Genome sequence of *Blochmannia pennsylvanicus* indicates parallel evolutionary trends among bacterial mutualists of insects

Patrick H. Degnan, Adam B. Lazarus and Jennifer J. Wernegreen

*Genome Res.* 2005 15: 1023-1033
Access the most recent version at doi:10.1101/gr.3771305
Genome sequence of *Blochmannia pennsylvanicus* indicates parallel evolutionary trends among bacterial mutualists of insects

Patrick H. Degnan, Adam B. Lazarus, and Jennifer J. Wernegreen

Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts 02543, USA

The distinct lifestyle of obligately intracellular bacteria can alter fundamental forces that drive and constrain genome change. In this study, sequencing the 792-kb genome of *Blochmannia pennsylvanicus*, an obligate endosymbiont of *Camponotus pennsylvanicus*, enabled us to trace evolutionary changes that occurred in the context of a bacterial-ant association. Comparison to the genome of *Blochmannia floridanus* reveals differential loss of genes involved in cofactor biosynthesis, the composition and structure of the cell wall and membrane, gene regulation, and DNA replication. However, the two *Blochmannia* species show complete conservation in the order and strand orientation of shared genes. This finding of extreme stasis in genome architecture, also reported previously for the aphid endosymbiont *Buchnera*, suggests that genome stability characterizes long-term bacterial mutualists of insects and constrains their evolutionary potential. Genome-wide analyses of protein divergences reveal 10- to 50-fold faster amino acid substitution rates in *Blochmannia* compared to related bacteria. Despite these varying features of genome evolution, a striking correlation in the relative divergences of proteins indicates parallel functional constraints on gene functions across ecologically distinct bacterial groups. Furthermore, the increased rates of amino acid substitution and gene loss in *Blochmannia* have occurred in a lineage-specific fashion, which may reflect life history differences of their ant hosts.

[Supplemental material is available online at www.genome.org. The complete, annotated genome sequence has been submitted to GenBank under accession no. CP000016.](http://www.ncbi.nlm.nih.gov/Genbank/)

Genome sequencing provides a rich data set to predict the metabolic capabilities of organisms, and comparative analyses among closely related species offer a powerful approach to examine mechanisms of genome flux. Since 2000, genomics has shed light on the metabolism and evolution of obligately intracellular, mutualistic bacteria that have coevolved with various insect groups for tens to hundreds of millions of years (Baumann et al. 2000). Fully sequenced genomes of *Buchnera* associated with aphids (Shigenobu et al. 2000; Tamas et al. 2002; van Ham et al. 2003), *Wigglesworthia* of bettwe flies (Akman et al. 2002), and *Blochmannia* associated with *Camponotus* (Gil et al. 2003) are extremely streamlined yet retain basic cellular processes and specific biosynthetic abilities required by the insect host. Genome comparisons within *Buchnera* have shown stability, with no gene acquisition, inversions, or translocations throughout 50–70 million years (Myr) of evolution within aphids (Tamas et al. 2002) and near-perfect synteny since the establishment of this association 150–200 million years ago (Mya) (van Ham et al. 2003). This exceptional stasis of genome architecture contrasts with lability of free-living and parasitic bacterial genomes. Genome stability may reflect the dearth of molecular tools for gene exchange (e.g., phage, certain rec genes, and repeated DNA sequences) in this mutualist and limited ecological opportunities to recombine with genetically distinct bacteria (Tamas et al. 2002; van Ham et al. 2003; Moran and Plague 2004). Such constraints on genome changes in stable mutualists may profoundly affect the evolutionary potential of these bacteria and their hosts. However, owing to the lack of multiple sequenced genomes within endosymbiont groups, genome stability in other long-term endosymbionts has remained untested.

In order to contribute to a more comprehensive model of genome evolution in ancient endosymbiotic associations, we have evaluated genome dynamics in *Blochmannia*, a bacterial genus that is closely related to *Buchnera* and has cospeciated with ants for ~30 Myr. The wide range of interactions between ants and other species, including plants, fungi, trophobionts, other insects, and diverse bacteria (Dasch et al. 1984; Currie 2001; Zientz et al. 2001), may explain the huge ecological success of ants, which play dominant roles in nutrient turnover in terrestrial ecosystems and include more than twice as many species as mammals (Hölldobler and Wilson 1990). *Blochmannia* is the most evolutionarily stable ant associate and lives exclusively within cells of the closely related genera *Polyrhachis*, *Colobopsis*, and *Camponotus* (Dasch et al. 1984; Schröder et al. 1996; Sameshima et al. 1999; Sauer et al. 2000; Degnan et al. 2004). *Blochmannia* has been studied most extensively in *Camponotus*, the second largest ant genus, with ~1000 species (Bolton 1995) ranging from omnivores to specialists on plant secretions and homopteran exudates (Dasch 1975; Hölldobler and Wilson 1990; Bolton 1995; Davidson 1997, 1998) and with nesting habitats including wood, soil beneath rocks, and the rainforest canopy (Bolton 1995).

The 706-kb sequence of *B. floridanus* (*Blochmannia* of host *Camponotus floridanus*) indicated this ant endosymbiont retains...
Degnan et al.

numerous metabolic pathways that may be involved in host nutrition, including nitrogen recycling and assimilation, biosynthesis of amino acids and fatty acids, and sulfate reduction (Gil et al. 2003). *Blochmannia* is thought to be important during host development (Sauer et al. 2002; Wolschin et al. 2004), but its specific roles in host physiology and ecology and its functional variation across ant species remain unclear. We have sequenced the 792-kb genome of *Blochmannia pennsylvanica* (*Blochmannia* associated with *Camponotus pennsylvanicus*) to examine genome changes since this lineage and *B. floridanus* diverged from a common ancestor ~16–20 Mya (Degnan et al. 2004). Comparing genome inventories and architectures of the two *Blochmannia* strains allowed us to trace functional and structural changes that occurred in the context of this bacterium–ant interaction and to contrast genome dynamics and protein evolution in two mutualist groups (*Blochmannia* and *Buchnera*).

Results

Genome features

The *B. pennsylvanica* genome consists of a 791,654-bp circular chromosome that we sequenced to 12 × coverage (Table 1; Fig. 1; Supplemental Fig. S1; GenBank entry CP000016). This AT-rich genome has a relatively low percentage of coding DNA (76.7%) and shows a GC skew that distinguishes the leading and the lagging strands of replication [(G – C)/(G+C) averaged across 2.5-kb sliding windows]. We identified predicted ribosomal binding sites (RBSSs) for 479 of 610 open reading frames (ORFs), but only lagging strands of replication [(G – C)/(G+C)] had detectable RBSSs.

Table 1. Comparison of general genome features among obligate insect mutualists

<table>
<thead>
<tr>
<th></th>
<th><em>B. pennsylvanica</em></th>
<th><em>B. floridanus</em></th>
<th><em>W. glossinidia</em></th>
<th><em>Buchnera</em>-BP</th>
<th><em>Buchnera</em>-SG</th>
<th><em>Buchnera</em>-APS</th>
<th><em>E. coli</em> K12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome, bp</td>
<td>791,654</td>
<td>705,557</td>
<td>697,724</td>
<td>615,980</td>
<td>641,454</td>
<td>640,681</td>
<td>4,639,221</td>
</tr>
<tr>
<td>Plasmids</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>G+C content total, %</td>
<td>29.6</td>
<td>27.38</td>
<td>22</td>
<td>25.3</td>
<td>26.3</td>
<td>26.2</td>
<td>50.8</td>
</tr>
<tr>
<td>Total gene number</td>
<td>658</td>
<td>636</td>
<td>677</td>
<td>553</td>
<td>629</td>
<td>621</td>
<td>4550</td>
</tr>
<tr>
<td>ORFs</td>
<td>610</td>
<td>590</td>
<td>621 (6)</td>
<td>507 (3)</td>
<td>554 (9)</td>
<td>571 (9)</td>
<td>4284</td>
</tr>
<tr>
<td>rRNAs</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>tRNAs</td>
<td>39</td>
<td>37</td>
<td>34</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>86</td>
</tr>
<tr>
<td>tRNAse</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>4</td>
<td>4</td>
<td>14</td>
<td>9</td>
<td>38</td>
<td>13</td>
<td>150</td>
</tr>
<tr>
<td>Chromosomal protein-coding regions, %</td>
<td>76.7</td>
<td>83.8</td>
<td>86.9</td>
<td>80.9</td>
<td>83.1</td>
<td>86.7</td>
<td>87.8</td>
</tr>
<tr>
<td>Average length CDS, bp</td>
<td>995</td>
<td>1,002</td>
<td>983</td>
<td>996</td>
<td>979</td>
<td>984</td>
<td>950</td>
</tr>
</tbody>
</table>

Total open reading frames (ORFs) include predicted genes located on plasmids (noted in parentheses). The repeated trpEG genes on *Buchnera*-APS and -SG plasmids were counted once. Minor discrepancies in gene numbers (ORFs and pseudogenes) from the original publications reflect our efforts to present the most current annotation data from the literature, available at GenBank (http://www.ncbi.nlm.nih.gov) and the Comprehensive Microbial Resource (CMR) database (http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl). Fusion of yidC in *B. pennsylvanica* counted as a single ORF.

Table 1. Comparison of general genome features among obligate insect mutualists

<table>
<thead>
<tr>
<th>Chromosome, bp</th>
<th><em>B. pennsylvanica</em></th>
<th><em>B. floridanus</em></th>
<th><em>W. glossinidia</em></th>
<th><em>Buchnera</em>-BP</th>
<th><em>Buchnera</em>-SG</th>
<th><em>Buchnera</em>-APS</th>
<th><em>E. coli</em> K12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plasmids</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>G+C content total, %</td>
<td>29.6</td>
<td>27.38</td>
<td>22</td>
<td>25.3</td>
<td>26.3</td>
<td>26.2</td>
<td>50.8</td>
</tr>
<tr>
<td>Total gene number</td>
<td>658</td>
<td>636</td>
<td>677</td>
<td>553</td>
<td>629</td>
<td>621</td>
<td>4550</td>
</tr>
<tr>
<td>ORFs</td>
<td>610</td>
<td>590</td>
<td>621 (6)</td>
<td>507 (3)</td>
<td>554 (9)</td>
<td>571 (9)</td>
<td>4284</td>
</tr>
<tr>
<td>rRNAs</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>tRNAs</td>
<td>39</td>
<td>37</td>
<td>34</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>86</td>
</tr>
<tr>
<td>tRNAse</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>4</td>
<td>4</td>
<td>14</td>
<td>9</td>
<td>38</td>
<td>13</td>
<td>150</td>
</tr>
<tr>
<td>Chromosomal protein-coding regions, %</td>
<td>76.7</td>
<td>83.8</td>
<td>86.9</td>
<td>80.9</td>
<td>83.1</td>
<td>86.7</td>
<td>87.8</td>
</tr>
<tr>
<td>Average length CDS, bp</td>
<td>995</td>
<td>1,002</td>
<td>983</td>
<td>996</td>
<td>979</td>
<td>984</td>
<td>950</td>
</tr>
</tbody>
</table>

Total open reading frames (ORFs) include predicted genes located on plasmids (noted in parentheses). The repeated trpEG genes on *Buchnera*-APS and -SG plasmids were counted once. Minor discrepancies in gene numbers (ORFs and pseudogenes) from the original publications reflect our efforts to present the most current annotation data from the literature, available at GenBank (http://www.ncbi.nlm.nih.gov) and the Comprehensive Microbial Resource (CMR) database (http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl). Fusion of yidC in *B. pennsylvanica* counted as a single ORF.
pattern at ytfM, ybiS, and gnuH, in which single frameshifts occur within poly(A) or poly(T) tracts. Poly(A) tracts are very common within Blochmannia ORFs (e.g., B. pennsylvanicus has 133 poly(A) or (T) tracts \( \geq 9 \) bp long) and likely reflect AT mutational bias and reduced selective pressures. The indels in ytfM, ybiS, hisH, and ubiF each show high coverage in the B. pennsylvanicus shotgun assembly (5–15 independent clones). In the case of hisH, the frameshift occurs across independent PCR products sequenced directly with varying types of sequencing chemistry.

Despite their consistency, we hesitate to interpret these frameshifts as firm evidence of pseudogenes. Notably, apart from a single frameshift, these genes would otherwise encode intact proteins that are relatively conserved between the two Blochmannia genomes, with an average protein divergence (0.67 \( \pm \) 0.16) that is only slightly above the average for other ORFs (0.565 \( \pm \) 0.309). The occurrence of frameshifts within homoplastic tracks is consistent with slippage during transcription (Wagner et al. 1990; Baranov et al. 2005) or during translation (Baranov et al. 2002; Gurvich et al. 2003) that could restore the full-length protein (see Discussion). If the frameshifts have, indeed, disrupted protein functions, these mutations must have occurred very recently since rapid mutation in Blochmannia (Degnan et al. 2004) is expected to erode pseudogenes quickly.

### Metabolic similarities of Blochmannia spp.

Analysis of B. pennsylvanicus and reanalysis of other insect mutualist genomes using MultiFun (Serres and Riley 2000) allowed more comprehensive metabolic comparisons across groups (see Fig. 2 and Supplemental Tables S1 and S2). The two Blochmannia genomes share nearly all metabolic pathways thought to contribute to host nutrition, including most biosynthetic functions, sulfate assimilation and metabolism, and hydrolysis of urea (see Gil et al. 2003).

### Metabolic differences between Blochmannia spp.

The 30 ORFs that distinguish the two Blochmannia genomes have a range of predicted cellular functions that may alter host-symbiont metabolic exchanges (Figs. 2 and 4). B. pennsylvanicus retains coaADE and dff for the biosynthesis of coenzyme A, an essential cofactor and substrate for the TCA cycle (Fig. 4A). In contrast, the deletion of these genes from B. floridanus implies that this endosymbiont does not require coenzyme A, uses other enzymes for its synthesis, or imports this cofactor from the host.

Several differences between the Blochmannia genomes involve the biosynthesis, transport, and mediation of cellular wall and membrane components. B. pennsylvanicus retains six distinct ORFs that contribute to the de novo synthesis of peptidoglycan
B. floridanus

B. pennsylvanicus

628 shared intact genes (a)

1 intact gene (proDF)

3 intact genes in B. florid, 
ψ in B. penns.

4 intact genes in B. penns., 
ψ in B. florid.

Figure 2. Comparison of B. pennsylvanicus and B. floridanus gene contents. Genes in the outer section of each circle are unique to one genome, while those listed in the intersection of the two circles are shared. Note that seven shared genes are apparently functional in one genome, but pseudogenes in the other. The truncation of B. floridanus dnaK (see text) is not included as a pseudogene here. The fusion of yidCD in B. pennsylvanicus is counted as two intact genes. (*) The 628 shared intact genes include genes that have a single frameshift within long poly(A) tracts but are otherwise intact (see text). These loci include B. pennsylvanicus hisH, ytfM, ubiB, and ybiS and B. floridanus yfmH, ybiS, and gmhB.

(murein), the major constituent of Gram-negative bacterial cell walls (Park 1996). In addition to Murl, it retains the complete pathway for the biosynthesis of isoprenoids (Fig. 4B), essential substrates for synthesis of peptidoglycan (El Ghachi et al. 2004), and several organic compounds including carotinoids, glycosyl carrier lipids, and the side chains of ubiquinone (Ogura et al. 1997; Kainou et al. 2001). In B. floridanus, the absence of murl and the interruption of isoprenoid biosynthesis implies this bacterium may import both D-glutamate and trans, trans-farnesyl diphosphate (isoprenoid precursor) from the host, might synthesize these components using other enzymes, or produce a cell wall lacking these structural elements (i.e., isoprenoid).

B. pennsylvanicus is the first fully sequenced insect mutualist that retains the entire sec-dependent secretory pathway, including the chaperonin SecB. This general pathway mediates the export and translocation of numerous proteins to the periplasm or inner membrane (Pugsley 1993). The other insect mutualist genomes lack certain components (typically secBDF) of this pathway (Shigenobu et al. 2000; Akman et al. 2002; Tamas et al. 2002; Gil et al. 2003; van Ham et al. 2003), although some short and medium-length spacers in B. floridanus are quite long (often >1 kb) in B. pennsylvanicus. Multiple gene-finding tools did not detect ORFs or recognizable pseudogenes in any spacers, which have lower GC contents (average of 20.0%) compared to ORFs (average of 32.1%) (Fig. 5B) and may represent eroded pseudogenes and/or regions involved in gene regulation.

Polymorphism within B. pennsylvanicus

Polymorphisms in the pooled symbiont population used for library construction were detectable as well-supported (Phred scores >40) discrepancies in the genome assembly that were represented by at least two independent clones. Nearly all (445/497) polymorphisms are single nucleotide polymorphisms (SNPs). Those located within ORFs occur primarily at third codon positions (Table 2). Although the majority of SNPs are located within ORFs, a disproportionate number of SNPs (35%) and indels (96% of insertions and/or deletions) occur within the intergenic regions (which comprise 23% of the B. pennsylvanicus genome). The two polymorphic indels within ORFs produce amino acid insertions/deletions in sucA and atpE.

Comparison of protein divergences

Wide variation in protein divergence across loci indicates variable functional constraint across Blochmannia proteins (Fig. 6; Supplemental Tables S3 and S5). Amino acid biosynthetic genes, ribosomal proteins, and particular chaperonins are exceptionally conserved (a mean protein divergence of 0.448, with a 99% confidence interval of 0.400–0.504), while surface structures [mean of 0.725 (0.667–0.7895)] and hypothetical or unclassified proteins [mean of 0.775 (0.633–0.991)] are quite different (Supplemental Table S5). The membrane proteins tolA and tonB are particularly divergent (differences of 1.74 and 2.64, respectively) (Supplemental Table S3).

Previous studies have demonstrated a negative relationship...
between GC content of endosymbiont genes and their level of divergence from free-living relatives (Herbeck et al. 2003; Banerjee et al. 2004), suggesting that amino acid changes in proteins under strong functional constraint are less severely affected by AT mutational bias. A strong negative association between GC content and protein divergence in *Blochmannia* (Supplemental Fig. S3) indicates this relationship also holds when protein divergences are estimated within a mutualist group (rather than to a more distant free-living relative). Protein divergences were, on average, \( \sim 1.88 \) times faster in *B. floridanus* lineage compared to *B. pennsylvanicus* lineage. Genes with particularly elevated rates in *B. floridanus* include *secG*, *rpsJ*, *rpsR*, and *rnpA*, each of which evolves more than 10-fold faster in *B. floridanus* compared to *B. pennsylvanicus* (Fig. 7; Supplemental Table S4). A 14-fold rate acceleration at *ispE* may reflect the loss or disruption of other genes involved in isoprenoid biosynthesis in the *B. floridanus* genome and relaxed selection at this remaining *isp* gene.

The *B. pennsylvanicus* genome offered the first opportunity to compare genome-wide patterns of protein evolution in the context of distinct endosymbiotic associations. A previous study showed accelerated evolution at 16S rDNA and at synonymous positions of select *Blochmannia* genes compared to enteric bacteria and even compared to the rapidly evolving *Buchnera* (Degnan et al. 2004). Here, we tested whether the *Blochmannia* genome also undergoes exceptionally fast rates of protein evolution. Strong correlations in divergences at homologous genes indicate parallel functional constraints in *Blochmannia* compared to *Buchnera* and *E. coli–Photorhabdus luminescens* (Fig. 6). However, notably higher protein divergences between *Blochmannia* spp., despite their relatively recent split, reflects a several-fold acceleration in absolute rates of protein evolution. Among the 302 genes for which RSD detected orthologs across *Blochmannia* and *Buchnera* and estimated divergences within each mutualist pair, the average substitution rate in *Blochmannia* was 0.0132 to 0.0166 amino acid substitutions/site per million years (based on a 20- to 16-Myr divergence, respectively). This is \( \sim 10 \) times faster than rates in *Buchnera* for the same gene set (0.00258–0.00362 amino acid substitutions/site per million years, based on a 70–50-Myr divergence). Previous studies have shown that nonsynonymous sites in *Buchnera* evolve an average of twofold (Canbäck et al. 2004) and up to 10-fold (Clark et al. 1999) faster than in the enterics. Likewise, we found an average \( \sim 50 \) times faster rate acceleration...
Degnan et al.

in *Blochmannia* compared to *E. coli–Salmonella typhimurium* across the 538 genes for which RSD detected orthologs (0.0142–0.0177 in *Blochmannia*, and 0.000245–0.000367 for *E. coli–S. typhimurium* assuming a divergence time of 150–100 Myr). Deviations from this general trend highlight genes that may experience different levels of functional constraint among genomes. For example, relative to other genes in the genome, amino acid biosynthetic genes evolve slightly slower in *Blochmannia* than in the enterics, while genes involved in translation (comprised largely of ribosomal proteins) evolve slightly faster than expected (Fig. 6).

Comparisons of average protein divergences within and between major functional categories confirmed many of these observations above (Supplemental Table S5). *Blochmannia*, *Buchnera*, and *E. coli–P. luminescens* show similar relative divergences across most functional categories, such as relatively high divergences of unclassified and hypothetical genes, loci for surface structures, cell membrane components; moderate divergences of genes encoding cofactor biosynthesis, information transfer, cell processes, metabolism, information transfer; and relatively low divergences of genes for nucleotide biosynthesis and amino acid biosynthesis. Striking differences among pairs include relatively high divergence of chaperonins and fatty acid biosynthetic genes within *Blochmannia* compared to *E. coli*, *P. luminescens*, and *S. graminum*–*S. graminum* pairs, which diverged 16–20 Mya. This exceptional genome stability, first demonstrated in *Buchnera* (Tamas et al. 2002; van Ham et al. 2003), suggests that *Blochmannia* genome stasis implies gene losses in either *Buchnera* lineage instead of in the *Buchnera*–*SG* association. Such constraints may constrain the evolutionary potential of this association. Such constraints have been proposed in *Buchnera* (Tamas et al. 2002), where the loss of genes for sulfur reduction and cysteine biosynthesis in *Buchnera–SG* may constrain the *S. graminum* host to its relatively cysteine-rich grass diet.

Certain aspects of *Blochmannia* metabolism remain unclear, owing to the uncertainty of whether or not single frameshifts in particular genes eliminate function. For example, if single indels within poly(A) tracts of hisH and ubiF are subject to correction of some type, the encoded proteins might be functional in *B. pennsylvaniaicus*. One possible mechanism for correction may be instability of such frameshifts during DNA replication, such that populations include heterogeneous genomes with different numbers of adenines in the homopolymeric repeats (e.g., Parkhill et al. 2000). However, among the many *C. pennsylvaniaicus* colonies used for symbiont library construction, we found no evidence for variation in lengths of these or any homopolymeric tracts. Alternatively, transcriptional slippage, or “stuttering” within repeat mononucleotides is well-documented in *E. coli* (Chamberlin and Berg 1962), where it often occurs within poly(A) or poly(T) tracts (Wagner et al. 1990). In a survey of published bacterial genomes, Baranov et al. (2005) identified several “pseudo pseudogenes,” for which transcriptional slippage could correct a frameshift within poly(A) or poly(T) tracts and restore uninterrupted ORFs. These authors describe a mechanism in which the RNA chain dissociates from the DNA template and reassociates in a new location. Third, functional proteins may be restored by frameshifting during translation, or “recoding,” a phenomenon that occurs in yeast (Hansen et al. 2003), archa (Cobucci-Ponzano et al. 2005), and *E. coli* (Gurvich et al. 2003) and can also follow poly(A) tracts, especially in the form of A_AAA_AAG (Baranov et al. 2005) and can also follow poly(A) tracts, especially in the form of A_AAA_AAG (Baranov et al. 2002, 2003). Such slippage has been proposed to explain aberrant indels in animal mitochondrial genes (Beckenbach et al. 2005), and in principle, might operate in endosymbionts. Given the multiple levels at which frameshifts within homopolymeric sequences may be corrected, we argue that such mutations should be interpreted cautiously, and with consideration of whether the gene otherwise encodes a full-length ORF. Although it is possible that such genes represent very recent loss of function, we take the approach of Baranov et al. (2005) in questioning whether such genes should be annotated as pseudogenes.

Conservation of genome architecture and overall similarity in gene content within *Blochmannia* contrasts with the exceptionally fast rates of sequence evolution observed in this group. Namely, protein divergences are higher for the two ant mutualists than within much older bacterial pairs. A previous study

**Figure 4.** Distinct biosynthetic capabilities of *B. pennsylvaniaicus* and *B. floridanus*. The two mutualistic genomes differ in genes encoded for the biosynthesis of (A) coenzyme A, and (B) isoprenoids. Gene names in gray are missing or pseudogenes (noted by ‘*‘).
A deviation from this pattern occurs in *Blochmannia*, where genes involved in translation are not the most conserved functional category relative to the enterics (Supplemental Table S5). We compared the ribosomal proteins that deviated from the best fit line in Figure 6 and those that showed a significant acceleration in the *B. floridanus* lineage (Fig. 7; Supplemental Table S4) to the detailed crystal structures of the 30S and 50S ribosomal subunits (Wimberly et al. 2000; Harms et al. 2001). These subunits assemble from 52 ribosomal proteins that form scaffolds around the 5S, 16S, and 23S ribosomal RNAs. Notably, many of the genes that show elevated substitution rates are positioned at the periphery of the ribosomal subunits, well away from the active regions at their interface. Furthermore, those proteins specifically involved the initiation and aggregation of the 30S and 50S ribosomes (RplL, RpsG) and those that have regulatory functions (RplADKT, RpsH), while exhibiting elevated rates were not among the genes with the highest divergences. These fast rates of substitution at ribosomal proteins might be influenced by several factors, including the effects of genome-wide patterns of nucleotide substitution; changes in regulatory interactions; compensatory changes due to the AT bias and consequently the secondary structure of the 16S, 23S, and 5S rRNAs; and the inactivation of the *rpmD* large ribosomal subunit locus. These amino acid substitutions in the ribosomal subunits might reduce the stability of the ribosome and thus decrease translation efficiency.

Within *Blochmannia*, faster divergence in the *B. floridanus* lineage at nearly all (~90%) proteins may reflect elevated mutation rates, reduced selective coefficients, or smaller effective population size of this symbiont and/or its host with associated increased genetic drift. Although data to distinguish these alternatives are limited, an analysis of four gene regions (*groEL*, *rpsB*, *atpB*, and *gidA*), all of which evolve faster in the *B. floridanus* than...
B. floridanus has a faster rate of mutations per replication. However, the year-round activity of C. floridanus and its relatives in the subgenus Myrmothrix contrasts with the winter dormancy of C. pennsylvanicus and related temperate species in the subgenus Camponotus. This activity may increase the number of host and bacterial generations per year and, consequently, the rate of mutations per unit time. Likewise, a combination of elevated mutation, relaxed selection, and/or increased genetic drift in the B. floridanus lineage may account for faster rates of gene loss compared to the B. pennsylvanicus lineage (see Lawrence and Roth [1999] for theoretical framework). However, given that current data cannot distinguish points along the B. floridanus lineage at which evolutionary rates accelerated or at which particular genes were lost, at this time we cannot link these changes to specific aspects of host ecology. More intensive sampling of Blochmannia across ecologically diverse hosts should allow such connections to be made.

Given the wide variation in Camponotus nutritional ecology, ranging from plant-specialists to omnivorous species, it seems unlikely that a single nutrient is lacking in the diet of all species that house Blochmannia. Rather than supplementing specific dietary deficiencies, nutritional functions of Blochmannia may play critical roles during two "starvation" phases of the host when metabolic demands exceed the available food supply—metamorphosis and colony founding (Wheeler and Martinez 1995). Recent work has shown that Blochmannia proliferate during pupation (Wolschin et al. 2004), a stage of metamorphosis when the host must construct all components of the adult body plan with no food intake (Wheeler and Martinez 1995). Genome sequence data provide a starting point for experimental analyses to clarify the functional significance of this mutualism, to explore the implications of genome variability on the physiology and ecology of both symbiotic partners, and to clarify the levels and timing of selection that shape this long-term bacterium–ant association.

**Methods**

**Blochmannia genome sequencing and assembly**

B. pennsylvanicus genomic DNA (gDNA) was prepared from worker and larvae C. pennsylvanicus collected from five colonies at two sites in Falmouth, Massachusetts, USA. The gDNA was either extracted directly from the agarose plugs containing the...
purified bacterial cells (Charles and Ishikawa 1999) or gel-purified from a chromosomal fragment resolved through Pulsed Field Gel Electrophoresis (PFGE) (Wernegreen et al. 2002). Short (1.5–2.5 kb) insert libraries were generated from hydrosheared DNA using a double adaptor kit (SeqWright Inc.) (Andersson et al. 1996). Plasmid clones were purified and bidirectionally sequenced using BigDye v3.0 chemistry on either an ABI3700 or an ABI3730xl (Applied Biosystems). Detailed methods for library construction and sequencing are provided in the Supplemental material.

Raw sequence data were analyzed by PHRED (Ewing and Green 1998; Ewing et al. 1998) (http://www.phrap.org/phredphrapconsed.html) and screened using BLASTN/X (Altschul et al. 1990) for ant host contamination. Sequence reads that were putatively identified as γ-Proteobacterial (E \leq 1^{-10}) were assembled using ARACHNE 2 (Jaffe et al. 2003) (http://www.genome.wi.mit.edu/wga/). The resulting contigs were analyzed by hand in CONSED (Gordon et al. 1998) and using BAMBUS (Pop et al. 2004). The *B. pennsylvanicus* assembly was aligned using LAGAN (Brudno et al. 2003) (http://lavan.stanford.edu/lavan_web/index.shtml) to the published *B. floridanus* genome (NC005061), which facilitated primer design for gap closure by PCR.

### Annotation and metabolic reconstruction

Open reading frames (ORFs) were identified iteratively using GLIMMER v2.10a (Delcher et al. 1999) (http://www.tigr.org/software/glimmer) and gene orthology predictions based on BLASTP sequence similarity to the NR, SWISS-PROT, and ECOIL databases; HMMR searches against Pfam_ls (Bateman et al. 2004); and identification of *E. coli* orthologs using the Reciprocal Sequence Distance (RSD) program (Wall et al. 2003) (details of the RSD method are noted below under “Comparison of protein divergences”). The few discrepancies among methods were limited to cases of differential loss of one gene from a pair of paralogs (*ilvBG*, *tuAB*, *argF*), the presence of gene fusions (*yidCD*) or split genes (*trpDG*), or failure of RSD to identify a given ortholog because of high sequence divergence.

Three pseudogenes (*uvrD*, *yqiC*, *rpmD*) were detected as regions with similarity to functional ORFs in other genomes, but with multiple indels and missense mutations resulting in stop codons throughout each gene. The two most degraded pseudogenes (*rpmD* and *yqiC*) were undetectable by BLASTX and were only identified because of the conservation of gene order between the two *Blochmannia* genomes. Among the three pseudogenes, *uvrD* retains the longest (107 amino acids) intact reading frame with similarity to functional orthologs, but even this region is just 15% of the length of UvrD in *E. coli* and other outgroups. In this sense, annotated pseudogenes clearly differ from truncated *B. pennsylvanicus* ORFs, which retain at least 60% of the length of orthologous proteins. The three *B. pennsylvanicus* pseudogenes also differ from genes with single frameshifts within homopolymeric regions (*hisH*, *ubiF*, *yfTM*, *yhiS*), since the latter would encode intact, relatively conserved proteins if the frameshift were corrected by slippage during transcription or translation (see Fig. 3, Results, and Discussion). ORFS and RNAs were manually curated using a Generic Model Organism Database (GMOD) Web browser. ORFs that lacked sequence similarity to any entry in GenBank or the Comprehensive Microbial Resource (Peterson et al. 2001) and lacked any predicted protein domains in Pfam_ls were excluded from the annotation. Functional and pseudo-transfer RNAs (tRNAs) were identified using tRNAscan-SE (http://selab.wustl.edu/cgi-bin/selab.pl?mode=software), and ribosomal RNAs and structural RNAs were identified by BLASTN searches of the intergenic regions. *Blochmannia* gene functions and interactions were inferred from orthologs of *E. coli* K12 MG1655 described in GenProtEC (Serres et al. 2004) (http://genprotec.mbl.edu) and characterized by MultiFun (Serres and Riley 2000), two resources that represent functions of >80% of the 4401 genes in *E. coli* K12. Metabolic pathways were evaluated using the reference pathways available for *E. coli* at EcoCyc (Karp et al. 2004) and KEGG (Kanehisa and Goto 2000). Genomes of other insect mutualists were reanalyzed in the same manner for a consistent metabolic comparison.

### Comparison of protein divergences

The RSD algorithm (Wall et al. 2003) was used to identify the reciprocal best BLAST hits (rhh) between translated ORFs of select *Buchnera* genomes. The program used BLAST to identify potential matches of a given translated gene, aligned all potential matches using CLUSTALW (Thompson et al. 1994), and calculated a maximum likelihood estimation of amino acid substitutions between proteins using PAML (Yang 1997). Protein divergences were based on an empirical amino acid substitution rate matrix (Jones et al. 1992) and accounted for variation in evolutionary rates among protein sites using a ω distribution with shape parameter α = 1.53 (as recommended by Nei et al. 2001). The protein with the lowest divergence was then BLASTed against the first genome, followed by the alignment and divergence calculations. If the protein match with the lowest divergence was the same as the original query sequence, the pair was considered orthologous and the divergence was retained in the output.

Such comparisons were performed within endosymbiont groups: *B. pennsylvanicus* versus *B. floridanus*; *Buchnera*-A. *pisum* versus *Buchnera*-S. *graminum*; and *Buchnera*-A. *pisum* versus *Buchnera*-Baizongia *pistaciae*. Divergences in endosymbionts were compared to the enterobacterial pairs *E. coli* versus *S. typhimurium*, and *E. coli* versus *P. luminescens*. Genomes were downloaded from NCBI in June 2004. All genomes were compared to *E. coli* using RSD for ortholog detection and MultiFun-based functional assignments (detailed in Supplemental material).

---

Figure 7. Accelerated rates of evolution in the lineage leading to *B. floridanus* compared to *B. pennsylvanicus*, since their divergence from a common ancestor. Rates of protein divergence were compared using a relative rates test with *E. coli* as an outgroup (see Methods). The analysis included 516 proteins for which RSD identified orthologs in the two genomes, and for which divergences between the two *Blochmannia* or between either endosymbiont and *E. coli* did not exceed 2.0. On average, proteins evolved 1.88 times faster in *B. floridanus* compared to *B. pennsylvanicus*. Only 50 of the proteins tested evolved more slowly in *B. pennsylvanicus* than in *B. floridanus* or at the same rate, with values ≤1 on the histogram, while 467 genes evolved faster in *B. floridanus*. Supplemental Table S4 lists proteins with particularly accelerated evolutionary rates in *B. floridanus*. 

---

Genome evolution in a bacterial mutualist of ants
Rates of protein divergence along the lineages leading to *B. pennsylvanicus* and *B. floridanus* were compared using *E. coli* as an outgroup. Relative rates were calculated as follows, with “B0” representing the common ancestor of the two *Blochmannia* lineages:

\[
\text{prot. div. (B0-Bt)} / \text{prot. div. (B0-Rec)} = \frac{[\text{prot. div. (B0-Ecoli)} + \text{prot. div. (B0-Bt)}]}{[\text{prot. div. (B0-Ecoli)} + \text{prot. div. (B0-Rec)}]}
\]

Pairwise \(dN/dS\) and \(dS\) values reported for certain genes [e.g., truncated genes or ORFs with apparent frameshifts within poly(A) tracts] were calculated using PAML (runmode = 2) (Yang 1997). In order to account for the extreme base composition of endosymbiont genes or ORFs with apparent frameshifts within poly(A) sequences using the program RevTrans (Wernersson and Andersson 2003).

### Acknowledgments

We thank Andrew McArthur, Margrethe Serres, Seth Bordenstein, Seth Kauppinen, and four anonymous reviewers for comments on the manuscript, and are grateful to Daniel Hahn, Diana Wheeler, Stefan Cover, and Diana Davidson for helpful discussion of *Camponotus* biology. Amy McCurley and Josh Larson contributed to the library construction, and Jeremy Brozek contributed to the JENA Prokaryotic Genome Viewer, Dennis P. Wall provided access to RSD, and D. Wall and Richard Fox offered assistance implementing this program. We thank Brian Dale for permission to collect *C. pennsylvanicus* for this study. This research was supported by grants to J.J.W. from NIH (RO1 GM62626-01) and the NSF (DEB 0089455). Additional support was provided by the NASA Astrobiology Institute (NCC2-1054 and NNA04CC04A). Template preparation and DNA sequencing was performed in the W.M. Keck Facility for Ecological Genomics at the Marine Biological Laboratory.

### References


---

**Web site references**

http://ecocyc.org; Encyclopedia of *Escherichia coli* K12 Genes and Metabolism.

http://genprotec.mbl.edu; GenProTEC E. coli genome and proteome database.

http://lagon stanford.edu/lagon_web/index.shtml; LAGAN.

http://selab.wustl.edu/cgi-bin/; tRNAscan-SE.


http://www.phrap.org/phredphrapconsem.html; PHRED, CONSED.

http://www.tigr.org/software/glimmer; GLIMMER.


Received February 3, 2005; accepted in revised form June 7, 2005.