horn M, Schmitz-Esser S. 2010. The genome of the amoeba symbiont "Candidatus *Amoebophilus asiaticus*" encodes an afp-like prophage possibly used for protein secretion. *Virulence* **1(6)**:541–545 DOI 10.4161/viru.1.6.13800.

Penz T, Schmitz-Esser S, Kelly SE, Cass BN, Müller A, Woyke T, Malfatti SA, Hunter MS, Horn M. 2012. Comparative genomics suggests an independent origin of cytoplasmic incompatibility in *Cardinium hertigii*. *PLOS Genetics* **8(10)**:e1003012 DOI 10.1371/journal.pgen.1003012.

Perlman SJ, Kelly SE, Hunter MS. 2008. Population biology of cytoplasmic incompatibility: maintenance and spread of cardinium symbionts in a parasitic wasp. *Genetics* **178(2)**:1003–1011 DOI 10.1534/genetics.107.083071.

Petersen TN, Brunak S, Von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8(10)**:785–786 DOI 10.1038/nmeth.1701.

Pilgrim J, Ander M, Garros C, Baylis M, Hurst GDD, Siozios S. 2017. Torix group *Rickettsia* are widespread in *Culicoides* biting midges (Diptera: Ceratopogonidae), reach high frequency and carry unique genomic features. *Environmental Microbiology* **19(10)**:4238–4255 DOI 10.1111/1462-2920.13887.

Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences of the United State of America* **98(24)**:13757–13762 DOI 10.1073/pnas.241370698.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the european molecular biology open software suite. *Trends in Genetics* **16(6)**:276–277 DOI 10.1016/S0168-9525(00)02024-2.

Rio RVM, Attardo GM, Weiss BL. 2016. Grandeur alliances: symbiont metabolic integration and obligate arthropod hematophagy. *Trends in Parasitology* **32(9)**:739–749 DOI 10.1016/j.pt.2016.05.002.

Ros VID, Breeuwer JAJ. 2009. The effects of and interactions between, *Cardinium* and *Wolbachia* in the doubly infected spider mite *Bryobia sarothamni*. *Heredity* **102(4)**:413–422 DOI 10.1038/hdy.2009.4.

Salminen MO, Carr JK, Burke DS, McCutchan FE. 1995. Identification of breakpoints in intergenotypic recombinants of HIV Type 1 by bootscanning. *AIDS Research and Human Retroviruses* **11(11)**:1423–1425 DOI 10.1089/aid.1995.11.1423.

Santos-Garcia D, Rollat-Farnier P-A, Beitia F, Zchori-Fein E, Vavre F, Mouton L, Moya A, Latorre A, Silva FJ. 2014. The genome of *Cardinium* cBtQ1 provides insights into genome reduction, symbiont motility, and its settlement in *Bemisia tabaci*. *Genome Biology and Evolution* **6(4)**:1013–1030 DOI 10.1093/gbe/evu077.

Sato K, Naito M, Yukitake H, Hirakawa H, Shoji M, McBride MJ, Rhodes RG, Nakayama K. 2010. A protein secretion system linked to bacteroidete gliding motility and pathogenesis. *Proceedings of the National Academy of Sciences of the United State of America* **107(1)**:276–281 DOI 10.1073/pnas.0912010107.

Schmitz-Esser S, Tischler P, Arnold R, Montanaro J, Wagner M, Rattei T, Horn M. 2010. The genome of the amoeba symbiont "*Candidatus Amoebophilus asiaticus*" reveals common mechanisms for host cell interaction among amoeba-associated bacteria. *Journal of Bacteriology* **192(4)**:1045–1057 DOI 10.1128/JB.01379-09.

Schön I, Kamiya T, Van den Berghe T, Van den Broecke L, Martens K. 2018. Novel *Cardinium* strains in non-marine ostracod (Crustacea) hosts from natural populations. *Molecular Phylogenetics and Evolution* **130**:406–415 DOI 10.1016/j.ympev.2018.09.008.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30(14)**:2068–2069 DOI 10.1093/bioinformatics/btu153.

Shen X-X, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution* **1(5)**:126 DOI 10.1038/s41559-017-0126.

Shimodaira H, Goldman N. 2002. An approximately unbiased test of phylogenetic tree selection. *Systematic Biology* **51(3)**:492–508 DOI 10.1080/10635150290069913.

Showmaker KC, Walden KKO, Fields CJ, Lambert KN, Hudson ME. 2018. Genome sequence of the soybean cyst nematode (*Heterodera glycines*) Endosymbiont "*Candidatus Cardinium hertigii*" Strain cHgTN10. *Genome Announcements* **6(26)**:e0062418 DOI 10.1128/genomeA.00624-18.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31(9)**:3210–3212 DOI 10.1093/bioinformatics/btv351.

Siozios S, Ioannidis P, Klasson L, Andersson SGE, Braig HR, Bourtzis K. 2013. The diversity and evolution of *Wolbachia* ankyrin repeat domain genes. *PLOS ONE* **8(2)**:e55390 DOI 10.1371/journal.pone.0055390.

Smith JM. 1992. Analyzing the mosaic structure of genes. *Journal of Molecular Evolution* **34(2)**:126–129 DOI 10.1007/BF00182389.

Smith KS, Ferry JG. 2000. Prokaryotic carbonic anhydrases. *FEMS Microbiology Reviews* **24(4)**:335–366 DOI 10.1111/j.1574-6976.2000.tb00546.x.

Smith KS, Jakubzick C, Whittam TS, Ferry JG. 1999. Carbonic anhydrase is an ancient enzyme widespread in prokaryotes. *Proceedings of the National Academy of Sciences of the United State of America* **96(26)**:15184–15189 DOI 10.1073/pnas.96.26.15184.

Steczkiewicz K, Muszewska A, Knizewski L, Rychlewski L, Ginalski K. 2012. Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily. *Nucleic Acids Research* **40(15)**:7016–7045 DOI 10.1093/nar/gks382.

Sudakaran S, Kost C, Kaltenpoth M. 2017. Symbiont acquisition and replacement as a source of ecological innovation. *Trends in Microbiology* **25(5)**:375–390 DOI 10.1016/j.tim.2017.02.014.

Toft C, Andersson SGE. 2010. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nature Reviews Genetics* **11(7)**:465–475 DOI 10.1038/nrg2798.

Tseng T-T, Tyler BM, Setubal JC. 2009. Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. *BMC Microbiology* **9(Suppl 1)**:S2 DOI 10.1186/1471-2180-9-S1-S2.

Ueda K, Nishida H, Beppu T. 2012. Dispensabilities of carbonic anhydrase in Proteobacteria. *International Journal of Evolutionary Biology* **2012**:e324549 DOI 10.1155/2012/324549.

Valdivia RH, Falkow S. 1997. Fluorescence-based isolation of bacterial genes expressed within host cells. *Science* **277(5334)**:2007–2011 DOI 10.1126/science.277.5334.2007.

Voth D. 2011. ThANKs for the repeat. *Cellular Logistics* **1(4)**:128–132 DOI 10.4161/cl.1.4.18738.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33(14)**:2202–2204 DOI 10.1093/bioinformatics/btx153.

Weeks AR, Marec F, Breeuwer JAJ. 2001. A mite species that consists entirely of haploid females. *Science* **292(5526)**:2479–2482 DOI 10.1126/science.1060411.

Weeks AR, Stouthamer R. 2004. Increased fecundity associated with infection by a *Cytophaga*–like intracellular bacterium in the predatory mite, *Metaseiulus occidentalis*. *Proceedings of the Royal Society of London Series B: Biological Sciences* **271(suppl_4)**:S193–S195 DOI 10.1098/rsbl.2003.0137.

Weinert LA, Araujo-Jnr EV, Ahmed MZ, Welch JJ. 2015. The incidence of bacterial endosymbionts in terrestrial arthropods. *Proceedings of the Royal Society B: Biological Sciences* **282(1807)**:20150249 DOI 10.1098/rspb.2015.0249.

Wheeler TJ, Clements J, Finn RD. 2014. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics* **15(1)**:7 DOI 10.1186/1471-2105-15-7.

Wu M, Sun LV, Vamathevan J, Riegler M, Deboy R, Brownlie JC, McGraw EA, Martin W, Esser C, Ahmadinejad N, Wiegand C, Madupu R, Beanan MJ, Brinkac LM, Daugherty SC, Durkin AS, Kolonay JF, Nelson WC, Mohamoud Y, Lee P, Berry K, Young MB, Utterback T, Weidman J, Nierman WC, Paulsen IT, Nelson KE, Tettelin H, O'Neill SL, Eisen JA. 2004. Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: a streamlined genome overrun by mobile genetic elements. *PLOS Biology* **2(3)**:e69 DOI 10.1371/journal.pbio.0020069.

Zchori-Fein E, Gottlieb Y, Kelly SE, Brown JK, Wilson JM, Karr TL, Hunter MS. 2001. A newly discovered bacterium associated with parthenogenesis and a change in host selection behavior in parasitoid wasps. *Proceedings of the National Academy of Sciences of the United State of America* **98(22)**:12555–12560 DOI 10.1073/pnas.221467498.

Zchori-Fein E, Perlman SJ. 2004. Distribution of the bacterial symbiont *Cardinium* in arthropods. *Molecular Ecology* **13(7)**:2009–2016 DOI 10.1111/j.1365-294X.2004.02203.x.

Zeng Z, Fu Y, Guo D, Wu Y, Ajayi OE, Wu Q. 2018. Bacterial endosymbiont *Cardinium c*Sfur genome sequence provides insights for understanding the symbiotic relationship in *Sogatella furcifera* host. *BMC Genomics* **19(1)**:688 DOI 10.1186/s12864-018-5078-y.