# Parallel genomic evolution and metabolic interdependence in an ancient symbiosis

John P. McCutcheon[†‡] and Nancy A. Moran[‡§]

[†]Center for Insect Science and [‡]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721-0088

Edited by E. Peter Greenberg, University of Washington School of Medicine, Seattle, WA, and approved October 19, 2007 (received for review September 18, 2007)

**Obligate symbioses with nutrient-provisioning bacteria have originated often during animal evolution and have been key to the ecological diversification of many invertebrate groups. To date, genome sequences of insect nutritional symbionts have been restricted to a related cluster within Gammaproteobacteria and have revealed distinctive features, including extreme reduction, rapid evolution, and biased nucleotide composition. Using recently developed sequencing technologies, we show that *Sulcia muelleri*, a member of the Bacteroidetes, underwent similar genomic changes during coevolution with its sap-feeding insect host (sharpshooters) and the coresident symbiont *Baumannia cicadellinicola* (Gammaproteobacteria). At 245 kilobases, *Sulcia*'s genome is approximately one tenth of the smallest known Bacteroidetes genome and among the smallest for any cellular organism. Analysis of the coding capacities of *Sulcia* and *Baumannia* reveals striking complementarity in metabolic capabilities.**

Bacteroidetes | insects | pyrosequencing | Sharpshooters | genome reduction

Members of the Bacteroidetes are widely distributed in nature, and have been reported as prominent members of environments as diverse as coastal marine waters (1), the human gut (2, 3), and dental plaques (4). In insects, a member of the Bacteroidetes called *Sulcia muelleri* (Fig. 1) has been shown to be an ancient symbiont of a large group of sap-feeding insects, in which the initial infection was acquired >260 million years ago (5). In addition to *Sulcia*, these insects have at least one other long-term heritable symbiont (5), exemplified by *Baumannia* in the case of sharpshooters. These symbioses are models of codiversification over long time periods: *Baumannia*, *Sulcia*, and their sharpshooter hosts seem to have diversified through strict vertical association during evolution of this insect group (6).

Sharpshooters feed exclusively on xylem sap, which is the most dilute and unbalanced food source used by herbivores (7, 8). Xylem composition varies depending on the plant assayed, but the primary components are typically a dilute mix of the amino acids glutamate, glutamine, aspartate, and asparagine; some simple organic acids (primarily malate); and various sugars (primarily glucose) (7, 8). The genome sequence of *Baumannia* was recently completed, along with fragments of the *Sulcia* genome, both from the invasive agricultural pest *Homalodisca vitripennis* (formerly *H. coagulata*, also known as the Glassy-Winged Sharpshooter) (9). Analysis of the *Baumannia* genome revealed that it primarily contributes vitamins and cofactors to the host, while encoding at least partial pathways for two of the 10 essential amino acids (9). The partial sequence obtained for *Sulcia* suggested that it is primarily responsible for essential amino acid biosynthesis (9).

To fully and unambiguously assess the role of *Sulcia* in the metabolism of this tripartite symbiosis, we sequenced the genome using pyrosequencing (454 Life Sciences/Roche Applied Science) (10). We attempted to enrich for *Sulcia* DNA in our sample through dissection of the appropriate bacteriome. Nonetheless, as in previous genome sequencing projects on noncultivable, host-associated microorganisms, we started with a complex sample containing a mixture of DNA from the insect host, *Sulcia*, and *Baumannia*, with the host DNA constituting the majority fraction of the sample by weight and the *Sulcia* DNA representing the majority of the bacterial fraction.

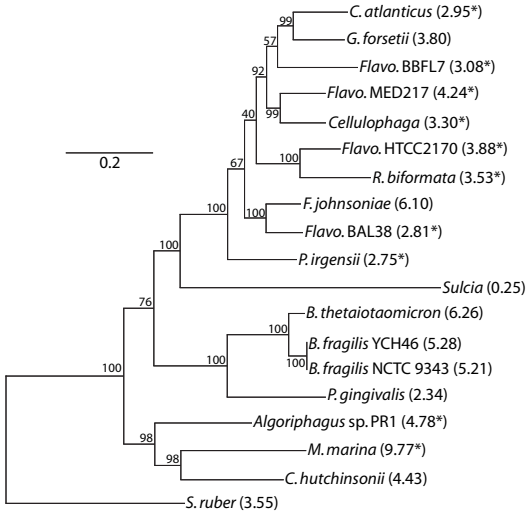


**Fig. 1.** Relationships of *Sulcia muelleri* to other sequenced members of the Bacteroidetes. Bootstrap values for 100 replicates are shown at the bifurcations. The genome sizes (in megabases) are given in parentheses after the organism name. Those sizes marked with an asterisk are from incomplete genome projects and should be considered approximate. (Scale bar: 0.2 changes per site.)

## Results

By maximizing the representation of *Sulcia* in our sample, we succeeded in obtaining deep coverage for this genome. Of the 416 contigs generated from the Newbler (10) assembly, 25 were cleanly separated by a greater average depth of coverage [supporting information (SI) Fig. 4]. Twenty-three of these contigs had gene contents suggesting they belonged to the *Sulcia* genome and were assembled into a complete circular genome based solely on data generated from the 454 run. This genome included almost all of the sequence contigs that had been assigned to *Sulcia* previously (9),

Author contributions: J.P.M. and N.A.M. designed research; J.P.M. performed research; J.P.M. and N.A.M. analyzed data; and J.P.M. and N.A.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequence of *Sulcia muelleri* has been deposited in the GenBank/EMBL/DDBJ database (accession no. CP000770).

[§]To whom correspondence should be addressed at: Department of Ecology and Evolutionary Biology, Biosciences West, Room 310, 1041 East Lowell Street, Tucson, AZ 85721-0088. E-mail: nmoran@email.arizona.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0708855104/DC1.

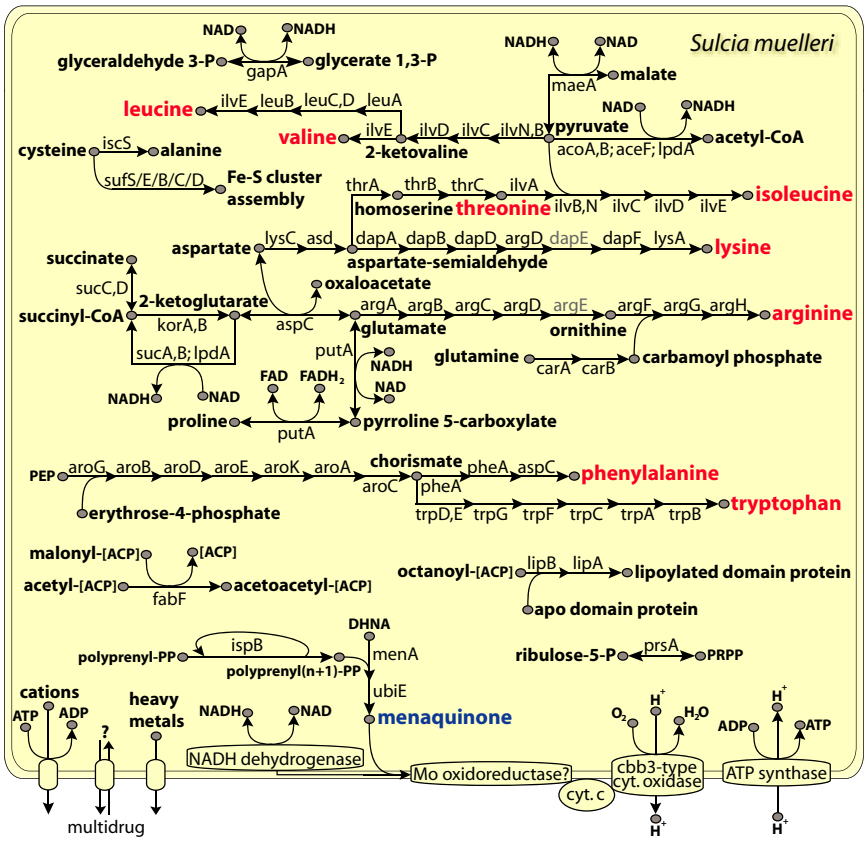


**Fig. 2.** The predicted metabolism of *Sulcia muelleri*. Essential amino acids are shown in red letters, and vitamins/cofactors are shown in blue. Gene names are written in plain letters, and names of selected compounds are written in bold and indicated by a small gray circle. Genes to which an ORF cannot be unambiguously assigned are written in gray.

indicating that the binning criteria were generally accurate but not perfect.

Because pyrosequencing has trouble accurately calling the number of bases in homopolymeric regions, especially for single-base runs longer than 5–8 bases (10, 11), and because *Sulcia*'s high A+T content results in numerous A and T homopolymeric tracts, we expected and found inaccuracies in those regions. Thirty homopolymeric regions that disagreed between the 454- and previously obtained Sanger-generated sequences available for portions of the *Sulcia* genome (9) were retested by PCR and Sanger sequencing; in all cases, the Sanger prediction was confirmed.

To resolve the homopolymeric regions in the *Sulcia* genome, 13,564,883 reads of 33 bases in length were generated from one partial run of an Illumina/Solexa Genome Analyzer, which affords many short, highly accurate reads (12). Mapping these reads onto the *Sulcia* genome corrected 125 additional homopolymeric regions and eliminated almost all frameshifts and in-frame stop codons in the predicted coding regions. However, an in-frame stop in a putative cytochrome *c* assembly protein was well supported by the Solexa data and was left in the genome, and two homopolymeric tracks (one in the tyrosyl tRNA synthetase *tyrS* and one in the valyl tRNA synthetase *valS*) had no Solexa coverage but were changed to extend the reading frame of a protein to match homologs in the sequence databases.

The *Sulcia* genome is 245,530 bp in length with G+C content of 22.4%. It encodes 228 protein genes, 31 tRNAs representing all 20 aa, one rRNA operon (23S, 16S, and 5S), and one tmRNA (SI Table 3). The coding density is 96.1%, with 61 genes overlapping at an average length of 11.8 bases. A COG-based assignment of functional groups reveals that 33.0% and 21.3% of all protein-coding genes are devoted to translation-related and amino acid biosynthesis functions, respectively (SI Fig. 5).

Analysis of *Sulcia*'s predicted metabolic map (Fig. 2) reveals an organism largely devoted to essential amino acid synthesis. We predict that pathways for leucine, valine, threonine, isoleucine, phenylalanine, and tryptophan are all complete. As observed in *Escherichia coli*, we predict that the DapC activity in lysine biosynthesis is fulfilled by ArgD (13). The lysine and arginine pathways show evolutionary homology (14), and one *Sulcia* gene cannot be definitely assigned as either *argE* or *dapE*. These two homologous enzymes perform the same chemical reaction (the cleavage of an amide bond) on related molecules, and this one gene may perform both roles in *Sulcia*. In addition, we predict that the *N*-acetyltransferase activity of ArgA is fused to the amino terminus of ArgG, not unlike other previous examples of the mobility of this activity (15).

*Sulcia* seems to be able to generate reducing power in the form of NADH, to aerobically transmit this energy to a *cbb*-3 type cytochrome *c* oxidase-terminated electron transport chain to generate a proton gradient, and to make ATP using this proton motive force. Whereas fragmentary compared with the corresponding gene sets in free-living relatives, *Sulcia* does have a minimal set of genes to transcribe RNA, translate protein, and replicate its genome, although its set of tRNA synthetases is incomplete (SI Table 4). Genes involved in DNA repair are limited to *mutL* and *mutS*. As previously observed for other obligate bacterial symbionts, *Sulcia* possesses surprisingly few transporters, especially considering the diversity of molecules that likely need to cross the cell membrane. Identifiable transporters are limited to a putative multiple antibiotic resistance-related transporter, an ATP-dependent cation transporter, and

EVOLUTION

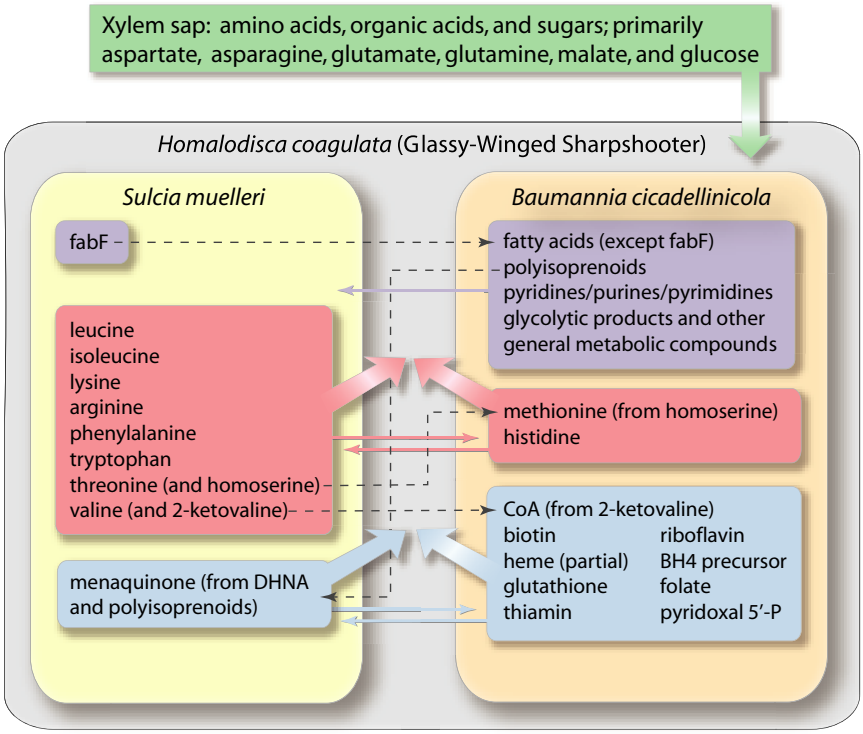


**Fig. 3.** The predicted metabolic capabilities of *Sulcia* and *Baumannia* are complementary. The major components of xylem sap are shown in green at the top of the figure. Compounds produced by the bacterial symbionts that are needed by the host are shown with large colored arrows, and those that are hypothesized to be shared between symbionts are indicated with small colored arrows. Compounds, processes, or genes shaded in red are involved in essential amino acid biosynthesis, those shaded in light blue are involved in vitamin/cofactor biosynthesis, and those shaded in purple are involved in other various metabolic functions. Gray dashed arrows indicate potential individual compounds or genes shared between the two bacterial symbionts.

a heavy metal ion transporter. It is particularly striking that no transporters for amino acids were found. Both the SecY and the twin-arginine protein translocation pathways are present and apparently functional, although in a minimal form compared with most Bacteria (16–18).

As suggested previously based on incomplete genomic sequence for *Sulcia* (9), the metabolic capabilities of *Baumannia* and *Sulcia* are broadly complementary in that *Sulcia* is primarily devoted to amino acid biosynthesis whereas *Baumannia* is primarily devoted to cofactor and vitamin synthesis. Our findings extend these earlier observations by demonstrating that the amino acid pathways found in *Sulcia* are complete (with the exception of one gene in the lysine or arginine pathways; see above), in striking contrast to the situation reported for *Carsonella ruddii*, a symbiont with a tiny genome but with only fragmentary amino acid biosynthetic pathways. In addition, by completing the *Sulcia* genome, we have confirmed that the predicted capabilities of *Baumannia* for vitamin biosynthesis are not present in *Sulcia*. Furthermore, amino acid and vitamin biosynthetic functions are not perfectly partitioned between the two genomes (Fig. 3), and our current data indicate that the complementarity in biosynthetic abilities extends to the exceptions to this general pattern. Thus, *Baumannia* has a complete pathway for histidine synthesis, and is able to make cysteine from homoserine, but is unable to make homoserine. In this regard, *Sulcia* shows remarkable complementarity in that it has no genes for the synthesis of either histidine or cysteine, but is able to make homoserine from aspartate. *Baumannia* has the capability of making CoA from 2-ketovaline, but cannot make 2-ketovaline itself, which can be produced instead by *Sulcia*, in the valine biosynthetic pathway.

The metabolic complementarity may extend to compounds that are not needed by the insect host. For example, *Baumannia* has all of the genes normally seen in a complete fatty acid biosynthetic pathway except the β-ketoacyl-ACP synthase II gene *fabF*, the sole enzymatic gene for fatty acid biosynthesis encoded in the *Sulcia* genome. (Also present is *acpP*, the inactive form of the acyl carrier protein of fatty acid biosynthesis.) *Baumannia* does encode *fabB*, the β-ketoacyl-ACP synthase I gene, the function of which partially overlaps *fabF*. Both *fabF* and *fabB* can perform all of the elongation steps in saturated fatty acid synthesis, but perform distinct reactions in unsaturated fatty acid biosynthesis (19). *Escherichia coli* carrying mutations in *fabF* show deficiencies in the temperature control of fatty acid composition (20), and *Haemophilus influenzae* Rd, which lacks *fabF*, is unable to alter the fatty acid content of its membranes over a wide range of growth temperatures (21). It is unclear how, if at all, *Sulcia* and *Baumannia* coordinate these distinct roles in fatty acid biosynthesis.

Additionally, whereas *Baumannia* is primarily responsible for vitamin and cofactor synthesis, it has no genes for the synthesis of ubiquinone or menaquinone. *Sulcia* has only two genes devoted to vitamin or cofactor synthesis (*menA* and *ubiE*); we hypothesize that both are for the production of menaquinone from polyprenyl diphosphate and 1,4-dihydroxy-2-napthoate (DHNA). *Baumannia* (and possibly the insect host) can make polyprenyl diphosphate, whereas the source of DHNA could be the plant (from the phylloquinone biosynthesis pathway) or the host.

One outstanding issue concerns the source of nitrogen for the entire three-member system (insect-*Baumannia*-*Sulcia*). Three possibilities were previously suggested (9): (*i*) the ammonium present in xylem or generated as waste from the metabolism of the host could be assimilated by *Sulcia*; (*ii*) the nonessential amino acids present in xylem sap could supply the needed nitrogen; or (*iii*) the insect genome could encode enzymes (e.g.,

**Table 1. General genomic properties of representative free-living and symbiotic Gammaproteobacteria and Bacteroidetes**

| | Gammaproteobacteria | | | | Bacteroidetes | | |
|---|---|---|---|---|---|---|---|
| | *Escherichia coli* | *Baumannia cicadellinicola* | *Buchnera* APS | *Carsonella ruddii* | *Bacteroides thetaiotaomicron* | *Porphyromonas gingivalis* | *Sulcia muelleri* |
| Genome size, bp | 4,639,675 | 686,194 | 640,681 | 159,662 | 6,260,361 | 2,343,476 | 245,530 |
| G + C, % | 50.8 | 33.2 | 26.3 | 16.6 | 42.8 | 48.3 | 22.4 |
| No. of genes | 4,418 | 651 | 607 | 213 | 4,864 | 2,015 | 263 |
| Coding density, % | 88.4 | 88.7 | 88.9 | 97.3 | 89.9 | 85.2 | 96.1 |
| Avg CDS length, bp | 954 | 986 | 989 | 826 | 1,174 | 1,014 | 1,006 |

The coding densities include both protein (CDS) and RNA genes. Values were calculated from the following GenBank accession files: *Escherichia coli* K-12 MG1655 (U00096.2), *Baumannia cicadellinicola* (CP000238.1), *Buchnera aphidicola* APS (BA000003.2), *Carsonella ruddii* (AP009180.1), *Bacteroides thetaiotaomicron* (AE015928.1), *Porphyromonas gingivalis* (AE015924.1), and *Sulcia muelleri* (CP000770.2).

glutamine synthase) that allowed incorporation of nitrogen in the form of ammonium. The complete *Sulcia* genome rules out ammonium assimilation from either of the bacterial symbionts and leaves both the nonessential amino acids and ammonium assimilation by the insect as the potential sources of nitrogen for the system.

## Discussion

All of the previously sequenced genomes from insect symbionts have been from members of the Gammaproteobacteria division of Bacteria (9, 22–29). These genomes share many features in common: small genome sizes, low G+C contents, and increased substitution rates compared with their free-living relatives. Our results show parallel evolution of these features in a symbiotic lineage outside the Gammaproteobacteria (Table 1). At ≈245 kb, the *Sulcia* genome is the second-smallest bacterial genome sequenced; correspondingly, its 22.4% G+C content is one of the most biased base compositions among bacterial genomes (9, 22–29). An increased rate of sequence evolution is also evident: phylogenetic analysis shows that *Sulcia* occurs on a long branch, indicating a higher rate of sequence evolution compared with its free-living relatives within the Bacteroidetes (Fig. 1).

Recent work has suggested that there are only 11 replication-related and 6 transcription-related genes universally conserved between genomes of symbionts and free-living Bacteria (30). Extension of these results to include three recent small genomes [*Buchnera aphidicola* Cc (25), *Carsonella ruddii* (28), and *Sulcia*] reduces this number even further (Table 2): only *gidA*, the glucose-inhibited division protein, and *dnaE*, the α-subunit of DNA polymerase III (the subunit that contains the polymerase activity), are universally conserved in replication. In RNA transcription, only the RNA polymerase core enzyme (*rpoA*, the α subunit; *rpoB*, the β subunit; and *rpoC*, the β′ subunit) and *rpoD*, the sigma 70 factor are universally conserved. Thus, the only replicative gene functions that seem to be universal in disparate Bacteria are the polymerization of dNTPs (DnaE) and the ability to turn this polymerization on and off (GidA), although the role of GidA replication control is controversial (31). In transcription, only the core polymerase (RpoA, RpoB, and RpoC) and its strongest binding partner, the sigma 70 factor (RpoD), seem to be universally distributed.

In contrast to all other insect symbiont genomes, *Sulcia* and *Carsonella* have incomplete sets of tRNA synthetases (SI Table 4). However, they both have complete sets of tRNA genes encoding all 20 aa (28 tRNAs in *Carsonella* and 31 in *Sulcia*). This situation is not unprecedented, as the Archaeon *Methanocaldococcus jannaschii* has no identifiable cysteinyl-tRNA synthetase and has alternative biochemical pathways for the synthesis of glutamine- and asparagine-charged tRNAs (32). There are a number of different mechanisms that *Sulcia* and *Carsonella* might use to overcome this deficiency: shuttling tRNAs out to be aminoacylated by *Baumannia* or the host; importing the needed tRNA synthetases into the cell; broadening the specificity of some tRNA synthetases to perform more than one aminoacylation reaction [again, this mechanism is not unprecedented; the prolyl-tRNA synthetase from *Deinococcus radiodurans* can charge both prolyl-tRNA with proline and cysteinyl-tRNA with cysteine (33)]; and using alternative biochemical pathways as in the case of *M. jannaschii* described above. Additionally, some of the hypothetical proteins in the *Sulcia* and *Carsonella* genomes could encode novel tRNA synthetases, as was demonstrated for lysyl-tRNA synthetase in the Archaeon *Methanococcus maripaludis* (32).

The highly reduced genomes of *Sulcia* and *Carsonella* (28) raise a number of interesting questions regarding the minimal gene content required for cellular life (34–36). *Sulcia* seems to be able to make NADH from NAD, and to use this reducing power to generate ATP. It has genes to replicate its genome, transcribe RNA, and translate this mRNA into protein, although it is missing many genes thought to be essential in these processes. Of course, *Sulcia* is not "free-living": it very likely cannot be cultured outside of the host, and it lacks most genes for membrane synthesis and cell division control [and has a strange elongated cell shape, sometimes reaching 80 μm (5)]. Nevertheless, it retains a stable and independently evolving genome that contains some genes for most required cell functions, and it displays corresponding compartmentalization of its cellular components. Perhaps the most interesting question regarding highly reduced symbiont genomes such as that of

**Table 2. Conserved genes for replication and transcription**

| | Replication | | | | | | | | | | | Transcription | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gene | *dnaE* | *gidA* | *dnaN* | *gyrA* | *gyrB* | *dnaB* | *dnaG* | *dnaX* | *polA* | *rnc* | *ssb* | *rpoA* | *rpoB* | *rpoC* | *rpoD* | *greA* | *nusA* | *nusG* |
| PS + B | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| *Buchnera* Cc | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| *Sulcia* | + | + | + | + | + | − | − | − | − | − | − | + | + | + | + | − | − | − |
| *Carsonella* | + | + | − | − | − | + | + | − | − | − | − | + | + | + | + | − | − | − |

The genes conserved among all previously sequenced symbiont genomes and free-living bacteria (30) are shown in the row labeled PS + B (previous symbionts + bacteria). +, homolog is present; −, homolog is absent.

EVOLUTION

*Sulcia* is the extent to which they are achieved through transferring genes to the host genome (coupled with importation of gene products back into the symbiont cell) versus through gene loss and modification of the retained genes to enable life with a very small but sufficient gene set.

## Materials and Methods

**DNA Preparation and Sequencing.** The "yellow" portion of the adult *H. vitripennis* bacteriome was dissected in 100% ethanol from frozen (−80°C) animals caught in June 2001 and June 2004 in a lemon orchard in Riverside, CA. Samples from the same collection were used previously to sequence the *Baumannia* genome (9). The bacteriomes were spun down briefly in a tabletop centrifuge, the ethanol was removed, and DNA was prepared by using Qiagen DNeasy Blood and Tissue Kits. For 454 sequencing, 5 $\mu$g of DNA was prepared as described in ref. 10 using a kit supplied by Roche Applied Science and sequenced on a Roche GS-FLX 100 using 454 technology. For Solexa sequencing, 5 $\mu$g of DNA was prepared in a kit supplied by Solexa/Illumina and sequenced as directed by the manufacturer. All Solexa/Illumina work was done at the Genome Sequencing Center at the Washington University School of Medicine.

**Genome Assembly.** The 454 sequencing was carried out at the Arizona Genomics Institute at the University of Arizona. The run generated 26,711,618 nt of sequence in 118,380 reads with an average read length of 225.6 nt. The Newbler (software version 1.1.01.20, standard running parameters with ace file output selected) (10) assembly of these data resulted in 416 contigs greater than 500 bp in length totaling 984,315 nt with an average G+C content of 31.0%. A summary of these data is presented in SI Fig 4.

Twenty-five contigs representing 245,748 nt of sequence with an average G+C content of 22.5% were clearly separated based on average depth. Twenty-three of these contigs generated high-scoring tblastx alignments to proteins from Bacteroidetes genomes. These 23 contigs could be linked up into a putative circular chromosome using the "_to" and "_from" information appended to the read name in the ace file generated from the Newbler assembly, which were visualized by using the HAWKEYE assembly viewer (37). [The two remaining high-depth contigs had no significant tblastx hits to any protein in the GenBank nonredundant protein database (nr).] These 23 contigs represented 243,948 nt of sequence with an G+C content of 22.1% and had an average depth of coverage of 26.6X.

The Solexa/Illumina Genome Analyzer System generated 14,035,386 reads of 33 nt for 463,167,738 nt of sequence from one partial run (three of eight channels) using the Solexa Analysis Pipeline version 0.2.2.5. After removing any read containing an N (3.4% of the reads), 13,564,883 reads were left for analysis. These reads were mapped onto the *Sulcia* genome using NCBI blastn (with the parameters −G 2 −E 1 −F F −e 1e-8 −W 5 −b 1 −v 1 −a 2). The average coverage of the genome was 132-fold in Solexa reads, although these reads were not distributed evenly with some small regions having no coverage and some regions showing very high coverage.

**Genome Annotation.** The *Sulcia* genome has no detectable GC skew nor a *dnaA* gene, two common ways of positioning the origin of replication. The putative origin of replication was therefore based on a weak transition in oligonucleotide skew using the originx (38) program.

Protein-coding genes were predicted by using the g3-iterated.csh script contained within version 3 of the GLIMMER (39) software package. [The new GLIMMER module that was developed to discriminate between host and symbiont DNA (39) was not needed because of the strong signal from the differences in the depth of coverage in the assembled contigs described above.] These predicted protein-coding genes were annotated by combining results from a NCBI blastp (parameters: −F "m S" −e 1 −b 5 −v 5 −a 2) search against the GenBank nonredundant database (downloaded May 15, 2007), an hmmpfam (HMMER version 2.3.2, default parameters, http://hmmer.janelia.org/) search against the Pfam 21.0 database (40), an hmmpfam search against the TIGRFAM 6.0 database (41), and a blastp search (parameters: −F "m S" −e 1 −b 5 −v 5) against the COG database (42) (downloaded on March 2, 2007).

tRNAs were identified with tRNAscan-SE (43), using the bacterial tRNA model. The 16S and 23S rRNAs were identified by using blastn against the GenBank nonredundant nucleotide database. The 5S rRNA was identified by using the web version of the profile stochastic context-free grammar search on the Rfam 8.0 database (44). The lone tmRNA (also known as 10Sa RNA) was identified by using BRUCE version 1.0 (45).

Repeat sequences were identified by using the web version of Tandem Repeats Finder (http://tandem.bu.edu/trf/trf.html) (46). Of the 126 repeats found using the program with default parameters, only 4 are annotated in the genome. No repeats with <5 copies were considered, and many were rejected because they were completely contained within genes, were very degenerate, or were highly AT-biased.

The preceding results were overlaid and organized by using the Artemis annotation tool (47) for the final annotation of the *Sulcia* genome.

**Metabolic Pathway Construction.** *Sulcia*'s metabolic pathways (Fig. 2) were built by hand, using Ecocyc (48) and Metacyc (49) as guides.

**Phylogenetic Tree Construction.** The tree in Fig. 1 was calculated from a concatenated set of 10 protein sequences (a subset of proteins suggested in ref. 50) that were first aligned using CLUSTALW (51) and then concatenated. All columns with a gap character were removed, leaving 5,520 usable characters. A maximum-likelihood tree was generated with proml from the PHYLIP package (52) using the JTT model of amino acid change. Bootstrap values for 100 replicates were calculated. The proteins used in the analysis were as follows: DNA polymerase III, $\alpha$ subunit; initiation factor IF-2; leucyl-tRNA synthetase; phenylalanyl-tRNA synthetase, $\beta$ subunit; valyl-tRNA synthetase; elongation factor Tu; RNA polymerase, $\beta$ subunit; and ribosomal proteins L2, S5, and S11. The GenBank accession numbers for the genomes are: *Croceibacter atlanticus* HTCC2559, AAMP00000000; *Gramella forsetii* KT0803, CU207366; *Flavobacteria* bacterium BBFL7, AAPD00000000; *Flavobacterium* sp. MED217, AANC00000000; *Cellulophaga* sp. MED134, AAMZ00000000; *Flavobacteriales* bacterium HTCC2170, AAOC00000000; *Robiginitalea biformata* HTCC2501, AAOI00000000; *Flavobacterium johnsoniae* UW101, CP000685; *Flavobacteria* bacterium BAL38, AAXX00000000; *Polaribacter irgensii* 23-P, AAOG00000000; *Bacteroides thetaiotaomicron* VPI-5482, AE015928; *Bacteroides fragilis* NCTC 9343, CR626927; *Bacteroides fragilis* YCH46, AP006841; *Porphyromonas gingivalis* W83, AE015924; *Algoriphagus* sp. PR1, AAXU00000000; *Microscilla marina* ATCC 23134, AAWS00000000; *Cytophaga hutchinsonii* ATCC 33406, CP000383; *Salinibacter ruber* DSM 13855, CP000159.

**COG Analysis.** COG categories were assigned to the *Sulcia* genome as part of the annotation process. The COG assignments for the *C. ruddii* genome were obtained from A. Nakabachi (personal communication), the values for *B. aphidicola* Cc were extracted from GenBank record CP000263, and the COG values for *Baumannia cicadellinicola*, *Escherichia coli* K12, *Bacteroides thetaiotaomicron*, and *Porphyromonas gingivalis* W83 were obtained from *.cog files at the National Center for Biotechnology

Information Bacterial genomes ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/).

We thank Vince Magrini, Matt Hickenbotham, and Elaine Mardis at the Washington University Genome Sequencing Center for the Solexa run; Yeisoo Yu and Rod Wing at the Arizona Genome Institute at the University of Arizona for the 454 sequencing run; Patrick Degnan for making the 454 library; and Becky Nankivell for help in making the figures. This work was funded by National Science Foundation Microbial Genome Sequencing award 0626716 (to N.A.M.). J.P.M. is funded by the University of Arizona's Center for Insect Science through National Institutes of Health Training Grant 1K 12 GM00708.

1. Alonso C, Warnecke F, Amann R, Pernthaler J (2007) *Environ Microbiol* 9:1253–1266.
2. Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) *Nature* 444:1022–1023.
3. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA (2005) *Science* 308:1635–1638.
4. Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent RL, Jr (1998) *J Clin Periodontol* 25:134–144.
5. Moran NA, Tran P, Gerardo NM (2005) *Appl Environ Microbiol* 71:8802–8810.
6. Takiya DM, Tran PL, Dietrich CH, Moran NA (2006) *Mol Ecol* 15:4175–4191.
7. Andersen PC, Brodbeck BV, Mizell RFI (1992) *J Insect Physiol* 38:611–622.
8. Redak RA, Purcell AH, Lopes JR, Blua MJ, Mizell RF, III, Andersen PC (2004) *Annu Rev Entomol* 49:243–270.
9. Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL, Khouri H, Tallon LJ, Zaborsky JM, Dunbar HE, Tran PL, *et al*. (2006) *PLoS Biol* 4:e188.
10. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, *et al*. (2005) *Nature* 437:376–380.
11. Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Folta KM, Soltis DE (2006) *BMC Plant Biol* 6:17.
12. Bentley DR (2006) *Curr Opin Genet Dev* 16:545–552.
13. Ledwidge R, Blanchard JS (1999) *Biochemistry* 38:3019–3024.
14. Velasco AM, Leguina JI, Lazcano A (2002) *J Mol Evol* 55:445–459.
15. Xu Y, Labedan B, Glansdorff N (2007) *Microbiol Mol Biol Rev* 71:36–47.
16. Lee PA, Tullman-Ercek D, Georgiou G (2006) *Annu Rev Microbiol* 60:373–395.
17. Osborne AR, Rapoport TA (2007) *Cell* 129:97–110.
18. Wickner W, Schekman R (2005) *Science* 310:1452–1456.
19. Campbell JW, Cronan JE, Jr (2001) *Annu Rev Microbiol* 55:305–332.
20. Garwin JL, Klages AL, Cronan JE, Jr (1980) *J Biol Chem* 255:3263–3265.
21. Wang H, Cronan JE (2003) *J Bacteriol* 185:4930–4937.
22. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H (2000) *Nature* 407:81–86.
23. Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ, Sandstrom JP, Moran NA, Andersson SG (2002) *Science* 296:2376–2379.
24. van Ham RC, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, Fernandez JM, Jimenez L, Postigo M, Silva FJ, *et al*. (2003) *Proc Natl Acad Sci USA* 100:581–586.
25. Perez-Brocal V, Gil R, Ramos S, Lamelas A, Postigo M, Michelena JM, Silva FJ, Moya A, Latorre A (2006) *Science* 314:312–313.
26. Gil R, Silva FJ, Zientz E, Delmotte F, Gonzalez-Candelas F, Latorre A, Rausell C, Kamerbeek J, Gadau J, Holldobler B, *et al*. (2003) *Proc Natl Acad Sci USA* 100:9388–9393.
27. Degnan PH, Lazarus AB, Wernegreen JJ (2005) *Genome Res* 15:1023–1033.
28. Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M (2006) *Science* 314:267.
29. Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M, Aksoy S (2002) *Nat Genet* 32:402–407.
30. Klasson L, Andersson SG (2004) *Trends Microbiol* 12:37–43.
31. Kinscherf TG, Willis DK (2002) *J Bacteriol* 184:2281–2286.
32. Tumbula D, Vothknecht UC, Kim HS, Ibba M, Min B, Li T, Pelaschier J, Stathopoulos C, Becker H, Soll D (1999) *Genetics* 152:1269–1276.
33. Zhang CM, Hou YM (2004) *RNA Biol* 1:35–41.
34. Koonin EV (2003) *Nat Rev Microbiol* 1:127–136.
35. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison CA, III, Smith HO, Venter JC (2006) *Proc Natl Acad Sci USA* 103:425–430.
36. Gil R, Silva FJ, Pereto J, Moya A (2004) *Microbiol Mol Biol Rev* 68:518–537.
37. Schatz MC, Phillippy AM, Shneiderman B, Salzberg SL (2007) *Genome Biol* 8:R34.
38. Worning P, Jensen LJ, Hallin PF, Staerfeldt HH, Ussery DW (2006) *Environ Microbiol* 8:353–361.
39. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) *Bioinformatics* 23:673–679.
40. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, *et al*. (2004) *Nucleic Acids Res* 32:D138–141.
41. Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC, Richter AR, White O (2007) *Nucleic Acids Res* 35:D260–264.
42. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, *et al*. (2003) *BMC Bioinformatics* 4:41.
43. Lowe TM, Eddy SR (1997) *Nucleic Acids Res* 25:955–964.
44. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A (2005) *Nucleic Acids Res* 33:D121–124.
45. Laslett D, Canback B, Andersson S (2002) *Nucleic Acids Res* 30:3449–3453.
46. Benson G (1999) *Nucleic Acids Res* 27:573–580.
47. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B (2000) *Bioinformatics* 16:944–945.
48. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD (2005) *Nucleic Acids Res* 33:D334–337.
49. Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, Kaipa P, Krummenacker M, Paley S, Pick J, Rhee SY, *et al*. (2006) *Nucleic Acids Res* 34:D511–516.
50. Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ (2001) *Nat Genet* 28:281–285.
51. Thompson JD, Higgins DG, Gibson TJ (1994) *Nucleic Acids Res* 22:4673–4680.
52. Felsenstein J (1989) *Cladistics* 5:164–166.