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 cCpun 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 cCpun from biting midges establishes *Cardinium* as a paraphyletic group, and reveals a novel gene family expansion in a symbiont

Stefanos Siozios Correspond. 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 cCpun, 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 cCpun 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 cCpun 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 cCpun 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 *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 *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 *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 *Cpun*, 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 *gyrB* 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 “*candidatus*” 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 (cEper1) of the parasitic wasp *Encarsia pergandiella* (Penz et al. 2012) and the *Cardinium* endosymbiont (cBtQ1) 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 cCpun, 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 cCpun 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 version 1.2.1 (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 cCpun 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 cCpun 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 cEper1 (Penz et al. 2012) and Cardinium hertigii cBtQ1 (Santos-Garcia et al. 2014), the Cardinium endosymbiont of the plant-parasitic nematode Heterodera glycines candidatus 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 cCpun, cEper1, cBtQ1, 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 cCpun 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 cCpun 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 cCpun draft genomes

The final assembly of the cCpun 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 cCpun genome shares many characteristics with those of the previously sequenced Cardinium strains cEper1, cBtQ1, 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 (cEper1 and cBtQ1) (Table 1, Supplementary Fig. S2). Nevertheless, we were able to detect several regions with sequence similarity to elements of the two plasmids found in cEper1 and cBtQ1. 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 cCpun 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). cCpun 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 cEper1 and Ca. P. endonii (cHgTN10) genomes, which suggests that cCpun is a near complete genome. Overall, the redundancy in cCpun 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 cCpun to be underestimated.
Phylogenomic analyses place cCpun 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) cCpun 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 cEper1 and cBtQ1 in exclusion of cCpun (\(p < 0.001\), Fisher’s exact test). In contrast, only 105 genes (31%) support the monophyletic grouping of cCpun with cEper1 and cBtQ1 while a small subset of genes (n=53; 16%) supports the monophyletic grouping of cCpun with Ca. P. endonii.

Genome content comparisons estimate both a core Cardinium genome, genes associated with an insect-symbiont lifestyle, and cCpun 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, cCpun, cEper1, and cBtQ1).

The four genomes share a core of 442 ortholog clusters of which 338 consist of single-copy
genes (Fig. 2). The cCpun genome codes for a substantial number of unique proteins (Fig. 2, Supplementary Table S2). Specifically, among the 812 ortholog clusters predicted for cCpun, 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 cCpun specific proteins, 156 (~65%), had no significant matches (E-value ≤ 10^{-10}) in the NCBI-nr database or functional domains and were assigned as hypothetical proteins. Amongst the remaining 85 predicted cCpun-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 cCpun 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 cCpun 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 ShIB (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 cCpun 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 (cCpun, cEper1 and cBtQ1), 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 cEper1 and cBtQ1 (Penz et al. 2012; Santos-Garcia et al. 2014), a Lysozyme M1 homolog, a nicotinamide mononucleotide transporter and a putative peptidase.

**cCpun 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 cCpun 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 cCpun, a system related to gliding motility and pathogenicity in several members of the phylum Bacteroidetes (McBride & Zhu 2013; McBride & Nakane 2015). cCpun 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 cCpun, A. asiaticus, and cBtQ1 while an additional three proteins with homology to the lipoproteins GldD, GldJ and GldH are uniquely shared between cCpun and A. asiaticus (Supplementary Table S3). More recently, core components of the T9SS secretion system were found on the plasmid of Cardinium cBtQ1 (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 cCpun as well as A. asiaticus supports an ancestral origin of the T9SS machinery which was subsequently lost from cEper1 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 *cEper1* 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 cCpun 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 cCpun 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 cCpun scaffolds and may be artefactually truncated. Thus our estimate of gene family size is conservative.

The members of the DUF1703 gene family display in cCpun 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 *cCpun* genome amongst sequenced genomes; *cEper1*, *cBtQ1* 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 *cCpun*.

Phylogenetic network analyses revealed several reticulation events within the DUF1703 gene family in *cCpun* 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 *cCpun* 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 *cCpun* paralogs. Overall, our results point to a HGT scenario for the origin of *Cardinium* DUF1703 gene family with subsequent expansion in the *cCpun* 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 cCpun 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 cCpun 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 cEper1 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 cCpun (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 wPip Wolbachia strain (CinB) (Beckmann et al. 2017).

**Horizontal gene transfer as a source of genes in the cCpun genome**
Horizontal gene transfer (HGT) has been previously reported as the source of several genes in *A. asiaticus*, *cEper1*, and *cBtQ1* (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 *cCpun* (Table 2). The majority of the accessory genes of *cCpun* 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 *cCpun*-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 anhydrase (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 *cCpun*-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 *cCpun* and *Ca. P. endonii* and not in other *Cardinium* strains nor *A. asiaticus*. Notably, the *cCpun* 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). cCpun 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 cCpun 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 cCpun 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 cCpun 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 (Ramirez-Puebla et al. 2015; Lindsey et al. 2016; Ramirez-Puebla et al. 2016). Wolbachia is currently defined as a single species named “Wolbachia pipiens” classified in at least 16 divergent supergroups (Glowska et
al. 2015), with this single species designation persisting despite the observation that some of these supergroups have been irreversibly separated suggesting that they might consist separate species (Ellegaard et al. 2013). Nakamura et al. 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 (Nakamura et al. 2009). 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 actually consist of 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 cCpun Cardinium strain with the RiCINE Rickettsia symbiont of C. 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 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 cEper1 genome, a biotin synthesis system was not observed in the cCpun 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 that selection for symbiont-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.
References


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


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 separately 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.
a) 

- **cBtQ1**
  - **cEper1**
  - **P. endonii**
    - **cCpun**
      - **A. asiaticus**
    - **C. marinum**
  - **M. tractuosa**

- Tree branches with support values: 100/100, 100/91, 100/100, 100/100

b) 

- ΔGLS distribution for different genes A, B, C

- Graph showing ΔGLS for genes A, B, C

- Legend: A, B, C

---

c) 

- ΔGLS distribution for different genes A, B, C

- Graph showing ΔGLS for genes A, B, C

- Legend: A, B, C
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 cCpun, cEper1, cBtQ1, 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 cCpun 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).
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).
DUF1703 gene family expansion in cCpun genome.

a) phylogenetic analysis of the cCpun 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 cCpun DUF1703 gene family within the *Amoebophilaceae*. c) Phylogenetic network showing the reticulated evolution of the cCpun DUF1703 paralogs.
Figure 5 (on next page)

Planet DUF1703.

Abundance and taxonomic distribution of DUF1703 proteins in PFAM database.
Table 1 (on next page)

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

* contigs >500 bp

** chromosome (plasmid)
Example of cCpun genes likely originated from HGTs.
<table>
<thead>
<tr>
<th>Gene id</th>
<th>Length (AA)</th>
<th>Annotation</th>
<th>Taxonomy of the Best BLAST hit, (GenBank Accession)</th>
<th>E-value</th>
<th>AA identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCPUN 00040</td>
<td>308</td>
<td>hypothetical protein, putative transposase</td>
<td>Rickettsia endosymbiont of Culicoides newsteadi, (WP_094649760)</td>
<td>2E-128</td>
<td>64%</td>
</tr>
<tr>
<td>CCPUN 00530</td>
<td>328</td>
<td>hypothetical protein, putative transposase</td>
<td>Rickettsia endosymbiont of Culicoides newsteadi, (WP_094649760)</td>
<td>3E-124</td>
<td>62%</td>
</tr>
<tr>
<td>CCPUN 01090</td>
<td>346</td>
<td>hypothetical protein, putative transposase</td>
<td>Rickettsiales bacterium, (PCJ29205)</td>
<td>6E-133</td>
<td>58%</td>
</tr>
<tr>
<td>CCPUN 02050</td>
<td>379</td>
<td>hypothetical protein, putative transposase</td>
<td>Rickettsiales bacterium, (PCJ24349)</td>
<td>5E-55</td>
<td>44%</td>
</tr>
<tr>
<td>CCPUN 04150</td>
<td>328</td>
<td>hypothetical protein, putative transposase</td>
<td>Rickettsia endosymbiont of Culicoides newsteadi, (WP_094649760)</td>
<td>9E-125</td>
<td>59%</td>
</tr>
<tr>
<td>CCPUN 04430</td>
<td>297</td>
<td>hypothetical protein, putative transposase</td>
<td>Rickettsiales bacterium, (PCJ25778)</td>
<td>9E-136</td>
<td>65%</td>
</tr>
<tr>
<td>CCPUN 00570</td>
<td>729</td>
<td>Lactococcin-G-processing and transport ATP-binding protein LagD</td>
<td>Crocinitomix algicola, (WP_066755554)</td>
<td>0E+00</td>
<td>62%</td>
</tr>
<tr>
<td>CCPUN 01020</td>
<td>280</td>
<td>D-alanyl-D-alanine dipeptidase</td>
<td>candidature division TM6 bacterium, (KKR96749)</td>
<td>7E-57</td>
<td>46%</td>
</tr>
<tr>
<td>CCPUN 01120</td>
<td>218</td>
<td>Carbonic anhydrase</td>
<td>Lysobacter sp. Root494, (WP_056131435)</td>
<td>2E-95</td>
<td>59%</td>
</tr>
<tr>
<td>CCPUN 01870</td>
<td>374</td>
<td>Capsule biosynthesis protein CapA</td>
<td>Crocinitomix sp. MedPE-Swsnd, (OIQ37660)</td>
<td>1E-112</td>
<td>51%</td>
</tr>
<tr>
<td>CCPUN 03570</td>
<td>551</td>
<td>DNA repair protein RecN</td>
<td>Rickettsiales bacterium, (PCJ29272)</td>
<td>2E-175</td>
<td>48%</td>
</tr>
<tr>
<td>CCPUN 03790</td>
<td>122</td>
<td>hypothetical protein</td>
<td>Flavobacterium branchiphilum, (OXA70659)</td>
<td>2E-46</td>
<td>62%</td>
</tr>
<tr>
<td>CCPUN 03800</td>
<td>900</td>
<td>DNA primase</td>
<td>Geofilum rubicundum, (WP_083985273)</td>
<td>0E+00</td>
<td>44%</td>
</tr>
<tr>
<td>CCPUN 03900</td>
<td>258</td>
<td>hypothetical protein, putative transposase</td>
<td>Candidatus Paracaedibacter acanthamoebae, (WP_038464592)</td>
<td>3E-114</td>
<td>67%</td>
</tr>
<tr>
<td>CCPUN 03960</td>
<td>111</td>
<td>HTH-type transcriptional regulator ImmR</td>
<td>Arachidiococcus rhizophaeae, (WP_091401557)</td>
<td>1E-51</td>
<td>75%</td>
</tr>
<tr>
<td>CCPUN 06490</td>
<td>469</td>
<td>Arginine/agmatine antiporter</td>
<td>Gammaproteobacteria bacterium 39-13, (OJV90723)</td>
<td>4E-112</td>
<td>43%</td>
</tr>
<tr>
<td>CCPUN 07130</td>
<td>156</td>
<td>hypothetical protein</td>
<td>Gammaproteobacteria bacterium, (OGT11002)</td>
<td>7E-47</td>
<td>57%</td>
</tr>
<tr>
<td>CCPUN 07910</td>
<td>266</td>
<td>Chromosome-partitioning protein Spo0J</td>
<td>Candidatus Jdiaibacter acanthamoebae, (WP_053332526)</td>
<td>9E-73</td>
<td>47%</td>
</tr>
<tr>
<td>CCPUN 07920</td>
<td>327</td>
<td>Sporulation initiation inhibitor protein Soj</td>
<td>Candidatus Jdiaibacter acanthamoebae, (WP_039455583)</td>
<td>2E-109</td>
<td>53%</td>
</tr>
<tr>
<td>CCPUN 08370</td>
<td>224</td>
<td>Biotin transport ATP-binding protein BioM</td>
<td>Erwinia amylovora, (WP_004170656)</td>
<td>5E-100</td>
<td>62%</td>
</tr>
<tr>
<td>CCPUN 08840</td>
<td>436</td>
<td>Folyipolyglutamate synthase</td>
<td>Wolbachia pipiens, (WP_010963010)</td>
<td>0E+00</td>
<td>76%</td>
</tr>
<tr>
<td>CCPUN 08880</td>
<td>242</td>
<td>Uridylate kinase</td>
<td>Sphingobacterium mizuitaii, (WP_093100754)</td>
<td>3E-72</td>
<td>50%</td>
</tr>
<tr>
<td>CCPUN 08910</td>
<td>340</td>
<td>hypothetical protein</td>
<td>Rickettsia felis, (WP_039595314)</td>
<td>2E-155</td>
<td>73%</td>
</tr>
<tr>
<td>CCPUN 08830</td>
<td>426</td>
<td>hypothetical protein</td>
<td>Aedes aegypti, (XP_001656120)</td>
<td>2E-60</td>
<td>39%</td>
</tr>
<tr>
<td>CCPUN_08280</td>
<td>1360</td>
<td>hypothetical protein</td>
<td>Aedes albopictus, (KXJ68548)</td>
<td>5E-72</td>
<td>27%</td>
</tr>
</tbody>
</table>