SUPPLEMENTARY RESEARCH DATA

DETAILED METHODS FOR LIBRARY CONSTRUCTION AND SEQUENCING

Insect collection, identification, and DNA extraction. We constructed two short insert shotgun libraries for *Blochmannia pennsylvanicus*, each of which required the purification of 25µg of genomic DNA (gDNA). To facilitate this process, entire nests of *Camponotus pennsylvanicus* were identified in the field and collected from Bebe Woods Town Forest and Goodwill State Park in Falmouth, MA, USA during May-July of 2001 and May-December of 2002. Nests that were collected during the summer were maintained at 25°C on a modified Bhatkar diet and supplemented with live or frozen insects (Hölldobler and Wilson 1990). Ants collected while they were over-wintering were kept at 4°C and required no feeding. Since *C. pennsylvanicus* is monogynous, we expect that *Blochmannia* in a single colony were transmitted by one queen. Nonetheless, due to fast rates of sequence evolution in this mutualist (Degnan et al. 2004), we expected some degree of polymorphism within the pooled symbiont DNA sample.

Because *Camponotus* lacks discrete bacteriomes that house the obligate mutualist *Blochmannia,* we utilized a whole insect endosymbiont isolation protocol. This process required the homogenization of entire insects followed by serial filtrations and separation of the bacterial cells on a Percoll density gradient (Charles and Ishikawa 1999; Wernegreen et al. 2002). In the construction of the initial library (termed BSLA), we obtained gDNA directly from the agarose plugs containing lysed bacterial cells by digestion with β-Agarase (New England Biolabs, Beverly, MA) following the manufacturer's instructions and followed by NH_3OAc/ isopropanol precipitation (Sambrook and Russell 2001).

The second library (BSLB) was constructed in an effort to generate a higher purity library. To this end we performed further endosymbiont preparations and subsequently digested the bacterial DNA embedded in the agarose plugs with *Asc I* (GG^CGCGCC; NEB)**,** a GC rich restriction endonuclease that cut the *Blochmannia* chromosome once to generate a single linear fragment. This fragment was then resolved with PFGE (Wernegreen et al. 2002), excised, and purified via dialysis (ref. Strong et al. 1997, with certain modifications) and NaCl/ isopropanol precipitation (Sambrook and Russell 2001). The gDNA was pooled from multiple replicates of this process.

Short Insert Library Construction and Sequencing. Two short insert shotgun libraries were constructed from the isolated *Blochmannia pennsylvanicus* gDNA using a kit that implements the "double-adaptor" method (Andersson et al. 1996). In brief, this technique utilizes specific adaptors with extended overhangs that are ligated to blunt-ended fragments and which complement the overhangs on the modified pUC18 vector. Construction of the libraries were completed using the manufacturer's instructions (SeqWright Inc., Houston TX) with several modifications noted as follows; (i) pooled gDNA for both libraries were sheared using a Hydroshear (Genomic Solutions, Ann Arbor, MI) to a size range with a median size of 2 kb (Speed Code: 7) (ii) QIAquick columns (Qiagen, Valencia, CA) were utilized in place of all phenol/ chloroform extractions and NaOAc precipitations, and (iii) the QIAquick Gel Purification kit was used to extract and purify the size-selected (1.5-2.5kb), adaptor ligated, insert DNA. Inserts were annealed to the vector and transformed into XL-1 Blue Supercompetent Cells (Stratagene, La Jolla, CA). Single colonies were picked and re-grown in 96 – deep well plates and prepared using a RevOrbit (Genomic Solutions). Paired-end sequence reads were generated from the plasmid clones on either the ABI3700 (BSLA) or ABI3730xl (BSLB) using the BigDye v3.0 chemistry (Applied Biosystems, Foster City, CA) and M13F-21 and M13R primers.

Closure sequencing. Based on the alignment of the 55 contigs and 20 supercontigs to the reference of sequence (NC005061), we designed oligonucleotide primers to span the remaining physical and sequence gaps. We used PCR conditions similar to those described in (Degnan et al. 2004) to amplify products from the same sheared pooled template gDNA used for the BSLB library construction. PCR products that yielded multiple products were cloned and sequenced to verify polymorphisms. Three sequence gaps were attributed to Big Dye sequencing stops due to single strand loop formation. We used a protocol developed for dGTP BigDye terminator v3.0 (Applied Biosystems) by Genome Sequencing Center at University of Washington, St. Louis (http://www.genome.wustl.edu/tools/protocols/) to successfully read through the final three gaps. In total, 12,210 sequence reads derived from 9,833 total recombinant clones and 20 closure PCR amplicons were used in the final assembly.

METHOD OF FUNCTIONAL ASSIGNMENTS FOR ANALYSIS OF PROTEIN DIVERGENCES

While most genes were cross-listed in two or more MultiFun categories, several comparisons required that each gene was represented only once. We identified six key functions of potential importance in endosymbionts and labeled them separately: biosynthesis of amino acids (MultiFun 1.5.1), nucleotides (1.5.2), cofactors (1.5.3) and fatty acids (1.5.4); chaperonins (2.3.4); translation (2.3.2, the majority of which were ribosomal proteins); and a cumulative category of surface structures such as antigens, flagella, and phage-related functions (1.6.12, 8.1, and other MultiFun categories). A gene belonging to more than one key category (e.g. some chaperonins are also considered surface proteins) was labeled based on the best estimation of the primary function, such as the primary annotation of that genome.

Second, our comparison of average divergences across functional groups accounted for the fact that numerous genes have two or more MultiFun assignments but enforced a trimming approach to reduce multiple comparisons involving the same gene. This approach involved the following steps: (i) Genes assigned to one of the six key categories were labeled as such, allowing for cross-listing between two or more key categories. (ii) Genes not assigned to any of the six categories were labeled by their major MultiFun category (e.g. metabolism, information transfer, regulation, transport, cell processes, cell structure) allowing for cross-listing among these categories. Under this approach, genes assigned to one or more key category were not listed in the broader categories, and conversely, broader categories excluded specific functions noted as key. Mean dN values within functional categories were determined using Excel, and 99% confidence intervals for each category were determined by resampling values with replacement 1000 times (Resampling Stats Excel add-in version 3.0). For each genome pair, we measured variation in divergences among categories as the sum of absolute deviations between the category means and the grand mean. We tested the null hypothesis of no difference among categories by calculating the same value across 1000 "shuffled" data sets (also created by resampling without replacement using Resampling Stats Excel add-in version 3.0). For all three genome pairs, the empirical sum of absolute deviations exceeded all 1000 values from shuffled data, indicating significant differences among category means at the $p<0.001$ level.

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. Comparison of GC skew in *B. pennsylvanicus* and *B. floridanus*. Outer two circles (blue and green) indicate ORFs on the + and - strands, respectively, of the two *Blochmannia* genomes. Central circle (red) shows the calculated GC skew (see legend of Figure 1). For both genomes, a shift in the GC skew is consistent with the proposed origin of replication near *gidA*. Interestingly, the *B. floridanus* genome shows an abrupt switch in the GC skew ~29 kb upstream of *gidA* and again ~180$^{\text{o}}$ opposite to that point, suggesting a possible shift in the origin of replication. This shift may be due to differences in the DNA replication machinery of the two *Blochmannia* strains (see text).

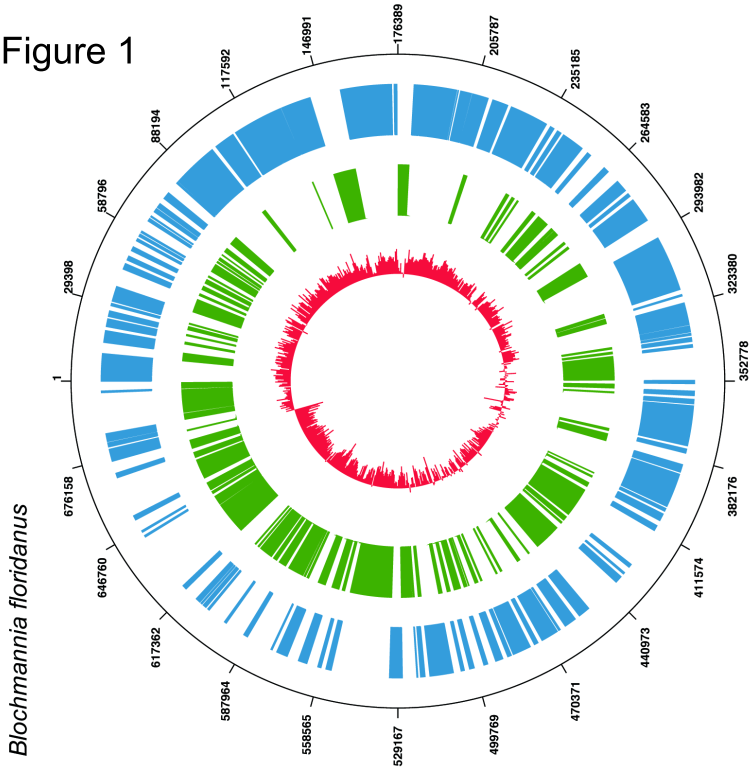
Figure S2. MultiFun categorization of gene functions. Comparison of the gene content of the six sequenced genomes of insect mutualists across functional categories. Endosymbionts show considerable differences in their abilities to biosynthesize amino acids, fatty acids, and cofactors. Differences with some previous genome comparisons (e.g., a greater number of endosymbiont genes are assigned to regulatory functions in this analysis) reflect our use of the MultiFun assignment schema instead of COGs. MultiFun assignments for all *B. pennsylvanicus* genes are listed in GenBank file CP000016. Online Table S1 details the number of genes per MultiFun category and subcategories for each of the six genomes.

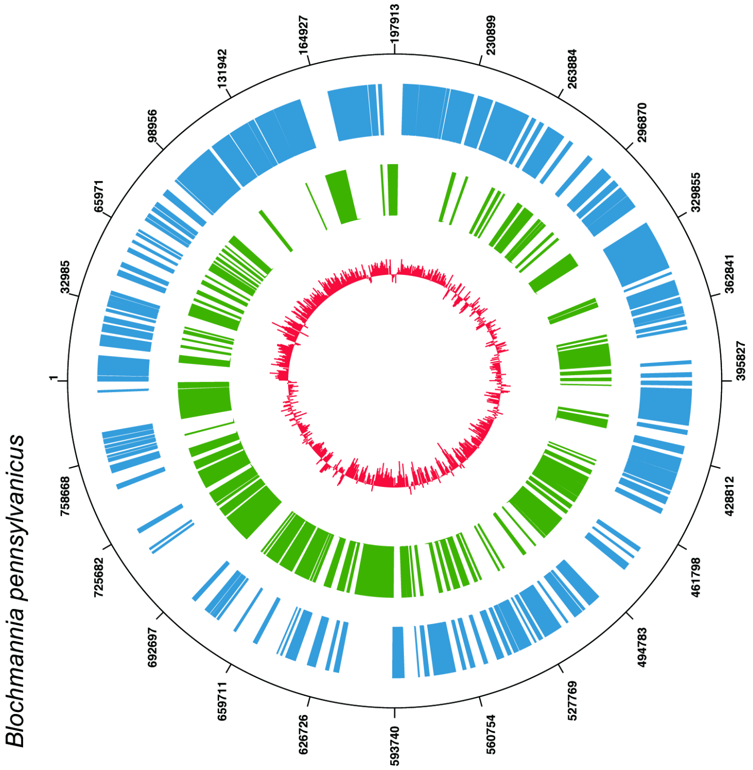
Figure S3. Divergence among *Blochmannia* proteins is negatively associated with GC content. This negative relationship indicates that selective constraint on proteins reduces the impact of AT mutational bias. Although divergence values above 2 are error prone due to

saturation, they are included here for comparative purposes. Select functional categories are labeled to illustrate different divergence rates among categories. The most conserved genes include ribosomal proteins, chaperonins, and certain biosynthetic functions, while the most divergent genes include surface factors and antigens. Online Table S3 lists pairwise divergence values.

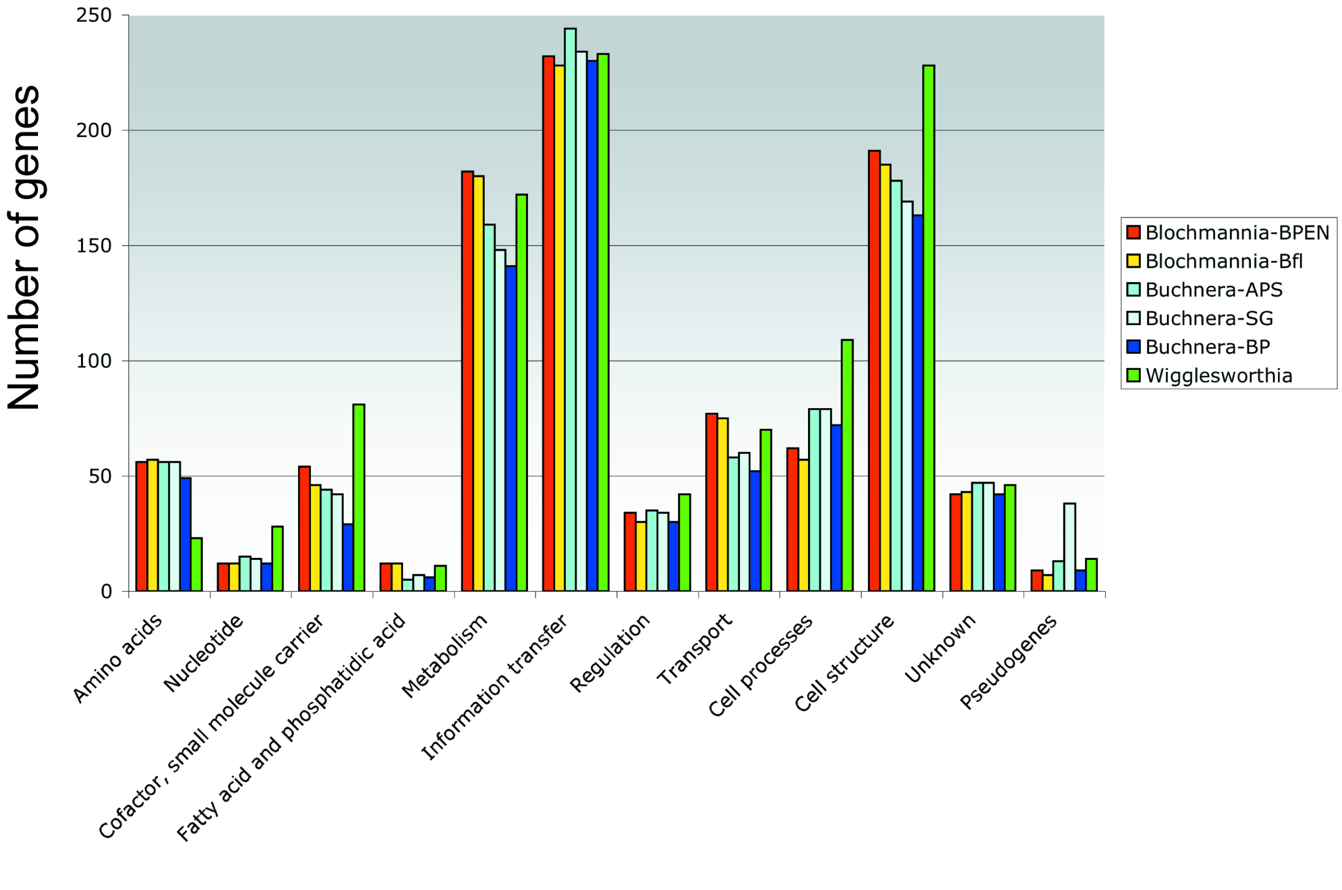
SUPPLEMENTARY FIGURES

Supplementary Figure 1

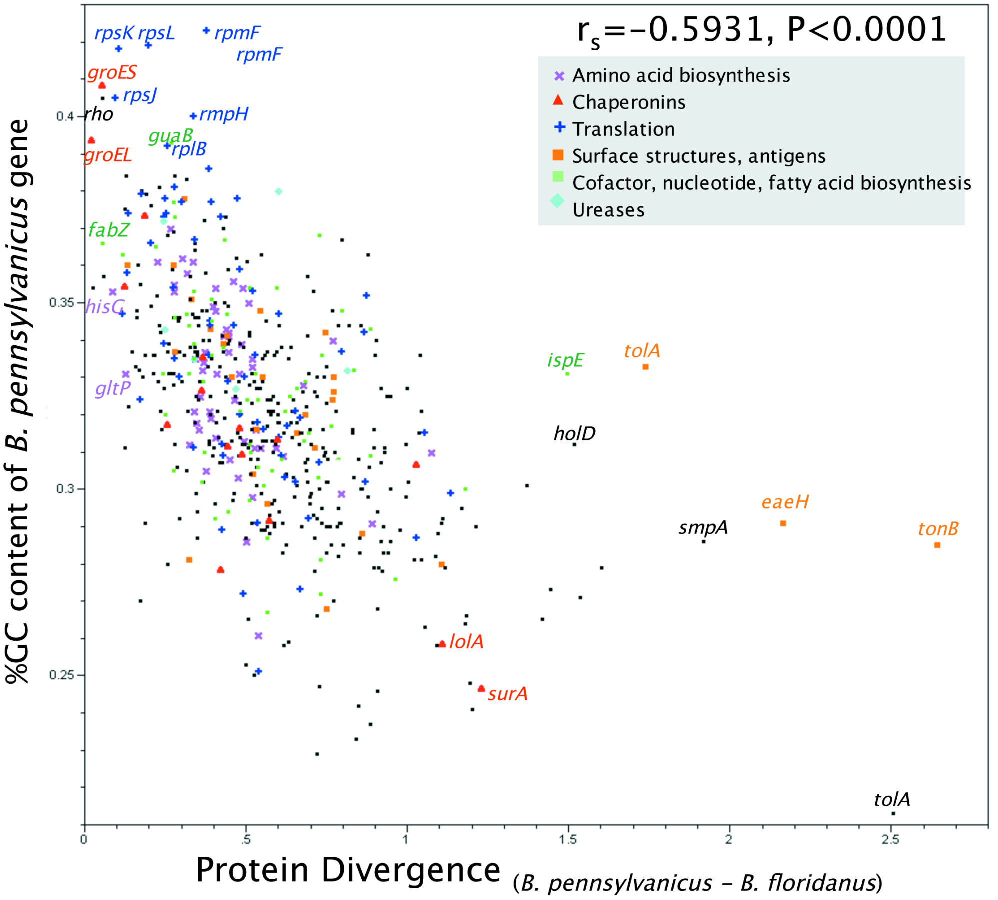




Supplementary Figure 2



Supplementary Figure 3



REFERENCES

Andersson, B., Wentland, M.A., Ricafrente, J.Y., Liu, W., and R.A. Gibbs. 1996. A "double adaptor" method for improved shotgun library construction. *Anal. Biochem.* **236:** 107 – 113.

Charles, H. and Ishikawa, H. 1999. Physical and genetic map of the genome of *Buchnera*, the primary endosymbiont of the pea aphid *Acyrthosiphon pisum. J. Mol. Evol.* **48:** 142 – 150.

Degnan, P.H., Lazarus, A.B., Brock, C., and Wernegreen, J.J. 2004. Host-symbiont stability and fast evolutionary rates in an ant-bacterium association: Cospeciation of *Camponotu*s species and their endosymbionts, *Candidatus* Blochmannia. *Syst. Biol.* **53:** 95 – 110.

Hölldobler, B. and Wilson, E.O. 1990. *The Ants*. Belknap Press of Harvard University Press, Cambridge, Mass.

Sambrook, J. and Russell, D.W. 2001. *Molecular cloning: A laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Strong, S.J., Ohta, Y., Litman, G.W., and Amemiya, C.T. 1997. Marked improvement of PAC and BAC cloning is achieved using electroelution of pulsed-field gel-separated partial digests of genomic DNA. *Nucleic Acids Res.* **25:** 3959 – 3961.

Wernegreen, J.J., Lazarus, A.B., and Degnan, P.H. 2002. Small genome of *Candidatus* Blochmannia, the bacterial endosymbiont of *Camponotus*, implies irreversible specialization to an intracellular lifestyle. *Microbiology* **148:** 2551 – 2556.