A peer-reviewed version of this preprint was published in PeerJ on 21 February 2019.

View the peer-reviewed version (peerj.com/articles/6448), which is the preferred citable publication unless you specifically need to cite this preprint.

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
https://doi.org/10.7717/peerj.6448

The draft genome of strain *c*Cpun from biting midges establishes *Cardinium* as a paraphyletic group, and reveals a novel gene family expansion in a symbiont

Stefanos Siozios [Corresp., 1], **Jack Pilgrim** [1], **Alistair C Darby** [1], **Matthew Baylis** [2, 3], **Gregory DD Hurst** [1]

[1] Institute of Integrative Biology, Faculty of Health and Life Sciences, University of Liverpool, Liverpool, United Kingdom

[2] Institute of Infection and Global Health, Faculty of Health and Life Sciences, University of Liverpool, Liverpool, United Kingdom

[3] NIHR Health Protection Research Unit in Emerging and Zoonotic Infections (HPRU-EZI), University of Liverpool, Liverpool, United Kingdom

Corresponding Author: Stefanos Siozios
Email address: siozioss@liverpool.ac.uk

Background: It is estimated that 13% of arthropod species carry the heritable symbiont *Cardinium hertigii*. 16S rRNA and gyrB sequence divides this species into three clades, with the A group infecting a range of arthropods, the B group infecting nematode worms, and the C group infecting *Culicoides* biting midges. To date, genome sequence has only been available for strains from clade 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 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 revealed that *Cardinium,* as currently considered, is paraphyletic, with strains of *Ca*. Paenicardinium endoni from nematodes 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, of which 24 had a common ancestor in the C group ancestor. 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 *Cardinium,* as currently conceived, to be paraphyletic. The draft genome further produces new hypotheses as to the interaction of the symbiont with the midge host, in particular the biological role of DUF1703 nuclease proteins that are predicted as being secreted by cCpun, but in contrast provides no support for a role for the symbiont in provisioning the host with B vitamins.

The draft genome of strain *c*Cpun from biting midges establishes *Cardinium* as a paraphyletic group, and reveals a novel gene family expansion in a symbiont.

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

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

Corresponding Author:
Stefanos Siozios[1]
Department of Evolution, Ecology and Behaviour, Institute of Integrative Biology, Biosciences Building, University of Liverpool, Liverpool L69 7ZB, United Kingdom
E-mail: siozioss@liverpool.ac.uk
Tel.: +44(0)1517954528

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 three clades, with the A group infecting a range of arthropods, the B group infecting nematode worms, and the C group infecting *Culicoides* biting midges. To date, genome sequence has only been available for strains from clade 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 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 revealed that *Cardinium,* as currently considered, is paraphyletic, with strains of *Ca*. Paenicardinium endoni from nematodes 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, of which 24 had a common ancestor in the C group ancestor. 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 *Cardinium,* as currently conceived, to be paraphyletic. The draft genome further produces new hypotheses as to the interaction of the symbiont with the midge host, in particular the biological role of DUF1703 nuclease proteins that are predicted as being secreted by cCpun, but in contrast provides no support for a role for the symbiont in provisioning the host with B vitamins.

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 sophisticated, these associations can have a major impact on their individual host, and represent major drivers of both ecological and evolutionary dynamics (McLean *et al.* 2016; Sudakaran *et al.* 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 *et al.* 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 et al. 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 and Frost, 2015). These effects have further prompted their application in vector and pest management (Iturbe-Ormaetxe *et al.* 2011).

Cardinium is a member of the *Bacteroidetes* group that is found in a wide range of arthropod species, and which has a wide variety of impacts on host individuals. 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). This symbiont is widely distributed in arthropods, but is heterogeneous in its incidence, with pronounced 'hotspots' in spiders, whiteflies and biting midges (Duron *et al.*, 2008; Zchori-Fein & Perlman 2004; Nakamura *et al.* 2009; Morag *et al.* 2012; Lewis *et al.* 2014; Mee *et al.* 2015). A related strain, commonly termed *Candidatus* Paenicardinium, was described from plant parasitic nematodes (Noel & Atibalentja 2006; Denver *et al.* 2016) with evidence of an additional divergent strain in copepods (Edlund *et al.* 2012). *Cardinium*/*Paenicardinium* form a monophyletic clade with sister relationship to the amoeba symbiont *Amoebophilus asiaticus* (Schmitz-Esser *et al.* 2010; Santos-Garcia *et al.* 2014).

The impact of *Cardinium* on its hosts has been investigated in a number of cases, and reveals reproductive manipulations including cytoplasmic incompatibility in parasitic wasps and several spider-mite species (Hunter *et al.* 2003; Gotoh *et al.* 2006; Perlman *et al.* 2008; Ros & Breeuwer 2009), parthenogenesis induction in parasitic wasps (Zchori-Fein *et al.* 2001) and feminization in

spider mites (Weeks *et al.* 2001; Groot & Breeuwer 2006). Moreover, direct evidence suggests that *Cardinium* may exert fitness effects on certain hosts including increased fecundity in the predatory mite *Metaseiulus occidentalis* (Weeks & Stouthamer 2004). Indirect evidence suggests that the microbe may supplement B-vitamin provision in parasitic wasps (Penz *et al.* 2012).

Phylogenetic analyses based on known *Cardinium 16S* rRNA and *gyr*B gene sequences suggested 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 (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 "c*andidatus"* Paenicardinium endonii, an endosymbiont of the soybean cyst nematode *Heterodera glycines* (Noel & Atibalentja 2006). 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 copepod *Nitocra spinipes* (Edlund *et al.* 2012).

To date, both phenotypic study and genomic characterization has been restricted to A-group *Cardinium* strains. It is in this group that reproductive manipulation phenotypes have been established, and it is from this group that the only two insect-associated *Cardinium* strains have been sequenced. These include the cytoplasmic incompatibility-inducing *Cardinium* endosymbiont (*c*Eper1) of the parasitic wasp *Encarsia pergandiella* (Penz *et al.* 2012) and the *Cardinium* endosymbiont (*c*BtQ1) of the whitefly *Bemisia tabaci* (Santos-Garcia *et al.* 2014). More recently, the genome sequence for B group *Paenicardinium* from *H. glycines* has been completed (Showmaker *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 to estimate the relationship between C clade *Cardinium* and those of A and B groups. 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 impact on the physiology of a blood sucking host, and (by parallel with *Wolbachia*) its vector competence. 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 *Amoebophilus 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. DNA was extracted from single individuals using the QIAGEN DNAeasy™ 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 (CGR), 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 2×125bp 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 and Fass 2011).

Identification and filtering of symbiont reads were performed using a similar approach as we saw before (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 (Supplementary 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 all available *Cardinium* genomes - cBtQ1 and cEper1 endosymbionts of the whitefly *Bemisia tabaci* and the parasitic wasp *Encarsia pergandiella* respectively (Santos-Garcia *et al.* 2014; Penz *et al.* 2012), that of *Ca*. P. endonii (cHgTN10) from *Heterodera glycines* (Showmaker *et al.* 2018) and the more distantly related *Acanthamoeba* endosymbiont *Amoebophilus 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 500bp 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 correctness 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). 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). Finally, an estimation of the repeat density (repeats ≥ 200bp and at least 95% identity) in the *c*Cpun genome was assessed using MUMmer-plots (Kurtz *et al.* 2004).

Ortholog identification, comparative and phylogenetic analyses

The genome sequences of the two available arthropod-associated *Cardinium* strains *Cardinium hertigii c*Eper1 (Penz *et al.* 2012) and *Cardinium hertigii c*BtQ1 (Santos-Garcia *et al.* 2014), the Cardinium endosymbiont of the plant-parasitic nematode *Heterodera glycines canditatus* Paenicardinium endonii (cHgTN10) (Showmaker *et al.* 2018) and the *Acanthamoeba* endosymbiont *Amoebophilus asiaticus* (Schmitz-Esser *et al.* 2010) were obtained from GenBank and used for comparative analyses (accession numbers GCF_000304455.1, GCF_000689375.1, GCA_003176915.1 and GCF_000020565.1 respectively). Finally, 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, *Ca*. P. endonii (cHgTN10) and *Amoebophilus asiaticus* using an all-vs-all BLAST search and 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 338 single copy core protein sequences identified between the four *Cardinium* genomes, the genome of *Amoebophilus asiaticus* and two free living *Bacteroides* species (*Cyclobacterium marinum* and *Marivirga tractuosa*) that were used as outgroup. To this end, a super-matrix was generated by concatenating the protein alignments of the 338 core proteins and trimmed with trimAl version 1.4 (Capella-Gutiérrez *et al.* 2009) using the "automated" option. The best substitution model (LG+F+R5) 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 1000 replicates (Hoang *et al.* 2018). Alternative phylogenetic hypotheses were tested by constrained tree searches using the approximately unbiased (AU) test (Shimodaira *et al.* 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 *et al.* 2017). Briefly, for each of the 338 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+R5). 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 49 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 analyse this expansion further, a protein sequence alignment of the DUF1703 gene family from *Cardinium* together with selected ORFs with sequence similarity retrieved as best BLAST hits form NCBI's NR database was performed using MAFFT v7 and default parameters (Katoh and Standley 2013). Ambiguously aligned positions were subsequently removed using trimAl version 1.4 and the "automated" option. A maximum likelihood (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 neighbour-net phylogenetic network was inferred from the translated nucleotide alignment of the cCpun 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 *et al.*, 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). RDP implements several methods for detecting recombination signals including MaxChi (Smith 1992), GENECONV (Padidam *et al.* 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 1000 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 analysed. To avoid such misalignment artefacts, the 25 complete DUF1703 paralogs were grouped into 3 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 *et al.*, 2006).

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 ≥ 1000 bp) with an average GC content of ~33% and an average depth of coverage 90X (Table 1, Supplementary Fig. S2). Overall, the *c*Cpun genome shares many characteristics with those of the previously sequenced *Cardinium* strains *c*Eper1, *c*BtQ1, and *Ca*. P. endonii (cHgTN10) including similar genome size of around 1 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, Supplementary 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, Supplementary Table S1). *c*Cpun harbours 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] (Supplementary Fig. S3) was comparable to that observed for the previously sequenced and complete *c*Eper1 and *Ca*. P. endonii (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 cBtQ1 previously described as highly repetitive (Santos-Garcia *et al.* 2014) (Supplementary Fig. S4). However, the draft nature of the assembly and the effect of repeat-collapsing during the assembly process may have led to the repeat-content obtained for *c*Cpun to be underestimated.

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

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 338 single copy core protein coding genes as well as a subset of 49 ribosomal protein genes shared between the five *Amoebophilaceae* genomes. The results of both analyses clearly support the position of the midge *Cardinium* clade as a sister group to both the other arthropod *Cardinium* and *Ca*. Paenicardinium nematode symbiont clade represented by cHgTN10 (Fig. 1a). *Cardinium* is thus paraphyletic, with *Ca.* P. endonii nested within the clade. Constrained tree tests for two alternative topologies (a) *Ca. Paenicardinium* as sister group of all other arthropod *Cardinium* and (b) *c*Cpun and *Ca. Paenicardinium* 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 (Fig. 1b and 1c). A total of 180 out of the 338 single copy core genes (53%) support the monophyletic grouping of *Ca.* P. endonii with *c*Eper1 and *c*BtQ1 in exclusion of *c*Cpun ($p < 0.001$, Fisher's exact test). In contrast, only 105 genes (31%) support the monophyletic grouping of *c*Cpun with *c*Eper1 and *c*BtQ1 while a small subset of genes (n=53; 16%) supports the monophyletic grouping of *c*Cpun with *Ca.* P. endonii.

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 2015 ortholog protein clusters across the five *Amoebophilaceae* genomes (*A. asiaticus*, *Ca*. P. endonii, *c*Cpun, *c*Eper1, and *c*BtQ1). The four genomes share a core of 442 ortholog clusters of which 338 consist of single-copy

genes (Fig. 2). The *c*Cpun genome codes for a substantial number of unique proteins (Fig. 2, Supplementary Table S2). Specifically, among the 812 ortholog clusters predicted for *c*Cpun, 224 clusters - including 241 protein coding genes - were assigned as strain-specific (Fig. 2). Of these genes, 43 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, 156 (~65%), had no significant matches (E-value $\leq 10^{-10}$) in the NCBI-nr database or functional domains and were assigned as hypothetical proteins. Amongst the remaining 85 predicted *c*Cpun-specific protein clusters, those with ankyrin-repeat domains were particularly well represented in the strain specific set (Supplementary Table S2). ANK repeat containing proteins have been long thought - and in a few cases shown - to be involved in symbiotic interactions due to their abundance, diversity and presumably their eukaryotic origin (Siozios *et al.* 2013; Nguyen *et al.* 2014; Voth 2011; Pan *et al.* 2008). Forty-six 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, 27 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, 18 were assigned as putative mobile elements (transposases), 4 putative transporters including the BioMN biotin transport module, a DNA repair protein RecN, two putative GNAT-family acetyltransferases and a homologue of the hemolysin transporter protein ShlB (Supplementary Tables 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 three arthropod-associated *Cardinium* strains (*c*Cpun, *c*Eper1 and *c*BtQ1), and not *Amoebophilus* and *Paenicardinium*. The three strains from whitefly, wasp and midge uniquely share 13 ortholog protein clusters (Fig. 2). Among them we found the virulence-associated E family protein previously detected in the

plasmids harboured by *c*Eper1 and *c*BtQ1 (Penz *et al.* 2012; Santos-Garcia *et al.* 2014), a Lysozyme M1 homolog, a nicotinamide mononucleotide transporter and a putative peptidase.

_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 *et al.* 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 *et al.* 2012, 2010; Hurst *et al.* 2007). This system was also observed in the *c*Cpun genome (Fig. 3) (Penz *et al.* 2010, 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 second *Cardinium* strain reported to retain components of the T9SS system (Santos-Garcia *et al.* 2014). 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*, and *c*BtQ1 while an additional three proteins with homology to the lipoproteins GldD, GldJ and GldH are uniquely shared between *c*Cpun and *A. asiaticus* (Supplementary 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 *Ca*. P. endonii. 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 Type IX secretion systems (which may have more than 18 components). The low number of genes identified may either reflect cooption of other (unidentified) genes into

the secretion process, or a function outside of secretion. 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 *Ca*. P. endonii, 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 observed in other *Cardinium* genomes, or other heritable microbes. 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 *c*Cpun scaffolds and may be artefactually truncated. Thus our estimate of gene family size is 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 (Supplementary 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 (Supplementary

Fig. S5). Phylogenetic analysis indicates that all but one of the *Cardinium* 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). The exception is the gene CCPUN_02500, which forms a distinct group with the only intact DUF1703 carrying homolog in *Ca*. P. endonii, 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).

The expansion of the DUF1703 gene family is unique to the *c*Cpun genome amongst sequenced genomes; *c*Eper1, *c*BtQ1 and *Ca*. P. endonii contain only a single gene homolog whilst no homologs were detected in *A. asiaticus* or free-living relatives (Fig. 4b). Our results suggest that the DUF1703 genes have originated in *Cardinium* after they diverged from *A. asiaticus*, presumably by HGT with later expansion in the lineage leading to *c*Cpun.

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 (Supplementary 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.

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 (data not shown). However, *Cardinium* 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 *Encarsia 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 (Supplementary Table S5) suggesting that they potentially secreted, acting against DNA/RNA outside of the symbiont. It is noteworthy that an intact DUF1703 homolog of bacterial origin has been reported as component of the Maternal-Effect Dominant Embryonic Arrest ("Medea") factor, a selfish genetic element reported in *Tribolium castaneum* (Lorenzen *et al.*, 2008). More recently, the DUF1703 PDDEXK_9 nuclease domain has been identified in one of the proteins likely associated with Cytoplasmic Incompatibility (CI) in *w*Pip *Wolbachia* strain (CinB) (Beckmann *et al.* 2017).

Horizontal gene transfer as a source of genes in the *c*Cpun genome

Horizontal gene transfer (HGT) 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 *Rickettsialles* order, consistent with HGT within the shared environment of the cell.

In line with the 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 to genes within the Proteobacteria having closest sequence similarity (Table 2). For *c*Cpun-specific genes, closest sequence matches lay within bacteria 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). Among these putatively horizontally exchanged genes were genes encoding for putative transposases, a carbonic anydrase (CA), an amino acid permease, a putative chromosome-partitioning protein and three transporters including homologs of the Biotin transport ATP-binding protein BioM and BioN permease protein which belong to the BioMNY biotin transport complex. Finally, two *c*Cpun-specific genes encoding hypothetical proteins had their closest homologs within *Aedes* mosquitoes (Table 2). 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 uncharacterised genomes.

The presence of carbonic anhydrase (CAs) gene is interesting. Among *Amoebophilaceae,* CA homologs were detected only in *c*Cpun and *Ca*. P. endonii and not in other *Cardinium* strains nor *A. asiaticus*, Notably, the *c*Cpun and *Ca*. P. endonii CA copies are not monophyletic, with *Ca*. P endonii homolog being more closely associated with a putative CA previously identified in the *Rickettsia* endosymbiont previously found in biting midges (Pilgrim *et al.* 2017) (Supplementary Fig. S6). Our results suggest that the *Cardinium* CA homologs have independent evolutionary histories and probably originated from independent horizontal transfer events into the two 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 *et al.* 2002). However, whilst CAs are common in many bacterial groups, they are less commonly observed in the genomes of obligate intracellular bacteria (Ueda *et al.* 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).

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). *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. Interestingly, the BioM and BioN transporters were likely derived by independent HGT events since no homologs were detected in the rest of the *Amoebophilaceae*. The BioM homolog shares 62% amino acid identities with *Erwinia amilovora* while BioN shares 41% identities with *Bartonella*.

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 nematode *Ca.* Paenicardinium symbiont within the symbionts derived from insect strains. This paraphyly resembles that for *Wolbachia*, where nematode *Wolbachia* strains are nested within a diverse set of arthropod *Wolbachia* strains. It is clear that heritable microbes occasionally switching between distant host phyla may be more common than previously considered. The pattern is seen in *Wolbachia* (nematode and arthropod infections), torix *Rickettsia* (leech and arthropod lineages) and here in Cardinium *sensu lato*.

The ordering of these strains, alongside complete or draft genomes, enables a more nuanced picture of evolution in the genus to be established. Comparison of the genome content between the three *Cardinium* strains as well as the genome of *A. asiaticus* revealed an extensive accessory genome associated with each *Cardinium* clade (group). Although the three *Cardinium* genomes contain similar number of coding sequences, their accessory genome differs considerably. Among them *c*Cpun contains the largest number of strain-specific genes. Notable are a greater number of genes in the ANK family of proteins compared to the other insect symbiotic strains, and the expansion of the DUF1703 nuclease family of genes in the *c*Cpun genome. The diversification of the DUF1703 gene family is evolutionarily old – notwithstanding two conserved motifs, the sequence similarity amongst the paralogs is low. The presence of a predicted signal peptide makes it likely these nuclease genes function in symbiosis within midges, but it is not clear what these functions might be.

An interesting question arising is 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. Recently, the validity of a species framework within *Wolbachia* clade has become the subject of considerable debate among the *Wolbachia* research community (Ramírez-Puebla *et al.* 2015; Lindsey *et al.* 2016; Ramírez-Puebla *et al.* 2016). *Wolbachia* is currently defined as a single species named "*Wolbachia pipientis*" classified in at least 16 divergent supergroups (Glowska *et*

539 *al.* 2015), with this single species designation persisting despite the observation that some of
540 these supergroups have been irreversibly separated suggesting that they might consist separate
541 species (Ellegaard *et al.* 2013). Nakamura *et al.* had previously proposed the use of the single
542 species name "*Candidatus* Cardinium hertigii" to describe the three *Cardinium* clades (A, B, C)
543 based on morphological similarities and comparable substitutions in the 16S rRNA gene with
544 other symbiotic bacteria (Nakamura *et al.* 2009). The paucity of *Cardinium* genomic data and the
545 complete absence of phenotypic information on all but clade-A suggest that is still early to apply
546 an accurate systematic framework. However, the extensive genomic diversity between
547 *Cardinium* clades suggest that *Cardinium* clades may actually consist of separate species. Future
548 genomic and phenotypic data will allow us to revise the taxonomy within *Cardinium* lineage.
549
550 The presence of *Rickettsia* alongside *Cardinium* in midges presents an opportunity to examine
551 whether the genomes show any convergent properties and if HGT has occurred. Comparison of
552 the gene content of the *c*Cpun *Cardinium* strain with the RiCINE Rickettsia symbiont of *C.*
553 *newsteadi* revealed some similarities. Expansion of the DUF1703 gene family and presence of a
554 carbonic anhydrase gene were notable. However, neither case reflects HGT in the intracellular
555 environment of midges, with the same pattern being independently derived. This separate
556 derivation indicates the possession of these genes may be biologically related to symbiotic life in
557 biting midge hosts, rather than HGT within a shared environment.
558
559 Finally, our data indicate that the *Cardinium* symbiont in biting midges is unlikely to serve as a
560 source of B vitamins to its haematophagous host. Contrary to the *c*Eper1 genome, a biotin
561 synthesis system was not observed in the *c*Cpun genome, and indeed the presence of a biotin
562 transporter system indicates the symbiont may in fact be an importer of biotin, and thus a B
563 vitamin sink rather than source. This result perhaps reflects the mixed trophic relationship of
564 biting midges, where larval phases are aquatic and detritivores, and the adult phase either
565 haematophagous (female) or reliant only on sugar sources (males). It is likely that B vitamins are
566 acquired heterotrophically in the larval phase in sufficient quantities that selection for symbiont-
567 mediated supplementation is low.

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 Georgette Kluiters for their support with the collection of midge samples.

572 **References**

573

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

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

581 Beckmann JF., Ronau JA., Hochstrasser M. 2017. A *Wolbachia* deubiquitylating enzyme
582 induces cytoplasmic incompatibility. *Nature Microbiology* 2:17007. DOI:
583 10.1038/nmicrobiol.2017.7.

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

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

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

591 Bruen TC., Philippe H., Bryant D. 2006. A Simple and Robust Statistical Test for Detecting the
592 Presence of Recombination. *Genetics* 172:2665–2681. DOI: 10.1534/genetics.105.048975.

593 Bryant D., Moulton V. 2004. Neighbor-Net: An Agglomerative Method for the Construction of
594 Phylogenetic Networks. *Molecular Biology and Evolution* 21:255–265. DOI:
595 10.1093/molbev/msh018.

596 Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden TL. 2009.
597 BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. DOI: 10.1186/1471-2105-
598 10-421.

599 Capella-Gutiérrez S., Silla-Martínez JM., Gabaldón T. 2009. trimAl: a tool for automated
600 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973. DOI:
601 10.1093/bioinformatics/btp348.

602 Chang J., Masters A., Avery A., Werren JH. 2010. A divergent Cardinium found in daddy long-
603 legs (Arachnida: Opiliones). *Journal of Invertebrate Pathology* 105:220–227. DOI:
604 10.1016/j.jip.2010.05.017.

605 Conway JR., Lex A., Gehlenborg N., Hancock J. 2017. UpSetR: an R package for the
606 visualization of intersecting sets and their properties. *Bioinformatics* 33:2938–2940. DOI:
607 10.1093/bioinformatics/btx364.

608 Dale C., Moran NA. 2006. Molecular Interactions between Bacterial Symbionts and Their Hosts.
609 *Cell* 126:453–465. DOI: 10.1016/j.cell.2006.07.014.

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

613 Denver DR., Brown AMV., Howe DK., Peetz AB., Zasada IA. 2016. Genome Skimming: A
614 Rapid Approach to Gaining Diverse Biological Insights into Multicellular Pathogens. *PLOS*
615 *Pathog* 12:e1005713. DOI: 10.1371/journal.ppat.1005713.

616 Domman D., Collingro A., Lagkouvardos I., Gehre L., Weinmaier T., Rattei T., Subtil A., Horn
617 M. 2014. Massive Expansion of Ubiquitination-Related Gene Families within the Chlamydiae.
618 *Molecular Biology and Evolution* 31:2890–2904. DOI: 10.1093/molbev/msu227.

619 Dunbar HE., Wilson ACC., Ferguson NR., Moran NA. 2007. Aphid Thermal Tolerance Is
620 Governed by a Point Mutation in Bacterial Symbionts. *PLOS Biology* 5:e96. DOI:
621 10.1371/journal.pbio.0050096.

622 Duron O., Hurst GDD., Hornett EA., Josling JA., Engelstädter J. 2008. High incidence of the
623 maternally inherited bacterium Cardinium in spiders. *Molecular Ecology* 17:1427–1437. DOI:
624 10.1111/j.1365-294X.2008.03689.x.

625 Edlund A., Ek K., Breitholtz M., Gorokhova E. 2012. Antibiotic-Induced Change of Bacterial
626 Communities Associated with the Copepod Nitocra spinipes. *PLOS ONE* 7:e33107. DOI:
627 10.1371/journal.pone.0033107.

628 Ellegaard KM., Klasson L., Näslund K., Bourtzis K., Andersson SGE. 2013. Comparative
629 Genomics of Wolbachia and the Bacterial Species Concept. *PLOS Genet* 9:e1003381. DOI:
630 10.1371/journal.pgen.1003381.

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

634 Ferrari J., Vavre F. 2011. Bacterial symbionts in insects or the story of communities affecting
635 communities. *Philosophical Transactions of the Royal Society B: Biological Sciences* 366:1389–
636 1400. DOI: 10.1098/rstb.2010.0226.

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

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

644 Glowska E., Dragun-Damian A., Dabert M., Gerth M. 2015. New Wolbachia supergroups
645 detected in quill mites (Acari: Syringophilidae). *Infection, Genetics and Evolution* 30:140–146.
646 DOI: 10.1016/j.meegid.2014.12.019.

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

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

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

654 Hansen AK., Vorburger C., Moran NA. 2012. Genomic basis of endosymbiont-conferred
655 protection against an insect parasitoid. *Genome Research* 22:106–114. DOI:
656 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: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* 35:518–522. 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: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: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* 270:2185–2190. DOI: 10.1098/rspb.2003.2475.

Hurst GDD., Frost CL. 2015. Reproductive Parasitism: Maternally Inherited Symbionts in a Biparental World. *Cold Spring Harbor Perspectives in Biology* 7:a017699. DOI: 10.1101/cshperspect.a017699.

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

Huson DH., Bryant D. 2006. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution* 23:254–267. 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* 12:508–518. 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) [Software]. Available at https://github.com/najoshi/sickle.

686 Kalyaanamoorthy S., Minh BQ., Wong TKF., Haeseler A von., Jermiin LS. 2017. ModelFinder:
687 fast model selection for accurate phylogenetic estimates. *Nature Methods* 14:587–589. DOI:
688 10.1038/nmeth.4285.

689 Katoh K., Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7:
690 Improvements in Performance and Usability. *Molecular Biology and Evolution* 30:772–780.
691 DOI: 10.1093/molbev/mst010.

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

695 Krzywinski M., Schein J., Birol İ., Connors J., Gascoyne R., Horsman D., Jones SJ., Marra MA.
696 2009. Circos: An information aesthetic for comparative genomics. *Genome Research* 19:1639–
697 1645. DOI: 10.1101/gr.092759.109.

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

701 Kurtti TJ., Munderloh UG., Andreadis TG., Magnarelli LA., Mather TN. 1996. Tick Cell Culture
702 Isolation of an Intracellular Prokaryote from the TickIxodes scapularis. *Journal of Invertebrate*
703 *Pathology* 67:318–321. DOI: 10.1006/jipa.1996.0050.

704 Kurtz S., Phillippy A., Delcher AL., Smoot M., Shumway M., Antonescu C., Salzberg SL. 2004.
705 Versatile and open software for comparing large genomes. *Genome Biology* 5:R12. DOI:
706 10.1186/gb-2004-5-2-r12.

707 Kusian B., Sültemeyer D., Bowien B. 2002. Carbonic Anhydrase Is Essential for Growth of
708 Ralstonia eutropha at Ambient CO2 Concentrations. *Journal of Bacteriology* 184:5018–5026.
709 DOI: 10.1128/JB.184.18.5018-5026.2002.

710 Laetsch, D.R. (2016) blobtools:blobtools v0.9.19.4. Available at:
711 http://doi.org/10.5281/zenodo.61799.

712 Langmead B., Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*
713 9:357–359. DOI: 10.1038/nmeth.1923.

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

717 Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin
718 R., Subgroup 1000 Genome Project Data Processing. 2009. The Sequence Alignment/Map
719 format and SAMtools. *Bioinformatics* 25:2078–2079. DOI: 10.1093/bioinformatics/btp352.

720 Lindsey ARI., Bordenstein SR., Newton ILG., Rasgon JL. 2016. Wolbachia pipientis should not
721 be split into multiple species: A response to Ramírez-Puebla et al., "Species in Wolbachia?
722 Proposal for the designation of 'Candidatus Wolbachia bourtzisii', 'Candidatus Wolbachia
723 onchocercicola', 'Candidatus Wolbachia blaxteri', 'Candidatus Wolbachia brugii', 'Candidatus
724 Wolbachia taylori', 'Candidatus Wolbachia collembolicola' and 'Candidatus Wolbachia
725 multihospitum' for the different species within Wolbachia supergroups." *Systematic and Applied*
726 *Microbiology* 39:220–222. DOI: 10.1016/j.syapm.2016.03.001.

727 Lo N., Casiraghi M., Salati E., Bazzocchi C., Bandi C. 2002. How Many Wolbachia Supergroups
728 Exist? *Molecular Biology and Evolution* 19:341–346.

729 Lorenzen MD., Gnirke A., Margolis J., Garnes J., Campbell M., Stuart JJ., Aggarwal R.,
730 Richards S., Park Y., Beeman RW. 2008. The maternal-effect, selfish genetic element Medea is
731 associated with a composite Tc1 transposon. *Proceedings of the National Academy of Sciences*
732 105:10085–10089. DOI: 10.1073/pnas.0800444105.

733 Mann E., Stouthamer CM., Kelly SE., Dzieciol M., Hunter MS., Schmitz-Esser S. 2017.
734 Transcriptome Sequencing Reveals Novel Candidate Genes for Cardinium hertigii-Caused
735 Cytoplasmic Incompatibility and Host-Cell Interaction. *mSystems* 2:e00141-17. DOI:
736 10.1128/mSystems.00141-17.

737 Martin D., Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences.
738 *Bioinformatics (Oxford, England)* 16:562–563.

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

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

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

745 McBride MJ., Zhu Y. 2013. Gliding Motility and Por Secretion System Genes Are Widespread
746 among Members of the Phylum Bacteroidetes. *Journal of Bacteriology* 195:270–278. DOI:
747 10.1128/JB.01962-12.

748 McLean AHC., Parker BJ., Hrček J., Henry LM., Godfray HCJ. 2016. Insect symbionts in food
749 webs. *Phil. Trans. R. Soc. B* 371:20150325. DOI: 10.1098/rstb.2015.0325.

750 Mee PT., Weeks AR., Walker PJ., Hoffmann AA., Duchemin J-B. 2015. Detection of Low-Level
751 Cardinium and Wolbachia Infections in Culicoides. *Applied and Environmental Microbiology*
752 81:6177–6188. DOI: 10.1128/AEM.01239-15.

753 Merlin C., Masters M., McAteer S., Coulson A. 2003. Why Is Carbonic Anhydrase Essential to
754 Escherichia coli? *Journal of Bacteriology* 185:6415–6424. DOI: 10.1128/JB.185.21.6415-
755 6424.2003.

756 Mitsuhashi S., Ohnishi J., Hayashi M., Ikeda M. 2003. A gene homologous to β-type carbonic
757 anhydrase is essential for the growth of Corynebacterium glutamicum under atmospheric
758 conditions. *Applied Microbiology and Biotechnology* 63:592–601. DOI: 10.1007/s00253-003-
759 1402-8.

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

763 Nakamura Y., Kawai S., Yukuhiro F., Ito S., Gotoh T., Kisimoto R., Yanase T., Matsumoto Y.,
764 Kageyama D., Noda H. 2009. Prevalence of Cardinium Bacteria in Planthoppers and Spider
765 Mites and Taxonomic Revision of "Candidatus Cardinium hertigii" Based on Detection of a New
766 Cardinium Group from Biting Midges. *Applied and Environmental Microbiology* 75:6757–6763.
767 DOI: 10.1128/AEM.01583-09.

768 Nguyen L-T., Schmidt HA., von Haeseler A., Minh BQ. 2015. IQ-TREE: A Fast and Effective
769 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and*
770 *Evolution* 32:268–274. DOI: 10.1093/molbev/msu300.

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

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

777 Noel GR., Atibalentja N. 2006. 'Candidatus Paenicardinium endonii', an endosymbiont of the
778 plant-parasitic nematode Heterodera glycines (Nemata: Tylenchida), affiliated to the phylum
779 Bacteroidetes. *International Journal of Systematic and Evolutionary Microbiology* 56:1697–
780 1702. DOI: 10.1099/ijs.0.64234-0.

781 Nurk S., Bankevich A., Antipov D., Gurevich A., Korobeynikov A., Lapidus A., Prjibelsky A.,
782 Pyshkin A., Sirotkin A., Sirotkin Y., Stepanauskas R., McLean J., Lasken R., Clingenpeel SR.,
783 Woyke T., Tesler G., Alekseyev MA., Pevzner PA. 2013. Assembling Genomes and Mini-
784 metagenomes from Highly Chimeric Reads. In: Deng M, Jiang R, Sun F, Zhang X eds. *Research*
785 *in Computational Molecular Biology*. Lecture Notes in Computer Science. Springer Berlin
786 Heidelberg, 158–170. DOI: 10.1007/978-3-642-37195-0_13.

787 Padidam M., Sawyer S., Fauquet CM. 1999. Possible Emergence of New Geminiviruses by
788 Frequent Recombination. *Virology* 265:218–225. DOI: 10.1006/viro.1999.0056.

789 Pan X., Lührmann A., Satoh A., Laskowski-Arce MA., Roy CR. 2008. Ankyrin Repeat Proteins
790 Comprise a Diverse Family of Bacterial Type IV Effectors. *Science* 320:1651–1654. DOI:
791 10.1126/science.1158160.

792 Penz T., Schmitz-Esser S., Kelly SE., Cass BN., Müller A., Woyke T., Malfatti SA., Hunter MS.,
793 Horn M. 2012. Comparative Genomics Suggests an Independent Origin of Cytoplasmic
794 Incompatibility in Cardinium hertigii. *PLOS Genet* 8:e1003012. DOI:
795 10.1371/journal.pgen.1003012.

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

799 Perlman SJ., Kelly SE., Hunter MS. 2008. Population Biology of Cytoplasmic Incompatibility:
800 Maintenance and Spread of Cardinium Symbionts in a Parasitic Wasp. *Genetics* 178:1003–1011.
801 DOI: 10.1534/genetics.107.083071.

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

804 Pilgrim J., Ander M., Garros C., Baylis M., Hurst GDD., Siozios S. 2017. Torix group Rickettsia
805 are widespread in Culicoides biting midges (Diptera: Ceratopogonidae), reach high frequency
806 and carry unique genomic features. *Environmental Microbiology* 19:4238–4255. DOI:
807 10.1111/1462-2920.13887.

808 Posada D., Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA
809 sequences: Computer simulations. *Proceedings of the National Academy of Sciences* 98:13757–
810 13762. DOI: 10.1073/pnas.241370698.

811 Ramírez-Puebla ST., Servín-Garcidueñas LE., Ormeño-Orrillo E., Vera-Ponce de León A.,
812 Rosenblueth M., Delaye L., Martínez J., Martínez-Romero E. 2016. A response to Lindsey et al.
813 "Wolbachia pipientis should not be split into multiple species: A response to Ramírez-Puebla et
814 al." *Systematic and Applied Microbiology* 39:223–225. DOI: 10.1016/j.syapm.2016.03.004.

815 Ramírez-Puebla ST., Servín-Garcidueñas LE., Ormeño-Orrillo E., Vera-Ponce de León A.,
816 Rosenblueth M., Delaye L., Martínez J., Martínez-Romero E. 2015. Species in Wolbachia?
817 Proposal for the designation of 'Candidatus Wolbachia bourtzisii', 'Candidatus Wolbachia
818 onchocercicola', 'Candidatus Wolbachia blaxteri', 'Candidatus Wolbachia brugii', 'Candidatus
819 Wolbachia taylori', 'Candidatus Wolbachia collembolicola' and 'Candidatus Wolbachia
820 multihospitum' for the different species within Wolbachia supergroups. *Systematic and Applied*
821 *Microbiology* 38:390–399. DOI: 10.1016/j.syapm.2015.05.005.

822 Ramulu HG *et al.* 2014. Ribosomal proteins: Toward a next generation standard for prokaryotic
823 systematics? Mol. Phylogenet. Evol. 75:103–117. doi: 10.1016/j.ympev.2014.02.013.

824 Rice P., Longden I., Bleasby A. 2000. EMBOSS: The European Molecular Biology Open
825 Software Suite. *Trends in Genetics* 16:276–277. DOI: 10.1016/S0168-9525(00)02024-2.

826 Rio RVM., Attardo GM., Weiss BL. 2016. Grandeur Alliances: Symbiont Metabolic Integration
827 and Obligate Arthropod Hematophagy. *Trends in Parasitology* 32:739–749. DOI:
828 10.1016/j.pt.2016.05.002.

829 Ros VID., Breeuwer J a. J. 2009. The effects of, and interactions between, Cardinium and
830 Wolbachia in the doubly infected spider mite Bryobia sarothamni. *Heredity* 102:413–422. DOI:
831 10.1038/hdy.2009.4.

832 Salminen MO., Carr JK., Burke DS., McCUTCHAN FE. 1995. Identification of Breakpoints in
833 Intergenotypic Recombinants of HIV Type 1 by Bootscanning. *AIDS Research and Human*
834 *Retroviruses* 11:1423–1425. DOI: 10.1089/aid.1995.11.1423.

835 Santos-Garcia D., Rollat-Farnier P-A., Beitia F., Zchori-Fein E., Vavre F., Mouton L., Moya A.,
836 Latorre A., Silva FJ. 2014. The Genome of Cardinium cBtQ1 Provides Insights into Genome
837 Reduction, Symbiont Motility, and Its Settlement in Bemisia tabaci. *Genome Biology and*
838 *Evolution* 6:1013–1030. DOI: 10.1093/gbe/evu077.

839 Sato K., Naito M., Yukitake H., Hirakawa H., Shoji M., McBride MJ., Rhodes RG., Nakayama
840 K. 2010. A protein secretion system linked to bacteroidete gliding motility and pathogenesis.
841 *Proceedings of the National Academy of Sciences* 107:276–281. DOI:
842 10.1073/pnas.0912010107.

843 Schmitz-Esser S., Tischler P., Arnold R., Montanaro J., Wagner M., Rattei T., Horn M. 2010.
844 The Genome of the Amoeba Symbiont "Candidatus Amoebophilus asiaticus" Reveals Common
845 Mechanisms for Host Cell Interaction among Amoeba-Associated Bacteria. *Journal of*
846 *Bacteriology* 192:1045–1057. DOI: 10.1128/JB.01379-09.

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

849 Shen X-X., Hittinger CT., Rokas A. 2017. Contentious relationships in phylogenomic studies can
850 be driven by a handful of genes. *Nature Ecology & Evolution* 1:0126. DOI: 10.1038/s41559-017-
851 0126.

852 Shimodaira H., Goldman N. 2002. An Approximately Unbiased Test of Phylogenetic Tree
853 Selection. *Systematic Biology* 51:492–508. 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 Announc.* 6:e00624-18. 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* 31:3210–3212. 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:e55390. DOI: 10.1371/journal.pone.0055390.

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

Smith KS., Ferry JG. 2000. Prokaryotic carbonic anhydrases. *FEMS Microbiology Reviews* 24:335–366. 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* 96:15184–15189. DOI: 10.1073/pnas.96.26.15184.

Sudakaran S., Kost C., Kaltenpoth M. 2017. Symbiont Acquisition and Replacement as a Source of Ecological Innovation. *Trends in Microbiology* 25:375–390. DOI: 10.1016/j.tim.2017.02.014.

Toft C., Andersson SGE. 2010. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nature Reviews Genetics* 11:465–475. 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* 9:1–9. 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* 277:2007–2011. DOI: 10.1126/science.277.5334.2007.

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

883 Weeks AR., Marec F., Breeuwer JAJ. 2001. A Mite Species That Consists Entirely of Haploid
884 Females. *Science* 292:2479–2482. DOI: 10.1126/science.1060411.

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

888 Weinert LA., Araujo-Jnr EV., Ahmed MZ., Welch JJ. 2015. The incidence of bacterial
889 endosymbionts in terrestrial arthropods. *Proc. R. Soc. B* 282:20150249. DOI:
890 10.1098/rspb.2015.0249.

891 Wu M., Sun LV., Vamathevan J., Riegler M., Deboy R., Brownlie JC., McGraw EA., Martin W.,
892 Esser C., Ahmadinejad N., Wiegand C., Madupu R., Beanan MJ., Brinkac LM., Daugherty SC.,
893 Durkin AS., Kolonay JF., Nelson WC., Mohamoud Y., Lee P., Berry K., Young MB., Utterback
894 T., Weidman J., Nierman WC., Paulsen IT., Nelson KE., Tettelin H., O'Neill SL., Eisen JA.
895 2004. Phylogenomics of the Reproductive Parasite Wolbachia pipientis wMel: A Streamlined
896 Genome Overrun by Mobile Genetic Elements. *PLOS Biology* 2:e69. DOI:
897 10.1371/journal.pbio.0020069.

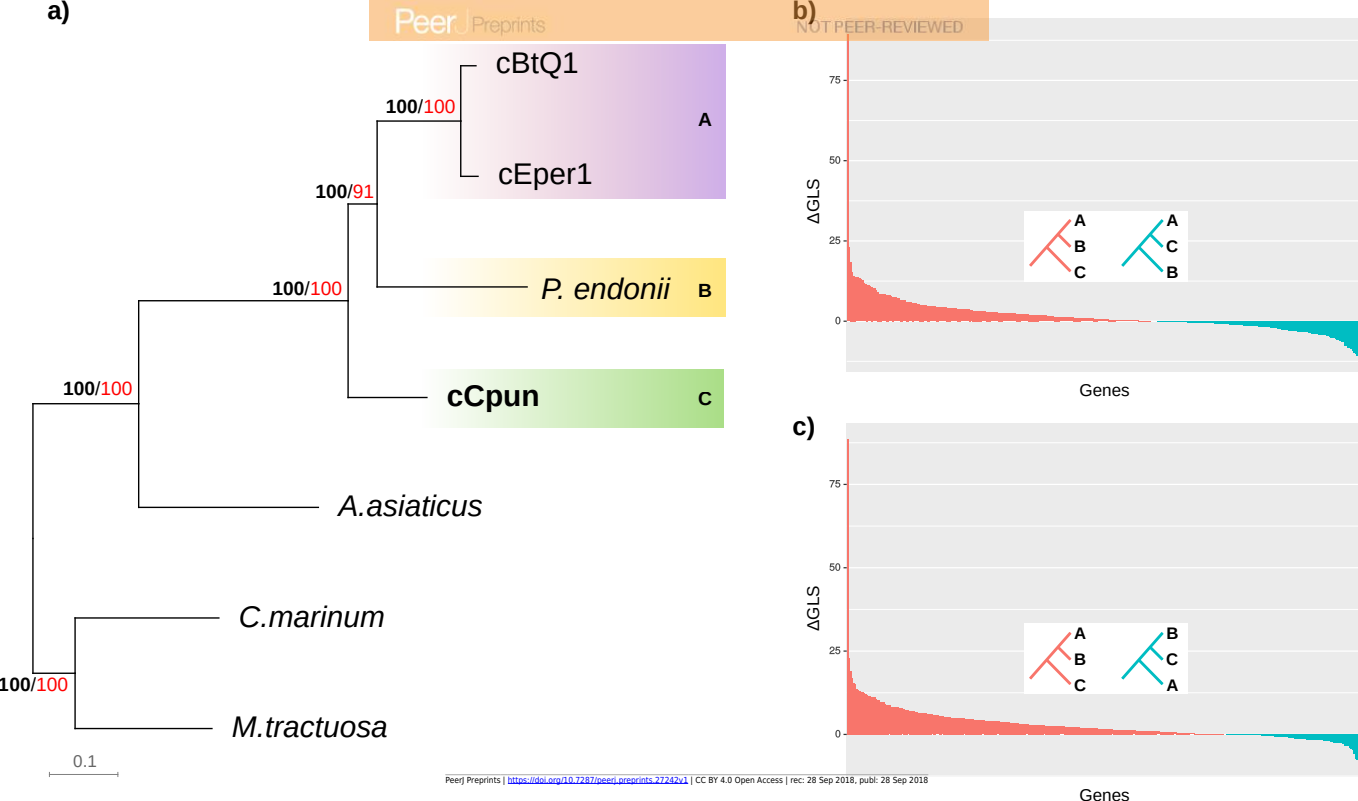
898 Zchori-Fein E., Gottlieb Y., Kelly SE., Brown JK., Wilson JM., Karr TL., Hunter MS. 2001. A
899 newly discovered bacterium associated with parthenogenesis and a change in host selection
900 behavior in parasitoid wasps. *Proceedings of the National Academy of Sciences* 98:12555–
901 12560. DOI: 10.1073/pnas.221467498.

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

Figure 1(on next page)

Phylogenetic relationships of *Cardinium* strains including *Ca*. Paenicardinium endonii.

a) The phylogenetic tree was inferred from the concatenated analysis of 338 single copy core proteins and seperately from a subset of 49 core ribosomal proteins using the Maximum likelihood method as implemented in IQTRE v1.6.6 (model: LG+F+R5). 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 1000 bootstrap replicates (black bold values: complete matrix; blue values: ribosomal dataset). The three major *Cardinium* groups A, B and C are denoted with different colour 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 *et al.* 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.

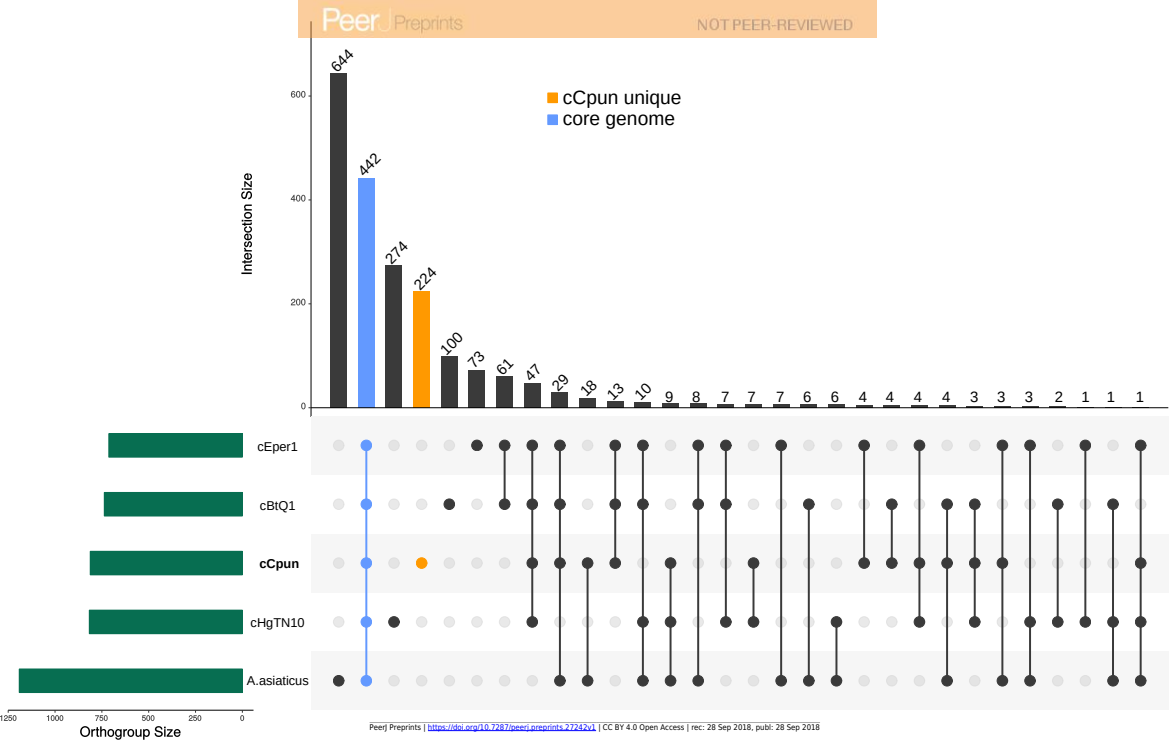


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27242v1 | CC BY 4.0 Open Access | rec: 28 Sep 2018, publ: 28 Sep 2018

Figure 2(on next page)

Genome content comparison across *the five Amoebophilaceae* genomes.

UpSet plot showing unique and overlapping protein ortholog clusters across the five Amoebophilaceae genomes *c*Cpun, *c*Eper1, *c*BtQ1, *Ca*. P. endonii (cHgTN10) and *Amoebophilus asiaticus*. The intersection matrix is sorted in descending order. Green bars on the left represent the orthogroup size for each genome. 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).

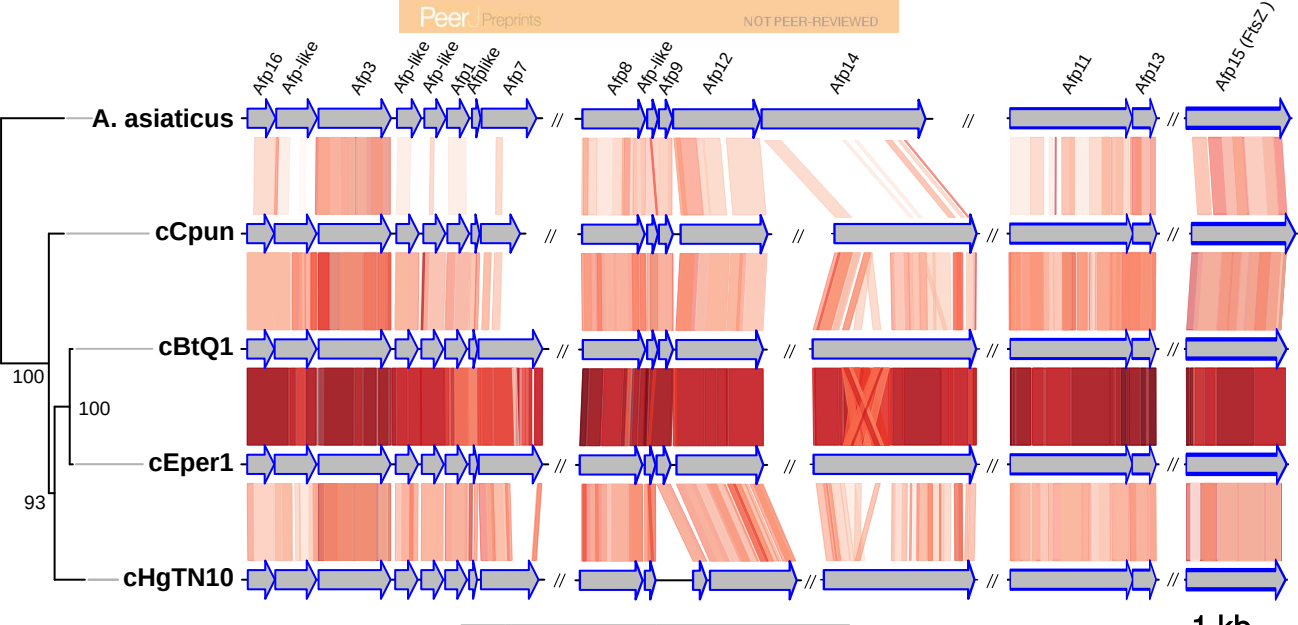


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27242v1 | CC BY 4.0 Open Access | rec: 28 Sep 2018, publ: 28 Sep 2018

Figure 3(on next page)

Organization and comparison of the antifeeding prophage (Afp-like) genes clusters in the five Amoebophilaceae genomes.

The phylogeny of the Afp-like secretion system was inferred with Maximum Likelihood based on the concatenated alignment of the 16 constituent protein sequences using IQTREE v1.6.6. Conserved regions are connected with a gradient of red shadings based on tblastx identities. The synteny and the phylogenetic tree of the Afp-like gene clusters were visualized using the genoPlotR package (Guy et al. 2010).

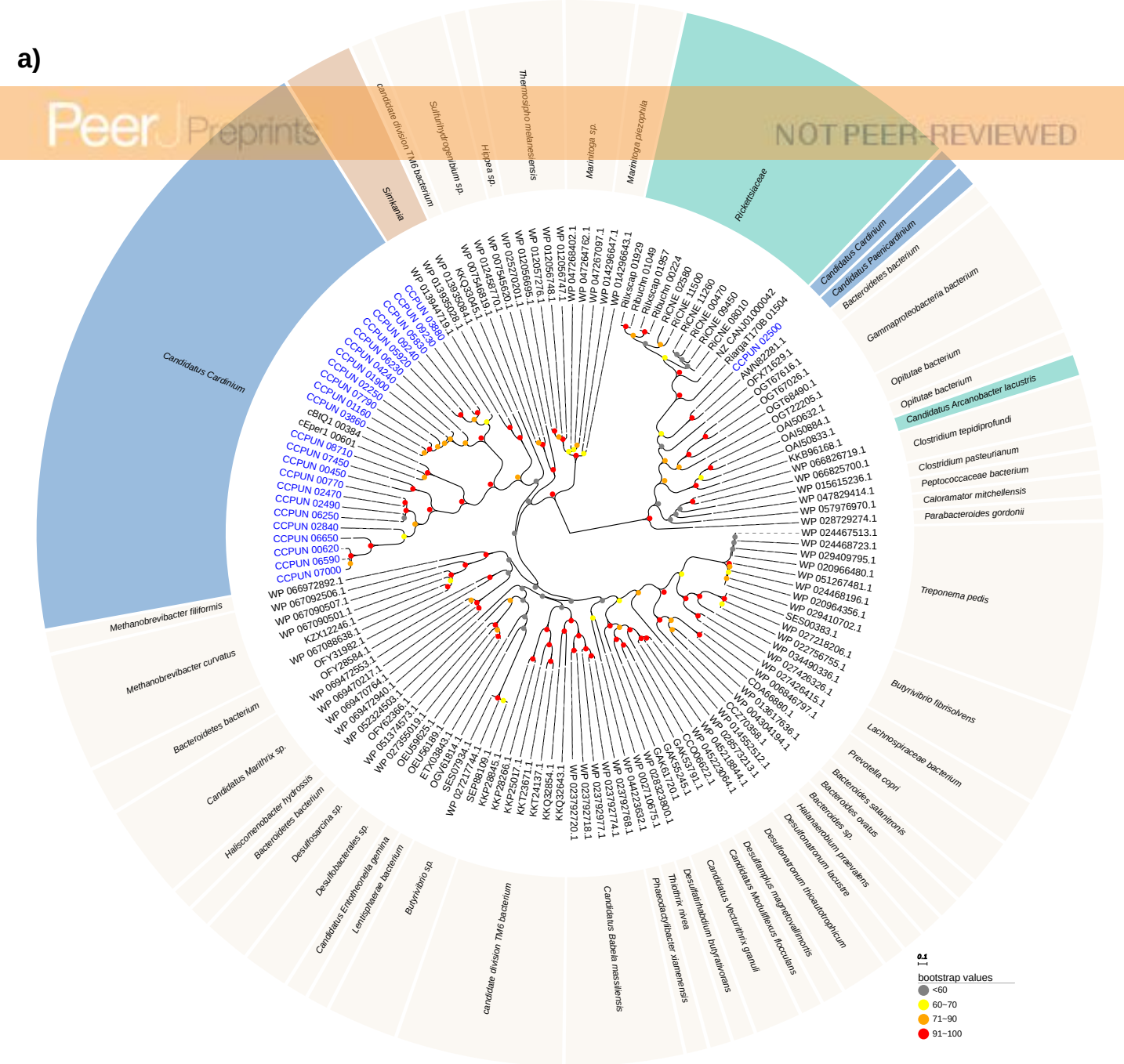


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27242v1 | CC BY 4.0 Open Access | rec: 28 Sep 2018, publ: 28 Sep 2018

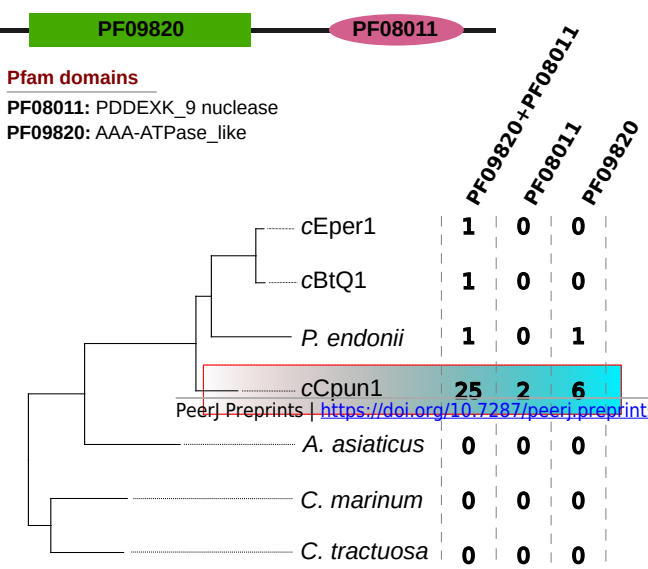
Figure 4(on next page)

DUF1703 gene family 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 139 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 *Amoebophilacheae.* c) Phylogenetic network showing the reticulated evolution of the *c*Cpun DUF1703 paralogs.



b)



c)

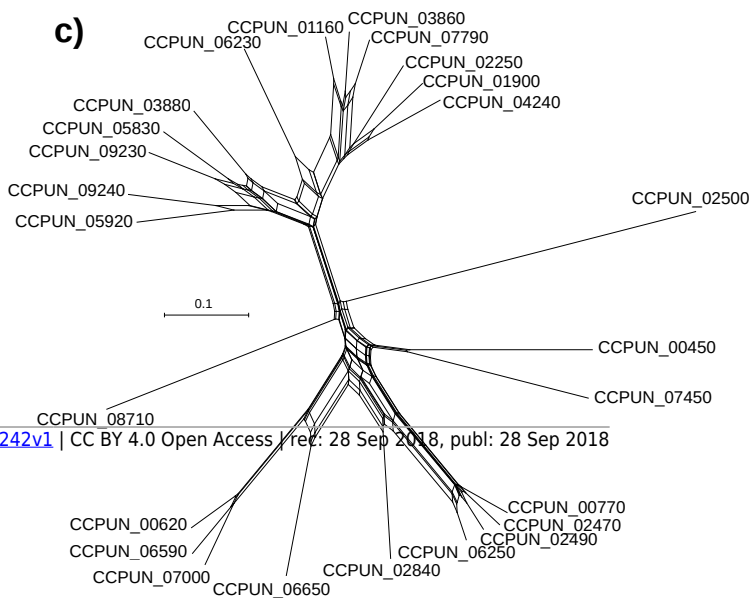
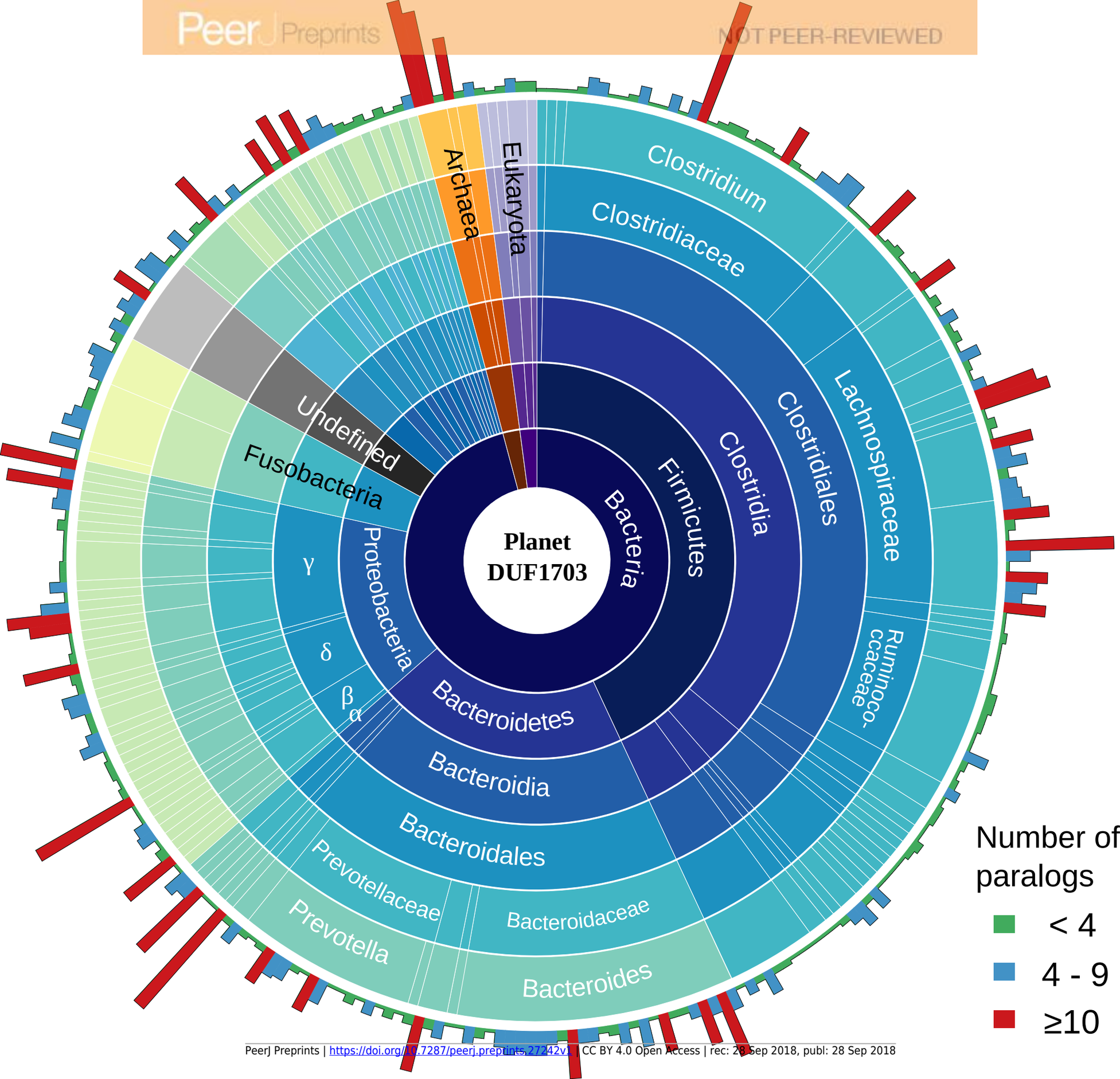


Figure 5(on next page)

Planet DUF1703.

Abundance and taxonomic distribution of DUF1703 proteins in PFAM database.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27242v1 | CC BY 4.0 Open Access | rec: 28 Sep 2018, publ: 28 Sep 2018

Table 1(on next page)

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

1

	cCpun	cEper1**	cBtQ1**	*Ca*. P. endonii (cHgTN10)	*A.asiaticus*
Number. of scaffolds	57*	1	11	1	1
Plasmids	0	1	1	0	0
Total size in kb	1137	887 (58)	1013 (52)	1193	1884
GC content (%)	33.7	36.6 (31.5)	35 (32)	38.2	35
CDS	917	841 (65)	709 (30)	974	1557
Avg. CDS length (bp)	993	911 (733)	1033 (1,389)	997	990
Coding density (%)	80	85.5 (82.1)	79.7 (80.1)	81.4	81.8
rRNAs	3	3	3	3	3
tRNAs	37	37	35	37	35
Ankyrin repeat proteins	46	18-19	26	27	54
Reference	this study	Penz et al. 2012	Santos-Garcia et al. 2014	Kurt et al. 2018	Schmitz-Esser et al. 2010

* contigs >500 bp

** chromosome (plasmid)

2
3

Table 2(on next page)

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

1

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_00570	729	Lactococcin-G-processing and transport ATP-binding protein LagD	*Crocinitomix algicola*, (WP_066755554)	0E+00	62%
CCPUN_01020	280	D-alanyl-D-alanine dipeptidase	candidate division TM6 bacterium, (KKR96749)	7E-57	46%
CCPUN_01120	218	Carbonic anhydrase	*Lysobacter sp*. Root494, (WP_056131435)	2E-95	59%
CCPUN_01870	374	Capsule biosynthesis protein CapA	*Crocinitomix sp*. MedPE-Swsnd, (OIQ37660)	1E-112	51%
CCPUN_03570	551	DNA repair protein RecN	Rickettsiales bacterium, (PCJ29272)	2E-175	48%
CCPUN_03790	122	hypothetical protein	*Flavobacterium branchiophilum*, (OXA70659)	2E-46	62%
CCPUN_03800	900	DNA primase	*Geofilum rubicundum*, (WP_083985273)	0E+00	44%
CCPUN_03900	258	hypothetical protein, putative transposase	*Candidatus* Paracaedibacter acanthamoebae, (WP_038464592)	3E-114	67%
CCPUN_03960	111	HTH-type transcriptional regulator ImmR	*Arachidicoccus rhizosphaerae*, (WP_091401557)	1E-51	75%
CCPUN_06490	469	Arginine/agmatine antiporter	Gammaproteobacteria bacterium 39-13, (OJV90723)	4E-112	43%
CCPUN_07130	156	hypothetical protein	Gammaproteobacteria bacterium, (OGT51102)	7E-47	57%
CCPUN_07910	266	Chromosome-partitioning protein Spo0J	*Candidatus* Jidaibacter acanthamoeba, (WP_053332526)	9E-73	47%
CCPUN_07920	327	Sporulation initiation inhibitor protein Soj	*Candidatus* Jidaibacter acanthamoeba, (WP_039455583)	2E-109	53%
CCPUN_08370	224	Biotin transport ATP-binding protein BioM	*Erwinia amylovora*, (WP_004170656)	5E-100	62%
CCPUN_08380	194	Energy-coupling factor transporter transmembrane protein BioN	Bartonella washoensis, (WP_006922939)	5E-39	39%
CCPUN_08840	436	Folylpolyglutamate synthase	Wolbachia pipientis, (WP_010963010)	0E+00	76%
CCPUN_08880	242	Uridylate kinase	Sphingobacterium mizutaii, (WP_093100754)	3E-72	50%
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	1360	hypothetical protein	Aedes albopictus, (KXJ68548)	5E-72	27%

2
3