Comparative Genomics of *Blattabacterium cuenoti*: The Frozen Legacy of an Ancient Endosymbiont Genome

Rafael Patiño-Navarrete[1], Andrés Moya[1,2,3], Amparo Latorre[1,2,3,*], and Juli Peretó[1,4,*]

[1]Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de València, València, Spain

[2]Departament de Genètica, Universitat de València, València, Spain

[3]Centre for Public Health Research (CSISP), València, Spain

[4]Departament de Bioquímica i Biologia Molecular, Universitat de València, València, Spain

*Corresponding authors: E-mail: amparo.latorre@uv.es; Juli.Pereto@uv.es.

Accepted: January 18, 2013

Data deposition: This project has been deposited at GenBank/DDJB/EMBL under the accession number CP003605 for the chromosome and CP003606 for the plasmid.

Abstract

Many insect species have established long-term symbiotic relationships with intracellular bacteria. Symbiosis with bacteria has provided insects with novel ecological capabilities, which have allowed them colonize previously unexplored niches. Despite its importance to the understanding of the emergence of biological complexity, the evolution of symbiotic relationships remains hitherto a mystery in evolutionary biology. In this study, we contribute to the investigation of the evolutionary leaps enabled by mutualistic symbioses by sequencing the genome of *Blattabacterium cuenoti*, primary endosymbiont of the omnivorous cockroach *Blatta orientalis,* and one of the most ancient symbiotic associations. We perform comparative analyses between the *Blattabacterium cuenoti* genome and that of previously sequenced endosymbionts, namely those from the omnivorous hosts the *Blattella germanica* (Blattelidae) and *Periplaneta americana* (Blattidae), and the endosymbionts harbored by two wood-feeding hosts, the subsocial cockroach *Cryptocercus punctulatus* (Cryptocercidae) and the termite *Mastotermes darwiniensis* (Termitidae). Our study shows a remarkable evolutionary stasis of this symbiotic system throughout the evolutionary history of cockroaches and the deepest branching termite *M. darwiniensis*, in terms of not only chromosome architecture but also gene content, as revealed by the striking conservation of the *Blattabacterium* core genome. Importantly, the architecture of central metabolic network inferred from the endosymbiont genomes was established very early in *Blattabacterium* evolutionary history and could be an outcome of the essential role played by this endosymbiont in the host's nitrogen economy.

Key words: *Blattabacterium* endosymbiont, *Blatta orientalis*, nitrogen metabolism, pan-genome, urease, genome reduction.

Introduction

Symbiotic associations between eukaryotes and prokaryotes are common in nature and have been described in all branches of the eukaryotic tree of life (Moya et al. 2008). Insects characterized by feeding upon unbalanced diets have established symbiotic associations with bacteria that provide nutrients lacking in the diet. This is the case of essential amino acid provision by *Buchnera aphidicola*, the primary endosymbiont of pea aphids feeding on phloem sap (Baumann 2005; Shigenobu and Wilson 2011), whereas the same metabolic function is performed by a consortium comprising the coprimaries *Buchnera aphidicola* and *Serratia symbiotica* in the case of the cedar aphid (Pérez-Brocal et al. 2006; Lamelas et al. 2011). However, in insects such as cockroaches, which feed on complex diets, the role of the obligatory endosymbiont, *Blattabacterium cuenoti*, was unclear before the genomes of *Blattella germanica* (BBge) (López-Sánchez et al. 2009) and *Periplaneta americana* (BPam) (Sabree et al. 2009) strains were sequenced.

The symbiosis between *Blattabacterium* and cockroaches may have become established between 300 Ma, the age of the first fossils of roaches in the Carboniferous, and 140 Ma, that is, before the diversification of extant cockroach families in the Cretaceous (Clark et al. 2001; Vrsansky et al. 2002; Lo et al. 2003). Phylogenetic analyses based on 16S ribosomal DNA have located *Blattabacterium* as members of the class Flavobacteria in the phylum Bacteroidetes (López-Sánchez et al. 2008). The presence of these endosymbiotic bacteria has been observed in all studied species, with the only known exception being that of cockroaches from the genus *Nocticola* (Lo et al. 2007). Moreover, in termites, a sister clade of cockroaches, only *Mastotermes darwiniensis* has retained a symbiont (Bandi et al. 1995). Recently, another two genomes of *Blattabacterium* strains have been sequenced from wood-feeding hosts, the subsocial roach *Cryptocercus punctulatus* (BCpu) (Neef et al. 2011), and the termite *M. darwiniensis* (BMda) (Sabree et al. 2012). All these four *Blattabacterium* genomes show typical features of other insect primary endosymbionts, such as low GC content, total absence of repeated elements, and a reduced genome size compared to their closest free-living relatives (McCutcheon and Moran 2012).

Flux balance analyses (FBA) of the reconstructed metabolic networks from BBge and BPam strains have highlighted the functional equivalence of both networks, despite a prolonged divergence time and a few remarkable enzymatic differences, such as the absence of the first three steps of the tricarboxylic acid cycle (TCA) cycle in BPam (González-Domenech et al. 2012). Analyses would also suggest the endosymbionts play a role in the host's nitrogen metabolism (Feldhaar et al. 2007) leading researchers to postulate a mechanism for the intriguing ammonotelism described in cockroaches (Mullins and Cochran 1972, 1976; Cochran 1985). In addition, metabolic inferences of *Blattabacterium* strains from *B. germanica* (BBge) and *P. americana* (BPam) suggest they are involved in supplying essential amino acids and cofactors to the host (López-Sánchez et al. 2009; Sabree et al. 2009). Moreover, *Blattabacterium* strains from all four studied species encode urease genes, and endosymbiont-enriched extracts from *P. americana* and *B. germanica* show urease activity (López-Sánchez et al. 2009), thus this activity may be essential for nitrogen recycling from urate deposits (González-Domenech et al. 2012). On the other hand, BCpu and BMda strains retain the enzymes required for nitrogen metabolism, even though they have lost pathways for the synthesis of several amino acids (Neef et al. 2011; Sabree et al. 2012).

In this work, we report on the genome of the *Blattabacterium* from *Blatta orientalis* (BBor), a sister clade of *P. americana* from the family Blattidae (Kambhampati 1995; Inward et al. 2007). Comparative genome analysis of all five genomes has shed light on the evolution of the primary endosymbiont because it became associated with a common ancestor of cockroaches and termites. Of particular interest is the extraordinary conservation of both genome architecture and gene content.

Materials and Methods

Cockroach Rearing, Dissections, and DNA Extraction

Blatta orientalis (Blattaria: Blattidae) were reared at the Cavanilles Institute for Biodiversity and Evolutionary Biology (University of Valencia, Spain) in compartments at 25 °C and fed with sucrose-enriched dog food. *Blatta orientalis* adults were killed by a 15–20 min treatment with ethyl acetate. The fat body was then separated, and the bacteriocytes were extracted as described previously (Gil et al. 2003; López-Sánchez et al. 2008). Genomic DNA from the bacteriocytes was obtained following the CTAB method (Murray and Thompson 1980).

Genome Sequencing and Assembly

DNA was sequenced using the Roche GS-FLX 100 pyrosequencing technology. For half plate, 133,562 reads were obtained with an average length of 235 bp. Those reads were assembled using the Roche Newbler software, obtaining 1,201 contigs with an average length of 846 bp; 39 of those contigs were greater than 500 bp in length and contained 745,669 nucleotides in 129,647 reads, which implied that 97% of reads were included in the 39 biggest contigs. Additionally, sequences obtained by the Sanger method were used to fill gaps and confirm some uncertainties. The data obtained from both methods were combined in a database as described by López-Sánchez et al. (2009). The GAP4 software included in the Staden Package (Staden et al. 2000) was used for the final assembly. Comparison with other previously sequenced strains using the Artemis Comparative tool (Carver et al. 2005) was useful for correct orientation of different contigs, due to the genomic stability demonstrated by these genomes.

Annotation

Protein-coding genes (CDS) were predicted by GLIMMER3 software (Delcher et al. 2007), using *Blattabacterium* symbiotic strains from *B. germanica* and *P. americana* as training set sequences. Then, they were manually curated with the genome viewer Artemis (Rutherford et al. 2000). CDS were annotated through basic local alignment search tool searches (Altschul et al. 1997) against the gene nonredundant Kyoto Encyclopedia of Genes and Genomes (KEGG) database (www.genome.jp, last accessed February 4, 2013) and the genomes of the previously annotated *Blattabacterium* strains. Genes from the five *Blattabacterium* strains were classified into Cluster of Orthologous Genes (COG) categories searching with BLASTP (*e* value: 0.001) against the cluster of orthologous groups database (Vasudevan et al. 2003), and a heat map was generated with the gplots library (Warnes 2011). RNA genes were annotated by comparing the genome sequence with RNA databases using the INFERNAL software (Nawrocki et al. 2009).

Similar to previously sequenced *Blattabacterium* strains, the strain from *Blatta orientalis* does not possess any features determining replication origin, with the exception of the GC-skew, which was determined with the OriginX software (Worning et al. 2006).

Graphic representation of genome-compared graphs between the five strains of *Blattabacterium* was obtained with the genoPlotR package (Guy et al. 2010).

Identification of Orthologous Genes

Orthologous genes present in all five *Blattabacterium* strains and the free-living Bacteriodete *Flavobacterium psychrophilum* were identified using the OrthoMCL algorithm (Chen et al. 2006). The bacterial genomes used were the five strains of *Blattabacterium* and *F. psychrophilum* JIP02/86 (accession numbers in supplementary table S1, Supplementary Material online). The protein-coding genes of all bacteria were compared, all-against-all using BLASTP, the minimum *E* value was established at 1e − 05, a 70% cutoff and 1.5 inflation value.

Construction of *Blattabacterium* Pan-Genome

Blattabacterium genomes were retrieved from their respective databases (see supplementary table S1, Supplementary Material online). Gene counts are based on orthology (supplementary table S2, Supplementary Material online). Pseudogenes are considered absent. Visual display of the pan-genome subspaces was done using the R custom modified drawVennDiagram function of package gplots (Warnes 2011).

Phylogenetic Analysis

Amino acid sequences from the orthologous genes present in all six genomes, mentioned earlier, were aligned with MAFFT (Katoh et al. 2005) and manually concatenated. ProtTest 2.4 (Abascal et al. 2005) was used to select the best evolutionary model for our data. Finally, the maximum likelihood best tree was obtained with RAxML (Stamatakis et al. 2005) with 100 bootstrap replicate and using the PROTGAMMA algorithm.

Testing the Molecular Clock Hypothesis

Nucleotide alignments from the protein-coding genes in the core from all five *Blattabacterium* strains were obtained from the MAFFT (Katoh et al. 2005) aligned protein sequences with the program Tranalign from the EMBOSS package (Rice et al. 2000). Maximum likelihood models are known to be robust to violation of the model, including divergence times and saturation. However, because of the large distances in the phylogeny and nucleotide biases, we minimized the underestimate of divergence levels owing to the problem of saturation of nucleotide sites by removing all third codon positions from downstream evolutionary sequence analyses. Moreover, first and second codon positions account for most of the nonsynonymous sites in the protein-coding genes, which are often subject to selection and are less prone to saturation. The best-fit evolutionary model for each alignment was selected with the jModeltest 2 (Darriba et al. 2012). Finally, a likelihood ratio test (LRT) was performed as implemented in the program baseml from the PAML package version 4.4 (Yang 2007) using for each gene the evolutionary model previously selected. Each gene in the core was tested under two models, one assuming homogenous rates (rooted model, $n-1$ branch length are estimated) and the other assuming different rates for each branch (unrooted model, $2n-3$ branch length should be estimated). The LRT compares the differences in the log likelihood (l), values for each model ($2\Delta l$) with a χ^2 distribution with $n-2$ degrees of freedom.

Core genes that do not reject the molecular clock hypothesis, possess an orthologous gene in the *F. psychrophilum* genome, and do not accumulate more than 2.5 nucleotide substitutions per site (supplementary table S3, Supplementary Material online) were used as data set to determine the date of the split between the strains BBor and BPam, as well as the split between the strains BMda and BCpu. Nucleotide alignments were obtained following the aforementioned procedure, also removing third codon positions. Then best-fit model and the molecular clock were tested. The divergence time between BBge and the rest of the strains estimated according the fossil record (140 Ma, Vrsansky et al. 2002; Sabree et al. 2010) was used as a calibration point.

Results

Genome Characteristics

The main features of the five *Blattabacterium* strains sequenced are summarized in table 1. The BBor genome has a total size of 638,184 bp (with a coverage of 41X) and is composed of a 634,449-bp chromosome and a 3,735-bp plasmid, with a GC content of 28.2% and 30.6%, respectively. Gene order is the same as in BPam and is highly conserved among all five *Blattabacterium* strains (fig. 1). In total, 620 putative genes were identified on the chromosome and seven on the plasmid, distributed as follows: 579 protein-coding genes, one operon coding for the three rRNAs, 33 tRNAs, one tmRNA, the RNA component for the signal recognition particle, and the RNase P. Nine genes were finally annotated as pseudogenes: *cysH*, *cysI*, *cysG*, *hemC*, *lpxP*, *hemD*, *dut*, as well as the genes coding for an hypothetical protein present also in BPam, BCpu, and BBge, which is included in the orthology group blb_0578 (see supplementary table S2, Supplementary Material online), and for the ABC transporter clustered in the orthologous group blb_0575. The genes *cysH* and *cysI* are both involved in sulfate reduction. The genes *hemC*, *hemD,* and *cysG* participate in biosynthesis of heme groups, the latter in siroheme biosynthesis. The genes *lpxP* and *dut* code for a thermonuclease family protein and a deoxyuridine triphosphatase, respectively. Finally, seven genes are duplicated in this genome: *rodA*, *uvrD*, *lpdA miaB argD*, *ppiC,* and *ser*C, the first five in all five genomes, whereas *ppiC* and *serC* have only one copy in BMda and BCpu, respectively. In addition, *dut* has a pseudogenized copy in BBor, whereas

Table 1

General Genomic Features of the Five Sequenced *Blattabacterium* Strains

Strain[a]	BBor	BPam	BCpu	BMda	BBge
GenBank accession number[b]	CP003605, CP003606	NC_013418.2, NC_0.13419	CP003015.1, CP003016.1	NC_016146.1, NC_016150.1	NC_013454.1, NC_015679.1
Genome size (bp)	638,184	640,442	609,561	590,554	640,335
Plasmids	1	1	1	1	1
Plasmid size (bp)	3,735	3,448	3,816	3,306	3,485
Chromosome size (bp)	634,449	636,994	605,745	587,248	636,850
GC content (%)	28.1	28.2	23.8	27.5	27.1
Total number of genes[c]	628 (7)	621 (4)	589 (4)	593 (4)	631 (4)
CDSs	579	582	548	544	590
rRNAs	3	3	3	3	3
tRNAs	33	33	32	34	34
Other ncRNAs	3	3	3	3	3
Duplicated genes	7	8	7	7	9
Pseudogenes	9	6	3	9	1

[a] BBor, *Blattabacterium* from *Blatta orientalis*; BPam, *Blattabacterium* from *Periplaneta americana*; BBge, *Blattabacterium* from *Blattella germanica*; BCpu, *Blattabacterium* from *Cryptocercus punctulatus,* and BMda for *Blattabacterium* from *Mastotermes darwiniensis*.

[b] Accession number for chromosome above, accession number for plasmid below.

[c] Number in parenthesis reflects the number of genes coded in the plasmid.

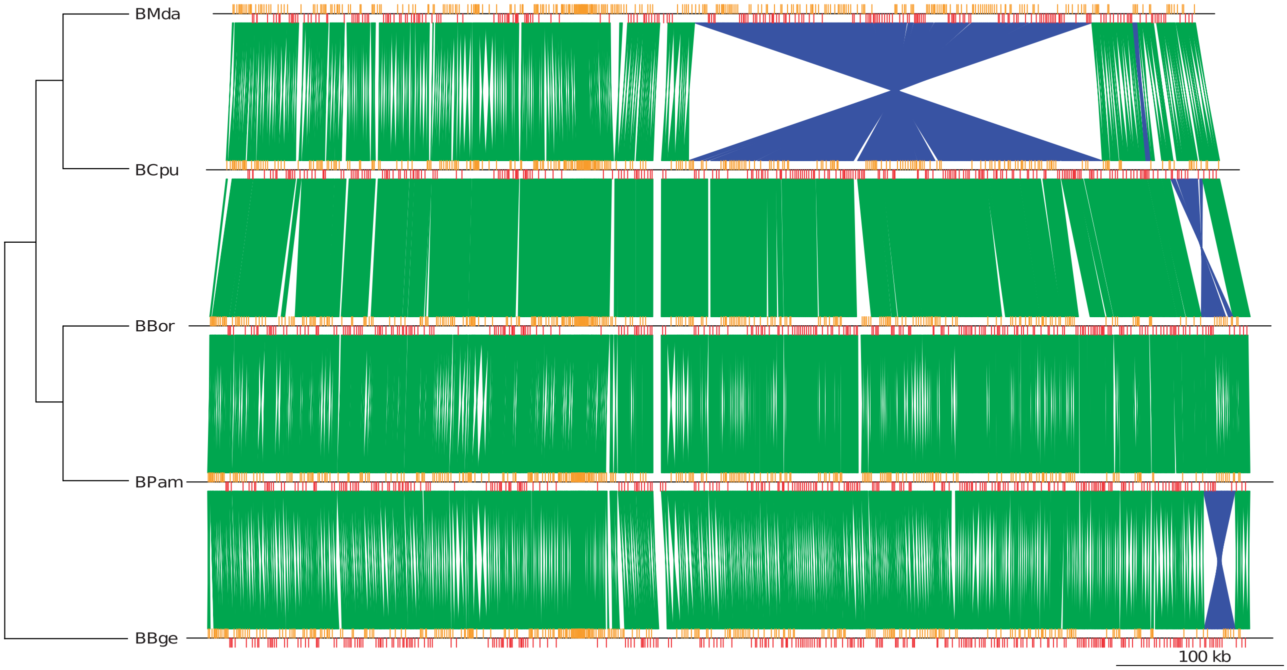


Fig. 1.—Gene order comparison between all *Blattabacterium* strains. Red shows CDS located in the leading strand, and orange indicates CDS coded on lagging strand. Lines between genomes connect orthologous genes in green if genes are in the same orientation, in blue if they are inverted. In this case, the first gene in all strains is *pdxJ*.

hemD maintains only one intact copy (but different paralogs) in all five genomes (supplementary table S4, Supplementary Material online).

Pan-Genome Reconstruction and Functional Profiles

The pan-genome for all five *Blattabacterium*-sequenced genomes possess 612 CDS, one ribosomal operon (the three rRNA genes, namely 16S, 23S, and 5S rRNA), 34 tRNA genes, and three ncRNA genes (fig. 2). The core shared by all strains accounts for 504 CDS, three rRNAs, 31 tRNA genes, and three ncRNAs. The accessory genome comprises 108 protein-coding genes and three tRNA genes, distributed as follows: 16 genes are strain specific, 13 are shared by two strains, 39 by three strains, and finally 43 by four strains

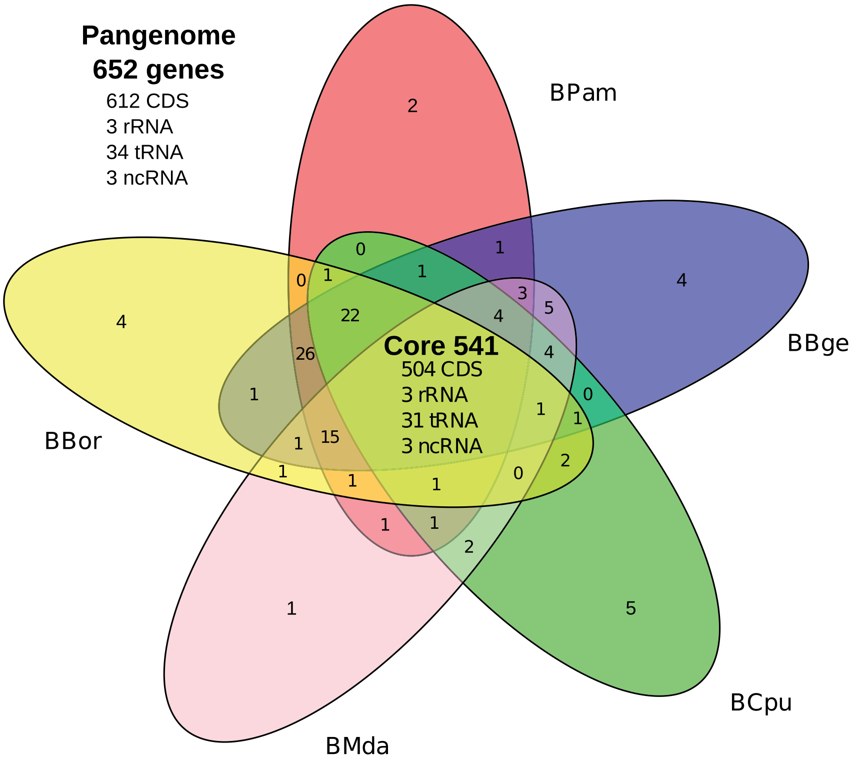


FIG. 2.—Venn diagram showing the genes encoded by each *Blattabacterium* strain. The core genes are those located at the intersection of five circles.

(fig. 2). Of the tRNA genes, one codes for proline (lost in BPam and BCpu; anticodon GGG), one for arginine (lost in BCpu; anticodon CCG), and one for valine (lost in BBor; anticodon TAC). It is worth mentioning that two annotated CDS embedded in the 23S rRNA gene in BMda (Sabree et al. 2012), namely MADAR_308 (cell wall-associated hydrolase) and MADAR_309 (hypothetical protein) (blb_589 and blb_611, respectively, in supplementary table S2, Supplementary Material online), have been removed from the accessory genome because they are false positive as the result of genome misannotations as shown by Tripp et al. (2011).

Protein-coding genes have been classified in COG categories for all five *Blattabacterium* strains (supplementary table S2, Supplementary Material online). According to the Kruskal–Wallis nonparametric test ($\chi^2 = 0.2244$, df = 4, *P* value = 0.9942), there were not any statistical differences in COG distribution among the five strains. Notwithstanding, a clustering diagram, with the functional profile of *F. psychrophilum* as outgroup, shows the functional profile of BBor is closer to that observed in the pan-genome and the other two omnivorous strains, BBge and BPam, whereas the strains from the two wood-feeding species, Cpu and Mda, cluster with the core genome (fig. 3). The most represented functional category in the endosymbiont genomes are genes involved in translation (J), accounting for 20% of all genes. The second most represented functional category, especially in the omnivorous species, contains genes involved in amino acid transport and metabolism (E) (13%) and shows the main gap between the wood-feeding species and the rest, because it harbors most of the gene losses described in these two strains (Neef et al. 2011; Sabree et al. 2012).

Differential Gene Losses and Evolutionary History of *Blattabacterium* Genomes

To obtain a reliable topology of the different *Blattabacterium* lineages and reconstruct the chronology of gene losses, a phylogenetic reconstruction was performed with the five *Blattabacterium* strains and the free-living bacterium *F. psychrophilum* as outgroup. The 464 protein-coding genes found to be orthologous among *Blattabacterium* strains and possessing orthologs in the genome of *F. psychrophilum* were used for the analysis, giving rise to a concatenated sequence of aligned proteins with 173,523 sites. The best evolutionary model for our data set, estimated with ProtTest (Abascal et al. 2005), was CpREV + G + F. The resulting phylogeny situates the two Blattidae endosymbionts as a sister clade to the one formed by BCpu and BMda, placing the BBge strain as the most basal lineage, thus corroborating previous analyses of host genes (Inward et al. 2007) (supplementary fig. S1, Supplementary Material online).

The phylogenetic reconstruction of the five *Blattabacterium* lineages, the corresponding pan-genome, and the set of retained genes in each genome were used to establish the evolutionary history of gene losses, basically following the same

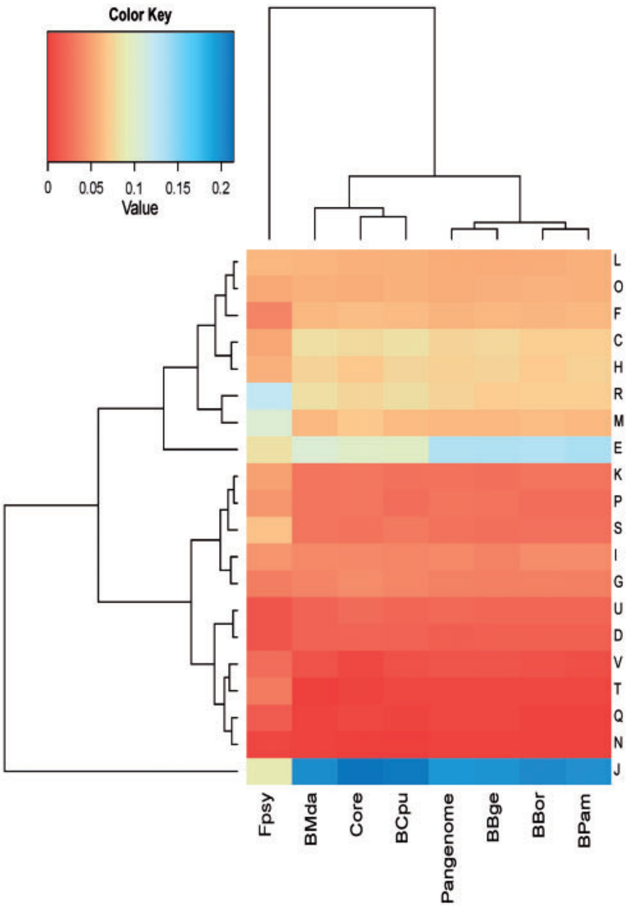


Fig. 3.—Heat map comparison of COG frequency profiles among different *Blattabacterium* strains with their pan-genome and core and the free-living Bacteroidete *F. psychrophilum*. J, translation; K, transcription; L, replication, recombination, and repair; D, cell cycle control; M, cell/wall membrane biogenesis; N, cell motility; O, posttranslational modification, protein turnover, chaperones; P, inorganic ion transport and metabolism; T, signal transduction mechanism; U, intracellular trafficking and secretion; V, defense mechanism; V, defense mechanism; C, energy production and conversion; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; Q, secondary metabolites biosynthesis, transport, and catabolism; R, general function prediction only; and S, function unknown.

strategy as previously reported (Lamelas, Gosalbes, et al. 2011). The underlying assumption to our analyses is that the endosymbiotic genomes have experienced no gene gain, thus the retained genes are only vertically inherited. Although this assumption is legitimate (e.g., endosymbiotic bacteria are housed in bacteriocytes and lack complete set of recombination genes), we nevertheless tested for possible events of horizontal transfer of genes from other bacteria to the endosymbiont and we found no evidence supporting such events (supplementary results, Supplementary Material online). In consequence, the status of each ancestral gene in each extant *Blattabacterium* genome was evaluated: unique or convergent losses and active state or pseudogenized sequence (table 2). Unique losses correspond to those affecting one specific strain or occurring before divergence of related strains (i.e., the loss took place in their most recent common ancestor). When unrelated lineages show specific pseudogenized sequences or the absence of a gene, we assume that convergent loss took place in those lineages. During the evolution of *Blattabacterium*, a total of 108 CDS and three tRNA genes have been lost in a total of 161 independent events, 77 of which were unique losses and 34 convergent losses, 18 genes have been lost twice, and another 16 have been lost three times (table 2). The number of specific losses in each lineage is indicated in figure 4, and a detailed description of all the gene losses undergone during *Blattabacterium* strain diversification is provided in supplementary results, Supplementary Material online. We would like to remark that some of the losses might be related to specific events during cockroach evolution, such as those that occurred before the split between the two Blattidae and the two wood-feeding lineages. Hence, the loss of the three first steps of the Krebs cycle may have taken place in the common ancestor of BPam and BBor, whereas the ancestor of BMda and BCpu lost the ability to synthesize several essential and nonessential amino acids (fig. 4). Also, it is worth mentioning that two convergent changes in sulfur source for metabolism, from sulfate to sulfide, have occurred in the ancestor of BPam and BBor and in the BCpu lineage (fig. 5).

Dating the Split Times

LRT supported the molecular clock-like evolution for 313 core CDS out of a total of 504 CDS. No functional COG category was found particularly enriched for genes evolving under the molecular clock hypothesis (χ^2 test, *P* value = 0.991). In this set of 313 core CDS, 290 had an ortholog in *F. pshychophilum*, of which for 281, the molecular clock hypothesis could not be rejected, and of the 281, 246 show less than 2.5 nucleotide substitutions per site. Taking into account that the split of *B. germanica* from the rest of roaches was dated in 140 Ma, the divergence times between BPam and BBor, and between BMda and BCpu were estimated as 12.9 ± 8.4 and 88.6 ± 19.8 Ma, respectively.

Discussion

The genome sequence of BBor has confirmed the extreme stability of the *Blattabacterium* genome architecture despite the fact extant cockroach families appeared more than 140 Ma (Sabree et al. 2010, 2012; Neef et al. 2011). The only rearrangements are a 20 kb inversion, which occurred in the Blattidae family lineage, and two inversions in the strain BMda, one of 242 kb and another of 2.9 kb, containing the genes *rffH*, *dut,* and *wzxC* (fig. 1). As in other endosymbiotic bacteria, synteny is highly conserved among the

Table 2
Classification of Gene Losses in *Blattabacterium* Strains

Losses		BBor	BPam	BMda	BCpu	BBge	*n*	Genes
Unique loss (77)		–	–	–	–	+	4	*sirBC*, BLBBGE_195, BLBBGE_159, BLBBGE_595
		+	+	+	+	–	1	blb_0523
		–	–	+	+	+	4	*Icd, acnA, gltA*, blb_0575
		–	+	+	+	+	4	*dut*, blb_0543, blb_544, tRNA-Val
		+	–	+	+	+	1	*secE*
		+	+	–	–	+	26	*ilvA, ilvB, ilvC, ilvD, ilvH, leuA, leuC, leuD, leuB, trpE, trpG, trpD, trpC, trpB, trpA, thrB, thrC, lysA, argH, metC, ygfA, asnC, ung*, blb_0546, blb_0547, blb_0548
		+	+	+	–	+	15	*metE, metF, cysE, cysK, serC, tgt, luxE, clpB, mdlA, msbA*, blb_0537, blb_0507, blb_0536, blb_0542, tRNA-Arg
		+	+	–	+	+	22	*ubiE, ribD, nadD, mvaK1, mvaD, efp, truB, trmH, nth, lolE, hemD, ppiC, clpX, dksA, secDF, era*, blb_0525, blb_0518, blb_0522, blb_0525, blb_0529, blb_0541
Convergent losses (34)								
	Twice	–	–	+	+	–	2	*ywrO*, blb_0586
		–	–	+	–	+	5	cysD, cysN, cysI, hemD, blb_0595,
		+	–	–	–	+	1	blb_0591
		–	+	–	–	+	1	blb_0596
		+	+	+	–	–	1	*desA*
		+	+	–	+	–	1	blb_0565
		–	+	+	+	–	1	*ccoS*
		–	+	+	–	+	3	*cysG, cysH, hemC*
		–	+	–	+	+	1	blb_0578
		+	–	–	+	+	1	*lolD*
		+	–	+	–	+	1	tRNA-Pro
	3-fold	+	–	–	–	–	4	BLBBOR_p001, BLBBOR_p002, BLBBOR_p007, BLBBOR_609
		–	+	–	–	–	2	*trpF*,[a] BPLAN_099
		–	–	+	–	–	1	MADAR_453
		–	–	–	+	–	5	BLBCPU_006, BLBCPU_149, BLBCPU_186, BLBCPU_463, BLBCPU_511
		–	+	+	–	–	1	*lpxP*
		+	–	+	–	–	1	blb_0585
		+	–	–	+	–	2	*RnpA*, blb_0587

Note.—Strain abbreviations as in table 1. +, gene present; –, gene absent or pseudogene. For genes annotated as hypothetical protein, the number of the orthologous cluster in which the gene is classified is indicated by the code blb_ followed by a number (see supplementary table S2, Supplementary Material online). In the case the hypothetical protein is present in only one strain, the locus tag of the GenBank file is indicated.

[a] This gene is fused to *trpB* in all other strains.

different strains; however, surprisingly, *Blattabacterium* has also maintained a very similar gene content, given 83.0% of the pan-genome genes are represented in the core (table 3). This conservation is even more striking when only the strains from omnivorous cockroaches are considered: 93.9% of the pan-genome genes are included in the core. Another example of endosymbiotic bacteria of omnivorous hosts is *Blochmannia* sp., primary endosymbiont of *Camponotus* sp. ants. In this case also, the number of genes is well conserved because 93.5% of the pan-genome genes are in the core (Williams and Wernegreen 2010). However, the divergence time among the three *Blochmannia* strains has been established at approximately 20 Ma (Degnan et al. 2004; Gómez-Valero et al. 2008), whereas the symbiosis between cockroaches and *Blattabacterium* originated at least 140 Ma (Clark et al. 2001; Lo et al. 2003). Comparison with a similarly ancient symbiotic association, similar to the one established between *Buchnera* and aphids between 86 and 164 Ma (von Dohlen and Moran 2000), and considering only true primary endosymbionts, that is, those without any known metabolic complementation with other symbionts (as is the case of the cedar and tuja aphids), the genetic conservation is much lower, because several gene losses have occurred in the different lineages. In particular, only 74% of the pan-genome genes are present in the core (table 3). All these data suggest that massive gene losses may have occurred in *Blattabacterium* genomes soon after the transition from the free-living state to the intracellular life style, establishing an optimal genome to fulfill the host requirements, albeit minimized. We postulate that this gene content stasis may be correlated with the role played by *Blattabacterium* in the nitrogen metabolism in cockroaches (González-Domenech et al. 2012). In fact, gene losses

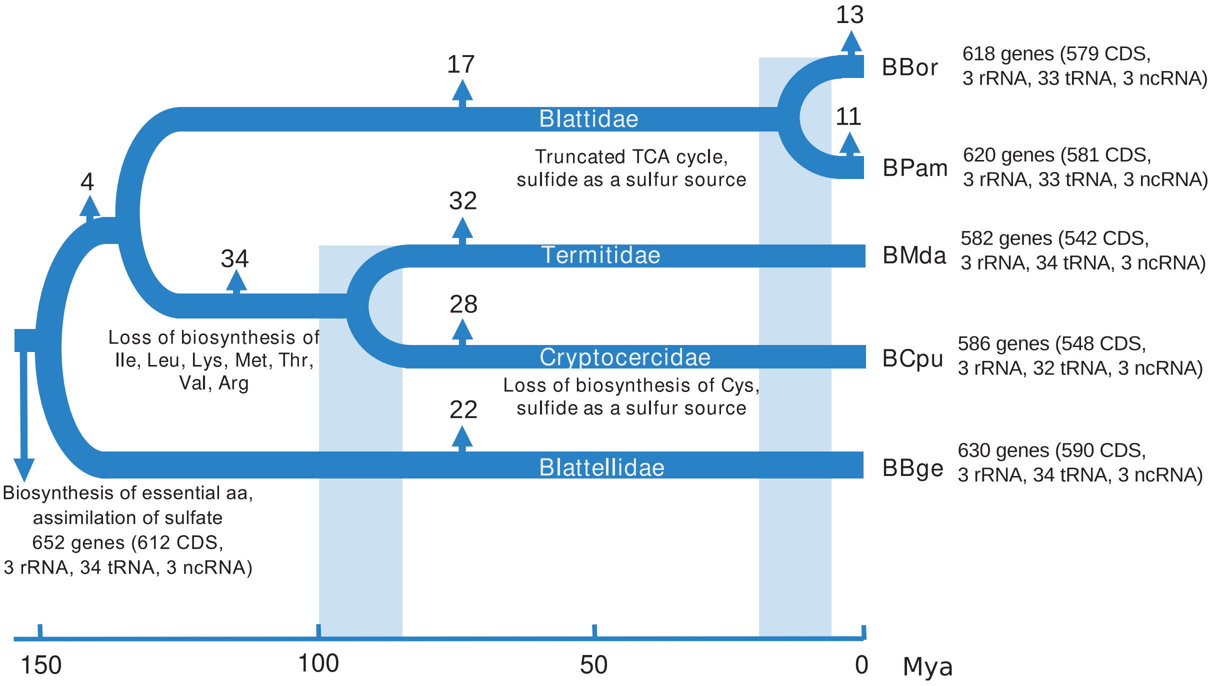
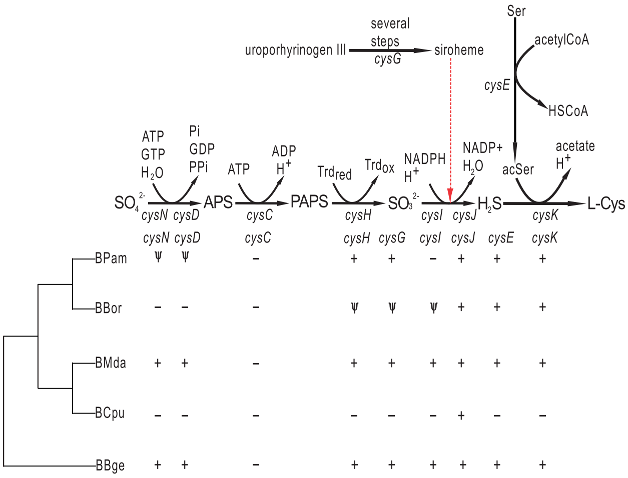

13
17
Blattidae
BBor
618 genes (579 CDS, 3 rRNA, 33 tRNA, 3 ncRNA)
11
BPam
620 genes (581 CDS, 3 rRNA, 33 tRNA, 3 ncRNA)
4
Truncated TCA cycle, sulfide as a sulfur source
32
34
Termitidae
BMda
582 genes (542 CDS, 3 rRNA, 34 tRNA, 3 ncRNA)
28
Loss of biosynthesis of Ile, Leu, Lys, Met, Thr, Val, Arg
Cryptocercidae
BCpu
586 genes (548 CDS, 3 rRNA, 32 tRNA, 3 ncRNA)
Loss of biosynthesis of Cys, sulfide as a sulfur source
22
Blattellidae
BBge
630 genes (590 CDS, 3 rRNA, 34 tRNA, 3 ncRNA)
Biosynthesis of essential aa, assimilation of sulfate
652 genes (612 CDS, 3 rRNA, 34 tRNA, 3 ncRNA)
150
100
50
0
Mya

Fig. 4.—Gene loss during *Blattabacterium* diversification. Number of genes in each strain is indicated. Numbers above the branches indicate gene loss events. Host family names are indicated on each branch. Abbreviations of *Blattabacterium* strains as in table 1.



	cysN	cysD	cysC	cysH	cysG	cysI	cysJ	cysE	cysK
BPam	ψ	ψ	–	+	+	–	+	+	+
BBor	–	–	–	ψ	ψ	ψ	+	+	+
BMda	+	+	–	+	+	+	+	+	+
BCpu	–	–	–	–	–	–	+	–	–
BBge	+	+	–	+	+	+	+	+	+

Fig. 5.—Sulfate assimilatory pathway genes in the different *Blattabacterium* sp. strains. +, gene present; –, gene absent; ψ, pseudogene. Abbreviations of *Blattabacterium* strains as in table 1.

observed in the Blattidae lineage only affect metabolism in peripheral activities (i.e., the change from sulfate to sulfide as sulfur source) or irrelevant metabolic steps in the Krebs cycle. Hence, one of the most remarkable losses in BBor is the absence of genes coding for the three first steps of the TCA cycle, citrate synthase (*gltA*), aconitate hydratase (*acnA*), and isocitrate dehydrogenase (*icd*). Nevertheless, this metabolic feature shared with BPam (Sabree et al 2009) has no effect on metabolic network functionality as shown by FBA (González-Domenech et al. 2012).

Most of the core genes in the *Blattabacterium* pan-genome (62.1%) follow a molecular clock, which allowed us to determine the split times between the Blattidae lineages and the Termitidae and Cryptocercidae lineages (fig. 4 and supplementary fig. S1, Supplementary Material online). The divergence time calculated for the wood-feeding lineages is closer to the split between *B. germanica* and the rest of the cockroach species (88.6 vs. 140 Ma). This indicates that the metabolic changes taking place in the ancestor of the endosymbionts of *M. darwiniensis* and *C. punctulatus,* mainly affecting the biosynthesis of essential and nonessential amino acids (fig. 4), took place very early in the evolutionary history of the lineage leading to Termitidiae and Cryptocercidae. Conversely, *Blatta orientalis* and *P. americana* are much more recent lineages (split occurring 12.9 Ma) and, as previously stated, the metabolic profile of their common ancestor indicates notable metabolic stasis.

Metabolic comparison among *Blattabacterium* strains highlights the role of the bacterial metabolic network in host physiology. For instance, in the case of nitrogen metabolism, urease corresponds to the *Blattabacterium* core genome, suggesting that the role of the endosymbiont in urate mobilization is ancestral and conserved in all studied *Blattabacterium* lineages. Hence, the combination of urease with the urea cycle (interrupted only in BCpu and BMda by the absence of argininosuccinate lyase [ASL]) is a metabolic network with the

Table 3

Number of Genes in the Pan-Genome and the Core in *Blattabacterium*, *Blochmannia* sp., and *Buchnera*

	Genes in Pan-Genome	Genes in Core (%)[a]
Blattabacterium		
Strains: BBge, BPam, BBor, BCpu, and BMda	652	541 (83.0)
Onmivorous strains: BBge, BPam, and BBor	644	605 (93.9)
Blochmannia		
Strains: Bfl, Bpen, and Bva	660	617 (93.5)
Buchnera		
Strains: BAp, BKo, BUa, BSg, and BBp	650	481 (74.0)

Note.—Strain abbreviations are as follows. *Blochmannia* strains: Bfl, *Blochmannia floridanus*; Bpen, *B. pennsylvanicus*; Bva, *B. vafer*. *Buchnera* strains: BAp, *Buchnera* from *Acyrtosiphon pisum*; BKo, *Buchnera* from *A. kondoi*; BUa, *Buchnera* from *Uroleucon ambrosiae*; BSg, *Buchnera* from *Schizaphis graminum*; BBp; *Buchnera* from *Baizongia pistaciae*.

[a]Proportion of the pan-genome present in the core genome.

potential to catabolize nitrogen compounds and generate ammonia (López-Sánchez et al. 2009; González-Domenech et al. 2012). Thus, there is a biochemical explanation for the classical model of urate deposits acting as nitrogen storage in the cockroach fat body (Mullins and Cochran 1974; Cochran et al. 1979; Cochran 1985) and the intriguing ammonotelism displayed by these insects (González-Domenech et al. 2012).

With respect to gene losses in the two wood-feeding hosts, a higher number of events have occurred although very early in their evolutionary history. Remarkably, these affected the biosynthesis of essential and nonessential amino acids before the split between Cryptocercidae and Termitidae, and the convergent loss of sulfate assimilation ability in the BCpu lineage. As stated earlier, the symbionts from the three omnivorous cockroaches are self-sufficient in terms of supplying their host with essential amino acids, whereas up to six pathways for essential amino acids have been lost in the wood-feeding hosts. On the basis of the phylogenetic analysis, we propose that a major reduction in amino acid biosynthetic ability took place in the common ancestor of the two wood-feeding lineages. It has been postulated that this metabolic impairment of the endosymbiont could have been compensated for by amino acid supply from the diet and/or their syntheses by gut microbiota (Neef et al. 2011; Sabree et al. 2012), comparable with the metabolic complementation observed in some bacterial consortia, like *Buchnera* and *Serratia* in cedar aphids (Lamelas et al. 2011). Additionally, the gene *argH* coding for ASL, the last step in the Arg biosynthesis pathway, has been lost in both BCpu and BMda strains, thus a host-encoded ASL may also have taken over Arg biosynthesis; nonetheless, the contribution of diet or gut microbiota cannot be ruled out. In fact, 8 of the 10 complete genomes from insects accessible in the KEGG database (http://www.genome.jp/kegg/, last accessed February 4, 2013) contain genes for ASL. The corresponding gene is absent from the pea aphid *Acyrthosiphon pisum* genome (but present in its primary endosymbiont *Buchnera aphidicola*) nor is it present in the human body louse *Pediculus humanus* genome (in this case, Arg may be supplied by the blood ingested by the louse). Studies of metabolic modeling supported by transcriptomic and proteomic data (Macdonald et al. 2012) have revealed the action of host cell enzymes at the end of amino acid biosynthesis in the pea aphid and *Buchnera aphidicola*, which could have important regulatory consequences.

There are some remarkable convergent metabolic traits, in particular, the capacity for assimilating sulfate. The symbionts BBge and BMda possess the genes coding for all enzymes involved in sulfate assimilation, with the exception of 5′-phosphosulfate kinase (encoded by *cysC*). Experimental data indicate that this pathway may be operative, because aposymbiotic individuals of *B. germanica* are unable to incorporate sulfate into cysteine and methionine (Block and Henry 1962). As this pathway is absent in the other *Blattabacterium* strains, it was most likely lost in at least two independent events, one in the lineage leading to BCpu and the other during the evolution of the endosymbionts of Blattidae species (fig. 5). The genes *cysN*, *cysD,* and *cysI* are found as pseudogenes in BPam (Sabree et al. 2009), whereas in BBor, they have been completely lost. With respect to the remaining genes of this pathway, *cysH* and *cysG* are pseudogenized in BBor but seem to be functional in BPam. However, gene *cysJ* is present in all sequenced *Blattabacterium* strains, even in those which are unable to assimilate sulfate. The product of this gene is a flavoprotein, which forms sulfite reductase with the hemoprotein coded by *cysI*. In this case, *cysJ* may have been recruited by other processes, because it has previously been described to work as FMN reductase (Covès et al. 1993).

The genomes of BBor, BCpu, and BMda retain seven of the nine duplicated genes found in BBge (López-Sánchez et al. 2009; Neef et al. 2011; Sabree et al. 2012), whereas eight of these genes are also maintained in BPam (Sabree et al. 2009). The presence of these duplicated genes is quite surprising in the context of a reduced genome, similar to that of *Blattabacterium* and other primary endosymbionts. The fact most of these genes are still present in all three strains points to their possible functional role in bacterium physiology and that these genes might be ecoparalogs (Sanchez-Perez et al. 2008).

In summary, the hosts harboring *Blattabacterium* strains have been evolving separately for a very long time; however, the genomic and metabolic architecture of these symbiotic bacteria remains strikingly stable. Thus, the basic genetic and metabolic traits of *Blattabacterium* were established in a short time period, when these bacteria infected the common ancestor of all cockroaches and termites.

Supplementary Material

Supplementary results, figure S1, and tables S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The sequencing service of the University of Valencia (SCSIE) is acknowledged, as well as the suggestions by two anonymous reviewers. The authors acknowledge Dr M.A. Fares for his helpful comments on the manuscript. This work was supported by the Ministerio de Ciencia e Innovación, Spain, BFU2009-12895-C02-01/BMC and BFU2012-39816-C02-01 to A.L., and the Generalitat Valenciana, Spain, Prometeo/2009/092 to A.M. R.P.-N. was recipient of a fellowship from the Ministerio de Ciencia e Innovación, Spain.

Literature Cited

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Bandi C, et al. 1995. The establishment of intracellular symbiosis in an ancestor of cockroaches and termites. Proc Biol Sci. 268:293–299.

Baumann P. 2005. Biology bacteriocyte-associated endosymbionts of plant sap-sucking insects. Annu Rev Microbiol. 59:155–189.

Block RJ, Henry SM. 1962. Metabolism of the sulphur amino acids and of sulphate in *Blattella germanica*. Nature 191:392–393.

Carver T, et al. 2005. ACT: the Artemis Comparison Tool. Bioinformatics 21:3422–3423.

Chen F, et al. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. 34: D363–D368.

Clark JW, et al. 2001. Coevolution between a cockroach and its bacterial endosymbiont: a biogeographical perspective. Proc Biol Sci. 268: 393–398.

Cochran D. 1985. Nitrogen excretion in cockroaches. Annu Rev Entomol. 30:29–49.

Cochran DG, Mullins DE, Mullins KJ. 1979. Cytological changes in the cat body of the American cockroach *Periplaneta americana* in relation to dietary nitrogen levels. Ann Entomol Soc Am. 72:197–205.

Covès J, Nivière V, Eschenbrenner M, Fontecave M. 1993. NADPH-sulfite reductase from *Escherichia coli*. A flavin reductase participating in the generation of the free radical of ribonucleotide reductase. J Biol Chem. 268:18604–18609.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest2: more models, new heuristics and parallel computing. Nat Methods. 9:772.

Degnan PH, Lazarus AB, Brock CD, Wernegreen JJ. 2004. Host-symbiont stability and fast evolutionary rates in an ant-bacterium association: cospeciation of *Camponotus* species and their endosymbionts, *candidatus Blochmannia*. Syst Biol. 53:95–110.

Delcher A, Bratke K, Powers E, Salzberg S. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23: 673–679.

Feldhaar H, et al. 2007. Nutritional upgrading for omnivorous carpenter ants by the endosymbiont *Blochmannia*. BMC Biol. 5:48.

Gil R, et al. 2003. The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. Proc Natl Acad Sci U S A. 100: 9388–9393.

Gómez-Valero L, et al. 2008. Patterns and rates of nucleotide substitution, insertion and deletion in the endosymbiont of ants *Blochmannia floridanus*. Mol Ecol. 17:4382–4392.

González-Domenech CM, et al. 2012. Metabolic stasis in an ancient symbiosis: genome-scale metabolic networks from two *Blattabacterium cuenoti* strains, primary endosymbionts of cockroaches. BMC Microbiol. 12:S5.

Guy L, Kultima JR, Andersson SG. 2010. genoPlotR: comparative gene and genome visualization in R. Bioinformatics 26:2334–2335.

Inward D, Beccaloni G, Eggleton P. 2007. Death of an order: a comprehensive molecular phylogenetic study confirms that termites are eusocial cockroaches. Biol Lett. 3:331–335.

Kambhampati S. 1995. A phylogeny of cockroaches and related insects based on DNA sequence of mitochondrial ribosomal RNA genes. Proc Natl Acad Sci U S A. 92:2017–2020.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33: 511–518.

Lamelas A, et al. 2011. *Serratia symbiotica* from the aphid *Cinara cedri*: a missing link from facultative to obligate insect endosymbiont. PLoS Genet. 7:e1002357.

Lamelas A, Gosalbes MJ, Moya A, Latorre A. 2011. New clues about the evolutionary history of metabolic losses in bacterial endosymbionts, provided by the genome of *Buchnera aphidicola* from the aphid *Cinara tujafilina*. Appl Environ Microbiol. 77:4446–4454.

Lo N, Bandi C, Watanabe H, Nalepa C, Beninati T. 2003. Evidence for cocladogenesis between diverse dictyopteran lineages and their intracellular endosymbionts. Mol Biol Evol. 20:907–913.

Lo N, Beninati T, Stone F, Walker J, Sacchi L. 2007. Cockroaches that lack *Blattabacterium* endosymbionts: the phylogenetically divergent genus *Nocticola*. Biol Lett. 3:327–330.

López-Sánchez MJ, et al. 2008. Blattabacteria, the endosymbionts of cockroaches, have small genome sizes and high genome copy numbers. Environ Microbiol. 10:3417–3422.

López-Sánchez MJ, et al. 2009. Evolutionary convergence and nitrogen metabolism in Blattabacterium strain Bge, primary endosymbiont of the cockroach *Blattella germanica*. PLoS Genet. 5:e1000721.

Macdonald SJ, Lin GG, Russell CW, Thomas GH, Douglas AE. 2012. The central role of the host cell in symbiotic nitrogen metabolism. Proc Biol Sci. 279:2965–2973.

McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol. 10:13–26.

Moya A, Peretó J, Gil R, Latorre A. 2008. Learning how to live together: genomic insights into prokaryote-animal symbioses. Nat Rev Genet. 9: 218–229.

Mullins DE, Cochran DG. 1972. Nitrogen excretion in cockroaches: uric acid is not a major product. Science 177:699–701.

Mullins DE, Cochran DG. 1974. Nitrogen metabolism in the American cockroach: an examination of whole body and fat body regulation of cations in response to nitrogen balance. J Exp Biol. 61: 557–570.

Mullins DE, Cochran DG. 1976. A comparative study of nitrogen excretion in twenty-three cockroach species. Comp Biochem Physiol A Comp Physiol. 53:393–399.

Murray MG, Thompson WF. 1980. Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. 8:4321–4325.

Nawrocki EP, et al. 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics 25:1335–1337.

Neef A, et al. 2011. Genome economization in the endosymbiont of the wood roach *Cryptocercus punctulatus* due to drastic loss of amino acid synthesis capabilities. Genome Biol Evol. 3:1437–1448.

Pérez-Brocal V, et al. 2006. A small microbial genome: the end of a long symbiotic relationship? Science 314:312–313.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16:276–277.

Rutherford K, et al. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16:944–945.

Sabree ZL, Degnan PH, Moran NA. 2010. Chromosome stability and gene loss in cockroach endosymbionts. Appl Environ Microbiol. 76: 4076–4079.

Sabree ZL, et al. 2012. Genome shrinkage and loss of nutrient-providing potential in the obligate symbiont of the primitive termite *Mastotermes darwiniensis*. Appl Environ Microbiol. 78:204–210.

Sabree ZL, Kambhampati S, Moran NA. 2009. Nitrogen recycling and nutritional provisioning by *Blattabacterium*, the cockroach endosymbiont. Proc Natl Acad Sci U S A. 106:19521–19526.

Sanchez-Perez G, Mira A, Nyiro G, Pasic L, Rodriguez-Valera F. 2008. Adapting to environmental changes using specialized paralogs. Trends Genet. 24:154–158.

Shigenobu S, Wilson AC. 2011. Genomic revelations of a mutualism: the pea aphid and its obligate bacterial symbiont. Cell Mol Life Sci. 68: 1297–1309.

Staden R, Beal K, Bonfield J. 2000. The Staden package, 1998. Methods Mol Biol. 132:115–130.

Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456–463.

Tripp HJ, Hewson I, Boyarsky S, Stuart JM, Zehr JP. 2011. Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. Nucleic Acids Res. 39:8792–8802.

Vasudevan S, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

von Dohlen C, Moran NA. 2000. Molecular data support a rapid radiation of aphids in the Cretaceous and multiple origins of host alternation. Biol J Linn Soc. 71:689–717.

Vrsansky P, Vishniakova VN, Rasnitsyn AP. 2002. Order Blattida Latreille, 1810. In: Rasnitsyn AP, Quicke DLJ, editors. History of insect orders. Dordrecht (The Netherlands): Kluwer Academic Publishers. p. 263–270.

Warnes G. 2011. gplots: Various R programming tools for plotting data. Available from: http://cran.r-project.org/web/packages/gplots//index.html (last accessed February 4, 2013).

Williams LE, Wernegreen JJ. 2010. Unprecedented loss of ammonia assimilation capability in a urease-encoding bacterial mutualist. BMC Genomics 11:687.

Worning P, et al. 2006. Origin of replication in circular prokaryotic chromosomes. Environ Microbiol. 8:353–361.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Associate editor: Richard Cordaux