Genomic Features of a Bumble Bee Symbiont Reflect Its Host Environment

Vincent G. Martinson, a,b* Tanja Magoc, d Hauke Koch, b,c Steven L. Salzberg, d Nancy A. Moran b,c

Center for Insect Science, University of Arizona, Tucson, Arizona, USA; Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut, USA; Department of Integrative Biology, University of Texas at Austin, Austin, Texas, USA; Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA

Here, we report the genome of one gammaproteobacterial member of the gut microbiota, for which we propose the name “Candidatus Schmidhempelia bombi,” that was inadvertently sequenced alongside the genome of its host, the bumble bee, Bombus impatiens. This symbiont is a member of the recently described bacterial order Orbales, which has been collected from the guts of diverse insect species; however, “Ca. Schmidhempelia” has been identified exclusively with bumble bees. Metabolic reconstruction reveals that “Ca. Schmidhempelia” lacks many genes for a functioning NADH dehydrogenase I, all genes for the high-oxygen cytochrome o, and most genes in the tricarboxylic acid (TCA) cycle. “Ca. Schmidhempelia” has retained NADH dehydrogenase II, the low-oxygen specific cytochrome bd, anaerobic nitrate respiration, mixed-acid fermentation pathways, and citrate fermentation, which may be important for survival in low-oxygen or anaerobic environments found in the bee hindgut. Additionally, a type 6 secretion system, a Flp pilus, and many antibiotic/multidrug transporters suggest complex interactions with its host and other gut commensals or pathogens. This genome has signatures of reduction (2.0 megabase pairs) and rearrangement, as previously observed for genomes of host-associated bacteria. A survey of wild and laboratory B. impatiens revealed that “Ca. Schmidhempelia” is present in 90% of individuals and, therefore, may provide benefits to its host.

AUTOTHOUSONOUS GUT MICROORGANISMS GREATLY INFLUENCE ANIMAL HEALTH BY PROVIDING A RANGE OF NUTRITIONAL, DEVELOPMENTAL, AND PROTECTIVE BENEFITS (E.G., ENERGY, VITAMINS, IMMUNE PRIMING, DETOXIFICATION, AND PATHOGEN EXCLUSION) TO THEIR HOSTS (1–3). HIGHLY CONSISTENT GUT-ASSOCIATED MICROBES ARE COMMON AMONG EU SOCIAL INSECT SPECIES (4–8). THESE MICROBES CAN BE ACTIVELY PASSED BETWEEN GENERATIONS THROUGH TROPHALLAXIS (MOUTH-TO-MOUTH OR ANUS-TO-MOUTH FOOD SHARING) OR PASSIVELY TRANSMITTED VIA A FECAL-ORAL ROUTE DUE TO COMMUNAL LIVING (9, 10). BENEFICIAL GUT BACTERIA ARE OFTEN SELECTIVELY TRANSMITTED BETWEEN GENERATIONS AND FORM WELL-ESTABLISHED INTERACTIONS WITH THEIR HOSTS (11); HOWEVER, MANY OF THE EU SOCIAL-INSECT-ASSOCIATED MICROBES HAVE UNKNOWN RELATIONSHIPS WITH THEIR HOSTS. GENOME SEQUENCING CAN PROVIDE INSIGHT INTO THE METABOLIC CAPABILITIES AND BIOLOGICAL SIGNIFICANCE OF THESE HOST-ASSOCIATED MICROBES.

GUTS OF BUMBLE BEE (Bombus sp.) ARE COMMONLY INHABITED BY TWO BACTERIAL SPECIES THAT ARE CLOSELY RELATED TO THE HONEY BEE (Apis mellifera)-ASSOCIATED LINEAGES OF Gilliamella apicola (previously called the “Gamma-1” phylotype, Gammaproteobacteria) and Snodgrassella alvi (previously called the “Beta” phylotype, Betaproteobacteria) (6, 12–14). METAGENOMIC SEQUENCING OF THE A. mellifera GUT BACTERIA REVEALED THAT Gilliamella and Snodgrassella contain genes that may contribute to pollen digestion, detoxification of mannone, and defense against pathogens (15). THE CLOSELY RELATED GUT MICROBIOTA OF Bombus terrestris WAS EXPERIMENTALLY DETERMINED TO PROVIDE AN EXTENDED-IMMUNE PHENOTYPE AGAINST THE TRYPANOSOMATID GUT PARASITE Crithidia bombi, YET THE MECHANISM OF THIS DEFENSE WAS NOT IDENTIFIED (16).

The Bombus impatiens genome sequencing project recovered the genome sequence of a gammaproteobacterium related to the Gilliamella apicola clade. This genome sequence provides insights into the phylogenetic relationships and lifestyle of Gilliamella-related bumble bee symbionts, as well as clues about the role of this gammaproteobacterium in B. impatiens biology. Here, we describe the first genome sequence from the gammaproteobacterial order Orbales and compare it to the metagenome recovered from the A. mellifera microbiota to identify basic metabolic and ecological attributes and potential effects that this symbiont may have on its host.

MATERIALS AND METHODS

Source DNA, sequencing, identification and assembly of bacterial reads. A single adult male Bombus impatiens approximately 24 h posteclosion was collected from a colony purchased from Biobest Biological Systems (Leamington, Ontario, Canada). DNA was extracted from the entire specimen using a standard phenol-chloroform preparation.

Three pair-end libraries were constructed with fragment sizes of 400, 4,000, and 8,000 bp. DNA was sequenced using Illumina GAIIx sequencing technology to generate eight lanes of 125-bp raw sequences. After error correction, the average read length was 105 bp, and the total number of reads used in assembled contigs was 150,442,748. Based on an estimated genome size of 250 megabases (Mb), this yields approximately 65-fold coverage of the B. impatiens genome. The reads were assembled using the CABOG assembler (17) modified to handle short Illumina reads. The draft assembly contained 69,944 contigs (including symbiont contigs) with a length greater than 100 bp of which 6,658 contigs were longer than 10,000 bp. In addition, the assembly contained 1,086,650 contigs longer than 10,000 bp.

Received 29 January 2014 Accepted 8 April 2014
Published ahead of print 18 April 2014

Copyright © 2014, American Society for Microbiology. All Rights Reserved.

doi:10.1128/AEM.00322-14

* Present address: Vincent G. Martinson, Department of Biology, University of Rochester, Rochester, New York, USA.
* Present address: Vincent G. Martinson, Department of Biology, University of Rochester, Rochester, New York, USA.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/AEM.00322-14. Journals.ASM.org on July 23, 2015 by Univ of Texas Austin

July 2014 Volume 80 Number 13 Applied and Environmental Microbiology p. 3793–3803 aem.asm.org 3793
degenerate contigs, primarily small repetitive sequences or contigs with very low quality. Another 60,355,858 reads remained as unassembled singletons.

To scan the 69,944 contigs for sequences of possible bacterial symbionts, we initially aligned them to the genomes of four strains of Wolbachia, which we considered the first candidates because Wolbachia bacteria are the most ubiquitous endosymbionts of invertebrates. We aligned all assembled contigs against genomes of each of these four strains using the promer program from the MUMmer package (18), which translates both the reference and the query sequences to amino acids in all six reading frames. This allowed a more sensitive alignment than a DNA-based approach. We identified multiple contigs that had strong homology to at least one of the Wolbachia species, indicating that bacterial sequences might be present in the whole-genome data. We extracted these contigs and used the BLASTX program to map them against the complete nonredundant protein database at NCBI in order to find bacteria that were closer to the symbiont in B. impatiens. The best hits from this mapping were to three Gammaproteobacteria species: Photobacterium profundum, Versicibacter eneca, and Proteus mirabilis, indicating that the bacterial sequences were from Entobacteriales (Gammaproteobacteria) and not Wolbachia (Alphaproteobacteria).

We then aligned all assembled contigs and all singleton reads to the complete genomes of each of these three bacteria. We ran DNA alignments and translated protein alignments using the nucl and promer programs from the MUMmer package and mapped all contigs against both the DNA and the protein sequences of the three bacterial genomes. The more sensitive protein-based alignments were compared to the annotated coordinates of the proteins in each of the bacterial genomes. Each contig that contained at least one complete protein was considered possibly bacterial. In addition, contigs not longer than 500 bp that contained at least a partial match at greater than 60% identity to any bacterial protein were also considered bacterial. This analysis identified 367 regular contigs, 1,129 degenerate contigs, and 255,589 singleton reads as possibly bacterial. In addition, contigs not longer than 500 bp that contained at least one complete protein was considered possibly bacterial. In addition, contigs not longer than 500 bp that contained at least a partial match at greater than 60% identity to any bacterial protein were also considered bacterial. This analysis identified 367 regular contigs, 1,129 degenerate contigs, and 255,589 singleton reads as possibly bacterial. In addition, contigs not longer than 500 bp that contained at least one complete protein was considered possibly bacterial. In addition, contigs not longer than 500 bp that contained at least a partial match at greater than 60% identity to any bacterial protein were also considered bacterial. This analysis identified 367 regular contigs, 1,129 degenerate contigs, and 255,589 singleton reads as possibly bacterial.

We used the CABOG assembler’s raw output files to locate all reads used to build these bacterial contigs and extracted these reads from the original sequence files with their paired-end mates. This resulted in 615,185 mate pairs from the 4-kb insert size library, 8,164 mate pairs from the 9-kb insert size library, and 121,568 unpaired reads. These reads were assembled de novo with the CABOG assembler. The final bacterial assembly contained 1,998,543 bp in 79 contigs. The longest contig contained 110,984 bp, and the assembly had an N50 size of 39,885 bp. The 79 contigs were combined into 33 scaffolds spanning 2,004,741 bp, with a scaffold N50 size of 98,624 bp and a maximum scaffold of 204,248 bp. The approximate average coverage of the genome is 37-fold.

Annotation of the bacterial genes. Gene annotation was completed in the automated Integrated Microbial Genomes Expert Review (IMG/ER; Joint Genome Institute) pipeline (19). Protein-coding sequences and RNA-coding genes were predicted within its framework using Prodigal and tRNAscan-SE (23). Functional annotations were assigned to genes based on protein domain characterization according to clusters of orthologous groups of proteins (COG) terms, Pfam, TIGRfam, InterPro domains, Gene Ontology (GO) terms, and KEGG Orthology (KO) terms with metabolic pathway maps. Additional manual assessment with KEGG (20), EcoCyc (21), and the MetaCyc Pathological program (22) was performed to check pathway completeness. Genes identified as missing from main pathways (e.g., tricarboxylic acid [TCA] cycle or NADH dehydrogenase I) were manually investigated using BLASTP searches against the B. impatiens symbiont genome with corresponding genes from Escherichia coli. A metabolic map was manually created for the B. impatiens gammaproteobacte rium (BiG). The annotated contigs for this genome are available at the IMG/ER website (https://img.jgi.doe.gov/cgi -bin/er/main.cgi) (proposed name, “Candidatus Schmidhempelia bombi Bimp”).

Core gene phylogeny and metagenome gene phylogenies. A set of 89 single-copy orthologous (SICO) genes was selected from an original set of 203 consistently present, single copy, non-horizontally transferred core genes (23); this set was used to reconstruct the phylogenetic placement of BiG. SICO genes were selected from 28 gammaproteobacterial genomes using SICO gene lists in MaGE (24) or using a cutoff of a bit-score ratio of >0.30 in a BLASTP search with the 203 SICO genes from E. coli (23). Inferred protein sequences of the 89 genes were individually aligned in MUSCLE (25) and concatenated together. The Gblocks (26) program was used to remove poorly aligned regions, and the resulting alignment consisted of 27,452 amino acid sites. Maximum-likelihood reconstruction was performed with RAxML (27) on 100 bootstrap replicates using the PROTGAMMA algorithm and the WAG substitution matrix, which were selected with ProtTest version 2.4 (28). Individual SICO gene trees were built with the same methods as the multiprotein data set and subsequently sorted with PhyloSort (29).

Gene content of the BiG genome was compared to the gene content of the metagenome of the A. mellifera gut microbiota, which was determined in a previous study (15). The comparison was made to the taxonomic bins from that study including the all-bacteria bin and the gammaproteobacteria (Gamma) bin using the COG (clusters of orthologous genes) annotations from IMG/ER (Joint Genome Institute). Of the 195 SICO genes present in BiG, a set of 193 was identified as being shared between BiG and the Gamma bin using BLASTP to identify pairwise protein sequence identities (SICOs identified with a bit-score ratio of >0.30 to the set found in Letar et al. [21]). The subset of 89 SICO genes used to construct the multiprotein phylogeny was amended with corresponding SICO genes from the Gamma bin, and individual gene trees were constructed using previously described methods. The phylogenetic relationships between the Gamma bin sequences, which contained genes from both G. apicola and Frischella perrara, and the BiG were collected for each tree.

16S rRNA gene phylogeny. The 16S rRNA gene sequence was used to reconstruct and refine the phylogenetic relationships of the BiG among a representative set of G. apicola, F. perrara (collected from the genera Apis and Bombus), and closely related sequences from the NCBI database. One of the four 16S rRNA gene copies in the BiG genome was selected and aligned with the 16S rRNA data set using Infernal (30). A maximum-likelihood phylogeny was constructed with RAxML using the GTRCAT parameter and 100 bootstrap replicates (27).

Putative HGT. The IMG annotation pipeline identified putative horizontally transferred genes. Further analysis of potential horizontal gene transfer (HGT) was assayed with a phylogenetic pipeline modified from Moustafa and Bhattacharya (29). Briefly, the pipeline identifies closely related genes in the NCBI database, aligns the amino acid sequences, and constructs phylogenies for each. PhyloSort (29) was used to find trees which indicated horizontal transfer from Firmicutes to BiG. The vast majority of genes from the BiG genome have best hits to Gammaproteobacteria. The potential HGT genes are the outstanding examples having best hits to and clustering in a phylogeny with Firmicutes. These potential HGT genes come from many different contigs, most of which are fairly long and are otherwise composed of genes with top hits to Gammaproteobacteria. Because the phylogenetically near neighbors to BiG have not been densely sampled, these computational findings may reflect sampling bias rather than true horizontal gene transfer, and further work will be needed to validate these findings.

PCR screen for BiG among B. impatiens individuals. Specific primers BombusG2F (5’-CCTGCTGTCGCTGGAGTATTGT-3’) and BombusG2r1 (5’-AGTTGCCTGCTCACTGCTG-3’) were used to search for BiG within B. impatiens individuals or gut organs from wild and laboratory individuals. Cycling was performed with an annealing temperature of 54°C for 35 cycles and a 1-min extension.
Nucleotide sequence accession number. The bacterial symbiont genome assembly was deposited in the GenBank database under accession number AWGA00000000.

RESULTS AND DISCUSSION

Retrieval of a nearly complete genome of a *B. impatiens* symbiont. Sequencing of the *B. impatiens* genome resulted in the by-product sequencing of a gammaproteobacterial genome corresponding to an organism present in the bee from which DNA was extracted. From the ~250 Mb of assembled sequence from the project, 79 contigs representing 2 Mb were assigned to the bacterial genome (see Materials and Methods). Contigs that represent the gammaproteobacterial genome harbor at least one open reading frame (ORF) per contig and have an \( N_{\text{seq}} \) length of 39.9 kb. To determine the completeness of the genome and the number of distinct bacterial genomes present, coverage of a presellected set of 203 single-copy, near-universal bacterial genes was assessed using BLASTP. We identified 195 of these 203 genes (96%), each with exactly one copy indicating the complete or nearly complete coverage of a single bacterial genome. Calculation of \( G + C \) content of the contigs produced a unimodal distribution with a mean of 36.6%, and the average depth of coverage (37-fold) was consistent across the contigs, providing further evidence for the retrieval of a single bacterial genome. Here, we refer to this organism as the *B. impatiens* gammaproteobacterium, or BiG.

The BiG genome is at least 1.99 Mb in size (Table 1). A total of 30 tRNA genes and 23 tRNA synthetase genes were identified, corresponding to all 20 amino acids. Altogether, 1,694 protein-coding genes were identified from the assembly, with 14% (236) of them having unknown functions. Roughly 72% had functions predicted as clusters of orthologous groups (COGs) (31). The largest COG categories represented were translation (9.7%), general function only (9.3%), amino acid transport and metabolism (8.6%), cell wall biogenesis (7.6%), replication (6.8%), coenzyme transport and metabolism (5.8%), and carbohydrate transport and metabolism (5.3%).

The small size and low \( G + C \) content resemble genomes of previously sequenced host-dependent commensals and pathogens (e.g., *Candidatus Hamiltonella defensa* [32] or *Histophilus somnius* [33]). Further, gut bacteria with strict host associations (e.g., *Helicobacter sp.*, *Lactobacillus reuteri*, and *Pasterurelalles* species) often have small genomes (34, 35), suggesting that BiG may have a restricted host distribution. Overall, the genome lacks large regions of chromosomal synteny with other genomes, even for regions conserved among many species of *Enterobacteriales* and *Pasteurelalles* (see Fig. S1 and S2 in the supplemental material for examples). However, contigs from the *A. mellifera* Gamma bin harbor regions with strikingly similar gene orders, even with interrupted operons for TCA cycle enzymes (\( \text{suca} \text{ABC}D \); only \( \text{sucD} \) are retained) and the NADH dehydrogenase I complex (\( \text{nuoA} \text{nuoN} \); only \( \text{nuoL} \text{nuoN} \) are retained) (see Fig. S1 and S2). The conserved synteny between BiG and sequences from the gut microbiota of *A. mellifera* confirms the presence of similar bacteria in both bee species and substantiates the robust assembly of the BiG contigs.

The majority of BiG protein-coding genes (98%, 1,659/1,694) were shared with the Gamma bin of the *A. mellifera* gut microbiota metagenome. Notably, genes for the four subunits of the respiratory nitrate reductase A are present in BiG but absent from the Gamma bin (see Table S1 in the supplemental material).

Consistent association of BiG with laboratory-raised and wild *Bombus impatiens* individuals. The BiG genome sequence was retrieved from a single male individual from a laboratory-raised *B. impatiens* colony. To understand the consistency of the association between BiG and *B. impatiens*, we undertook a PCR survey of workers, males, and queens from natural and laboratory environments. Specific primers that differentiated between BiG and the closely related *Gilliamella* revealed that BiG is nearly ubiquitous (90%, 18/20) among laboratory adults and larvae and wild adults (see Fig. S3 in the supplemental material). The bacterium was detected in all gut organs (crop, midgut, ileum, and rectum), as well as from the leg DNA extraction. The leg sample may have been surface contaminated with feces because the source colonies lacked locations to remove excrement, possibly increasing fecal contamination within the hive relative to normal conditions.

The presence of this bacterium in both wild and laboratory-raised *B. impatiens* bees implies that BiG is passed from one colony to the next, as documented for members of the closely related genus *Gilliamella* in honey bees and bumble bees (14, 16). Queens were shown to harbor this bacterium, thus affirming a transmission link between the annual disintegration of *Bombus* colonies in the fall and the founding of new colonies in the spring. Overall, BiG is a common associate of *B. impatiens* and potentially many other bumble bee species.

BiG is a relative of *Gilliamella apicola* in the newly described order *Orbales*. Our 16S rRNA gene tree placed BiG as a member of *Orbales*, a bacterial order previously recovered from numerous honey bee and bumble bee species (13). BiG clusters among sequences collected from native and commercially reared bumble bee species from around the world (Fig. 1b; see also Fig. S4 in the supplemental material). A survey of the bacterial associates of bumble bees (36) shows that BiG clusters within a separate clade from the genus *Gilliamella* (found in honey bees and bumble bees [13]), *Frischella perrara* (Gamma-2 of honey bees [37]), and other *Orbales* species, with strong bootstrap support (95%) (Fig. 1b; see also Fig. S4). The BiG sequences were identified in geographically and phylogenetically diverse bumble bees, but their 16S rRNA sequences are very similar (>98% identical). Therefore, BiG is a member of a geographically widespread bacterial clade that is strictly associated with bumble bees, based on surveys conducted to date.

Our concatenated, multiprotein phylogeny retrieved high bootstrap support for previously established evolutionary relationships between the bacterial orders of *Gammaproteobacteria*

### TABLE 1 General features of the BiG genome

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value for the parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome length (bp)</td>
<td>&gt;1,999,325</td>
</tr>
<tr>
<td>Extrachromosomal elements</td>
<td>Presence unknown</td>
</tr>
<tr>
<td>GC content (%)</td>
<td>36.6</td>
</tr>
<tr>
<td>Total no. of predicted CDSs</td>
<td>1,770</td>
</tr>
<tr>
<td>Coding density (%)</td>
<td>82</td>
</tr>
<tr>
<td>Average CDS length (bp)</td>
<td>954</td>
</tr>
<tr>
<td>No. of tRNA operons</td>
<td>55 tRNA</td>
</tr>
<tr>
<td></td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>16S rRNA</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>23S rRNA</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td>No. of tRNAs</td>
<td>50</td>
</tr>
</tbody>
</table>

* a CDS, coding sequence.
FIG 1  (a) Phylogenetic placement of BiG as a singleton clade among five orders of Gammaproteobacteria. Maximum-likelihood reconstruction inferred from 89 concatenated SiCO genes (27,452 aligned amino acid sites).  (b) Location of BiG among members of the insect gut-associated Orbulales. The tree is based on maximum likelihood with the 16S rRNA gene. (c) Proportions of the 89 individual SiCO gene trees that returned each phylogenetic pattern with their average (Avg) bootstrap support at the indicated node and average gene length. (d) Proportions of individual SiCO gene trees that united the BiG and the Apis mellifera Gamma bin (Gbin) gene copies as a monophyletic clade and average bootstrap support. Asterisks represent bootstrap support values of 100, and values below 50 are not shown. Str, strain; SS, secondary symbiont.

included in the analysis (Enterobacteriales, Pasteurellales, Vibrionales, Alteromonadales, and Pseudomonadales) (23, 38, 39) (Fig. 1a). In this tree, BiG is sister to Enterobacteriales, which supports the previous placement of Orbulales (Fig. 1a)(13). Trees created with individual SICO genes varied in their support, mainly due to their differing sequence complexities (46 to 1,407 amino acids in length) (Fig. 1c). The majority (61/89) of individual SICO genes resulted in a tree topology uniting BiG, Enterobacteriales, and Pasteurellales. Nearly three-fourths of those trees (44/61) placed BiG sister to, or within, the Enterobacteriales, with an average bootstrap support of >80% (Fig. 1c).

To identify the relationship between BiG and the sequences from the Gamma bin of the A. mellifera metagenome, which included sequences from the related genera Gilliamella and Frischella, the corresponding Gamma bin SICO sequences were analyzed with each of the 89 SICO genes. Many strains of Gilliamella and Frischella are present within the metagenomic data set for the A. mellifera microbiota, and the number of copies corresponding to the 89 BiG SICOs varies among the loci. The majority (92%, 81/89) of SICOs retrieves the Gamma bin sequences and the BiG as a singleton clade (Fig. 1d). The BiG gene copy usually branches basal to the Gamma bin sequences (69%, 61/81), with a high average bootstrap support (90%) for this placement (Fig. 1d), which adds further support that BiG represents a new genus of Orbulales and not a member of Gilliamella or Frischella. This data set of protein-coding genes supports the sister relationship between the Orbulales (i.e., BiG and the Gamma bin clade) and Enterobacteriales, with high support (71% average bootstrap value; 36/57 trees) for this pattern, rather than Pasteurellales (50% average bootstrap value; 21/57 trees).
heme and siroheme, and ubiquinone assembly (Fig. 3). Biosynthesis, lipopolysaccharide production, peptidoglycan fabrication, the full pentose phosphate pathway, nucleotide metabolism of genes in predicted metabolic pathways for glycolysis, gluconeogenesis, and cytochrome bd (cydAB), which have optimal conditions at low oxygen levels and have been shown to be critical for colonization of the hypoxic mouse gut by Escherichia coli (40, 41). In E. coli this cytochrome has its maximal level of expression and optimal conditions at low oxygen concentrations; additionally cytochrome bd reinforces the hypoxic environment by scavenging the remaining oxygen.

Additionally, BiG encodes genes for anaerobic respiration. Together the periplasmic respiratory nitrate reductase A (narGHI) and the nitrate-induced formate dehydrogenase N (fhdGHI) can produce a respiratory chain, resulting in a proton motive force and cytoplasmic nitrite (42, 43). The NADH-dependent nitrite reductase (nirBD) detoxifies resulting nitrite to ammonia (44). Notably, the narGHI genes are not present in the A. mellifera gut microbiota metagenome Gamma bin, which further suggests that this Gilliamella-like bacterium is a divergent lineage among this group of common bee associates (see Table S1 in the supplemental material). However, the absence of the narGHI genes in the A. mellifera metagenome could reflect incomplete sampling of the gammaproteobacterial genomes.

BiG has retained genes for citrate fermentation and for several branches of mixed acid fermentation. A Na/citrate symporter is adjacent to the complete citrate fermentation gene locus (citABCDEFGX), encoding capabilities for transporting citrate into the cell, sensing citrate in the environment, and degrading citate to oxaloacetate and acetate via citrate lyase. Mixed-acid fermentation genes are present for lactate fermentation (ldha), pyruvate cleavage to formate and acetyl-CoA (pflB), ATP generation via acetate formation (pta and ackA), and NAD+ generation through ethanol production (adhE and adhP). Fermentation results in the excreted metabolites of ethanol, formate, and the short-chain fatty acids acetate and lactate. Short-chain fatty acids are the bulk of carbon and energy sources of ruminant animals (45). This raises the possibility that the bacterium is providing its host a nutritional benefit through biosynthesis of needed compounds (46). However, biosynthetic contributions (e.g., amino acids or vitamins) may not be very significant because the bee diet is composed of both easily accessible mono- and disaccharides and protein-rich pollen (47).

A primary constituent of the bee diet is pollen, which presents plant macromolecules that are difficult to degrade and that form barriers surrounding the pollen germ, the primary source of protein for bees. Some strains of G. apicola from A. mellifera encode pectate lyase and are able to degrade pectin, a function that may aid their host in nutrient acquisition by releasing pollen contents (15). This capability is not encoded within the BiG genome; however, BiG has transporters (bgIF and beta-glucosidases (bgLA, glucohydrolase family 4) (48) that may confer the ability to import and metabolize some of the products of cellulase activity (e.g., cellulbiose and cellotriose) potentially produced by other bacteria present in the gut or present in nectar or pollen. Metabolic scavenging of these compounds could provide the majority of energy for this organism since glucose and fructose (the major sugars in honey) are absorbed rapidly in the midgut and are not abundant in the hindgut (49, 50).

The pattern of missing genes implies that BiG may inhabit a low-oxygen niche within the bumble bee gut, which is consistent with the cultivation conditions for related members of Orbales, both microaerophilic (Gilliamella apicola) and anaerobic (Frisch-
ella perrara) (13, 37). Recent *in situ* analysis of the *A. mellifera* microbiota showed that the majority of the bacteria (including *G. apicola*) reside within the hindgut (14). Oxygen levels could govern colonization of the gut organs (i.e., anaerobic hindgut and aerobic foregut), restricting members of the *Orbales* to low-oxygen or anaerobic environments. Fluorescent *in situ* microscopy of the *Bombus* microbiota and microsensor surveys (i.e., O2) of the *Apis* and *Bombus* gut are needed to test this hypothesis and determine the breadth of this pattern in corbiculate bees (51).

Candidate horizontal gene transfers from *Firmicutes* to *BiG*. A total of 54 genes were identified by IMG as putatively horizontally transferred from *Firmicutes* to *BiG*. Of these genes, PhyloSort supported horizontal transfer from *Firmicutes* for 39 genes (Table 2). Because *BiG* is a novel genus, phylogenetic placement of its genes may not be as reliable as with a more thoroughly sequenced clade, and these 39 genes should be considered candidates worthy of further scrutiny rather than confirmed horizontal gene transfers. Closer analysis of potentially transferred genes identified several sugar uptake and degradation genes, including the previously mentioned *bglA* (Table 2). These genes may enable *BiG* to utilize the numerous sugars found in nectar that cannot be metabolized by *B. impatiens* or that are abundant in the gut. An intact operon for the uptake of mannose (phosphotransferase system [PTS]) may have been transferred from a species related to *Bacillus* (Table 2). The mannose
PTS has been shown to have an extensive history of horizontal transfer in bacteria and is mostly found in bacteria associated with animal guts (52). Mannose is toxic to honey bees and bumble bees; therefore, microbial assimilation of this sugar could protect the host from small amounts of mannose, which is often present in nectar (53, 54). Alternatively, these transport systems can often act on a broad range of substrates (52). Evidence that mannose PTSs may be linked to the bee gut environment also comes from the finding that they are overrepresented in the A. mellifera gut metagenome relative to other gut metagenomes (15); however, this overrepresentation may merely reflect the taxonomic composition of the bacterial community.

**Potential interactions with the host and other gut microorganisms.** Several putative host interaction factors were present in the BiG genome, including Sell repeat proteins, bacterial Ig-like domains, and bacterial α_{5}-macroglobulins; these could be critical for recognition of this bacterial strain by the host epithelium (55). A full Fli plus gene set was present; this apparatus is known to be critical for adhesion and biofilm formation (56). Strains of Gilliamella associated with the honey bee have been shown to form thick biofilms within the ileum (14), and genes involved in biofilm formation/adhesion were overrepresented in the A. mellifera Gamma bin (15). More generally, adhesion to the gut wall may play a critical role for insect gut associates because, unlike mammals, insects do not secrete a mucus layer that facilitates microbial residence (57).

Similar to the A. mellifera Gamma bin of the metagenome, BiG has a marked abundance of antibiotic/multidrug resistance transporters, including several ABC, drug metabolic transporter (DMT), multidrug/oligosaccharidyl-lipid/poly-saccharide (MOP), membrane fusion protein (MFP), Eam/Emr, and arabinose efflux pumps. As mentioned in Engel et al. (15), the bee gut is exposed to plant defense compounds during pollen/nectar foraging and ingestion. These compounds could be a selective force for mechanisms enabling efficient elimination of these toxins in bee gut bacteria.

Complex interactions with the host epithelium are additionally supported by the presence of multiple secretion systems (types I, V, and VI) (Fig. 3). The recently described type VI secretion system (T6SS) evolved from viral tail fibers into a syringe-like effector delivery mechanism that is present in many bacteria and is critical for discriminating and attaching to target cells (56). The T6SS, is nearly complete, as indicated by the unimodal G + C content, the presence of a complete tRNA synthetase complement, exactly one copy of 96% of a defined set of single-copy genes, and a consistent coverage of 37-fold across all contigs. Previous genome sequencing projects have been shown to produce data sets corresponding to symbiotic microorganisms (71); however, most of these symbionts do not assemble as well as the BiG genome retrieved in our study (79 contigs), indicating the presence of a clonal or near-clonal bacterial population in the source DNA sample. The loss of central metabolism components (i.e., TCA cycle) and the lack of synteny with other bacterial genomes suggest that this genome and genomes of closely related organisms (Orbales) underwent rearrangement, reduction, and specialization to the host environment similar to the processes observed in other symbiotic genomes (72, 73).

**Conclusions.** As a member of the Orbales, the BiG genome represents a recently described order of Gammaproteobacteria that is found in honey bees and bumble bees and that has been repeatedly collected from other insects (6, 12, 13, 36, 70). Our analyses show that the BiG genome, sequenced concurrently with the genome of B. impatiens, is nearly complete, as indicated by the unimodal G + C content, the presence of a complete tRNA synthetase complement, exactly one copy of 96% of a defined set of single-copy genes, and a consistent coverage of 37-fold across all contigs. Previous genome sequencing projects have been shown to produce data sets corresponding to symbiotic microorganisms (71); however, most of these symbionts do not assemble as well as the BiG genome retrieved in our study (79 contigs), indicating the presence of a clonal or near-clonal bacterial population in the source DNA sample. The loss of central metabolism components (i.e., TCA cycle) and the lack of synteny with other bacterial genomes suggest that this genome and genomes of closely related organisms (Orbales) underwent rearrangement, reduction, and specialization to the host environment similar to the processes observed in other symbiotic genomes (72, 73).

**“Candidatus Schmidhempelia bombi.”** We propose the name “Candidatus Schmidhempelia bombi” for the bacterium identified within the bumble bee Bombus impatiens and several other bumble bee species (36). Phylogenetic reconstruction places this bacterium within the Orbales, a recently described family of Gammaproteobacteria that has been identified nearly exclusively within insect guts. “Ca. Schmidhempelia” is a distinct clade (95% bootstrap support) of Orbales that is separate from the named genera Gilliamella and Frischella that are symbiotic in honey bees and bumble bees, as well as from other non-bee-associated members of the Orbales (Fig. 1b; see also Fig. S4 in the supplemental material). This bacterium has an average of 5% 16S rRNA gene sequence divergence from Gilliamella apicola sequences and has been identified in several bumble bee species from around the world, yet it has not been found in the thoroughly surveyed honey bee microbiota (36). The proposed genus for “Ca. Schmidhempelia bombi” refers to the Swiss evolutionary parasitologist Paul Schmid-Hempel, who has studied the evolutionary ecology of bumble bee species and associated organisms, while the specific epithet reflects that this bacterium resides within bumble bees. Unique features of this organism include its apparent restriction to bumble bee hosts and the 16S rRNA gene sequence 5’-TTTAA AACTGGTCGTCGAGTATTGT-3’ (positions 636 to 662 of the 16S rRNA gene, with E. coli numbering).
<table>
<thead>
<tr>
<th>IMG gene object ID*</th>
<th>IMG locus tag</th>
<th>Gene product name</th>
<th>Length (aa)</th>
<th>Scaffold IMG accession no.</th>
<th>Scaffold length (bp)</th>
<th>Contig GC%</th>
<th>COG</th>
<th>Genome source</th>
</tr>
</thead>
<tbody>
<tr>
<td>2505924216</td>
<td>GBI_0012.00000200</td>
<td>Arabinose efflux permease</td>
<td>405</td>
<td>GBI_ctg820001289386</td>
<td>38219</td>
<td>0.39</td>
<td>COG2814</td>
<td>Staphylococcus carnosus subsp. carnosus TM300</td>
</tr>
<tr>
<td>2505924217</td>
<td>GBI_0012.00000210</td>
<td>Uridine phosphorylase (EC 2.4.2.3)</td>
<td>251</td>
<td>GBI_ctg820001289386</td>
<td>38219</td>
<td>0.39</td>
<td>COG2820</td>
<td>Lactobacillus rhamnosus Lc 705</td>
</tr>
<tr>
<td>2505924690</td>
<td>GBI_0029.00000900</td>
<td>Peptide methionine sulfoxide reductase (EC 1.8.4.11, EC 1.8.4.12, EC 1.8.4.11)</td>
<td>177</td>
<td>GBI_ctg820001289403</td>
<td>102,764</td>
<td>0.35</td>
<td>COG0225</td>
<td>Listeria grayi DSM 20601</td>
</tr>
<tr>
<td>2505924734</td>
<td>GBI_0036.00000050</td>
<td>Nitrous oxide-stimulated promoter.</td>
<td>112</td>
<td>GBI_ctg820001289410</td>
<td>14,474</td>
<td>0.35</td>
<td></td>
<td>Bacillus coagulans 2-6</td>
</tr>
<tr>
<td>2505924830</td>
<td>GBI_0041.00000930</td>
<td>Peptide methionine sulfoxide reductase (EC 1.8.4.11, EC 1.8.4.12, EC 1.8.4.11)</td>
<td>124</td>
<td>GBI_ctg820001289416</td>
<td>38,071</td>
<td>0.35</td>
<td>COG4687</td>
<td>Enterococcus faecium E980</td>
</tr>
<tr>
<td>2505924831</td>
<td>GBI_0041.00000140</td>
<td>PTS, mannose/fructose/N-acetylgalactosamine-specific component IID</td>
<td>302</td>
<td>GBI_ctg820001289416</td>
<td>38,071</td>
<td>0.35</td>
<td>COG0225</td>
<td>Listeria grayi DSM 20601</td>
</tr>
<tr>
<td>2505924832</td>
<td>GBI_0041.00000150</td>
<td>PTS, mannose/fructose/N-acetylgalactosamine-specific component IIC</td>
<td>267</td>
<td>GBI_ctg820001289416</td>
<td>38,071</td>
<td>0.35</td>
<td>COG3715</td>
<td>Bacillus thuringiensis subsp. israelensis ATCC 35646</td>
</tr>
<tr>
<td>2505924833</td>
<td>GBI_0041.00000160</td>
<td>PTS, mannose/fructose/N-acetylgalactosamine-specific component IIB (EC 2.7.1.69)</td>
<td>323</td>
<td>GBI_ctg820001289416</td>
<td>38,071</td>
<td>0.35</td>
<td>COG4444</td>
<td>Bacillus thuringiensis IBL200</td>
</tr>
<tr>
<td>2505924834</td>
<td>GBI_0041.00000170</td>
<td>Uncharacterized protein conserved in bacteria</td>
<td>244</td>
<td>GBI_ctg820001289416</td>
<td>38,071</td>
<td>0.35</td>
<td>COG1636</td>
<td>Enterococcus faecium E980</td>
</tr>
<tr>
<td>2505924868</td>
<td>GBI_0043.00000060</td>
<td>Predicted hydrolases of the HAD superfamily</td>
<td>257</td>
<td>GBI_ctg820001289418</td>
<td>10,510</td>
<td>0.37</td>
<td>COG0561</td>
<td>Listeria welshimeri serovar 6b, SLCC5334</td>
</tr>
<tr>
<td>2505924928</td>
<td>GBI_0048.00000120</td>
<td>Adenine-specific DNA methylase Mod (EC 2.1.1.72)</td>
<td>628</td>
<td>GBI_ctg820001289423</td>
<td>59,508</td>
<td>0.38</td>
<td>COG2189</td>
<td>Clonobacterium fuscus DM1</td>
</tr>
<tr>
<td>2505924929</td>
<td>GBI_0048.00000130</td>
<td>Restriction endonuclease (EC 3.1.21.5)</td>
<td>1,035</td>
<td>GBI_ctg820001289423</td>
<td>59,508</td>
<td>0.38</td>
<td>COG3587</td>
<td>Clonobacterium fuscus DM1</td>
</tr>
<tr>
<td>2505924938</td>
<td>GBI_0048.00000220</td>
<td>Pyrimidine-nucleoside phosphorylase (EC 2.4.2.2)</td>
<td>433</td>
<td>GBI_ctg820001289423</td>
<td>59,508</td>
<td>0.38</td>
<td>COG2013</td>
<td>Bacillus atrophaeus 1942</td>
</tr>
<tr>
<td>2505924940</td>
<td>GBI_0048.00000240</td>
<td>Nucleoside permease</td>
<td>398</td>
<td>GBI_ctg820001289423</td>
<td>59,508</td>
<td>0.38</td>
<td>COG1972</td>
<td>Bacillus pumilus SAFR-032</td>
</tr>
<tr>
<td>2505925004</td>
<td>GBI_0051.00000020</td>
<td>Predicted membrane protein</td>
<td>173</td>
<td>GBI_ctg820001289427</td>
<td>39,885</td>
<td>0.37</td>
<td>COG4905</td>
<td>Eubacterium cylindroides T2-87</td>
</tr>
<tr>
<td>2505925044</td>
<td>GBI_0052.00000060</td>
<td>Putative glucose uptake permease</td>
<td>291</td>
<td>GBI_ctg820001289428</td>
<td>34,555</td>
<td>0.36</td>
<td>COG4975</td>
<td>Listeria welshimeri serovar 6b, SLCC5334</td>
</tr>
<tr>
<td>2505925059</td>
<td>GBI_0052.00000200</td>
<td>Type I site-specific restriction-modification system, R (restriction) subunit and related helicases (EC 3.1.21.3)</td>
<td>1,026</td>
<td>GBI_ctg820001289428</td>
<td>34,555</td>
<td>0.36</td>
<td>COG0610</td>
<td>Bacillus coagulans 36D1</td>
</tr>
<tr>
<td>2505925060</td>
<td>GBI_0052.00000210</td>
<td>Restriction endonuclease S subunits (EC 3.1.21.3)</td>
<td>417</td>
<td>GBI_ctg820001289428</td>
<td>34,555</td>
<td>0.36</td>
<td>COG0732</td>
<td>Bacillus coagulans 36D1</td>
</tr>
<tr>
<td>2505925061</td>
<td>GBI_0052.00000220</td>
<td>Type I restriction-modification system methyltransferase subunit (EC 2.1.1.72)</td>
<td>859</td>
<td>GBI_ctg820001289428</td>
<td>34,555</td>
<td>0.36</td>
<td>COG0286</td>
<td>Bacillus coagulans 36D1</td>
</tr>
<tr>
<td>2505925076</td>
<td>GBI_0054.00000020</td>
<td>Chorismate mutase (EC 5.4.99.5)</td>
<td>88</td>
<td>GBI_ctg820001289430</td>
<td>20,279</td>
<td>0.36</td>
<td>COG1605</td>
<td>Anaerofustis stercorihominis DSM 17244</td>
</tr>
<tr>
<td>2505925077</td>
<td>GBI_0054.00000030</td>
<td>Hypothetical protein</td>
<td>77</td>
<td>GBI_ctg820001289430</td>
<td>20,279</td>
<td>0.36</td>
<td></td>
<td>Desulfotobacterium hafniense DCB-2</td>
</tr>
</tbody>
</table>
| Genbank ID | Description                                                                 | Accession | Start | Length | COG
|------------|------------------------------------------------------------------------------|-----------|-------|--------|------
| GBI_0059.00000040 | Transposase                                                                 | GBi_ctg820001289435 | 37,254 | 0.37   | Staphylococcus aureus subsp. aureus ST398
| GBI_0060.000000290 | Predicted branched-chain amino acid permease (azaleucine resistance)    | GBi_ctg820001289436 | 77,226 | 0.37   | Listeria welshimeri serovar 6b, SLCC3534
| GBI_0061.000000300 | Predicted membrane protein                                                                 | GBi_ctg820001289436 | 77,226 | 0.37   | Listeria welshimeri serovar 6b, SLCC3534
| GBI_0061.000000100 | PTS beta-glucoside-specific II A component, Gk family (TC 4.A.1.2.5) | GBi_ctg820001289437 | 27,350 | 0.37   | Bacillus subtilis subsp. spizizenii W23
| GBI_0061.000000110 | Beta-glucosidase/6-phospho-beta-glucosidase/beta-galactosidase (EC 3.2.1.86) | GBi_ctg820001289437 | 27,350 | 0.37   | Clastidium bartlettii DSM 16795
| GBI_0062.000000040 | Transcriptional regulator                                                                 | GBi_ctg820001289444 | 65,738 | 0.37   | Syntrophomonas wofei subsp. wofei strain Goettingen, DSM 2245B
| GBI_0062.000000050 | Methionine synthase (B12-independent) (EC 2.1.1.14)                         | GBi_ctg820001289444 | 65,738 | 0.37   | Paenibacillus curdlanolyticus YK9
| GBI_0063.000000060 | Tagtose-1,6-bisphosphate aldolase (EC 4.1.2.40)                               | GBi_ctg820001289444 | 65,738 | 0.37   | Enterococcus gallinarum EG2
| GBI_0064.000000020 | Fructose-2,6-bisphatase (EC 5.4.2.1)                                         | GBi_ctg820001289445 | 47,248 | 0.37   | Listeriaeae bacterium TTU MI-001
| GBI_0065.000000020 | Amidases related to nicotinamidase                                             | GBi_ctg820001289446 | 64,926 | 0.37   | Geobacillus sp. strain WCH70
| GBI_0065.0000000210 | Serine/threonine exchange transporter, LAT family (TC 2.A.3.8.12)            | GBi_ctg820001289448 | 31,379 | 0.36   | Bacillus cereus subsp. cytotoxins NVH 391-98
| GBI_0066.000000160 | Predicted membrane protein                                                                 | GBi_ctg820001289453 | 45,947 | 0.35   | Leptotrichia goodfellowii F024
| GBI_0066.000000170 | Acetyltransferases (EC 2.3.1.-)                                               | GBi_ctg820001289454 | 45,947 | 0.35   | Desulfitobacterium dehalogenans JW/1/DC1, ATCC 51507
| GBI_0066.000000300 | Predicted hydrolases of the HAD superfamily                                  | GBi_ctg820001289453 | 45,947 | 0.35   | Listeria seeligeri serovar 1/2b, SLCC3954
| GBI_0070.0000000300 | Predicted HD superfamily hydrolase                                           | GBi_ctg820001289456 | 51,284 | 0.34   | Pediococcus pentosaceus ATCC 25745
| GBI_0082.0000000300 | Predicted glutamine amidotransferase involved in pyridoxine biosynthesis     | GBi_ctg820001289460 | 6,516  | 0.34   | Bacillus subtilis subsp. subtilis strain 168
| GBI_0082.0000000400 | Predicted glutamine amidotransferase involved in pyridoxine biosynthesis     | GBi_ctg820001289460 | 6,516  | 0.34   | Bacillus subtilis subsp. subtilis strain 168
| GBI_0082.0000000500 | Predicted glutamine amidotransferase involved in pyridoxine biosynthesis     | GBi_ctg820001289460 | 6,516  | 0.34   | Listeria monocytogenes 08-5923

*ID, identifier.
ACKNOWLEDGMENTS

We thank Ellen O. Martinson for her helpful discussion and comments on the manuscript and Jennifer H. Wiseacre for bioinformatics assistance. V.G.M. was supported by the Center for Insect Science (University of Arizona) and a National Science Foundation award to N.A.M. (NSF 1046153). Sequencing and analysis were supported by NIH Director’s Pioneer Award 1DP1OD006416-01 (G. E. Robinson) and by NIH grant R01-HG006677 (S.L.S.). H.K. was supported by Swiss National Science Foundation grants 140157 and 147881.

REFERENCES


