Genome evolution in an ancient bacteria-ant symbiosis: Convergent gene loss among *Blochmannia* spanning the origin of the ant tribe Camponotini

Laura E Williams, Jennifer J Wernegreen

Stable associations between bacterial endosymbionts and insect hosts provide opportunities to explore genome evolution in the context of established mutualisms and assess the roles of selection and genetic drift across host lineages and habitats. *Blochmannia,* obligate endosymbionts of ants of the tribe Camponotini, have coevolved with their ant hosts for ~40 MY. To investigate early events in *Blochmannia* genome evolution across this ant host tribe, we sequenced *Blochmannia* from two divergent host lineages, *Colobopsis obliquus* and *Polyrhachis turneri*, and compared them with four published genomes from *Blochmannia* of *Camponotus sensu stricto*. Reconstructed gene content of the last common ancestor (LCA) of these six *Blochmannia* genomes is reduced (690 protein coding genes), consistent with rapid gene loss soon after establishment of the symbiosis. Differential gene loss among *Blochmannia* lineages has affected cellular functions and metabolic pathways, including DNA replication and repair, vitamin biosynthesis and membrane proteins. *Blochmannia* of *P. turneri* (i.e., *B. turneri*) encodes an intact DnaA chromosomal replication initiation protein, demonstrating that loss of *dnaA* was not essential for establishment of the symbiosis. Based on gene content, *B. obliquus* and *B. turneri* are unable to provision hosts with riboflavin. Of the six sequenced *Blochmannia*, *B. obliquus* is the earliest diverging lineage (i.e., the sister group of other *Blochmannia* sampled) and encodes the fewest protein-coding genes and the most pseudogenes. We identified 55 genes involved in convergent gene loss, including glutamine synthetase, which may participate in nitrogen recycling. Pathways for biosynthesis of coenzyme A, terpenoids and riboflavin were lost in multiple lineages, suggesting relaxed selection on the pathway after inactivation of one component. Analysis of Illumina read datasets did not detect evidence of plasmids encoding missing functions, nor the presence of coresident symbionts other than *Wolbachia*. Although gene order is strictly conserved in four *Blochmannia* of *Camponotus sensu stricto*, comparisons with deeply divergent lineages revealed inversions in eight genomic regions, indicating ongoing recombination despite ancestral loss of *recA*. In sum, the addition of two *Blochmannia* genomes of divergent host lineages enables reconstruction of early events in evolution of

this symbiosis and suggests that *Blochmannia* lineages may experience distinct, host-associated selective pressures. Understanding how evolutionary forces shape genome reduction in this system may help to clarify forces driving gene loss in other bacteria, including intracellular pathogens and free-living marine bacteria.

1   Laura E. Williams[1,2] and Jennifer J. Wernegreen[1,3,*]

2   [1] Duke Center for Genomic and Computational Biology, Duke University, Durham, NC

3   [2] Current address: Department of Biology, Providence College, Providence, RI

4   [3] Nicholas School of the Environment, Duke University, Durham, NC

5   [*] Corresponding author: Jennifer J. Wernegreen, 101 Science Drive, Box 3382, Durham, NC,

6   27708, 919-681-0331, j.wernegreen@duke.edu

Introduction

8   The evolution of stable mutualisms between bacteria and insects has occurred many times

9   and involves phylogenetically diverse lineages (Moran, McCutcheon & Nakabachi, 2008; Moya

10   et al., 2008; Kikuchi, 2009; Douglas, 2014). In many of these symbioses, the limited diets of

11   insect hosts are supplemented by an intracellular bacterial partner. Variation in dietary

12   requirements among insect hosts likely results in differing selective pressure on endosymbiont

13   genes, which in turn impacts endosymbiont genome evolution. For example, *Buchnera*

14   *aphidicola* supplement the carbohydrate-rich plant sap diet of aphids with amino acids

15   (Shigenobu et al., 2000), whereas *Wigglesworthia* species supplement the blood diet of tsetse

16   flies with vitamins and other cofactors (Akman et al., 2002). By contrast, cockroaches and ants of

17   the tribe Camponotini are generally considered omnivores with complex diets. Their bacterial

18   partners, *Blattabacterium* and *Blochmannia*, respectively, synthesize essential amino acids and

19   participate in nitrogen recycling (Gil et al., 2003; Lopez-Sanchez et al., 2009; Sabree,

20   Kambhampati & Moran, 2009), although the nitrogen recycling pathway is not complete in some

21   *Blochmannia* species (Williams & Wernegreen, 2010). In the case of *Blochmannia,* its nutritional

22   role may be most important at particular stages of the host's lifecycle (Zientz et al., 2006;

23   Feldhaar et al., 2007; Stoll et al., 2010).

24   Genomes of established endosymbionts like *Buchnera*, *Wigglesworthia* and *Blochmannia*

25   are typically characterized by high AT content, high mutation rates and extreme stability of gene

26   order (Moran, McCutcheon & Nakabachi, 2008). Recombination-related genes are often lost

27   during endosymbiont evolution, and very few recombination events are evident in endosymbiont

28   lineages. Genome reduction in established endosymbionts occurs via degradation and loss of

29   individual genes. Gene loss is likely shaped by both relaxed selective pressure due to the stable

30   intracellular niche and genetic drift due to small effective population sizes and bottlenecks during

31   vertical transmission of symbionts. The balance between these forces may shift over time for a

32   given association, from the initial acquisition event through the ongoing evolution of the

33   symbiosis.

34   To explore genome reduction and evolution in the context of a long-term endosymbiosis,

35   we sequenced two deeply divergent lineages of Candidatus *Blochmannia*, which are obligate

36   bacterial endosymbionts of ants of the tribe Camponotini (Sauer et al., 2000). *Blochmannia* are

37   closely related to free-living Enterobacteriaceae such as *Escherichia coli* and form a clade with

38   other obligate endosymbionts including *Baumannia*, *Sodalis* and *Wigglesworthia* (Herbeck,

39   Degnan & Wernegreen, 2005; Husnik, Chrudimsky & Hypsa, 2011). The presence of

40   *Blochmannia* in multiple extant ant genera of Camponotini points to a single colonization event

41   in the ancestral lineage as the origin of the symbiosis (Sameshima et al., 1999; Wernegreen et al.,

42   2009). Based on phylogenetic evidence, the ancestor of *Blochmannia* may have been a facultative

43   symbiont of insects (Herbeck, Degnan & Wernegreen, 2005; Wernegreen et al., 2009).

44   The ant hosts of the four previously sequenced *Blochmannia* (*Camponotus chromaiodes*,

45   *C. floridanus*, *C. pennsylvanicus* and *C. vafer*) belong to *Camponotus sensu stricto*. Whereas

46   these host species span the origin of *Camponotus* ~16-20 MY, the association between

47   *Blochmannia* and the tribe Camponotini is at least twice that old, on the order of 40 MY. To

48   reconstruct earlier events in the evolution of this symbiosis, we sequenced the genomes of

49   *Blochmannia* from two divergent lineages in the tribe Camponotini: the genus *Polyrhachis* and

50   the *Colobopsis* lineage.

51   Though *Colobopsis* is formally considered a subgenus of *Camponotus*, phylogenetic

52   evidence has revealed it is a separate lineage from *Camponotus* (Brady et al., 2006; Moreau &

53   Bell, 2013). These studies demonstrated that *Colobopsis* diverged early in the evolution of the

54   tribe Camponotini, and results are consistent with *Colobopsis* being the sister group of all other

55   Camponotini sampled, although relationships in this part of the tree were difficult to resolve

56   conclusively. Subsequent work has further suggested that *Colobopsis* is likely the sister group of

57   all other extant camponotines (pers. comm., P. S. Ward, 2014). In contrast to *Camponotus* species

58   with published *Blochmannia* genomes, *C. obliquus* lives in small twigs and branches, often in the

59   canopy. *Polyrhachis* diverged later than *Colobopsis* and is found in the Old World. The deep

60   evolutionary divergence of these host lineages, as well as differences in their geographic range

61   and habitats, provide a valuable opportunity to investigate the evolutionary trajectories of

62   *Blochmannia* across the Camponotini and to clarify ancient events that shaped this 40 MY old

63   ant-bacterial partnership.

64   **Materials and Methods**

65   **Preparation of genomic DNA**

66   A single colony of *C. obliquus* was collected near Morehead City, North Carolina, USA

67   by B. Guénard, and *P. turneri* was collected near Townsville, Australia by S. K. A. Robson (see

68   Acknowledgments). Voucher specimens were deposited in the Bohart Museum of Entomology,

69   University of California, Davis (UCDC), corresponding to voucher ID numbers

70   CASENT0221021 (*C. obliquus*) and CASENT0220426 (*P. turneri*). We used the Qiagen DNeasy

71   Blood and Tissue Kit to prepare genomic DNA from a pooled sample of seven eggs, two larvae,

72   five pupae, eight minor workers, six major workers and five female alates for *C. obliquus* and a

73   pooled sample of three worker gasters for *P. turneri*.

Sequencing and assembly of *Blochmannia* genomes

74

75   *C. obliquus* gDNA was sequenced on an Illumina HiSeq to generate 100 bp paired end

76   reads. We modified filter_reads.py (https://github.com/nickloman/xbase/blob/master/short-read-

77   assembly/filter_reads.py) for Sanger FASTQ format and used the script to remove paired reads

78   with any bases of quality score <30, which retained 6,184,892 read pairs. We assembled this

79   filtered read dataset using Velvet v1.2.07 (Zerbino & Birney, 2008) with a hash length of 61,

80   exp_cov 200, cov_cutoff 20 and scaffolding turned off. This generated 566 contigs, one of which

81   aligned to *B. pennsylvanicus* using MAUVE (Darling, Mau & Perna, 2010). We observed an

82   overlap of 116 bp between the contig ends, which suggested that *de novo* assembly had produced

83   a closed genome. To test this, we used Mosaik (Lee et al., 2014) to align the filtered read dataset

84   against a 1260 bp sequence encompassing the joined contig ends and flanking regions. This

85   alignment produced no zero coverage regions, which confirmed that the single contig was the

86   closed *B. obliquus* genome.

87   To finish the genome, we used a two-step process applying different alignment programs

88   to confirm the majority genotype, or the base represented by the majority of reads at each

89   position. In the first step, we used Mosaik, which allows the user to set the stringency of

90   mismatch tolerance. We aligned the filtered read dataset against the closed genome sequence

91   with a maximum mismatch threshold of 12 and then removed duplicate read pairs with the

92   DupSnoop module. Using Consed (Gordon, Abajian & Green, 1998), we generated a

93   questionable consensus bases report and a highly discrepant indels report for the resulting

94   alignment, which did not identify any positions that needed editing.

95   The second step in our finishing process invokes the IndelRealigner module of the

96   Genome Analysis Toolkit (GATK) to ensure accurate identification of indels (DePristo et al.,

97   2011). We aligned the filtered read dataset against the closed genome sequence using BWA (Li

98   & Durbin, 2009) and then processed the resulting alignment with RealignerTargetCreator and

99   IndelRealigner from the GATK package. We removed duplicate reads using Picard

100 MarkDuplicates (http://broadinstitute.github.io/picard). Finally, we analyzed the processed

101 alignment using VarScan (Koboldt et al., 2009). We did not identify any positions that needed

102 editing, confirming that the single contig was the closed, finished *B. obliquus* genome. Genome

103 coverage for this alignment averaged 558x. This sequence is deposited in GenBank as accession

104 number CP010049. The Illumina read dataset is deposited in the NCBI Short Read Archive

105 (SRA) as SRP050154.

106   *P. turneri* gDNA was sequenced on an Illumina Genome Analyzer II (GAIIx) to generate

107 150 bp paired end reads. We used DynamicTrim.pl and LengthSort.pl from the SolexaQA

108 package (Cox, Peterson & Biggs, 2010) to generate trimmed reads of at least 80 bp with quality

109 score >30 for each base. The resulting trimmed read dataset included 9,590,066 read pairs. We

110 assembled this filtered read dataset using Velvet with a hash length of 41, exp_cov 200,

111 cov_cutoff 20 and scaffolding turned off. This generated 8,275 contigs, four of which aligned to

112 *B. pennsylvanicus* using MAUVE. To close the four gaps, we generated Sanger sequencing reads

113 and used Phred/Phrap/Consed to assemble and manually examine the sequence.

114   To finish the *B. turneri* genome, we followed the two-step process described above for *B.*

115 *obliquus*. In the first step, we corrected nine positions based on alignment of the full read dataset

116 against the closed genome sequence using Mosaik. In the second step, we aligned the full read

117 dataset against the corrected sequence using BWA followed by processing with GATK. We

118 analyzed the processed alignment with VarScan and corrected a single position. Genome

119 coverage for this alignment averaged 1223x. The closed, finished genome sequence is deposited

120   in GenBank as accession number CP010048. The Illumina read dataset is deposited in the NCBI

121   Short Read Archive (SRA) as SRP050161.

122   **Annotation of *Blochmannia* genomes**

123   We used an annotation engine hosted by the Institute for Genome Sciences (IGS) at the

124   University of Maryland School of Medicine to generate an automated annotation of each genome

125   sequence (Galens et al., 2011), which we then manually curated within the MANATEE

126   framework (http://manatee.sourceforge.net/igs/). Protein-coding genes predicted by the

127   annotation engine were removed if they lacked a Blast-Extend-Repraze (BER) alignment score

128    $<10^{-5}$ to a protein-coding gene from outside of *Blochmannia*. We manually examined possible

129   frameshifted genes flagged by the annotation engine. For genes with frameshifts in homopolymer

130   tracts, we included the likely position of the frameshift in the GenBank annotation. Because the

131   frameshifts may be corrected by polymerase slippage, we consider these to be intact genes.

132   We curated start sites using BER alignments to *Blochmannia* and closely related species.

133   When possible, we used the gene name and symbol listed in SwissProt for the homologous gene

134   in *E. coli* to maintain consistency with existing proteobacterial annotations. For conserved

135   hypothetical proteins or proteins with similarity to a protein family but not a specific family

136   member, we did not assign a gene name and refer to them using the locus tag (for example,

137   BTURN675_020).

138   After curating the annotations, we analyzed intergenic regions in each genome with

139   RFAM (Burge et al., 2013) and BLASTX (Altschul et al., 1990) to identify uncalled genes and

140   pseudogenes. In both *B. obliquus* and *B. turneri*, RFAM identified three RNA-coding genes (*ffs*,

141   *rnpB* and tmRNA). To identify protein-coding genes, we aligned intergenic regions to the

142   GenBank non-redundant database using BLASTX with the low complexity filter turned off. We

143   manually examined hits with e-value <10^{-5}. Pseudogenes were identified by multiple nonsense

144   mutations, frameshifts and/or gaps. We annotated pseudogene coordinates using the boundaries

145   of the BLASTX alignments. Analysis of *B. obliquus* intergenic regions detected two genes (*cyoD*

146   and *sdhD*) and 15 pseudogenes (*dnaA*, *engD*, *glnA*, *pdxA*, *pdxB*, *pdxJ*, *ribA*, *ribB*, *ribC*, *ribD*,

147   *secD*, *secF*, *topA*, *yigB* and *uvrD*). Analysis of *B. turneri* intergenic regions detected five genes

148   (*cyoD*, *infA*, *rpmJ*, *ycaR* and *yidD*), one frameshifted gene (*ybeY*) and one pseudogene of a

149   hypothetical protein (BTURN675_514).

150   **Phylogenetic analysis**

151   We classified genes into six MultiFun categories (cell processes, cell structure,

152   information transfer, metabolism, regulation, and transport) by searching for the gene name in the

153   EcoCyc database. Some genes are assigned to more than one MultiFun category, and we included

154   all categories listed for each gene. If the gene had no associated MultiFun terms in EcoCyc or if it

155   had no gene name, such as BPEN_040, we considered the gene unclassified.

156   To construct a phylogeny, we chose *Baumannia* (NC_007984), *Hamiltonella*

157   (NC_012751) and *Sodalis* (NC_007712) as outgroups. We identified orthologs of *Blochmannia*

158   genes in these genomes using the Reciprocal Smallest Distance (RSD) algorithm (Wall, Fraser &

159   Hirsh, 2003) with default values for divergence (0.8) and e-value (10^{-5}). For each of the six

160   MultiFun categories, we randomly selected five genes present in all taxa. There were no

161   duplicates in the resulting set of 30 genes. We excluded *B. chromaiodes* from the phylogeny

162   because its genome sequence is 98.0% identical to that of *B. pennsylvanicus* (Williams &

163   Wernegreen, 2013). We used TranslatorX (Abascal, Zardoya & Telford, 2010) and MAFFT

164   (Katoh et al., 2005) to construct a multiple sequence alignment for each gene, which we then

165   trimmed with ZORRO (Wu, Chatterji & Eisen, 2012). We concatenated the trimmed amino acid

166   alignments and used MrBayes v3.2.1 (Ronquist & Huelsenbeck, 2003) to construct a majority

167   rule consensus tree.

168   **GC skew and DnaA box motif search**

169   For *B. pennsylvanicus* (NC_007292), *B. obliquus,* and *B. turneri*, we used DNAPlotter

170   (Carver et al., 2009) to construct plots of GC skew with 500 bp window size and 50 bp step size.

171   We used Pattern Locator (Mrazek & Xie, 2006) to search these three genome sequences for the

172   consensus DnaA box motif TTWTNCACA.

173   **BLAST analysis**

174   Because we prepared genomic DNA from whole ants or gasters, the Illumina read datasets

175   include coverage of genomes other than *Blochmannia*, such as the ant host nuclear and

176   mitochondrial genomes. To determine if the *de novo* assemblies included contigs from potential

177   *Blochmannia* plasmids or other bacterial symbionts, we used BLASTN to align all ≥500 bp

178   contigs to the GenBank non-redundant database, limited to bacteria (taxid 2). For contigs with at

179   least one hit of >30% coverage and an evalue of <10^{-5}, we aligned each contig against the full

180   non-redundant database and examined the top hits.

181   To test whether the Illumina read datasets contain evidence of riboflavin biosynthesis

182   genes, we constructed BLAST databases of the *C. obliquus* and *P. turneri* read datasets. For *C.*

183   *obliquus*, we built a BLAST database using the unaligned reads file generated by Mosaik during

184   the initial finishing step (see above), thereby excluding most reads originating from the *rib*

185   pseudogenes within the *B. obliquus* genome. We compiled a set of query genes representing

186   orthologs of conspicuously absent genes, including *ribABDEF* from *B. pennsylvanicus*

187   (NC_007292), *Baumannia* (NC_007984) and *E. coli* MG1655 (NC_000913), *ribH* from *B.*

188   *pennsylvanicus* and *Baumannia, ribC* from *E. coli* and the *Camponotus floridanus* gene for EF-

189   1alpha-F2 (EFN72500). For analysis of *C. obliquus* reads, we also included *ribF* and *rib*

190   pseudogenes from *B. obliquus* in the query set. We used BLASTN to align the query genes

191   against both BLAST databases and examined the output for alignments with e-values $<10^{-5}$.

192   Certain gene distribution patterns may be explained by either convergent gene loss or

193   acquisition via horizontal gene transfer. To test the hypothesis of horizontal gene transfer, we

194   aligned each of the 55 genes with such patterns against the GenBank non-redundant database

195   using BLAST. We first used BLASTN, and if this search did not return significant hits, we used

196   BLASTX. We examined the taxonomic assignments of the top hits to identify genes with high

197   scoring alignments to bacteria outside of the Enterobacteriaceae, which may support the

198   hypothesis of horizontal gene transfer.

199   **Results**

200   ***B. obliquus* has the fewest protein-coding genes and most pseudogenes**

201   The size and GC content of the *B. obliquus* and *B. turneri* genomes are within the ranges

202   observed for *Blochmannia* of *Camponotus sensu stricto* (Table 1). Although *B. obliquus* is on the

203   upper end of the size range, it has the fewest protein-coding genes of the six sequenced

204   *Blochmannia*. We also detected 15 pseudogenes in *B. obliquus*, which is an unusually high

205   number for *Blochmannia* (Table S1). In both *B. obliquus* and *B. turneri*, we identified genes that

206   have frameshifts in homopolymer tracts but otherwise are expected to encode a full-length

207 protein. We consider these genes to be functional, because they may be corrected by polymerase

208 slippage during transcription and expressed as full-length proteins (Tamas et al., 2008;

209 Wernegreen, Kauppinen & Degnan, 2010).

210 ***B. obliquus* is the earliest diverging *Blochmannia* lineage sequenced**

211 To determine the evolutionary relationships of the *Blochmannia* lineages, we constructed

212 a phylogeny of 30 randomly selected protein-coding genes shared among *Blochmannia* genomes

213 and the outgroups *Baumannia*, *Hamiltonella* and *Sodalis*. The resulting phylogeny has 100%

214 posterior probability at all nodes (Fig. 1). In this phylogeny, *B. turneri* and *Blochmannia* of

215 *Camponotus sensu stricto* are more closely related to each other than to *B. obliquus*. This

216 topology is congruent with recent phylogenies of the ant host taxa that show *Colobopsis* as a

217 separate lineage rather than a subgenus of *Camponotus* (Brady et al., 2006). Hereafter, we use the

218 phrase "*Blochmannia* of *Camponotus*" to mean *Blochmannia* of *Camponotus sensu stricto*, which

219 does not include *Colobopsis*.

220 **Assemblies show evidence of *Wolbachia* but no *Blochmannia* plasmids or other bacterial**

221 **symbionts**

222 Because we prepared genomic DNA from whole ants or gasters, the Illumina read datasets

223 include coverage of genomes other than *Blochmannia*. To determine whether the assemblies

224 generated by Velvet included contigs from other symbionts or possible plasmids, we aligned all

225 contigs ≥500 bp against the GenBank non-redundant database using BLASTN. We limited our

226 first search to bacteria (taxid 2). For contigs with hits of evalue $<10^{-5}$ and >30% query coverage,

227 we aligned them against the full database. The top hits from this search are shown in Table S2.

228   Analysis of the *C. obliquus* assembly identified 15 contigs with BLAST hits that met the

229   criteria outlined above. Three of these are from the ant nuclear genome and one is from the ant

230   mitochondrial genome (Table S2). The best-scoring hits for the remaining 11 contigs are all

231   matches to *Wolbachia*, an endosymbiont found in many insect species (Werren, Baldo & Clark,

232   2008), including ants (Russell et al., 2012). We did not detect evidence of any other symbiont or

233   any *Blochmannia* plasmids.

234   Analysis of the *P. turneri* assembly identified eight contigs with BLAST hits that met the

235   criteria outlined above. Four of these are from the ant nuclear genome and three are from the ant

236   mitochondrial genome (Table S2). The best-scoring hit for the remaining contig is to *B.*

237   *pennsylvanicus*. To examine this further, we aligned the contig to the *B. turneri* genome

238   sequence, which produced a much better alignment with 100% coverage and 85% identity. The

239   contig spans the 3' end of *prfB*, a 19 bp intergenic region and the 5' end of *lysS*. Coverage of this

240   contig averages 54x, whereas coverage of the same region in the *B. turneri* genome sequence

241   averages 1380x. Given its high nucleotide identity and synteny with the *B.* turneri genome and its

242   low coverage, we consider this contig an artifact of assembly rather than evidence of a plasmid or

243   other symbiont.

244   **Gene content of reconstructed Last Common Ancestor is highly conserved in divergent**

245   ***Blochmannia* lineages**

246   Using the six sequenced *Blochmannia* genomes, we reconstructed the likely gene content

247   of the Last Common Ancestor (LCA) of these lineages. Horizontal gene transfer (HGT) is

248   considered unlikely in *Blochmannia* due to the isolated intracellular niche of these bacteria.

249   Conservation of gene order in *Blochmannia* genomes supports the hypothesis that HGT is very

250   rare. Additionally, we used BLAST to identify the best hits for genes with distribution patterns

251   consistent with either independent loss in multiple lineages or acquisition via HGT, and we did

252   not find evidence of HGT (see below). For these reasons, we expect very little, if any, gain of

253   genes in *Blochmannia* lineages, and our reconstruction of the LCA includes any intact gene found

254   in at least one of the sequenced genomes.

255   By this definition, the genome of the LCA consists of 690 genes (Fig. 2). Most of these

256   genes are retained in the *Blochmannia* genomes, with 568 genes, or 82% of the gene content of

257   the LCA, found in all six species. The complete ortholog table can be found as Table S3. Our

258   reconstruction may underestimate LCA gene content, because it is possible that genes were lost

259   independently from all lineages at some point after divergence from the LCA. For example, *B.*

260   *chromaiodes*, *B. pennsylvanicus* and *B. obliquus* encode a pseudogene of *uvrD*, but no sequenced

261   *Blochmannia* has an intact *uvrD*. This is the only *Blochmannia* gene found solely as a

262   pseudogene. This gene may have been intact and functional in the LCA, but it is not represented

263   in our reconstruction here. Sequencing of additional taxa, including *Blochmannia* from other

264   deeply divergent ant host genera such as *Opisthopsis*, will further refine reconstruction of the

265   ancestral lineage.

266   **Gene content differences among divergent *Blochmannia* lineages include important cellular**

267   **functions**

268   DNA replication and repair

269   The *dnaA* chromosomal replication initiation protein is intact in *B. turneri*, detectable as a

270   pseudogene in *B. obliquus* and missing in all four sequenced *Blochmannia* of *Camponotus* (Fig.

271   2). In other gamma-proteobacteria, DnaA initiates replication by binding to 9-bp sequences called

272   DnaA boxes within the origin of replication (Zakrzewska-Czerwinska et al., 2007). GC skew

273   analysis predicts that the origin of replication is adjacent to *mnmG* in most *Blochmannia* species

274   (Fig. 3), so we searched the intergenic regions flanking *mnmG* in *B. obliquus*, *B. pennsylvanicus*

275   and *B. turneri* for the consensus DnaA box motif TTWTNCACA (Schaper & Messer, 1995).

276   We found two matches in *B. turneri*, none in *B. obliquus* and one match in *B.*

277   *pennsylvanicus*. By comparison, *Buchnera aphidicola*, which retained *dnaA*, has two DnaA boxes

278   in *oriC* (Mackiewicz et al., 2004); therefore, it is possible that DnaA can initiate replication in *B.*

279   *turneri* with only two consensus DnaA boxes. However, the presence of a DnaA box in *B.*

280   *pennsylvanicus*, which lacks *dnaA*, suggests that these motifs are not necessarily associated with

281   DnaA function in *Blochmannia*. In fact, when we searched the entire length of each genome, we

282   found multiple matches to the consensus DnaA box motif, which are unlikely to be involved in

283   DnaA function as described for gamma-proteobacteria such as *E. coli* (Hansen et al., 2006). More

284   work is needed to understand how the differential distribution of DnaA in *Blochmannia* species

285   affects control of DNA replication and whether DnaA boxes play a role in replication initiation in

286   *B. turneri*.

287   Loss of DNA repair mechanisms is a common characteristic of obligate intracellular

288   symbionts of insects. In the six sequenced *Blochmannia*, two genes involved in base excision

289   repair, *mutM* and *mutY*, are differentially distributed. *B. obliquus* encodes *mutM* but is missing

290   *mutY*, whereas *B. turneri* and the four *Blochmannia* of *Camponotus* are missing *mutM* and

291   encode *mutY* (Fig. 2). Both protein products act on 8-oxoG, which can mispair with adenine

292   (Michaels et al., 1992). MutM excises 8-oxoG when it is paired with cytosine, thereby initiating

293   base excision repair. If 8-oxoG is not removed prior to replication, MutY removes the mispaired

294   adenine, enabling repair. Inactivation of either gene leads to an increase in GC-to-TA

295 transversions, which may contribute to the reduced genomic GC content observed in

296 *Blochmannia* and other insect endosymbionts (Lind & Andersson, 2008).

297 Vitamin biosynthesis

298   Pyridoxal 5'-phosphate (PLP), the catalytically active form of vitamin B6, is an important

299 enzyme cofactor. With very few exceptions, insects and other animals cannot synthesize vitamin

300 B6 (Tanaka, Tateno & Gojobori, 2005). Biosynthesis of PLP in *E. coli* and other gamma-

301 proteobacteria occurs via a pathway encoded by seven genes (Mukherjee et al., 2011). This

302 pathway is conserved in *B. turneri* and *Blochmannia* of *Camponotus*, with the exception of

303 *epd/gapB*, which is missing from all six sequenced *Blochmannia* (Fig. 4). In *B. obliquus*, only

304 *dxs* and *serC* are conserved, whereas *pdxABHJ* are missing. We detected pseudogenes of three of

305 these genes. Based on these gene losses, *B. obliquus* appears unable to provision its ant host with

306 vitamin B6.

307   Riboflavin (vitamin B2) biosynthesis is another vitamin synthesis pathway with

308 differential gene distribution in these *Blochmannia* lineages. Riboflavin is essential for synthesis

309 of the cofactors flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD) (Abbas &

310 Sibirny, 2011). *Blochmannia* of *Camponotus* encode five riboflavin biosynthesis genes

311 (*ribABCDH*), which comprise the complete pathway in *E. coli*. By contrast, these genes are

312 missing in both *B. obliquus* and *B. turneri.* We detected pseudogenes of *ribABCD*, but not *ribH*,

313 in *B. obliquus.* The *ribF* gene, which encodes an enzyme that synthesizes FMN and FAD from

314 riboflavin, is conserved in all six sequenced *Blochmannia* species.

315   It is possible that riboflavin biosynthesis genes are encoded on a plasmid, by a secondary

316 symbiont (Lamelas et al., 2011), or even within the ant host nuclear genome (Husnik et al.,

317   2013). As discussed above, we detected *Wolbachia* but no plasmids or other symbionts in the

318   assemblies. To investigate possible alternative sources of riboflavin, we analyzed the Illumina

319   read datasets. Using BLASTN, we aligned a query set of *rib* genes from *B. pennsylvanicus*,

320   *Baumannia* and *E. coli* against the reads. We also included the elongation factor alpha F2 (EF-

321   1alpha-F2) gene from *Camponotus floridanus* in the query set to assess representation of the ant

322   host nuclear genome in the read datasets. Because *B. obliquus* has pseudogenes of four riboflavin

323   biosynthesis genes, we used the subset of reads that did not align to the *B. obliquus* genome to

324   build the *C. obliquus* BLAST database.

325   We found no evidence of intact riboflavin biosynthesis genes in the read datasets. For

326   reads from *P. turneri* gasters, we detected no BLASTN hits with e-values $<10^{-5}$ to the query

327   riboflavin biosynthesis genes. For reads from *C. obliquus* ants, some reads aligned to the *B.*

328   *obliquus* pseudogenes. These are likely low quality reads originating from these pseudogenes.

329   When we considered only reads that did not align by BLASTN to the *rib* pseudogenes, we

330   detected only three reads aligning to the query riboflavin biosynthesis genes with evalues $<10^{-5}$.

331   A single read, but not its mate, aligned to *ribE* from *B. pennsylvanicus* with only 58% coverage

332   and an evalue of $\sim10^{-6}$, and a single read pair aligned to *ribA* from *E. coli* with 100% identity and

333   100% coverage, which may be due to contamination or presence of gut-associated bacteria. These

334   single reads contrast with the 558x coverage of the *B. obliquus* genome. For both read datasets,

335   multiple reads aligned to the *C. floridanus* EF-1alpha gene with evalues $<10^{-20}$, confirming that

336   the reads include coverage of the host nuclear genome. Our analysis of the Illumina read datasets

337   did not reveal how these symbiotic systems compensate for the loss of riboflavin biosynthesis in

338   the *Blochmannia* partners.

339   Membrane proteins

340   All six sequenced *Blochmannia* encode *tolA*, an inner membrane protein and a component

341   of the Tol-Pal system; however, there are structural differences among the species. In *E. coli*,

342   TolA has three domains: a transmembrane domain that anchors the protein in the cytoplasmic

343   membrane, a central periplasmic domain with high alpha-helix structure and a globular

344   periplasmic domain (Godlewska et al., 2009). In *Blochmannia*, five of the six species have a

345   stretch of 80-100 bp tandem repeats in *tolA*, which varies in length and repeat sequence among

346   species (Table 2). By contrast, *B. obliquus* shows no evidence of this repeat region, which is also

347   not found in the *E. coli* homolog. *E. coli tolA* has a repeat region, but it occurs in a distinct

348   location in the protein and consists of shorter repeats (Zhou et al., 2012). In addition, the length

349   of *B. obliquus tolA* (771 bp) is shorter than *tolA* in the other *Blochmannia* species (990-1305 bp)

350   and *E. coli* (1266 bp).

351   The function of *tolA* in an insect endosymbiont such as *Blochmannia* is unknown. In *E.*

352   *coli*, the Tol-Pal system is thought to interact with phage particles and colicins (Godlewska et al.,

353   2009). Some structural features described for *E. coli* TolA are conserved in *Blochmannia*, such as

354   the N-terminal transmembrane domain. Both the TMHMM server v 2.0 and Phobius predicted

355   one transmembrane helix within the first 50 amino acids of TolA from each sequenced

356   *Blochmannia* genome. Six of the seven genes comprising the two Tol-Pal operons (*ybgC-tolQ-*

357   *tolR-tolA* and *tolB-pal-ybgF*) in *E. coli* are conserved in all six sequenced *Blochmannia*, with

358   only *ybgC* missing. The Tol-Pal system may be important in *Blochmannia* for host-endosymbiont

359   interactions.

360   **Phylogenetic framework reveals convergent gene losses**

361   We placed gene losses in a phylogenetic context to identify convergent losses (Fig. 5),

362   which we define as independent loss of the same gene in multiple lineages separated by a lineage

363   that retained the gene. By this definition, we identified 55 genes that were subject to convergent

364   loss in these *Blochmannia* lineages. An alternative explanation for the distribution patterns of

365   these 55 genes is acquisition by horizontal gene transfer. To test this explanation, we aligned each

366   of the 55 genes to the GenBank non-redundant database using BLAST and identified only two

367   genes (*icd* and *yqiC*) for which the top hits were to bacteria outside of the Enterobacteriaceae.

368   The top BLASTN hit for *icd* was to Candidatus *Profftella*, a beta-proteobacterial endosymbiont

369   of the Asian citrus psyllid. BLASTN did not return any significant hits for *yqiC*, but the top

370   BLASTX hit was to *Vibrio litoralis*. Although it is possible that these genes were acquired by

371   *Blochmannia* via horizontal gene transfer, for the purposes of our analysis we consider

372   convergent gene loss a more likely explanation for their distribution patterns.

373   Because of the branching order in the *Blochmannia* phylogeny (Fig. 1), the 15 genes

374   unique to *B. turneri* must have been independently lost in both *B. obliquus* and *Blochmannia* of

375   *Camponotus* (Fig. 5), assuming that gene content differences reflect gene loss rather than HGT

376   (see above). Four of these 15 genes are involved in DNA replication and repair, including the

377   chromosomal replication initiation protein *dnaA,* the DNA mismatch repair protein *mutS* and

378   DNA topoisomerases *topA* and *topB*. Three of the 15 genes (*acnB*, *gltA* and *icd*) encode enzymes

379   of the tricarboxylic acid (TCA) cycle and complete this pathway in *B. turneri*. By contrast, *B.*

380   *obliquus* and *Blochmannia* of *Camponotus* retained only TCA genes involved in energy

381   generation. In general, convergent loss of these 15 unique *B. turneri* genes in *B. obliquus* and

382   *Blochmannia* of *Camponotus* may indicate differing selective pressures on *B. turneri* and its ant

383   host.

384   Multiple genes involved in metabolic pathways were subject to convergent loss in the six

385   sequenced *Blochmannia*. Glutamine synthetase was independently lost in both *B. vafer* and *B.*

386   *obliquus*. In other *Blochmannia*, this enzyme may play an important role in nitrogen recycling for

387   the ant host (Feldhaar et al., 2007). Convergent loss of *glnA* suggests that this enzyme is not

388   essential to the symbiosis and its function can be fulfilled by an alternative pathway (Williams &

389   Wernegreen, 2010). A few convergent loss events involved genes from the same metabolic

390   pathway, such as *coaADE* and *dfp* (coenzyme A biosynthesis), *ispADFGH* (terpenoid backbone

391   biosynthesis) and *ribABCDH* (riboflavin biosynthesis). Loss of these genes may reflect relaxed

392   selection on components of a pathway after inactivation of a gene within that pathway.

393   In addition to genes involved in metabolism, genes encoding outer membrane and

394   secretory proteins, including *ompA, oprC, secBDF* and *tonB*, were lost independently in multiple

395   lineages, possibly affecting communication between endosymbiont and host cells. Finally,

396   convergent loss of five hypothetical proteins with unknown function emphasizes the need to

397   better characterize these proteins and understand their contributions to the symbiosis.

398   Maintenance of these hypothetical proteins in particular lineages suggests that they are under

399   selective pressure specific to host lineages.

400   **_Blochmannia_ lineages experienced multiple inversions**

401   Although gene order is conserved among the four sequenced *Blochmannia* of

402   *Camponotus*, comparisons including *B. obliquus* and *B. turneri* revealed inversions in eight

403   genomic regions involving between two and 34 genes (Figs. 3 and 6). Two of these regions

404   (labeled D and G in Figs. 3 and 6) show evidence of multiple separate inversion events (Table

405   S3), suggesting that they may be inversion "hotspots". To determine which lineage likely

406   experienced inversion events, we compared gene order in the eight regions to that of close

407   relatives strain HS (CP006569), which is a member of a *Sodalis*-allied clade, and *Baumannia*

408   *cicadellinicola* str. Hc (NC_007984). For some regions, these close relatives provide information

409   on the probable gene order of the ancestral *Blochmannia* lineage. Based on these comparisons,

410   we hypothesize that two inversions occurred in the lineage leading to *B. obliquus* (labeled A and

411   B), one occurred in the lineage leading to *B. turneri* (labeled F), one occurred in the ancestral

412   lineage of *B. turneri* and *Blochmannia* of *Camponotus* (labeled C), and one occurred in the

413   ancestral lineage of the four sequenced *Blochmannia* of *Camponotus* (labeled E). For the

414   remaining three regions, our comparisons were inconclusive.

415   **Discussion**

416   Comparative genomics of *Blochmannia* lineages spanning the origin of the ant tribe

417   Camponotini allows us to reconstruct early events in the evolution of this symbiosis. Previous

418   phylogenetic analyses identified a clade of secondary endosymbionts of mealybugs as the closest

419   relatives of *Blochmannia*, suggesting *Blochmannia* may have originated from an ancestor of this

420   group (Wernegreen et al., 2009). Reconstruction of the last common ancestor of the *Blochmannia*

421   analyzed here showed most genes in the LCA (82%) have been retained in all six *Blochmannia*

422   species sequenced to date. This is consistent with the hypothesis that the ancestor was an

423   endosymbiont with an already reduced genome. Alternatively, *Blochmannia* may have undergone

424   a similar trajectory as that proposed for *Blattabacterium*, which may have originated as a free-

425   living associate that experienced substantial and rapid gene loss after acquiring an endosymbiotic

426   lifestyle but before diverging into extant lineages (Patino-Navarrette et al., 2013). Sequencing the

427   genomes of *Blochmannia* of other deeply divergent lineages, such as *Opisthopsis*, and closely

428 related lineages outside of *Blochmannia* may provide data to distinguish these two hypotheses.

429   We found that gene content differences among *Blochmannia* lineages involve several key

430 functions, including information transfer, metabolism, and cell-cell communication, which may

431 affect the functioning of this mutualism across ant host lineages. Regarding metabolism, the

432 biosynthesis pathways for two vitamins, riboflavin and vitamin B6, vary among the sequenced

433 *Blochmannia* lineages. Plants and most bacteria encode a riboflavin biosynthesis pathway, but

434 animals lack this pathway. Many insect endosymbionts, such as *Baumannia, Blattabacterium*,

435 *Buchnera*, *Hamiltonella*, *Sodalis* and *Wigglesworthia*, synthesize riboflavin. Young aphids

436 depend on the supply of riboflavin from *Buchnera* for growth and development (Nakabachi &

437 Ishikawa, 1999). Our analyses showed that *B. obliquus* and *B. turneri* have both lost the ability to

438 synthesize riboflavin, whereas *Blochmannia* of *Camponotus* have retained this pathway and can

439 provision the ants with this vitamin. These differences in riboflavin biosynthetic functionality

440 illustrate how the nutritional roles of *Blochmannia* in this mutualism change during co-evolution

441 with hosts.

442   In other endosymbiont systems, loss of riboflavin biosynthesis is compensated by either a

443 secondary symbiont or transfer of the genes to the host nuclear genome. In the aphid *Cinara*

444 *cedri*, the primary endosymbiont *Buchnera* lacks riboflavin biosynthesis genes, but a more

445 recently integrated *Serratia symbiotica* associate retains this pathway (Perez-Brocal et al., 2006;

446 Lamelas et al., 2011). In the mealybug *Planococcus citri*, the symbiont *Moranella endobia*,

447 which is nested within cells of the endosymbiont *Tremblaya princeps*, encodes two riboflavin

448 biosynthesis genes. Two other genes are encoded by the host nuclear genome and appear to have

449 been transferred from facultative symbionts during past colonizations (Husnik et al., 2013). In

450 contrast to the above symbioses, our analysis of Illumina read datasets generated from genomic

451 DNA of *C. obliquus* ants and *P. turneri* gasters showed no evidence that other bacterial associates

452 or the ant host encode intact riboflavin biosynthesis genes. Rather, the ant hosts may acquire

453 riboflavin via their diet, or it is possible that members of the gut microbiome synthesize this

454 vitamin for the host. Gut bacteria are not represented at high coverage in our read datasets.

455   As with riboflavin, insects lack a pathway for biosynthesis of vitamin B6 (Tanaka, Tateno

456 & Gojobori, 2005) and thus rely on their diet or symbionts to supply this essential cofactor. Five

457 of the six sequenced *Blochmannia* encode the 'DXP dependent' pathway for vitamin B6

458 biosynthesis characteristic of *E. coli* and other gamma-proteobacteria (Mukherjee et al., 2011)

459 The *B. obliquus* lineage has lost all but two of the seven genes in this pathway. In the other

460 *Blochmannia* lineages, these genes are scattered along the genome rather than adjacent to each

461 other, suggesting loss due to relaxed selective pressure on the pathway instead of a large deletion

462 event affecting multiple genes. This explanation is supported by retention of *dxs* and *serC* in *B.*

463 *obliquus.* These genes are also involved in amino acid metabolism and terpenoid backbone

464 biosynthesis pathways, respectively, which are conserved in *Blochmannia*. The two genes are

465 likely under selective pressure due to their roles in these other metabolic pathways, which

466 prevented their loss in the *B. obliquus* lineage.

467   In addition to metabolism, *Blochmannia* lineages show differences in replication and

468 repair genes. Specifically, the chromosomal replication initiation protein *dnaA* was lost in all but

469 one of the sequenced *Blochmannia* lineages. Some obligate bacterial endosymbionts of insects,

470 such as *Buchnera* species, have retained *dnaA* despite severe genome reduction. By contrast, loss

471 of *dnaA* has occurred in a few insect endosymbionts, including *Baumannia, Blattabacterium*,

472 *Carsonella, Sulcia* and *Wigglesworthia* (Akman et al., 2002; Nakabachi et al., 2006; Wu et al.,

473 2006; Lopez-Sanchez et al., 2009). *DnaA* is also missing from a bacterial endosymbiont of

474   protists found in termite guts (Hongoh et al., 2008). Previously, it was hypothesized that loss of

475   *dnaA* was necessary for establishing a stable symbiosis between insects and bacterial

476   endosymbionts located in the cytosol, because it may enable direct control of symbiont DNA

477   replication by the host (Gil et al., 2003). The presence of intact *dnaA* in *B. turneri* and a *dnaA*

478   pseudogene in *B. obliquus* demonstrates that loss of *dnaA* was not required for establishment of a

479   stable symbiosis between *Blochmannia* and camponotines. However, the precise function of

480   DnaA in *B. turneri* and the mechanisms for controlling initiation of chromosome replication in

481   different *Blochmannia* lineages are unclear. In addition, it remains untested whether divergent

482   *Blochmannia* lineages live in the cytosol like *B. floridanus* or, alternatively, occupy host-derived

483   vacuoles.

484   The DNA repair genes *mutM* and *mutY* are also differentially distributed in the

485   *Blochmannia* lineages, with either one or the other retained. Loss of *mutY* appears to be more

486   common in intracellular bacteria (Garcia-Gonzalez, Rivera-Rivera & Massey, 2012).

487   Overexpression of *mutM* can "rescue" inactivation of *mutY* in *E. coli* (Michaels et al., 1992),

488   which implies that retention of *mutM* may be favored more strongly by selection. However, some

489   bacteria, including obligate intracellular *Rickettsia* species, lack both *mutM* and *mutY* (Garcia-

490   Gonzalez, Rivera-Rivera & Massey, 2012). Additional sequencing of *Blochmannia* from diverse

491   ant hosts may reveal the evolutionary trajectories leading to differential loss of *mutM* and *mutY* in

492   *Blochmannia*.

493   Obligate bacterial endosymbionts of insects are characterized by a high degree of genome

494   stability, with strictly conserved gene order observed among sequenced representatives of some

495   endosymbiont groups, such as *Carsonella* and *Sulcia*. Comparative genomics of other

496   endosymbiont genera, including *Blattabacterium* and *Buchnera*, have identified typically three or

497   fewer inversions. Recently, an exception to this extreme conservation of genome architecture was

498   described in *Portiera*, an endosymbiont of whiteflies (Sloan & Moran, 2013). Analysis of

499   *Portiera* genomes from divergent whitefly host genera predicted at least 17 inversion events, with

500   most occurring in one lineage that also had a high prevalence of tandem repeats.

501   Previously, the *Blochmannia* genome dataset sampled only species from *Camponotus*

502   hosts, and these four sequenced species shared strictly conserved gene order. By sequencing

503   *Blochmannia* from ant hosts on divergent branches of the tribe Camponotini, our analysis

504   revealed that inversions have occurred throughout the evolution of this endosymbiont group.

505   Comparisons with close relatives suggest that inversions are not limited to a particular

506   *Blochmannia* lineage, but rather may have occurred along all four major branches of the

507   phylogeny. We identified gene losses in some of the regions involved in inversion events; it is

508   possible that changes in mutational pressure arising from strand switch contributed to degradation

509   and eventual loss of these genes (Williams & Wernegreen, 2012).

510   By sequencing genomes from *Blochmannia* of divergent ant host lineages, we expanded

511   the available *Blochmannia* genome dataset beyond *Camponotus* hosts and reconstructed

512   evolutionary trajectories of *Blochmannia* that likely span the origin of the tribe Camponotini.

513   Analysis of deep branches in symbiont groups addresses questions surrounding the origin of

514   symbioses and the mechanisms involved in establishment of stable associations. Although

515   divergent *Blochmannia* genomes share much of their gene content, differential gene losses across

516   key functional categories are likely to impact the host-bacterial partnership. It remains

517   challenging to distinguish if different losses reflect selective fine-tuning across distinct ant hosts,

518   stochastic gene deletions or a combination of the two. However, our results, particularly the

519   numerous instances of convergent gene loss, hint that the strength or efficacy of selection to

520   maintain gene functions has varied across ant host lineages and contributed to observed genome

521   variation. These findings contribute to a broader understanding of processes shaping genome

522   reduction in insect endosymbionts and potentially in other bacteria, including intracellular

523   pathogens and free-living marine bacteria.

524 **Acknowledgments**

525 The authors gratefully acknowledge Simon K. A. Robson for providing specimens of *P. turneri*

526 and Benoit Guénard for providing *C. obliquus*. We thank Yongliang Fan for preparing genomic

527 DNA and Lisa Bukovnik, Olivier Fedrigo and Fangfei Ye for Illumina sequencing through the

528 core facility in the Center for Genomic and Computational Biology (GCB) at Duke University.

529 We are grateful to Julie Yi for assistance with genome gap closing and analysis of *tolA*. We thank

530 Michelle G. Giglio at the Institute for Genome Sciences in the University of Maryland School of

531 Medicine for assistance with the MANATEE annotation framework. We thank Philip S. Ward for

532 helpful discussion and for archiving voucher specimens at the Bohart Museum of Entomology,

533 University of California, Davis (UCDC).

534 **Funding**

535 This work was supported by grants to JJW from the National Institutes of Health

536 (R01GM062626) and the National Science Foundation (MCB-1103113).

537 **Figure Legends**

538 **Figure 1. *Blochmannia* phylogeny.** The phylogeny was constructed by Bayesian analysis of

539 concatenated amino acid sequence alignments for 30 genes. The phylogeny is artificially rooted

540 on the branch leading to *Hamiltonella.* All nodes have 100% posterior probability. The scale bar

541 shows amino acid substitutions.

542 **Figure 2. Gene content of three divergent *Blochmannia* lineages.** The total number of intact

543 genes, including genes that have frameshifts in homopolymer tracts, is shown in bold in each

544 section. *Camponotus Bloch LCA* includes any intact gene found in at least one of the four

545 sequenced *Blochmannia* of *Camponotus sensu stricto*. Note that *coaBC* is one gene. Although

546 *yidC* and *yidD* are separate genes in *B. turneri* and *B. floridanus*, these two genes are fused in the

547 other sequenced genomes and therefore counted as one gene here. Similarly, BOBLI757_064 and

548 BOBLI757_065 are counted as one gene because they encode the two domains of bifunctional

549 protein *hldE.*

550 **Figure 3. *Blochmannia* genome plots.** Plots of *B. obliquus* and *B. turneri* (sequenced in this

551 study) and *B. pennsylvanicus* (as a representative of *Blochmannia* of *Camponotus sensu stricto*)

552 were constructed with DNAPlotter. Position zero is set to the ATG start of *mnmG* for each

553 genome. Major and minor tick marks on the outer circle show 100 kbp and 20 kbp increments,

554 respectively. Tracks 1 and 2 show CDS in blue on the forward and reverse strands, respectively.

555 Track 3 shows pseudogenes in red. Track 4 shows the eight genomic regions that experienced

556 inversions in at least one of the lineages in grey. These regions are labeled A-H for consistency

557   with Figure 6. Track 5 shows GC skew calculated using 500 bp window size and 50 bp step size.

558   Green shading above the line indicates GC skew greater than the genome average, whereas

559   purple shading below the line indicates GC skew smaller than the genome average.

560   **Figure 4. Vitamin B6 synthesis encoded by six *Blochmannia* genomes.** The pathway for

561   synthesis of vitamin B6 is annotated with gene distribution in the six sequenced *Blochmannia*.

562   *Camp Bloch* refers to all four sequenced *Blochmannia* of *Camponotus sensu stricto*, which have

563   the same gene content for this pathway. A plus sign indicates an intact gene, a minus sign

564   indicates a missing gene, and a psi symbol indicates a pseudogene.

565   **Figure 5. Convergent gene losses.** The distribution of genes lost in multiple independent events

566   (i.e., convergent gene losses) are shown in a phylogenetic context. A plus sign indicates an intact

567   gene, which includes frameshifted genes, a psi symbol indicates a pseudogene and a minus sign

568   indicates that the gene is absent. The gene content of *B. chromaiodes* and *B. pennsylvanicus* is

569   identical; therefore, only *B. pennsylvanicus* is shown here. Acquisition via horizontal gene

570   transfer could, in principle, explain some observed patterns; however, such transfer is unlikely in

571   *Blochmannia*, and we found no evidence for HGT of genes listed here (see text).

572   **Figure 6. Inversions in divergent *Blochmannia* lineages.** Whole genome alignment of *B.*

573   *pennsylvanicus* (as a representative of *Blochmannia* of *Camponotus sensu stricto*), *B. turneri* and

574   *B. obliquus* with progressive MAUVE shows inversions as colored blocks below the midline for

575   each genome. The eight genomic regions that experienced inversions are labeled A-H with the

576   number of genes involved in each also shown.

577   **Supplemental Material**

578   **Supplemental Table S1. Pseudogene statistics.**

579   To examine gene degradation, we used TBLASTN to align *Blochmannia* pseudogenes to protein

580   sequences of intact genes from the most closely related lineage.

581   **Supplemental Table S2. BLASTN analysis of contigs from *de novo* assemblies.**

582   **Supplemental Table S3. Ortholog table.**

583   "pan-Camp" includes genes found in at least one of the four *Blochmannia* of *Camponotus sensu*

584   *stricto*. Genomic regions where inversion events occurred in at least one lineage are highlighted

585   in grey. For regions where comparisons to close relatives indicate which lineage likely

586   experienced the inversion event, only that lineage's genes are highlighted. When these

587   comparisons are inconclusive, all three lineages are highlighted. Note that BOBLI757_064 and

588   BOBLI757_065 are domain 1 and domain 2, respectively, of bifunctional protein HldE.

589 **References**

590 Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide

591   sequences guided by amino acid translations. *Nucleic Acids Research* 38:W7-13.

592 Abbas CA, Sibirny AA. 2011. Genetic control of biosynthesis and transport of riboflavin and

593   flavin nucleotides and construction of robust biotechnological producers. *Microbiology*

594   *and Molecular Biology Reviews* 75:321-360.

595 Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M, Aksoy S. 2002. Genome

596   sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*.

597   *Nature Genetics* 32:402-407.

598 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool.

599   *Journal of Molecular Biology* 215:403-410.

600 Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids*

601   *Research* 27:573-580.

602 Brady SG, Schultz TR, Fisher BL, Ward PS. 2006. Evaluating alternative hypotheses for the

603   early evolution and diversification of ants. *Proceedings of the National Academy of*

604   *Sciences of the United States of America* 103:18172-18177.

605 Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP,

606   Bateman A. 2013. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research*

607   41:D226-232.

608 Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. 2009. DNAPlotter: circular and linear

609   interactive genome visualization. *Bioinformatics* 25:119-120.

610 Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: At-a-glance quality assessment of Illumina

611   second-generation sequencing data. *BMC Bioinformatics* 11:485.

612   Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene

613   gain, loss and rearrangement. *PLoS One* 5:e11147.

614   DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del

615   Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY,

616   Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation

617   discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*

618   43:491-498.

619   Douglas AE. 2015. Multiorganismal insects: Diversity and function of resident microorganisms.

620   *Annual Review of Entomology* 60:2.1-2.18.

621   Feldhaar H, Straka J, Krischke M, Berthold K, Stoll S, Mueller MJ, Gross R. 2007. Nutritional

622   upgrading for omnivorous carpenter ants by the endosymbiont *Blochmannia*. *BMC*

623   *Biology* 5:48.

624   Galens K, Orvis J, Daugherty S, Creasy HH, Angiuoli S, White O, Wortman J, Mahurkar A,

625   Giglio MG. 2011. The IGS standard operating procedure for automated prokaryotic

626   annotation. *Standards in Genomic Sciences* 4:244-251.

627   Garcia-Gonzalez A, Rivera-Rivera RJ, Massey SE. 2012. The presence of the DNA repair genes

628   *mutM*, *mutY*, *mutL*, and *mutS* is related to proteome size in bacterial genomes. *Frontiers*

629   *in Genetics* 3:3.

630   Gil R, Silva FJ, Zientz E, Delmotte F, Gonzalez-Candelas F, Latorre A, Rausell C, Kamerbeek J,

631   Gadau J, Holldobler B, van Ham RC, Gross R, Moya A. 2003. The genome sequence of

632   *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proceedings of the*

633   *National Academy of Sciences of the United States of America* 100:9388-9393.

634   Godlewska R, Wisniewska K, Pietras Z, Jagusztyn-Krynicka EK. 2009. Peptidoglycan-associated

635 lipoprotein (Pal) of Gram-negative bacteria: function, structure, role in pathogenesis and

636 potential application in immunoprophylaxis. *FEMS Microbiology Letters* 298:1-11.

637 Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. *Genome*

638 *Research* 8:195-202.

639 Hansen FG, Christensen BB, Nielsen CB, Atlung T. 2006. Insights into the quality of DnaA

640 boxes and their cooperativity. *Journal of Molecular Biology* 355:85-95.

641 Herbeck JT, Degnan PH, Wernegreen JJ. 2005. Nonhomogeneous model of sequence evolution

642 indicates independent origins of primary endosymbionts within the Enterobacteriales

643 (gamma-proteobacteria). *Molecular Biology and Evolution* 22:520-532.

644 Hongoh Y, Sharma VK, Prakash T, Noda S, Taylor TD, Kudo T, Sakaki Y, Toyoda A, Hattori

645 M, Ohkuma M. 2008. Complete genome of the uncultured Termite Group 1 bacteria in a

646 single host protist cell. *Proceedings of the National Academy of Sciences of the United*

647 *States of America* 105:5555-5560.

648 Husnik F, Chrudimsky T, Hypsa V. 2011. Multiple origins of endosymbiosis within the

649 Enterobacteriaceae (gamma-proteobacteria): convergence of complex phylogenetic

650 approaches. *BMC Biology* 9:87.

651 Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, Tanaka M, Satoh N, Bachtrog D,

652 Wilson AC, von Dohlen CD, Fukatsu T, McCutcheon JP. 2013. Horizontal gene transfer

653 from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis.

654 *Cell* 153:1567-1578.

655 Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of

656 multiple sequence alignment. *Nucleic Acids Research* 33:511-518.

657 Kikuchi Y. 2009. Endosymbiotic bacteria in insects: their diversity and culturability. *Microbes*

658   *and Environments* 24:195-204.

659   Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson

660   RK, Ding L. 2009. VarScan: variant detection in massively parallel sequencing of

661   individual and pooled samples. *Bioinformatics* 25:2283-2285.

662   Lamelas A, Gosalbes MJ, Manzano-Marin A, Pereto J, Moya A, Latorre A. 2011. *Serratia*

663   *symbiotica* from the aphid *Cinara cedri*: a missing link from facultative to obligate insect

664   endosymbiont. *PLoS Genetics* 7:e1002357.

665   Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. 2014. MOSAIK: a hash-

666   based algorithm for accurate next-generation sequencing short-read mapping. *PLoS One*

667   9:e90581.

668   Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.

669   *Bioinformatics* 25:1754-1760.

670   Lind PA, Andersson DI. 2008. Whole-genome mutational biases in bacteria. *Proceedings of the*

671   *National Academy of Sciences of the United States of America* 105:17878-17883.

672   Lopez-Sanchez MJ, Neef A, Pereto J, Patino-Navarrete R, Pignatelli M, Latorre A, Moya A.

673   2009. Evolutionary convergence and nitrogen metabolism in *Blattabacterium* strain Bge,

674   primary endosymbiont of the cockroach *Blattella germanica*. *PLoS Genetics* 5:e1000721.

675   Mackiewicz P, Zakrzewska-Czerwinska J, Zawilak A, Dudek MR, Cebrat S. 2004. Where does

676   bacterial replication start? Rules for predicting the oriC region. *Nucleic Acids Research*

677   32:3781-3791.

678   Michaels ML, Cruz C, Grollman AP, Miller JH. 1992. Evidence that MutY and MutM combine

679   to prevent mutations by an oxidatively damaged form of guanine in DNA. *Proceedings of*

680   *the National Academy of Sciences of the United States of America* 89:7022-7025.

681   Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and evolution of heritable bacterial

682   symbionts. *Annual Review of Genetics* 42:165-190.

683   Moreau CS, Bell CD. 2013. Testing the museum versus cradle tropical biological diversity

684   hypothesis: phylogeny, diversification, and ancestral biogeographic range evolution of the

685   ants. *Evolution* 67:2240-2257.

686   Moya A, Pereto J, Gil R, Latorre A. 2008. Learning how to live together: genomic insights into

687   prokaryote-animal symbioses. *Nature Reviews: Genetics* 9:218-229.

688   Mrazek J, Xie S. 2006. Pattern locator: a new tool for finding local sequence patterns in genomic

689   DNA sequences. *Bioinformatics* 22:3099-3100.

690   Mukherjee T, Hanes J, Tews I, Ealick SE, Begley TP. 2011. Pyridoxal phosphate: biosynthesis

691   and catabolism. *Biochimica et Biophysica Acta: Protein Structure and Molecular*

692   *Enzymology* 1814:1585-1596.

693   Nakabachi A, Ishikawa H. 1999. Provision of riboflavin to the host aphid, *Acyrthosiphon pisum*,

694   by endosymbiotic bacteria, *Buchnera*. *Journal of Insect Physiology* 45:1-6.

695   Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M. 2006. The

696   160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* 314:267.

697   Patino-Navarrete R, Moya A, Latorre A, Pereto J. 2013. Comparative genomics of

698   *Blattabacterium cuenoti*: the frozen legacy of an ancient endosymbiont genome. *Genome*

699   *Biology and Evolution* 5:351-361.

700   Perez-Brocal V, Gil R, Ramos S, Lamelas A, Postigo M, Michelena JM, Silva FJ, Moya A,

701   Latorre A. 2006. A small microbial genome: the end of a long symbiotic relationship?

702   *Science* 314:312-313.

703   Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed

704   models. *Bioinformatics* 19:1572-1574.

705 Russell JA, Funaro CF, Giraldo YM, Goldman-Huertas B, Suh D, Kronauer DJ, Moreau CS,

706   Pierce NE. 2012. A veritable menagerie of heritable bacteria from ants, butterflies, and

707   beyond: broad molecular surveys and a systematic review. *PLoS One* 7:e51027.

708 Sabree ZL, Kambhampati S, Moran NA. 2009. Nitrogen recycling and nutritional provisioning

709   by *Blattabacterium*, the cockroach endosymbiont. *Proceedings of the National Academy*

710   *of Sciences of the United States of America* 106:19521-19526.

711 Sameshima S, Hasegawa E, Kitade O, Minaka N, Matsumoto T. 1999. Phylogenetic comparison

712   of endosymbionts with their host ants based on molecular evidence. *Zoological Science*

713   16:993-1000.

714 Sauer C, Stackebrandt E, Gadau J, Holldobler B, Gross R. 2000. Systematic relationships and

715   cospeciation of bacterial endosymbionts and their carpenter ant host species: proposal of

716   the new taxon Candidatus *Blochmannia* gen. nov. *International Journal of Systematic and*

717   *Evolutionary Microbiology* 50 Pt 5:1877-1886.

718 Schaper S, Messer W. 1995. Interaction of the initiator protein DnaA of *Escherichia coli* with Its

719   DNA target. *Journal of Biological Chemistry* 270:17622-17626.

720 Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. 2000. Genome sequence of the

721   endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* 407:81-86.

722 Sloan DB, Moran NA. 2013. The evolution of genomic instability in the obligate endosymbionts

723   of whiteflies. *Genome Biology and Evolution* 5:783-793.

724 Stoll S, Feldhaar H, Fraunholz MJ, Gross R. 2010. Bacteriocyte dynamics during development of

725   a holometabolous insect, the carpenter ant *Camponotus floridanus*. *BMC Microbiology*

726   10:308.

727 Tamas I, Wernegreen JJ, Nystedt B, Kauppinen SN, Darby AC, Gomez-Valero L, Lundin D,

728 Poole AM, Andersson SG. 2008. Endosymbiont gene functions impaired and rescued by

729 polymerase infidelity at poly(A) tracts. *Proceedings of the National Academy of Sciences*

730 *of the United States of America* 105:14934-14939.

731 Tanaka T, Tateno Y, Gojobori T. 2005. Evolution of vitamin B6 (pyridoxine) metabolism by gain

732 and loss of genes. *Molecular Biology and Evolution* 22:243-250.

733 Wall DP, Fraser HB, Hirsh AE. 2003. Detecting putative orthologs. *Bioinformatics* 19:1710-

734 1711.

735 Wernegreen JJ, Kauppinen SN, Brady SG, Ward PS. 2009. One nutritional symbiosis begat

736 another: phylogenetic evidence that the ant tribe Camponotini acquired *Blochmannia* by

737 tending sap-feeding insects. *BMC Evolutionary Biology* 9:292.

738 Wernegreen JJ, Kauppinen SN, Degnan PH. 2010. Slip into something more functional: selection

739 maintains ancient frameshifts in homopolymeric sequences. *Molecular Biology and*

740 *Evolution* 27:833-839.

741 Werren JH, Baldo L, Clark ME. 2008. *Wolbachia*: master manipulators of invertebrate biology.

742 *Nature Reviews: Microbiology* 6:741-751.

743 Williams LE, Wernegreen JJ. 2010. Unprecedented loss of ammonia assimilation capability in a

744 urease-encoding bacterial mutualist. *BMC Genomics* 11:687.

745 Williams LE, Wernegreen JJ. 2012. Purifying selection, sequence composition, and context-

746 specific indel mutations shape intraspecific variation in a bacterial endosymbiont.

747 *Genome Biology and Evolution* 4:44-51.

748 Williams LE, Wernegreen JJ. 2013. Sequence context of indel mutations and their effect on

749 protein evolution in a bacterial endosymbiont. *Genome Biology and Evolution* 5:599-605.

750   Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL, Khouri H, Tallon LJ, Zaborsky JM,

751   Dunbar HE, Tran PL, Moran NA, Eisen JA. 2006. Metabolic complementarity and

752   genomics of the dual bacterial symbiosis of sharpshooters. *PLoS Biology* 4:e188.

753   Wu M, Chatterji S, Eisen JA. 2012. Accounting for alignment uncertainty in phylogenomics.

754   *PLoS One* 7:e30288.

755   Zakrzewska-Czerwinska J, Jakimowicz D, Zawilak-Pawlik A, Messer W. 2007. Regulation of the

756   initiation of chromosomal replication in bacteria. *FEMS Microbiology Reviews* 31:378-

757   387.

758   Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn

759   graphs. *Genome Research* 18:821-829.

760   Zhou K, Vanoirbeek K, Aertsen A, Michiels CW. 2012. Variability of the tandem repeat region

761   of the *Escherichia coli* tolA gene. *Research in Microbiology* 163:316-322.

762   Zientz E, Beyaert I, Gross R, Feldhaar H. 2006. Relevance of the endosymbiosis of *Blochmannia*

763   *floridanus* and carpenter ants at different stages of the life cycle of the host. *Applied and*

764   *Environmental Microbiology* 72:6027-6033.

Genome statistics of six sequenced *Blochmannia*.

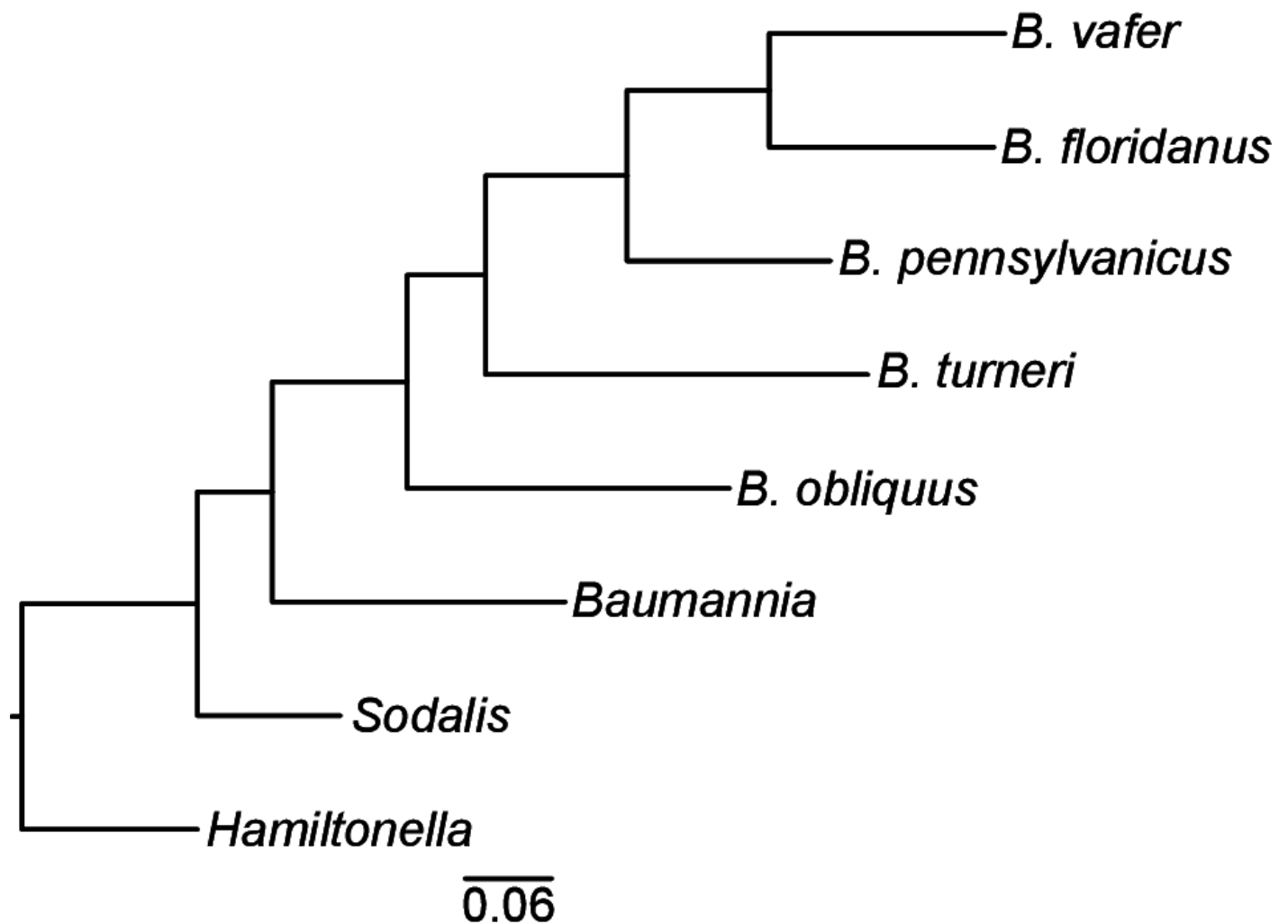
Table 1. Genome statistics of six sequenced *Blochmannia*.

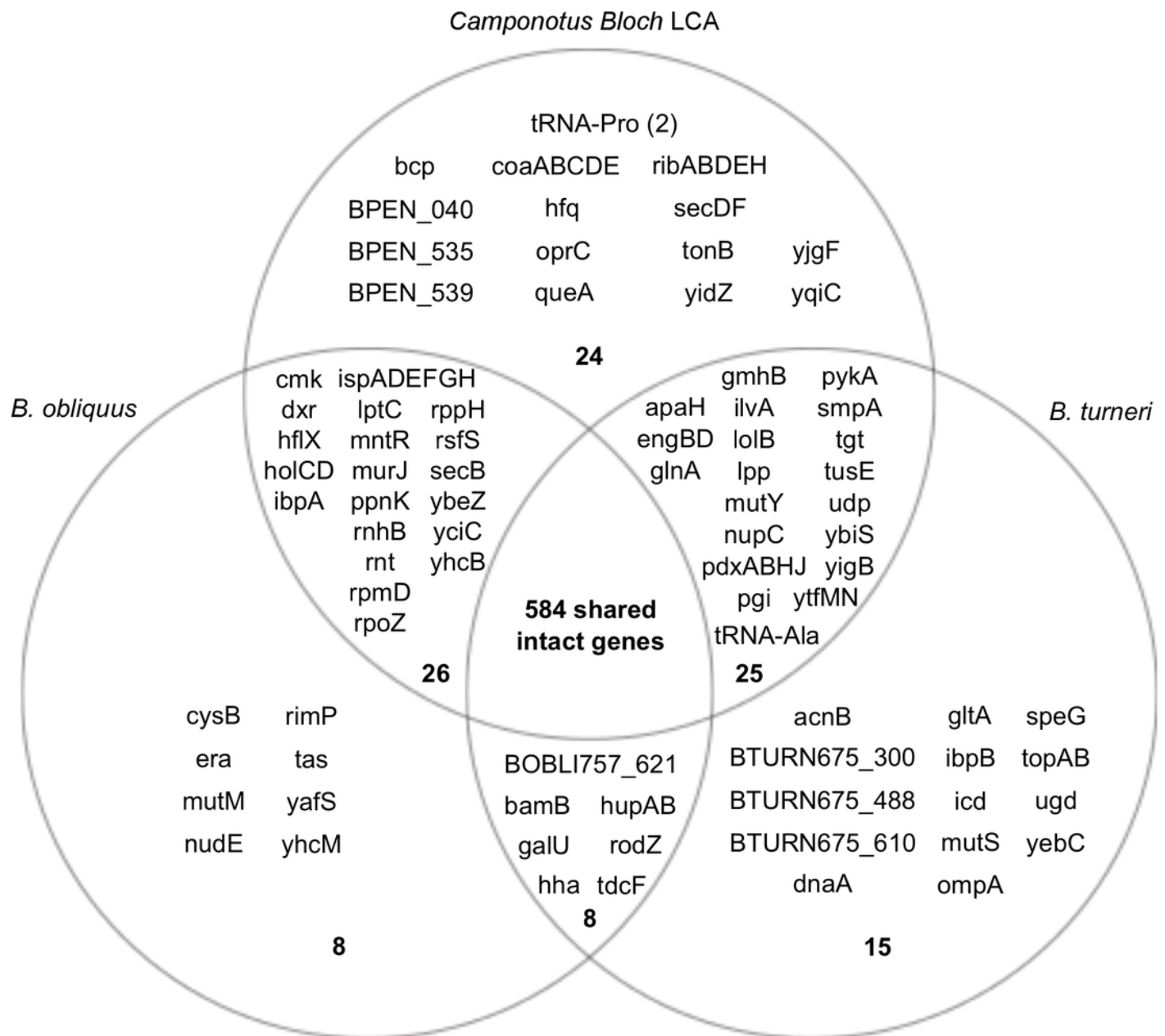
Genome	Size	GC content (%)	Genes (inc. pseudo)	Protein coding	tRNA	rRNA	other RNA	pseudo genes	frameshifted genes
B. obliquus	773,940	27.4	642	584	37	3	3	15	4
B. turneri	749,321	29.1	634	589	38	3	3	1	7
B. chromaiodes	791,219	29.5	658	609	40	3	3	3	4
B. pennsylvanicus	791,654	29.6	658	609	40	3	3	3	4
B. floridanus	705,557	27.4	637	590	37	3	3	4	4
B. vafer	722,585	27.5	631	587	37	3	2	2	8

1

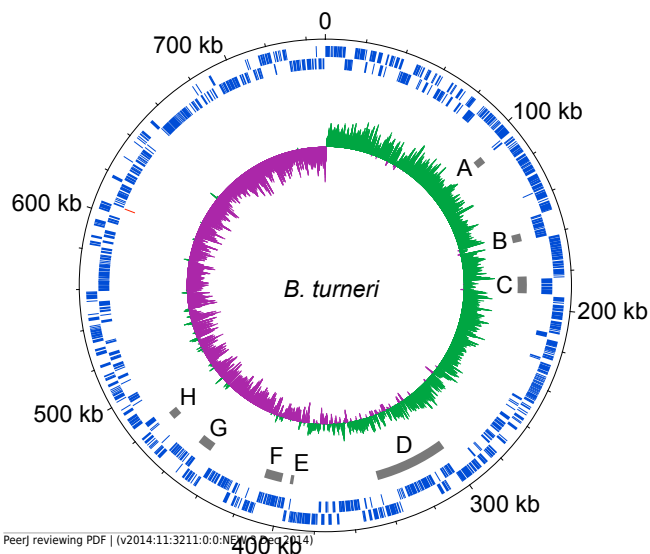
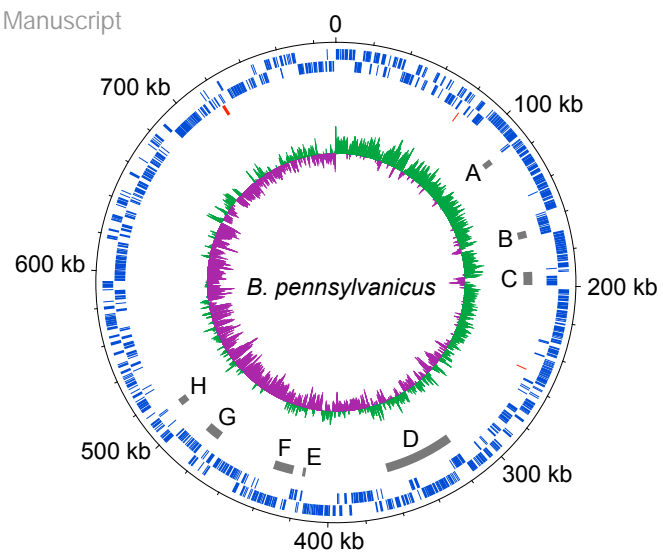
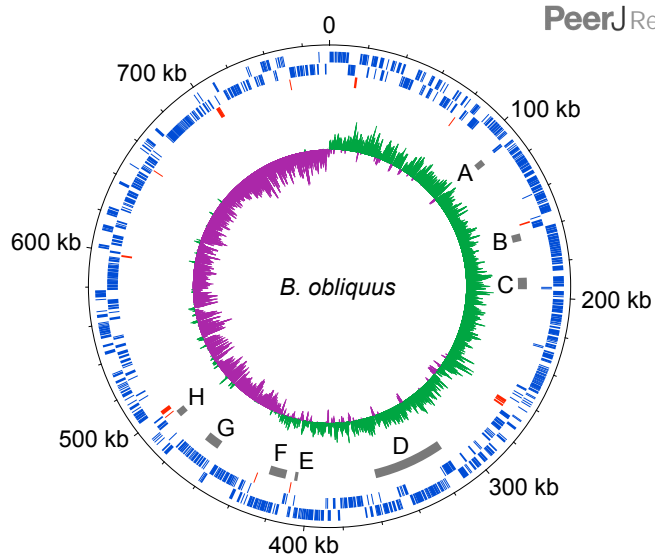
Blochmannia phylogeny.

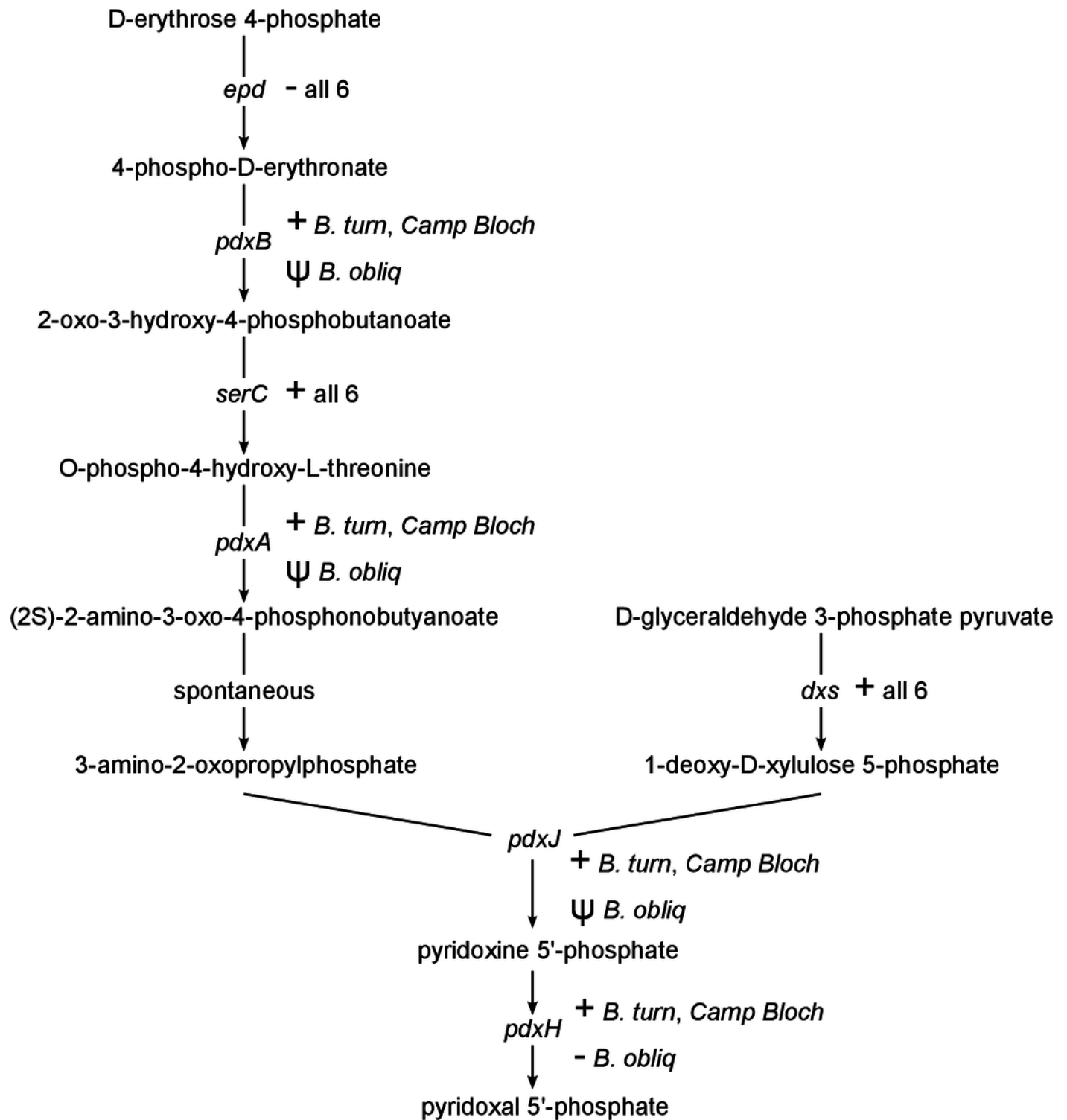
The phylogeny was constructed by Bayesian analysis of concatenated amino acid sequence alignments for 30 genes. The phylogeny is artificially rooted on the branch leading to *Hamiltonella.* All nodes have 100% posterior probability. The scale bar shows amino acid substitutions.

2

Gene content of three divergent *Blochmannia* lineages.

The total number of intact genes, including genes that have frameshifts in homopolymer tracts, is shown in bold in each section. *Camponotus Bloch LCA* includes any intact gene found in at least one of the four sequenced *Blochmannia* of *Camponotus sensu stricto*. Note that *coaBC* is one gene. Although *yidC* and *yidD* are separate genes in *B. turneri* and *B. floridanus*, these two genes are fused in the other sequenced genomes and therefore counted as one gene here. Similarly, BOBLI757_064 and BOBLI757_065 are counted as one gene because they encode the two domains of bifunctional protein *hldE*.

Blochmannia genome plots.

Plots of *B. obliquus* and *B. turneri* (sequenced in this study) and *B. pennsylvanicus* (as a representative of *Blochmannia* of *Camponotus sensu stricto*) were constructed with DNAPlotter. Position zero is set to the ATG start of *mnmG* for each genome. Major and minor tick marks on the outer circle show 100 kbp and 20 kbp increments, respectively. Tracks 1 and 2 show CDS in blue on the forward and reverse strands, respectively. Track 3 shows pseudogenes in red. Track 4 shows the eight genomic regions that experienced inversions in at least one of the lineages in grey. These regions are labeled A-H for consistency with Figure 6. Track 5 shows GC skew calculated using 500 bp window size and 50 bp step size. Green shading above the line indicates GC skew greater than the genome average, whereas purple shading below the line indicates GC skew smaller than the genome average.

4

Vitamin B6 synthesis encoded by six *Blochmannia* genomes.

The pathway for synthesis of vitamin B6 is annotated with gene distribution in the six sequenced *Blochmannia*. *Camp Bloch* refers to all four sequenced *Blochmannia* of *Camponotus sensu stricto*, which have the same gene content for this pathway. A plus sign indicates an intact gene, a minus sign indicates a missing gene, and a psi symbol indicates a pseudogene.

Tandem repeats within *tolA* of *Blochmannia* and *E. coli*.

Data from Tandem Repeats Finder (Benson, 1999).* Four repeat patterns were identified. The top-scoring pattern is reported here.

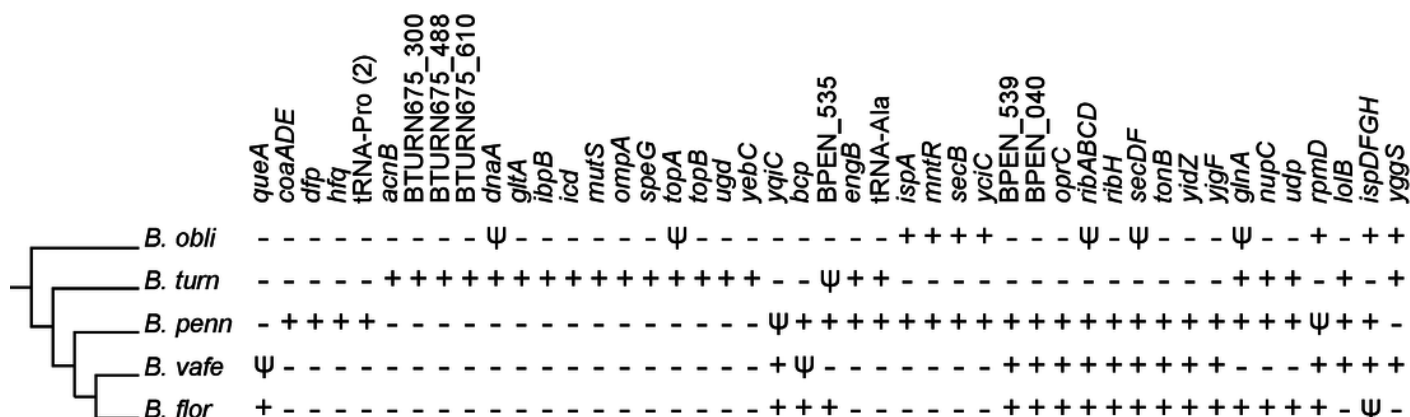
Table 2. Tandem repeats within *tolA* of *Blochmannia* and *E. coli*.

Genome	Length of *tolA* (bp)	Length of consensus repeat pattern (bp)	# of repeat copies	Alignment score	Coordinates of repeat region
B. obliquus	771	17	2.1	52	81-114
B. turneri	990	81	4.4	699	205-558
B. chromaiodes	1185	87	6.0	1030	199-722
B. pennsylvanicus	1185	87	6.0	1046	200-722
B. floridanus	1305	99	6.5	1260	215-853
B. vafer	1173	102	5.1	1038	208-726
E. coli MG1655*	1266	79	2.0	239	671-829

Data from Tandem Repeats Finder (Benson, 1999)
*Four repeat patterns were identified. The top-scoring pattern is reported here.

5

Convergent gene losses.

The distribution of genes lost in multiple independent events (i.e., convergent gene losses) are shown in a phylogenetic context. A plus sign indicates an intact gene, which includes frameshifted genes, a psi symbol indicates a pseudogene and a minus sign indicates that the gene is absent. The gene content of *B. chromaiodes* and *B. pennsylvanicus* is identical; therefore, only *B. pennsylvanicus* is shown here. Acquisition via horizontal gene transfer could, in principle, explain some observed patterns; however, such transfer is unlikely in *Blochmannia*, and we found no evidence for HGT of genes listed here (see text).

Inversions in divergent *Blochmannia* lineages.

Whole genome alignment of *B. pennsylvanicus* (as a representative of *Blochmannia* of *Camponotus sensu stricto*), *B. turneri* and *B. obliquus* with progressive MAUVE shows inversions as colored blocks below the midline for each genome. The eight genomic regions that experienced inversions are labeled A-H with the number of genes involved in each also shown.