Cell

Horizontal Gene Transfer from Diverse Bacteria to an Insect Genome Enables a Tripartite Nested Mealybug Symbiosis

Filip Husnik,[1] Naruo Nikoh,[2] Ryuichi Koga,[3] Laura Ross,[4] Rebecca P. Duncan,[5] Manabu Fujie,[6] Makiko Tanaka,[7] Nori Satoh,[7] Doris Bachtrog,[8] Alex C.C. Wilson,[5] Carol D. von Dohlen,[9] Takema Fukatsu,[3] and John P. McCutcheon[10,*]

[1]Faculty of Science, University of South Bohemia and Institute of Parasitology, Biology Centre ASCR, České Budějovice 370 05, Czech Republic

[2]Department of Liberal Arts, The Open University of Japan, Chiba 261-8586, Japan

[3]National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba 305-8566, Japan

[4]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

[5]Department of Biology, University of Miami, Coral Gables, FL 33146, USA

[6]DNA Sequencing Section

[7]Marine Genomics Unit

Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan

[8]Department of Integrative Biology, University of California Berkeley, Berkeley, CA 94720, USA

[9]Department of Biology, Utah State University, Logan, UT 84322, USA

[10]Division of Biological Sciences, University of Montana, Missoula, MT 59812, USA

*Correspondence: john.mccutcheon@umontana.edu

http://dx.doi.org/10.1016/j.cell.2013.05.040

SUMMARY

The smallest reported bacterial genome belongs to *Tremblaya princeps*, a symbiont of *Planococcus citri* mealybugs (PCIT). *Tremblaya* PCIT not only has a 139 kb genome, but possesses its own bacterial endosymbiont, *Moranella endobia*. Genome and transcriptome sequencing, including genome sequencing from a *Tremblaya* lineage lacking intracellular bacteria, reveals that the extreme genomic degeneracy of *Tremblaya* PCIT likely resulted from acquiring *Moranella* as an endosymbiont. In addition, at least 22 expressed horizontally transferred genes from multiple diverse bacteria to the mealybug genome likely complement missing symbiont genes. However, none of these horizontally transferred genes are from *Tremblaya*, showing that genome reduction in this symbiont has not been enabled by gene transfer to the host nucleus. Our results thus indicate that the functioning of this three-way symbiosis is dependent on genes from at least six lineages of organisms and reveal a path to intimate endosymbiosis distinct from that followed by organelles.

INTRODUCTION

Bacterial genomes range in size over two orders of magnitude, from approximately 0.14 to 14 Mb pairs in length (Chang et al., 2011; López-Madrigal et al., 2011; McCutcheon and von Dohlen, 2011). Those at the small end of the spectrum typically come from bacteria that reside exclusively in eukaryotic host cells, and the tiniest genomes—those less than 0.5 Mb in length—are thus far exclusively from bacteria that are nutritional endosymbionts of sap-feeding insects (McCutcheon and Moran, 2012). These symbionts play critical roles in the biology of their host insects by synthesizing nutrients, such as essential amino acids and vitamins, that the insects cannot make on their own and that are limiting in their plant sap diets (Baumann, 2005; Douglas, 1989; Moran, 2007). Typically, these tiny symbiont genomes retain few genes outside of pathways involved in DNA replication, transcription, translation, and nutrient provisioning to their hosts (McCutcheon, 2010; McCutcheon and Moran, 2012). The most severely reduced of these genomes are missing genes widely considered to be essential, making it unclear how they continue to function (Keeling, 2011; McCutcheon and Moran, 2012).

The smallest bacterial genome so far reported is from *Candidatus* Tremblaya princeps, an endosymbiont of the mealybug *Planococcus citri* (hereafter referred to as *Tremblaya* PCIT for simplicity) (López-Madrigal et al., 2011; McCutcheon and von Dohlen, 2011). The *Tremblaya* PCIT genome is only 139 kilobase pairs (kb) in length, encodes approximately 120 protein-coding genes, and is missing several essential translation-related genes. For example, *Tremblaya* PCIT encodes no functional aminoacyl-tRNA synthetases and lacks functional homologs for both bacterial translational release factors, elongation factor EF-Ts, ribosome recycling factor, and peptide deformylase. This extreme genome degeneracy is highly unusual in bacteria, evidenced by the fact that all other reduced symbiont genomes retain these translation-related gene homologs (although some do not code for complete sets of aminoacyl-tRNA synthetases [McCutcheon, 2010; McCutcheon and Moran, 2012]). The genome of *Tremblaya* PCIT is striking in its degeneracy not only for the genes it is missing but also for its low coding density

(López-Madrigal et al., 2011; McCutcheon and von Dohlen, 2011). Although other highly reduced bacterial genomes are extremely gene dense, the *Tremblaya* PCIT genome has a coding density of only 73% and contains approximately 19 detectable pseudogenes. These features strongly suggest that *Tremblaya* PCIT has undergone a relatively recent environmental or ecological shift, in which selection on some genes has been relaxed due to redundancy from another source.

The unusual nature of the mealybug symbiosis is the most obvious explanation for the extreme degeneracy of the *Tremblaya* PCIT genome: residing in *Tremblaya*'s cytoplasm is another organism, the gammaproteobacterium *Candidatus* Moranella endobia (hereafter referred to simply as *Moranella*) (von Dohlen et al., 2001). At 538 kb in length, the *Moranella* genome is almost four times larger than the *Tremblaya* PCIT genome, and its 406 protein-coding genes include all the critical translation-related genes missing or pseudogenized in *Tremblaya* PCIT (McCutcheon and von Dohlen, 2011). This suggests that much of the genomic erosion in *Tremblaya* might be explained by the incorporation of *Moranella* into its cytoplasm. However, other symbionts lacking intracellular bacteria also show highly reduced genomes, making it plausible that the severe gene loss observed in *Tremblaya* PCIT occurred before the acquisition of *Moranella.*

There are therefore several possible mechanisms—none mutually exclusive—that could allow *Tremblaya* PCIT to continue functioning: (1) the lost *Tremblaya* PCIT genes may have been transferred to the host mealybug nucleus, with their products imported back into the cell; (2) the lost *Tremblaya* PCIT genes may be compensated by host gene products of eukaryotic origin that are transported into the cell; (3) the lost *Tremblaya* PCIT genes may be compensated by bacterial genes that are the result of horizontal transfer from unrelated bacteria to the host genome (Nikoh and Nakabachi, 2009; Nikoh et al., 2010); and (4) *Tremblaya* PCIT may somehow acquire gene products directly from *Moranella*, as previously suggested (Koga et al., 2013; McCutcheon and von Dohlen, 2011). Defining the relative roles of each of these four processes is important, as possibilities (1) and (2) would parallel events that took place during organelle (mitochondria and chloroplast) formation (Keeling and Palmer, 2008; Timmis et al., 2004), scenario (3) would provide the first data suggesting heterologous complementation for a lost activity in a reduced symbiotic genome, and (4) would clarify the unique nature of this three-way nested symbiosis.

Gene retention patterns in essential amino acid biosynthesis pathways—the raison d'être for *Tremblaya* PCIT and *Moranella*, at least from the perspective of the mealybug host—offer some clues to the mechanisms enabling genome reduction of *Tremblaya* PCIT. While all ten essential amino acid biosynthesis pathways are incomplete when the contributions from *Tremblaya* PCIT and *Moranella* are analyzed independently, several pathways become complete when the inferred gene homologs from *Tremblaya* PCIT and *Moranella* are considered together with putative contributions from the host (McCutcheon and von Dohlen, 2011). These complementary gene retention patterns suggest but do not prove that gene products or metabolites for essential amino acid biosynthesis are shared between the two bacterial symbionts and indicate that the loss of critical genes in *Tremblaya* PCIT may be supplemented by *Moranella* gene products. However, the host clearly plays a large role in the functioning of the symbiosis because production of several amino acids seems to require chemistries carried out by host-encoded enzymes (McCutcheon and von Dohlen, 2011), similar to what has been hypothesized to occur in the pea aphid (International Aphid Genomics Consortium, 2010; Wilson et al., 2010). The available data therefore point to a potentially complex solution to the loss of essential genes in *Tremblaya* PCIT.

Adding to the complexity is the possibility that genes resulting from horizontal gene transfer (HGT) play a role in the functioning of the *Pl. citri* symbiosis. A number of HGT cases from microorganisms to animals have been reported recently, including several examples from insects (Acuña et al., 2012; Aikawa et al., 2009; Altincicek et al., 2012; Danchin et al., 2010; Doudoumis et al., 2012; Gladyshev et al., 2008; Grbić et al., 2011; Dunning Hotopp et al., 2007; Klasson et al., 2009; Kondo et al., 2002; Moran and Jarvik, 2010; Nikoh and Nakabachi, 2009; Nikoh et al., 2010; 2008; Werren et al., 2010; Woolfit et al., 2009). Although most transferred DNA is probably nonfunctional in the host genome (Dunning Hotopp et al., 2007; Kondo et al., 2002; Nikoh et al., 2008), a growing list of apparently functional transferred genes have been identified. These genes are expressed in tissue-specific patterns, subject to purifying selection, and/or explain well-known ecological traits (Acuña et al., 2012; Danchin et al., 2010; Grbić et al., 2011; Klasson et al., 2009; Moran and Jarvik, 2010; Nikoh and Nakabachi, 2009; Nikoh et al., 2010; Woolfit et al., 2009). In a few cases, the transferred genes have been shown to provide a clear and specific function in the biology of the animal (Acuña et al., 2012; Danchin et al., 2010). The taxonomic origins of these functional transfer events are diverse (Gladyshev et al., 2008) and include fungi (Altincicek et al., 2012; Grbić et al., 2011; Moran and Jarvik, 2010) and various groups of bacteria such as Bacilli (Acuña et al., 2012; Grbić et al., 2011), Actinobacteria (Danchin et al., 2010), and perhaps most commonly in insects, Alphaproteobacteria (Dunning Hotopp et al., 2007; Klasson et al., 2009; Nikoh and Nakabachi, 2009; Nikoh et al., 2010; Werren et al., 2010; Woolfit et al., 2009). Much of the DNA transferred from alphaproteobacterial sources is presumed to be from the reproductive manipulator *Wolbachia* or close relatives (Dunning Hotopp, 2011).

The role of lateral gene transfer in the functioning of symbioses involving bacteria with highly degenerate genomes such as *Tremblaya* PCIT is presently unclear. The best-studied and most relevant example for the mealybug system is the pea aphid, *Acyrthosiphon pisum*, and its bacterial endosymbiont *Buchnera aphidicola* (International Aphid Genomics Consortium, 2010; Nikoh et al., 2010; Shigenobu et al., 2000). Although *Buchnera* is a stably associated, long-term nutritional endosymbiont, its 641 kb genome encodes 574 protein-coding genes and so is relatively more complete compared to the degenerate genome of *Tremblaya* PCIT. When the pea aphid genome was analyzed for potential HGT events originating from *Buchnera*, two independent transfers were found, although both encoded nonfunctional gene products (Nikoh et al., 2010). This shows that HGT between insect nutritional symbionts and their hosts is possible but that it has not resulted in the acquisition of functional genes in

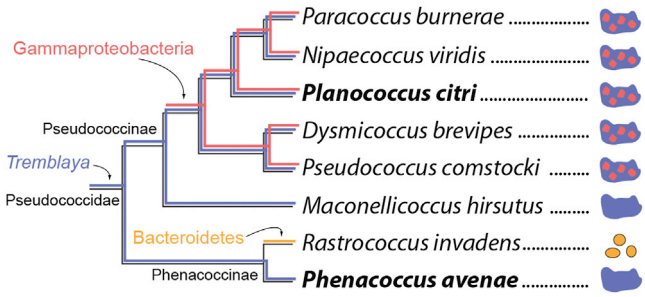

Figure 1. Cladogram of Selected Mealybugs and Their Obligate Symbionts

Tremblaya is the sole symbiont in some lineages of mealybugs (e.g., *Ph. avenae*); it was replaced with a symbiont from the Bacteroidetes in some lineages (e.g., *Rastrococcus invadens*; yellow line) and was itself infected with gammaproteobacteria in other lineages of mealybugs (red lines; e.g., with *Moranella endobia* in *Pl. citri*). This figure is a composite from previous work (Buchner, 1965; Gruwell et al., 2010; Hardy et al., 2008; Thao et al., 2002).

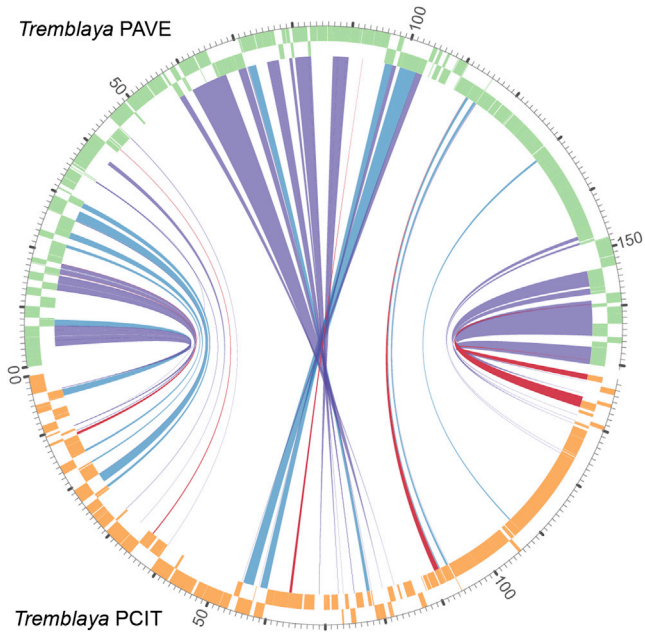

Figure 2. The *Tremblaya* PCIT Genome Is Largely a Subset of the *Tremblaya* PAVE Genome

The coding regions of *Tremblaya* PAVE (green boxes, top) and *Tremblaya* PCIT (orange boxes, bottom) are shown around the perimeter of the circle. Purple bands connect genes retained in *Tremblaya* PAVE to their presumed former positions in *Tremblaya* PCIT. Blue bands connect functional genes retained in *Tremblaya* PAVE to those that are present but pseudogenized in *Tremblaya* PCIT. Red bands connect genes retained in *Tremblaya* PCIT to their presumed former positions in *Tremblaya* PAVE. Of the 121 genes retained in *Tremblaya* PCIT, 110 are also present in *Tremblaya* PAVE. *Tremblaya* PCIT encodes 11 genes not present in *Tremblaya* PAVE; *Tremblaya* PAVE encodes 65 genes not present in *Tremblaya* PCIT. See Table S1 for a comparison of the general features of these genomes.

the pea aphid. Understanding the role that horizontal gene transfer has played in the evolution of insect endosymbionts is of great interest because many of these symbionts show nontrivial overlap with organelles in terms of genome size and organismal integration (Keeling, 2011; McCutcheon and Moran, 2012).

Here we take a comparative genomic and transcriptomic approach to disentangle the mechanisms used by *Tremblaya* PCIT to function in the mealybug symbiosis. To provide data on the role of *Moranella* in the biology of *Tremblaya*, we have sequenced a complete genome for *Tremblaya* from *Phenacoccus avenae* (PAVE), a species of mealybug possessing *Tremblaya* as its sole symbiont (Figure 1). To assess the role of the insect host in the functioning of *Tremblaya*, we performed RNA-seq on both the *Pl. citri* bacteriome (the symbiotic organ housing *Tremblaya* PCIT and *Moranella*) as well as whole animals to identify genes that are preferentially expressed in tissue relevant to the symbiosis. To verify the origin of the expressed genes found by our transcriptional work, we determined a draft insect genome for *Pl. citri*. Our results suggest a large role for *Moranella* gene products in the functioning of *Tremblaya* PCIT and uncover a surprising number of expressed genes transferred from heterologous bacterial sources (i.e., neither from *Tremblaya* nor *Moranella*) to the insect genome, which are involved in nutrient biosynthesis and bacterial cell wall maintenance. Because we find no clear functional gene transfer events from *Tremblaya* PCIT to the host genome, our data show that this organism is not progressing along an evolutionary path analogous to mitochondria and chloroplasts in their transition from endosymbiont to organelle, a process that included extensive gene transfer to the host nuclear genome.

RESULTS

The *Tremblaya* Genome from *Phenacoccus avenae* Is Much Less Degenerate Than in PCIT

Genome sequencing revealed that the gene set of *Tremblaya* PCIT is an almost perfect subset of *Tremblaya* PAVE (Figure 2 and Table S1 available online). The genome of *Tremblaya* PAVE is 170,756 bps and very gene dense (93.5% coding density), and it has few pseudogenes, making it similar to other tiny symbiont genomes such as *Hodgkinia cicadicola* (144 kb) (McCutcheon et al., 2009), *Carsonella ruddii* (158–166 kb) (Nakabachi et al., 2006; Sloan and Moran, 2012), and *Zinderia insecticola* (210 kb) (McCutcheon and Moran, 2010). It is colinear with *Tremblaya* PCIT with the exception of one large inversion and one unusual plasmid containing only two ribosomal genes (Figure 2 and Table S1). Importantly, many of the genes present in *Tremblaya* PAVE but missing in *Tremblaya* PCIT are the translation-related genes found in other highly reduced genomes (Figure 3), although like some other tiny genomes (McCutcheon, 2010; McCutcheon and Moran, 2012) *Tremblaya* PAVE does not encode a complete set of aminoacyl-tRNA synthetases.

The Sole PAVE Symbiont Encodes the Same Essential Amino Acid Pathways as the Dual PCIT Symbionts

As the sole nutritional symbiont for its insect host, *Tremblaya* PAVE retains exactly the same genes for essential amino acid biosynthesis as are collectively retained in the dual *Tremblaya* PCIT-*Moranella* symbiosis (Figure 3). This striking result is consistent with recent data showing that related species of

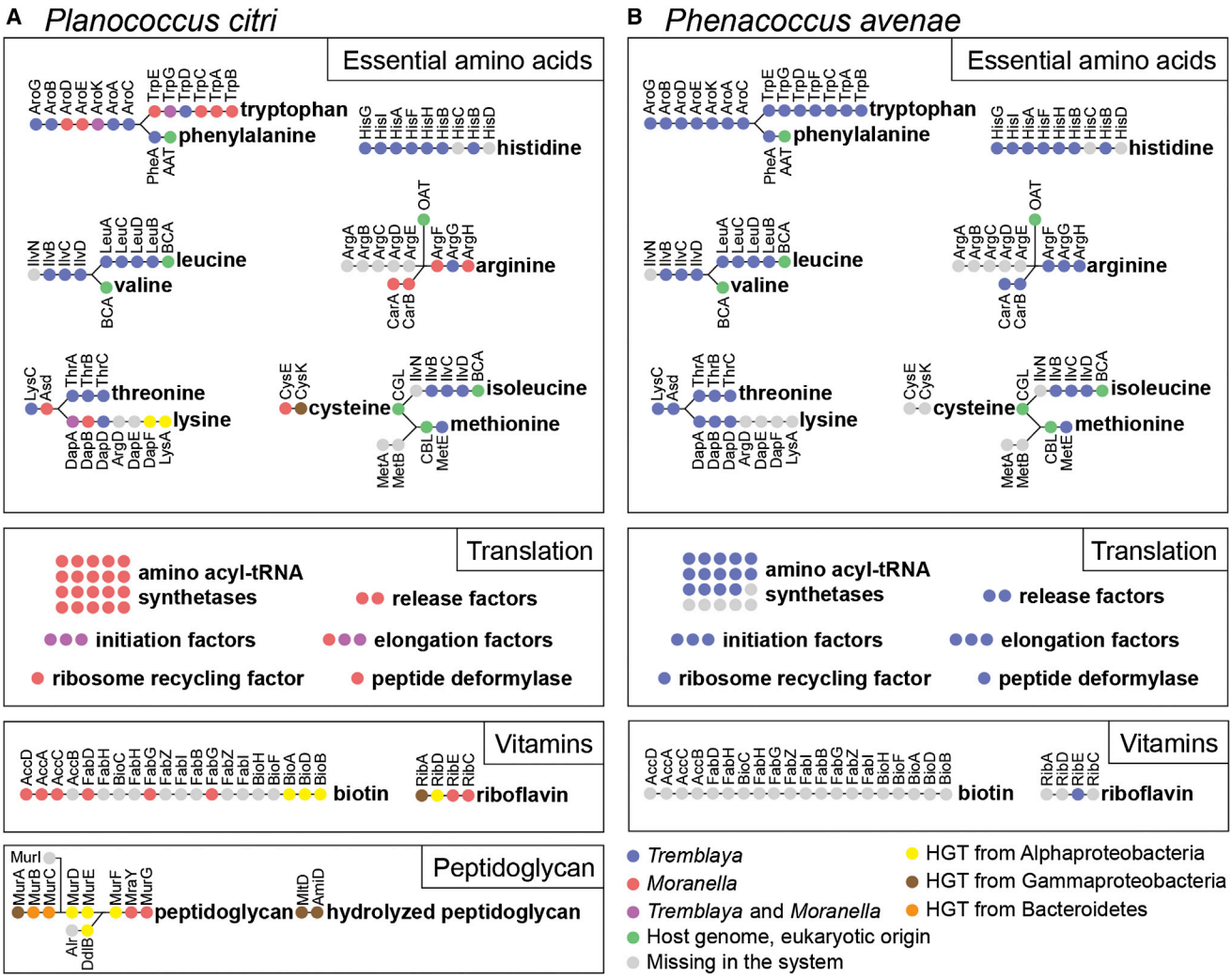

Figure 3. Symbiont Gene Retention and HTG Expression Patterns for the *Pl. citri* and *Ph. avenae* Symbioses

(A and B) We assume that because AAT, BCA, OAT, CGL, and CBL were found overexpressed in aphids (Hansen and Moran, 2011) and *Pl. citri*, they are also present and expressed in *Ph. avenae*; no direct data support the expression of these genes in *Ph. avenae*. See Table S2 for RT-qPCR verification that the ExHTGs shown here are expressed.

mealybugs with *Tremblaya* as the sole symbiont thrive on the same host plant as mealybugs with dual nested symbionts (Koga et al., 2013). These results indicate that both single- and dual-bacterial symbioses fulfill the same essential amino acid needs of their host insects. The single disparity in the *Pl. citri* and *Ph. avenae* symbiont pathways reflects a phylogenetic difference in tryptophan synthesis between the Betaproteobacteria and Gammaproteobacteria. In Betaproteobacteria, the indole-3-glycerol phosphate synthase (TrpC) and phosphoribosylanthranilate isomerase (TrpF) activities are encoded on separate proteins. In Gammaproteobacteria, activities are fused into one protein (TrpC).

We were struck by the observation that the histidine and lysine pathways remained incomplete in *Tremblaya* from both *Pl. citri* and *Ph. avenae*, with both genomes missing the same genes (*argD*, *dapE*, *dapF*, and *lysA* in lysine biosynthesis; *hisC* and *hisD* in histidine biosynthesis) (Figure 3). That identical gene retention patterns occur in symbionts of substantially diverged mealybugs strongly suggests that these pathways are actively maintained by selection in this incomplete state and indicates that the required intermediates or enzymes are somehow made available in both systems. We considered these pathway holes as prime candidates to be filled by genes acquired through HGT, and these enzymatic gaps in part motivated our search for genes horizontally transferred from *Tremblaya*, *Moranella*, or other unrelated bacteria to the insect host genome.

Transcriptomics Reveals Several Bacteria-to-Mealybug Horizontal Gene Transfer Events

We found at least 22 expressed horizontally transferred genes (ExHTGs) of bacterial origin on the *Pl. citri* nuclear genome (Table 1). This is a conservative estimate, as we considered only those

Table 1. The Expressed Horizontally Transferred Genes Found in This Work

Description (EC number)	Gene Name	Bacteriome Expression	Whole-Body Expression	Expression Ratio	Phylogenetic Origin
ExHTGs verified with phylogenetic analyses					
Cysteine synthase (EC: 2.5.1.47)	*cysK*	706.9	28.4	24.9	Gammaproteobacteria: Enterobacteriales
Tryptophan 2-monooxygenase oxidoreductase (EC: 1.13.12.3)	*tms1*	227.8	68.4	[3.3]	Gammaproteobacteria or Betaproteobacteria
Diaminopimelate decarboxylase (EC: 4.1.1.20)	*lysA*	204.4	9.4	21.7	Alphaproteobacteria: Rickettsiales
Fused deaminase/reductase (EC: 4.1.1.20)	*ribD*	174.2	7.9	21.9	Alphaproteobacteria: Rickettsiales
GTP cyclohydrolase (EC: 3.5.4.25)	*ribA*	142.2	3.8	37.5	Gammaproteobacteria: Enterobacteriales
Biotin synthase (EC: 2.8.1.6)	*bioB*	121.9	24.1	5.1	Alphaproteobacteria: Rickettsiales
Dethiobiotin synthase (EC: 6.3.3.3)	*bioD*	81.7	4.4	18.8	Alphaproteobacteria: Rickettsiales
Diaminopimelate epimerase (EC: 5.1.1.7)	*dapF*	74.3	2.3	32.6	Alphaproteobacteria: Rickettsiales
Adenosylmethionine-8-amino-7-oxononanoate transaminase (EC: 2.6.1.62)	*bioA*	74.3	2.9	25.4	Alphaproteobacteria: Rickettsiales
D-alanine-D-alanine ligase (EC: 6.3.2.4)	*ddlB*	49.9	1.6	31.8	Alphaproteobacteria: Rickettsiales
Beta-lactamase domain-containing protein	N/A	47.3	16.4	2.9	Gammaproteobacteria: Enterobacteriales
RNA methyltransferase (rlmI-like) (EC: 2.1.1.191)	*rlmI*	36.9	1.4	26.4	Gammaproteobacteria: Enterobacteriales
UDP-N-acetylglucosamine 1-carboxyvinyltransferase (EC: 2.5.1.7)	*murA*	21.3	0.9	23.6	Gammaproteobacteria: Enterobacteriales
UDP-n-acetylmuramate–L-alanine ligase (EC: 6.3.2.8)	*murC*	15.9	5.2	[3.1]	Bacteroidetes
UDP-N-acetylmuramoylalanyl-D-glutamyl diaminopimelate-D-alanyl-D-alanyl ligase (EC: 6.3.2.10)	*murF*	15.8	0.6	28.7	Alphaproteobacteria: Rickettsiales
UDP-N-acetylmuramoylalanine–D-glutamate ligase (EC: 6.3.2.9)	*murD*	13.6	1.7	7.8	Alphaproteobacteria: Rickettsiales
UDP-n-acetylmuramoylalanyl-D-glutamate diaminopimelate ligase (EC: 6.3.2.13)	*murE*	11.5	0.5	25.6	Alphaproteobacteria: Rickettsiales
UDP-N-acetylenolpyruvoylglucosamine reductase (EC: 1.1.1.158)	*murB*[a]	7.0	0.5	12.9	Bacteroidetes
Urea amidolyase [urea carboxylase/ allophanate hydrolase (EC: 6.3.4.6/ 3.5.1.54)]	DUR1,2	5.1	1.9	[2.7]	Gammaproteobacteria: Enterobacteriales
Lytic murein transglycosylase (EC: 3.2.1.-)	*mltB*	3.8	0.3	12.5	Gammaproteobacteria: Enterobacteriales
Glutamate-cysteine ligase-like protein	N/A	2.1	0.3	6.6	Gammaproteobacteria: Enterobacteriales
N-acetylmuramoyl-L-alanine amidase (EC: 3.5.1.28)	*amiD*	2.0	0.1	14.6	Gammaproteobacteria: Enterobacteriales
ExHTGs unverified by phylogenetic analyses					
AAA-type ATPase	N/A	102.3	2.9	35.2	Alphaproteobacteria: Rickettsiales[b]
Type III effector (skwp4/xopAD)	N/A	14.2	6.2	[2.3]	Betaproteobacteria or Gammaproteobacteria[b]
Ankyrin repeat domain protein	N/A	2.4	0.6	4.2	Alphaproteobacteria: Rickettsiales[b]

ExHTGs are ranked by their expression values in bacteriome tissue from highest to lowest. Expression information is included only for those transcripts meeting our criteria (blastx e-values less than 1×10^{-6} to a protein in GenBank nonredundant protein database (nr), FPKM values greater than 1 in bacteriome tissue, and expression ratios greater than 2); some transcripts showed evidence of either transcriptional isoforms or expression of paralogs but were excluded for clarity. Expression ratio refers to the ratio that the transcript showed in bacteriome tissue versus that found in the whole insect; those ratios determined not to be significantly different are shown in brackets.

[a]The terminal part of the *murB* transcript was broken in two sequences by the Trinity assembler.

[b]The bacterial nature of these transcripts was based only on sequence similarity, and they should therefore only be considered provisional HGT events. Transcripts for these three genes were present in many copies in the transcriptome, contain many repetitive sequences, and had poor assembly quality, so reliable phylogenetic analysis was not possible. See also Table S3.

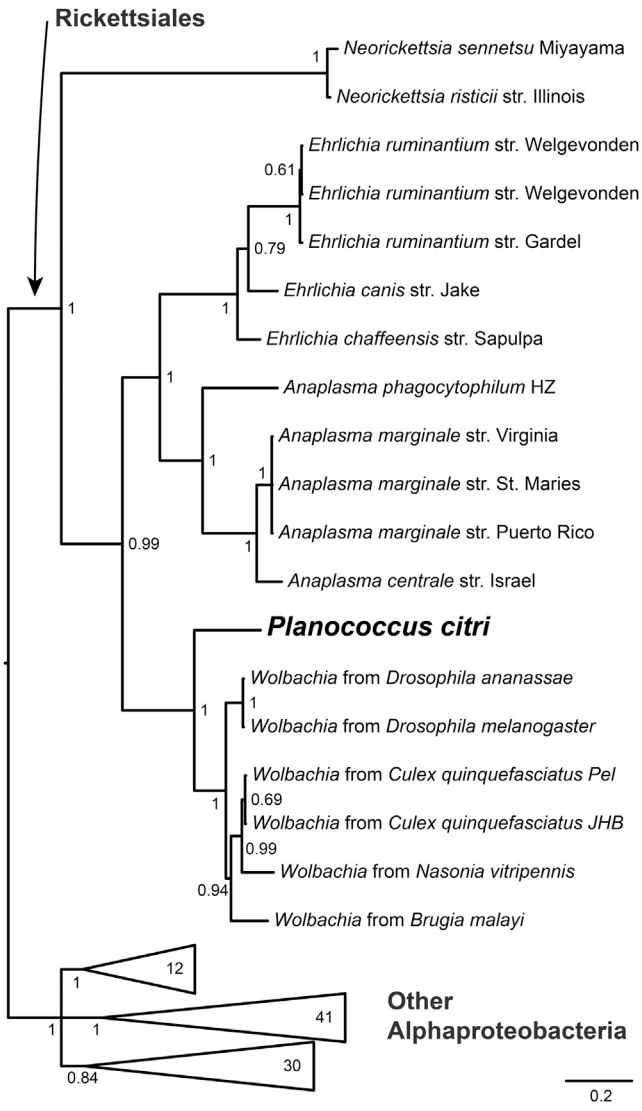

Figure 4. A Representative Phylogenetic Tree Confirming that RibD Is the Result of HGT

Posterior probabilities calculated from Markov chain Monte Carlo simulations on trees estimated using Bayesian inference methods are shown at each node. Collapsed branches are shown as triangular wedges with the number of sequences shown inside the wedge. Phylogenetic trees for the 21 other ExHTGs can be found in Data S1.

genes that had bacteriome FPMK expression values (fragments per kilobase of transcript per million fragments mapped [Trapnell et al., 2010]) greater than one to eliminate false positive reads (Ramsköld et al., 2009). We also required at least a two-fold greater expression value in the bacteriome tissue over the whole insect sample for a gene to be considered overexpressed. Although we did discover two ExHTGs related to lysine biosynthesis that appear to complement genes missing in the PCIT symbiotic system (*dapF* and *lysA*; Figure 3), we also found an unexpectedly large number of ExHTGs involved in the biosynthesis of other nutrients as well as in bacterial cell wall maintenance. Remarkably, the majority of these ExHTGs seem to complement genes that have been lost in *Tremblaya* and *Moranella*, and in some cases these ExHTGs complete biosynthetic pathways partially retained by *Moranella* (Figure 3). One ExHTG is involved in nonessential amino acid biosynthesis (*cysK*) and may complement *Moranella* in the two-step cysteine biosynthetic pathway; this gene could also take part in methionine synthesis by providing a substrate for insect cystathionine gamma-lyase (CGL). Five ExHTGs are involved in vitamin biosynthesis and together with genes retained in *Moranella* fill several gaps in the pathways for the production of riboflavin and biotin. Finally, five ExHTGs seem to complement the two retained functional genes and one pseudogene (*murC*) in *Moranella* involved in peptidoglycan biosynthesis, and two others are involved in peptidoglycan recycling. The expression of all 22 transcripts found by RNA-seq were verified by RT-qPCR (Table S2).

Phylogenetic Analyses Suggest the Source of Most ExHTGs Are Facultative Symbionts

The inferred phylogenetic positions of these ExHTGs suggest that facultative symbionts—i.e., bacteria that are not required for host survival—have been involved in HGT to the insect genome (*ribD* is shown in Figure 4; the remaining trees are shown in Data S1 in the order they are introduced in this paragraph). Six ExHTGs cluster within Rickettsiales (Alphaproteobacteria) as sister taxa to *Wolbachia* (*ribD*, *murDF*, and *ddlB*) or *Rickettsia* (*dapF*, and *murE*) clades. Two ExHTGs (*murBC*) cluster with *Cardinium* (Bacteroidetes), one (*cysK*) with *Sodalis* (Gammaproteobacteria), and one (GshA-like protein) with *Serratia symbiotica* (Gammaproteobacteria). The *bioABD* ExHGTs cluster with both Rickettsiales and with *Cardinium* species, consistent with previous work showing exchange of biotin genes between these two lineages (Penz et al., 2012); more thorough taxon sampling than currently available would be needed to determine which lineage acted as a donor of these genes in *Pl. citri*. Three other ExHGTs group with facultative symbionts from enterobacterial genera *Arsenophonus* (*ribA*, *amiD*) and *Sodalis* (*murA*) but are somewhat more distant, preventing us from making any deductions of their origins. Three ExHGTs (*mltB*, *rlmI*, and the beta-lactamase domain-containing protein) were identified as members of Enterobacteriaceae and one ExHGT was identified as a member of Rickettsiales (*lysA*), but their exact position could not be determined. The last two ExHTGs do not cluster with bacteria currently known to be facultative symbionts. These include DUR1,2 clustering within the enterobacterial genus *Pantoea* and *tms1* clustering with the proteobacterial genera *Pseudomonas* and *Ralstonia*. As none of the ExHTGs cluster confidently with Betaproteobacteria (*tms1* seems to have had a history of HGT between Gammaproteobacteria and Betaproteobacteria, preventing us from confidently inferring its phylogenetic origin), we conclude that *Tremblaya* has not been a major source of functional HGT to the mealybug nucleus. The *cysK* transfer groups with *Sodalis*, the closest sequenced relative of *Moranella*, indicating it is possible that this gene came from *Moranella*, but we lack the resolution to establish its origin at this time. We note that none of the putative source facultative symbionts are known to reside in the mealybug population used for RNA-seq (C.D.v.D., unpublished data) and thus seem to be signatures of historical, transient infections.

Verification that ExHTGs Are Encoded on the Insect Genome

Previous symbiont genome sequencing from *Pl. citri* bacteriomes found no other bacteria aside from *Tremblaya* and *Moranella* in the tissue at any appreciable level (McCutcheon and von Dohlen, 2011), suggesting that contamination is not a likely source of expression of the ExHTGs we find here. However, to provide stronger evidence that the ExHTGs we observed in the transcriptome data are encoded on the *Pl. citri* genome, we determined a rough low-pass insect draft genome of *Pl. citri*, using a line of insects isolated independently from the colony used for RNA-seq experiments (the transcriptome work was performed on insects from a greenhouse in Utah, USA, and the line used for the genome was isolated in London, England). With an average depth of coverage of 9.5 in k-mers (which corresponds to a base coverage of about 18× [Zerbino and Birney, 2008]), a scaffold N50 of 5,114, and a maximum scaffold size of 79,414 nts, the assembly was low quality but nevertheless confirmed that the ExHTGs we observed in the transcriptome assembly were very likely encoded on the insect genome.

That these scaffolds are from the insect genome and not from contaminating bacteria is supported by several lines of evidence (Table S3). First, 10 of the 22 ExHTGs are on scaffolds that include regions of sequence most closely resembling genes from other insects. Second, aligning the transcripts to the draft *Pl. citri* genome clearly showed that 9 of the 22 ExHTGs contain spliced canonical eukaryotic GT-AG introns. Interestingly, in five cases the introns are just upstream of the ExHTG open reading frame. Introns located immediately 5′ of start codons have been shown to increase gene expression in several eukaryotes (Rose et al., 2011), although it is unclear what function these introns have in this system. In all, 15 of the 22 ExHTGs are either coassembled with a putative insect gene, or found on a transcript that has functional introns (or in four cases, both). The remaining seven ExHTGs are found on scaffolds ranging in size from 1,938 to 10,645 bps in length, which do not encode any other bacterial open reading frame other than the ExHTG (in some cases, tandem duplicates of the gene are clearly present, see Table S3). A typical bacterial genome encodes approximately one gene per kilobase (Ochman and Davalos, 2006), so in most of these cases if the scaffold was from a bacterial contaminant it would be expected to encode at least one other bacterial gene. Thus, we conclude that most, if not all, of the ExHTGs we find in our transcriptomic experiments are encoded on the mealybug genome.

Probable but Unconfirmed ExHTGs

We found several transcripts for three protein families containing highly repetitive sequences: ankyrin repeat domain proteins (ANK), ATPases associated with various cellular activities (AAA-ATPases), and type III effector proteins (Table 1). These transcripts all show sequence similarity to bacterial proteins, but their low-complexity repetitive regions made conclusive phylogenetic proof of HGT difficult. We therefore consider these probable but unconfirmed HGTs.

In general, the discovery of such a large number of bacterial genes expressed from the *Pl. citri* genome implies that it may also encode several HGT relics because it is likely that the majority of HGT events result in the transfer of nonfunctional DNA that is not expressed and not subject to purifying selection. Because our genome assembly is not yet of sufficient quality to fully describe the transfer events that have occurred in *Pl. citri*, it is important to note that we are likely underestimating the level of bacteria-to-mealybug HGT that has occurred in this system.

DISCUSSION

The Role of *Moranella* in *Tremblaya*'s Extreme Genome Degeneracy

We hypothesized that if missing genes in *Tremblaya* PCIT are primarily complemented from gene products of the insect host, then *Tremblaya* from mealybug lineages lacking *Moranella* should have a similarly degenerate genome to *Tremblaya* PCIT. Conversely, if missing genes are primarily complemented by *Moranella* in the *Pl. citri* symbiosis, we hypothesized that *Tremblaya* from mealybug lineages lacking *Moranella* should have a more robust genome, perhaps similar in gene density and coding capacity to those found in other symbionts. By completing a *Tremblaya* genome from *Phenacoccus avenae*, a lineage lacking the intrabacterial symbiont *Moranella*, we have shown that genome reduction in *Tremblaya* occurs to a degree consistent with other previously reported tiny symbiont genomes when present as the sole symbiont. We also show that *Tremblaya* PCIT is an almost perfect subset of *Tremblaya* PAVE. These results suggest that much of the reductive genome evolution observed in *Tremblaya* (down to approximately 170 kb) occurred before the acquisition of *Moranella* in the common ancestor of *Pl. citri* and *Ph. avenae* and that the extreme genomic degeneracy observed in *Tremblaya* PCIT (from 170 kb to 140 kb) was likely due to the acquisition of *Moranella* by *Tremblaya* at some point in the lineage leading to *Pl. citri.* This scenario is consistent with studies showing that massive and rapid gene loss can occur in bacteria that transition to a symbiotic lifestyle (Mira et al., 2001; Moran and Mira, 2001; Nilsson et al., 2005), after which gene loss slows, and gross genomic changes become infrequent, even over hundreds of millions of years (McCutcheon and Moran, 2010; Tamas et al., 2002; van Ham et al., 2003). Assuming this model, the acquisition of *Moranella* would break *Tremblaya*'s genomic stability by relaxing selection on genes redundant with *Moranella*; this would allow further genomic erosion in *Tremblaya* and would account for its large number of pseudogenes and unusually small gene set. Our results suggest that the primary driving force shaping *Tremblaya* PCIT's extreme genomic degeneracy—for example, the loss of all aminoacyl-tRNA synthetases and its unusually low coding density—was the acquisition of *Moranella* into its cytoplasm. However, these comparative genomic data do not speak to the role of the host in the maintenance of this symbiosis, and they do not directly prove that symbiont genes have not been transferred to the host genome.

We took a transcriptomic approach to address the role of the host in the PCIT symbiosis and to test for expressed genes resulting from bacteria-to-insect transfer events. Although the vast majority of microorganism-to-animal HGT events have been discovered through genome sequencing projects, an interesting counterexample comes from the pea aphid, where early transcriptomic experiments, using only 2,600 expressed

Table 2. Expression Values for Selected Insect Transcripts

Description (EC number)	Gene Name	Bacteriome Expression	Whole-Body Expression	Expression Ratio
Cystathionine beta-lyase, cystathionine gamma-lyase (4.4.1.8/4.4.1.10)	CBL, CGL	2553.3	114.3	22.3
Glutamine synthetase (6.3.1.2)	GS	1567.3	229.4	6.8
Kynurenine-oxoglutarate transaminase (2.6.1.7)	KAT	666.6	74.9	8.9
Aspartate aminotransferase (2.6.1.1)	AAT	427.9	85.44	5.0
Phosphoserine aminotransferase (2.6.1.52)	PSAT	366.6	69.2	5.3
Branched-chain amino acid aminotransferase (2.6.1.42)	BCA	363.0	25.4	14.3
Homocysteine S-methyltransferase (2.1.1.10)	HMT	210.7	34.4	6.1
Glutamine oxoglutarate aminotransferase (1.4.1.13)	GOGAT	85.7	17.2	5.0
Putative riboflavin transporter	N/A	57.6	6.4	9.0

Transcripts are ranked by their expression values in bacteriome tissue from highest to lowest. Expression information is included only for those transcripts meeting our criteria (blastx e-values less than 1×10^{-6} to a protein in nr, FPKM values greater than 1, and expression ratios greater than 2); some copies of transcripts showing evidence of either transcriptional isoforms or expression of paralogs were excluded for clarity. Expression ratio refers to the ratio that the transcript showed in bacteriome tissue versus that found in the whole insect.

sequence tags (ESTs), uncovered two genes of bacterial origin in the aphid genome that were upregulated in aphid bacteriomes, *ldcA*, and *rplA* (Nakabachi et al., 2005). When the pea aphid genome was sequenced more recently (International Aphid Genomics Consortium, 2010), eight apparently functional genes of alphaproteobacterial origin were found (*ldcA*, *amiD*, *bLys*, and five copies of *rplA*), although only *ldcA*, *amiD*, and *rplA1-5* were found to be upregulated in bacteriocytes (Nikoh et al., 2010). Thus, as a very low level of transcriptome sequencing found two of three functional bacterial gene families that were expressed in aphid bacteriocytes, we reasoned that a high-throughput transcriptomics experiment would uncover most or all of the ExHTGs that are supporting the *Pl. citri* symbiosis. We note that none of the horizontally transferred and expressed genes discovered in the pea aphid system seem to directly support the symbiotic role of *Buchnera*—i.e., nutrient production—but two genes, *ldcA* and *amiD*, are possibly involved in peptidoglycan recycling (Nikoh and Nakabachi, 2009; Nikoh et al., 2010). The *amiD* transfer we find in *Pl. citri* was independent of the aphid event as the donor bacteria are from different phylogenetic groups.

Several Pathways Are Composed of Genes from Multiple Phylogenetic Sources

Previous work has shown that bacteria from the class Alphaproteobacteria are common donors of HTGs in insects (Dunning Hotopp, 2011). Our results are consistent with these findings, with ten ExHTGs grouping closely with other alphaproteobacterial sequences in phylogenetic trees (Figure 4 and Data S1). However, we also find nine ExHTGs from Gammaproteobacteria, two from Bacteroidetes, and one that is phylogenetically unresolved (Data S1). At least six distinct lineages of organisms therefore contribute to the *Pl. citri* symbiosis: the mealybug itself; *Moranella*; *Tremblaya* PCIT; and, through HGT, various bacteria in the Alphaproteobacteria, Gammaproteobacteria, and Bacteroidetes. Remarkably, these genes of diverse phylogenetic origins, now encoded on three different genomes, seem to be used in concert in some metabolic pathways (Figure 3). For example, the production and recycling of peptidoglycan uses three ExHTGs of gammaproteobacterial origin (*murA*, *mltD*, and *amiD*), four ExHTGs of alphaproteobacterial origin (*murDEF* and *ddlB*), two ExHTGs from Bacteroidetes (*murBC*), and two genes encoded on the *Moranella* genome (*mraY* and *murG*). Similarly, riboflavin biosynthesis requires two *Moranella* genes (*ribE* and *ribC*), an ExHTG of gammaproteobacterial origin (*ribA*), and an ExHTG of alphaproteobacterial origin (*ribD*). Although we do not have direct proof that these nutrients are produced by the metabolic mosaic shown in Figure 3, we do find an insect riboflavin transporter significantly upregulated in bacteriome tissue (Table 2), suggesting that the symbiosis is producing and utilizing riboflavin. Coincidentally, this riboflavin transporter happens to be encoded on a 32 kb scaffold containing the ExHTG *cysK*.

Of note, our results point to several interesting metabolic similarities and differences with other insect symbioses. As in the pea aphid system (Hansen and Moran, 2011; Wilson et al., 2010), *Pl. citri* may use homocysteine S-methyltransferase (2.1.1.10) to produce S-adenosylhomocysteine and methionine and uses glutamine synthetase and glutamine oxoglutarate aminotransferase (6.3.1.2/1.4.1.13, the GS/GOGAT cycle) for recycling ammonia into glutamate; glutamate could then be used by host aminotransferases to incorporate ammonium-derived nitrogen into symbiont-synthesized carbon skeletons of Phe, Leu, Ile, Val, and possibly Lys and His (Hansen and Moran, 2011). Interestingly, one of the ExHTG candidates is urea amidolyase, or DUR1,2 (Table 1), an enzyme that degrades urea into ammonia and CO_2. This suggests that, contrary to the single-step cleavage of urea by ATP-independent urease in the symbionts of cockroaches and carpenter ants (Gil et al., 2003; López-Sánchez et al., 2009; Sabree et al., 2009), mealybugs use the ATP-dependent route catalyzed by DUR1,2. Thus, like the cockroach and carpenter ant systems, mealybugs may have the ability to recycle urea but through a different pathway resulting from a horizontal gene transfer. In all three systems, toxic ammonium can then be recycled by glutamine synthetase (Table 2) into amino acids.

Host Genes of Eukaryotic Origin Overexpressed in Bacteriome Tissue

Reduced genomes of insect symbionts often encode metabolic pathways missing one or two gene homologs (McCutcheon, 2010; McCutcheon and Moran, 2012; Zientz et al., 2004). The loss of an essential biosynthetic gene in an otherwise conserved symbiont pathway is commonly explained by the presence of a host homolog, or by another promiscuous symbiotic/host gene that can compensate for the missing activity. In the pea aphid-*Buchnera* system, the role of the host in supplementing missing *Buchnera* activities was recently corroborated by transcriptomic and proteomic work (Hansen and Moran, 2011; Macdonald et al., 2012; Poliakov et al., 2011); our data from the mealybug system strongly support intimate host-symbiont cooperation in mealybugs, and suggest that it is a general feature of plant-sap-feeding insect symbioses. Accordingly, host enzymes originally hypothesized to complement missing symbiotic genes in production of essential amino acids (McCutcheon and von Dohlen, 2011)—BCA (2.6.1.42), AAT (2.6.1.1), OAT (2.6.1.13), CGL (4.4.1.1), and CBL (4.4.1.8)—are all significantly upregulated in mealybug bacteriocytes (Table 2). As in the *Buchnera*-pea aphid system (Hansen and Moran, 2011), TDH (4.3.1.19) activity was found not to be upregulated in mealybug bacteriocytes. It therefore seems likely that the source of 2-oxobutanoate, the metabolite required for isoleucine biosynthesis originally predicted to be produced by TDH (McCutcheon and von Dohlen, 2011), is available in both aphids and mealybugs from the activity of CGL (4.4.1.1), which is overexpressed in bacteriome tissue in both aphids (Hansen and Moran, 2011; Poliakov et al., 2011) and mealybugs (Table 2).

As our work did not identify any ExHTGs for four of six genes missing in lysine (*argD* and *dapE*) and histidine (*hisC* and *hisD*) biosynthetic pathways, these remaining enzymatic holes are candidates for complementation by host-encoded enzymes of eukaryotic origin. Two of the missing genes (*argD* and *hisC*) are aminotransferases, a class of enzymes that display remarkable plasticity in the reactions they catalyze (Carbonell et al., 2011; Rothman and Kirsch, 2003) and that play crucial roles in the *Buchnera*-aphid symbiosis (Hansen and Moran, 2011; Macdonald et al., 2012; Poliakov et al., 2011; Wilson et al., 2010). As there is only one aminotransferase gene retained in the *Moranella* genome (*serC*), and none in *Tremblaya* PCIT, this particular enzymatic activity has probably been largely taken over by the insect. We therefore hypothesize that ArgD and HisC activities can be compensated by one (or more) of several host aminotransferases that are upregulated in bacteriocytes (Table 2). Similarly, HisD is an NAD-like dehydrogenase, and this activity may also be replaceable by host dehydrogenases, although no obvious candidate is clear from our work. Finally, the *dapE* (N-succinyl-L-diaminopimelate desuccinylase) gene homolog has also been lost from several other symbiotic genomes (e.g., from *Sulcia* and its cosymbionts [McCutcheon and Moran, 2010]), although, like previous work, our data do not point to an obvious candidate enzyme that carries out this chemistry.

The overall picture of amino acid biosynthesis in mealybugs implies that the host insect is directly involved in production of phenylalanine, leucine, valine, isoleucine, lysine, methionine, and possibly histidine. Remarkably, only tryptophan and threonine are produced from pathways independent of host-derived gene products.

Host Control of Peptidoglycan Biosynthesis and Its Relation to *Moranella*

The presence of a large number of ExHTGs involved in peptidoglycan production and recycling (Figure 3 and Table 1) is consistent with the hypothesis that cell lysis is the mechanism used to share gene products between *Moranella* and *Tremblaya* PCIT (Koga et al., 2013; McCutcheon and von Dohlen, 2011). This idea was initially suggested based on a lack of transporters encoded on the *Moranella* genome combined with the large number of gene products or metabolites involved in essential amino acid biosynthesis and translation that would need to pass between *Moranella* and *Tremblaya* PCIT for the symbiosis to function (McCutcheon and von Dohlen, 2011). Subsequent electron microscopy on mealybugs closely related to *Pl. citri* showed that although most gammaproteobacterial cells infecting the *Tremblaya* cytoplasm were rod shaped, some were amorphous blobs seemingly in a state of degeneration (Koga et al., 2013). Our results suggest a plausible mechanism for how the insect host controls this process: by differentially controlling the expression of the horizontally transferred *murABCDEF* and *mltD/amiD* genes, the host could regulate the cell wall stability of *Moranella*. Increasing the expression of *murABCDE* genes would increase the integrity of *Moranella*'s cell wall, and increasing the expression of *mltD/amiD* would tend to decrease *Moranella*'s cell wall strength. As *Tremblaya* PCIT encodes no cell-envelope-related genes and likely uses host-derived membranes to define its cytoplasm, it would be unaffected by changes in gene expression related to peptidoglycan biosynthesis. This hypothesis is testable, because the levels of *Tremblaya* and *Moranella* are uncoupled in mealybugs closely related to *Pl. citri*; in males in particular, *Moranella* levels drop to undetectable levels while *Tremblaya* persists (Kono et al., 2008). In situations where *Moranella* is reduced with respect to *Tremblaya*, we would expect low expression of *murABCDEF* and increased expression of *mltD/amiD*. Interestingly, we find that of the five ExHTGs with recognizable eukaryotic signal peptides, four are involved in peptidoglycan metabolism (*amiD*, *mltD*, *murF*, and *murD*; the other ExHTG with a signal peptide is *rlmI*).

Tremblaya's Extreme Genomic Degeneracy and Its Implications for Understanding Intimate Mutualisms

The smallest reported bacterial genomes, which are all from nutritional symbionts of sap-feeding insects, are indistinguishable from organelles when considered only in terms of genome size and gene number (McCutcheon and Moran, 2012). Unlike organelles, however, they tend to retain a certain set of the most critical genes involved in DNA replication, transcription, and translation (McCutcheon, 2010). *Tremblaya* PCIT is strikingly different, as it has lost many genes involved in translation that are retained in other highly reduced genomes (López-Madrigal et al., 2011; McCutcheon and von Dohlen, 2011). This degeneracy, along with its extensive interdependency on *Moranella* and the insect host, makes it difficult to apply an appropriate label to *Tremblaya* PCIT—is it still a bacterium or has it transitioned to something more akin to an organelle? This labeling problem is

complicated by the lack of a generally accepted definition of "organelle" (Keeling, 2011; Keeling and Archibald, 2008; Theissen and Martin, 2006). In any case, more important than applying an appropriate label to *Tremblaya* is understanding how the *Pl. citri* symbiosis came to be and how it currently works, as this may provide insight on how host-organelle relationships formed in the general sense of being highly integrated mosaic organisms.

Here, we show that the extreme genomic degeneracy of *Tremblaya* PCIT—that is, its low coding density and loss of critical translation-related genes—is largely the result of the presence of *Moranella* in its cytoplasm. These results are consistent with the hypothesis that *Moranella* is providing many gene products or metabolites to *Tremblaya* PCIT, including those involved in essential amino acid production and translation. Our data also show the *Pl. citri* symbiosis is reliant on a mosaic of gene products from no fewer than six distinct organisms: the mealybug itself, *Tremblaya* PCIT, *Moranella*, and at least three bacterial groups that were donors of HTGs residing on the insect nuclear genome. Importantly, we did not find evidence of functional HGT events from *Tremblaya* PCIT to the host insect genome. Thus, genome reduction in *Tremblaya* was not associated with functional transfer of its genes to the host nucleus and therefore has not paralleled processes that have occurred in the evolution of organelles.

EXPERIMENTAL PROCEDURES

Additional information on the computational and experimental methods used here can be found in the Extended Experimental Procedures available online.

Insect Strains, DNA and RNA Isolation, and Sequencing

For sequencing the *Tremblaya* PAVE genome, DNA was isolated from the bacteriome of a laboratory-maintained individual and was amplified using phi29-based rolling circle amplification and subjected to 454 library creation and sequencing (see Figure S1 for the Southern blot of the PAVE plasmid-like molecule). For bacteriome mRNA-seq, total RNA was extracted from 20 dissected mealybug bacteriomes and whole female bodies as reported previously (McCutcheon and von Dohlen, 2011) and was subjected to Illumina library creation and sequencing. For *Pl. citri* draft genome sequencing, DNA was isolated from a single adult female from a colony that had undergone several rounds of inbreeding. The *Pl. citri* strain used in RNA-seq was from a greenhouse colony in Logan, UT, USA, and the *Pl. citri* strain used to generate the draft genome was from a colony in London, England, UK. As a result, the transcriptome and draft genome show some sequence divergence.

ACCESSION NUMBERS

The GenBank accession numbers for the Tremblaya PAVE genome reported in this paper are CP003982 (main chromosome) and CP003983 (plasmid). The GenBank Sequence Read Archive number for the raw transcriptome and genome reads is SRP021919. The GenBank accession numbers for the assembled ExHTG and host transcriptome contigs listed in Tables 1 and 2 are KF021954–KF021987, and KF021932–KF021953 for the associated ExHTG genome scaffolds.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, one figure, three tables, and one supplemental data file and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2013.05.040.

ACKNOWLEDGMENTS

We thank Dan Vanderpool, Yu Matsuura, Kaoru Nikoh, and Dionna Norris for technical and experimental assistance, Minyong Chung for facilitating genome sequencing, Jesse Johnson for access to computing resources, and Nobuo Sawamura and Junko Makino for insect samples. The authors declare no conflicts of interest in this work. F.H. was supported by the Grant Agency of the Czech Republic (P505/10/1401 and 13-01878S). T.F. and N.N. were supported by the Program for Promotion of Basic and Applied Researches for Innovations in Bio-oriented Industry (BRAIN) and by KAKENHI (22128001 and 22128007). L.R. was supported by the Royal Society and Somerville College, University of Oxford. R.P.D. was supported by an NSF Graduate Research Fellowship. C.D.v.D. was supported by the Utah Agricultural Experiment Station. D.B. was supported by NIH grants (R01GM076007 and R01GM093182) and a Packard Fellowship. A.C.C.W. was supported by University of Miami start-up funds and NSF award IOS-1121847. J.P.M. was supported by NSF award IOS-1256680, the Montana NSF-EPSCoR award EPS0701906, and is an Associate in the Integrated Microbial Biodiversity Program of the Canadian Institute for Advanced Research.

Received: January 14, 2013
Revised: May 1, 2013
Accepted: May 22, 2013
Published: June 20, 2013

REFERENCES

Acuña, R., Padilla, B.E., Flórez-Ramos, C.P., Rubio, J.D., Herrera, J.C., Benavides, P., Lee, S.J., Yeats, T.H., Egan, A.N., Doyle, J.J., and Rose, J.K. (2012). Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. Proc. Natl. Acad. Sci. USA *109*, 4197–4202.

Aikawa, T., Anbutsu, H., Nikoh, N., Kikuchi, T., Shibata, F., and Fukatsu, T. (2009). Longicorn beetle that vectors pinewood nematode carries many *Wolbachia* genes on an autosome. Proc. Biol. Sci. *276*, 3791–3798.

Altincicek, B., Kovacs, J.L., and Gerardo, N.M. (2012). Horizontally transferred fungal carotenoid genes in the two-spotted spider mite *Tetranychus urticae*. Biol. Lett. *8*, 253–257.

Baumann, P. (2005). Biology bacteriocyte-associated endosymbionts of plant sap-sucking insects. Annu. Rev. Microbiol. *59*, 155–189.

Buchner, P. (1965). Endosymbiosis of animals with plant microorganisms (New York: John Wiley & Sons).

Carbonell, P., Lecointre, G., and Faulon, J.L. (2011). Origins of specificity and promiscuity in metabolic networks. J. Biol. Chem. *286*, 43994–44004.

Chang, Y.J., Land, M., Hauser, L., Chertkov, O., Del Rio, T.G., Nolan, M., Copeland, A., Tice, H., Cheng, J.F., Lucas, S., et al. (2011). Non-contiguous finished genome sequence and contextual data of the filamentous soil bacterium Ktedonobacter racemifer type strain (SOSP1-21). Stand. Genomic Sci. *5*, 97–111.

Danchin, E.G.J., Rosso, M.N., Vieira, P., de Almeida-Engler, J., Coutinho, P.M., Henrissat, B., and Abad, P. (2010). Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. Proc. Natl. Acad. Sci. USA *107*, 17651–17656.

Doudoumis, V., Tsiamis, G., Wamwiri, F., Brelsfoard, C., Alam, U., Aksoy, E., Dalaperas, S., Abd-Alla, A., Ouma, J., Takac, P., et al. (2012). Detection and characterization of *Wolbachia* infections in laboratory and natural populations of different species of tsetse flies (genus *Glossina*). BMC Microbiol. *12*(*Suppl 1*), S3.

Douglas, A.E. (1989). Mycetocyte symbiosis in insects. Biol. Rev. Camb. Philos. Soc. *64*, 409–434.

Dunning Hotopp, J.C. (2011). Horizontal gene transfer between bacteria and animals. Trends Genet. *27*, 157–163.

Dunning Hotopp, J.C., Clark, M.E., Oliveira, D.C., Foster, J.M., Fischer, P., Muñoz Torres, M.C., Giebel, J.D., Kumar, N., Ishmael, N., Wang, S., et al. (2007).

Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science *317*, 1753–1756.

Gil, R., Silva, F.J., Zientz, E., Delmotte, F., González-Candelas, F., Latorre, A., Rausell, C., Kamerbeek, J., Gadau, J., Hölldobler, B., et al. (2003). The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. Proc. Natl. Acad. Sci. USA *100*, 9388–9393.

Gladyshev, E.A., Meselson, M., and Arkhipova, I.R. (2008). Massive horizontal gene transfer in bdelloid rotifers. Science *320*, 1210–1213.

Grbić, M., Van Leeuwen, T., Clark, R.M., Rombauts, S., Rouzé, P., Grbić, V., Osborne, E.J., Dermauw, W., Ngoc, P.C., Ortego, F., et al. (2011). The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature *479*, 487–492.

Gruwell, M.E., Hardy, N.B., Gullan, P.J., and Dittmar, K. (2010). Evolutionary relationships among primary endosymbionts of the mealybug subfamily phenacoccinae (hemiptera: Coccoidea: Pseudococcidae). Appl. Environ. Microbiol. *76*, 7521–7525.

Hansen, A.K., and Moran, N.A. (2011). Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. Proc. Natl. Acad. Sci. USA *108*, 2849–2854.

Hardy, N.B., Gullan, P.J., and Hodgson, C.J. (2008). A subfamily-level classification of mealybugs (Hemiptera: Pseudococcidae) based on integrated molecular and morphological data. Syst. Entomol. *33*, 51–71.

International Aphid Genomics Consortium. (2010). Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biol. *8*, e1000313.

Keeling, P.J. (2011). Endosymbiosis: bacteria sharing the load. Curr. Biol. *21*, R623–R624.

Keeling, P.J., and Archibald, J.M. (2008). Organelle evolution: what's in a name? Curr. Biol. *18*, R345–R347.

Keeling, P.J., and Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. *9*, 605–618.

Klasson, L., Kambris, Z., Cook, P.E., Walker, T., and Sinkins, S.P. (2009). Horizontal gene transfer between *Wolbachia* and the mosquito *Aedes aegypti*. BMC Genomics *10*, 33.

Koga, R., Nikoh, N., Matsuura, Y., Meng, X.Y., and Fukatsu, T. (2013). Mealybugs with distinct endosymbiotic systems living on the same host plant. FEMS Microbiol. Ecol. *83*, 93–100.

Kondo, N., Nikoh, N., Ijichi, N., Shimada, M., and Fukatsu, T. (2002). Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. Proc. Natl. Acad. Sci. USA *99*, 14280–14285.

Kono, M., Koga, R., Shimada, M., and Fukatsu, T. (2008). Infection dynamics of coexisting beta- and gammaproteobacteria in the nested endosymbiotic system of mealybugs. Appl. Environ. Microbiol. *74*, 4175–4184.

López-Madrigal, S., Latorre, A., Porcar, M., Moya, A., and Gil, R. (2011). Complete genome sequence of "Candidatus Tremblaya princeps" strain PCVAL, an intriguing translational machine below the living-cell status. J. Bacteriol. *193*, 5587–5588.

López-Sánchez, M.J., Neef, A., Peretó, J., Patiño-Navarrete, R., Pignatelli, M., Latorre, A., and Moya, A. (2009). Evolutionary convergence and nitrogen metabolism in *Blattabacterium* strain Bge, primary endosymbiont of the cockroach *Blattella germanica*. PLoS Genet. *5*, e1000721.

Macdonald, S.J., Lin, G.G., Russell, C.W., Thomas, G.H., and Douglas, A.E. (2012). The central role of the host cell in symbiotic nitrogen metabolism. Proc. Biol. Sci. *279*, 2965–2973.

McCutcheon, J.P. (2010). The bacterial essence of tiny symbiont genomes. Curr. Opin. Microbiol. *13*, 73–78.

McCutcheon, J.P., and Moran, N.A. (2010). Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. Genome Biol. Evol. *2*, 708–718.

McCutcheon, J.P., and Moran, N.A. (2012). Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. *10*, 13–26.

McCutcheon, J.P., and von Dohlen, C.D. (2011). An interdependent metabolic patchwork in the nested symbiosis of mealybugs. Curr. Biol. *21*, 1366–1372.

McCutcheon, J.P., McDonald, B.R., and Moran, N.A. (2009). Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. PLoS Genet. *5*, e1000565.

Mira, A., Ochman, H., and Moran, N.A. (2001). Deletional bias and the evolution of bacterial genomes. Trends Genet. *17*, 589–596.

Moran, N.A. (2007). Symbiosis as an adaptive process and source of phenotypic complexity. Proc. Natl. Acad. Sci. USA *104*(*Suppl 1*), 8627–8633.

Moran, N.A., and Jarvik, T. (2010). Lateral transfer of genes from fungi underlies carotenoid production in aphids. Science *328*, 624–627.

Moran, N.A., and Mira, A. (2001). The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. Genome Biol. *2*, 0054.

Nakabachi, A., Shigenobu, S., Sakazume, N., Shiraki, T., Hayashizaki, Y., Carninci, P., Ishikawa, H., Kudo, T., and Fukatsu, T. (2005). Transcriptome analysis of the aphid bacteriocyte, the symbiotic host cell that harbors an endocellular mutualistic bacterium, *Buchnera*. Proc. Natl. Acad. Sci. USA *102*, 5477–5482.

Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H.E., Moran, N.A., and Hattori, M. (2006). The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. Science *314*, 267.

Nikoh, N., and Nakabachi, A. (2009). Aphids acquired symbiotic genes via lateral gene transfer. BMC Biol. *7*, 12.

Nikoh, N., Tanaka, K., Shibata, F., Kondo, N., Hizume, M., Shimada, M., and Fukatsu, T. (2008). Wolbachia genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes. Genome Res. *18*, 272–280.

Nikoh, N., McCutcheon, J.P., Kudo, T., Miyagishima, S.Y., Moran, N.A., and Nakabachi, A. (2010). Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. PLoS Genet. *6*, e1000827.

Nilsson, A.I., Koskiniemi, S., Eriksson, S., Kugelberg, E., Hinton, J.C., and Andersson, D.I. (2005). Bacterial genome size reduction by experimental evolution. Proc. Natl. Acad. Sci. USA *102*, 12112–12116.

Ochman, H., and Davalos, L.M. (2006). The nature and dynamics of bacterial genomes. Science *311*, 1730–1733.

Penz, T., Schmitz-Esser, S., Kelly, S.E., Cass, B.N., Müller, A., Woyke, T., Malfatti, S.A., Hunter, M.S., and Horn, M. (2012). Comparative genomics suggests an independent origin of cytoplasmic incompatibility in *Cardinium hertigii*. PLoS Genet. *8*, e1003012.

Poliakov, A., Russell, C.W., Ponnala, L., Hoops, H.J., Sun, Q., Douglas, A.E., and van Wijk, K.J. (2011). Large-scale label-free quantitative proteomics of the pea aphid-*Buchnera* symbiosis. Mol. Cell. Proteomics *10*. M110, 007039. Published online March 18, 2012.

Ramsköld, D., Wang, E.T., Burge, C.B., and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput. Biol. *5*, e1000598.

Rose, A.B., Emami, S., Bradnam, K., and Korf, I. (2011). Evidence for a DNA-Based Mechanism of Intron-Mediated Enhancement. Front Plant Sci *2*, 98.

Rothman, S.C., and Kirsch, J.F. (2003). How does an enzyme evolved in vitro compare to naturally occurring homologs possessing the targeted function? Tyrosine aminotransferase from aspartate aminotransferase. J. Mol. Biol. *327*, 593–608.

Sabree, Z.L., Kambhampati, S., and Moran, N.A. (2009). Nitrogen recycling and nutritional provisioning by *Blattabacterium*, the cockroach endosymbiont. Proc. Natl. Acad. Sci. USA *106*, 19521–19526.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature *407*, 81–86.

Sloan, D.B., and Moran, N.A. (2012). Genome reduction and co-evolution between the primary and secondary bacterial symbionts of psyllids. Mol. Biol. Evol. *29*, 3781–3792.

Tamas, I., Klasson, L., Canbäck, B., Näslund, A.K., Eriksson, A.S., Wernegreen, J.J., Sandström, J.P., Moran, N.A., and Andersson, S.G. (2002). 50

million years of genomic stasis in endosymbiotic bacteria. Science *296*, 2376–2379.

Thao, M.L., Gullan, P.J., and Baumann, P. (2002). Secondary (gamma-Proteobacteria) endosymbionts infect the primary (beta-Proteobacteria) endosymbionts of mealybugs multiple times and coevolve with their hosts. Appl. Environ. Microbiol. *68*, 3190–3197.

Theissen, U., and Martin, W. (2006). The difference between organelles and endosymbionts. Curr. Biol. *16*, R1016–R1017, author reply R1017–R1018.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., and Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. *5*, 123–135.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

van Ham, R.C., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U., Fernández, J.M., Jiménez, L., Postigo, M., Silva, F.J., et al. (2003). Reductive genome evolution in *Buchnera aphidicola*. Proc. Natl. Acad. Sci. USA *100*, 581–586.

von Dohlen, C.D., Kohler, S., Alsop, S.T., and McManus, W.R. (2001). Mealybug beta-proteobacterial endosymbionts contain gamma-proteobacterial symbionts. Nature *412*, 433–436.

Werren, J.H., Richards, S., Desjardins, C.A., Niehuis, O., Gadau, J., Colbourne, J.K., Werren, J.H., Richards, S., Desjardins, C.A., Niehuis, O., et al.; Nasonia Genome Working Group. (2010). Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. Science *327*, 343–348.

Wilson, A.C., Ashton, P.D., Calevro, F., Charles, H., Colella, S., Febvay, G., Jander, G., Kushlan, P.F., Macdonald, S.J., Schwartz, J.F., et al. (2010). Genomic insight into the amino acid relations of the pea aphid, *Acyrthosiphon pisum*, with its symbiotic bacterium *Buchnera aphidicola*. Insect Mol. Biol. *19*(*Suppl 2*), 249–258.

Woolfit, M., Iturbe-Ormaetxe, I., McGraw, E.A., and O'Neill, S.L. (2009). An ancient horizontal gene transfer between mosquito and the endosymbiotic bacterium *Wolbachia pipientis*. Mol. Biol. Evol. *26*, 367–374.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Zientz, E., Dandekar, T., and Gross, R. (2004). Metabolic interdependence of obligate intracellular bacteria and their insect hosts. Microbiol. Mol. Biol. Rev. *68*, 745–770.

EXTENDED EXPERIMENTAL PROCEDURES

Phenacoccus avenae Genome Sequencing and Annotation

A young adult individual of laboratory-maintained *Ph. avenae* was dissected in PBS [137 mM NaCl, 8.1 mM Na_2HPO_4, 2.7 mM KCl, 1.5 mM KH_2PO_4 (pH 7.5)] with fine forceps and needles, and total DNA was extracted from the isolated oval bacteriome by using a conventional SDS-phenol method. The extracted DNA was amplified using GenomiPhi V2 DNA Amplification Kit (GE Healthcare Life Science) according to the manufacture's protocol and then was purified by QIAamp DNA Mini kit (QIAGEN). Two independent samples were sequenced by GS FLX system (Roche) at the OIST Sequencing Section, Okinawa, Japan. The 454 sequencing resulted in an sff data file of 141,681 reads totaling 45,763,075 bases.

Tremblaya PAVE genome assembly was carried out by GS De Novo Assembler v2.5.3 (Margulies et al., 2005) using default settings for read quality trimming and genome assembly. Assembled contigs were filtered based on average coverage, GC content and BLASTX v2.2.17 results against the GenBank nonredundant protein database (nr, posted January 18, 2011). The *Tremblaya* genome assembled into 11 contigs with an average coverage from 103.1 to 273.4X along with three short contigs with an average coverage more than six times higher than the rest of the genome (184 bp, 1825X; 225 bp, 1511.4X; 309 bp, 2121.3X). "To" and "from" information appended to the read name in the ACE file generated from the assembly and gene synteny to the *Tremblaya* PCIT genome was used to order and orient the contigs. Genome gaps were closed by PCR and Sanger sequencing to a single circular molecule.

The three short high-coverage contigs were not incorporated into gaps of the closed genome sequence, and the ACE file info suggested that these sequences might form a plasmid-like circular molecule. The three contigs were successfully joined by PCR and Sanger sequencing and the plasmid presence was confirmed by Southern blot analysis (Figure S1). For Southern blots, genomic DNA preparations of *Ph. avenae* were digested with restriction endonuclease HindIII (which does not cut the plasmid) and MunI (which cuts the plasmid at one location), and electrophoresed in agarose gels with an uncut DNA preparation as control. The separated DNA fragments were transferred to nylon membranes by a standard capillary blotting procedure, and fixed by UV crosslinking. Hybridization and detection of the probe were performed by using the DIG Detection Kit (Roche) according to manufacturer's instructions. The probe was generated by PCR (primer sequences GCATCTGACGATGTGAACAACCTT and CAGAATTAGAAAGGTGTTGCTTCTTC). The single band in the MunI lane agrees with the estimated size of the plasmid from genome assembly (744 bps). We attribute the larger sizes in the Uncut and HindIII lanes to the presence of concatenated circular molecules.

The *Tremblaya* genome was annotated as described previously (McCutcheon and Moran, 2007), except that Prodigal v1.20 (Hyatt et al., 2010) was used for gene prediction, RNAmmer v1.2 (Lagesen et al., 2007) was used to identify rRNAs and Rfam v10.1 (Gardner et al., 2009) was used to localize transfer-messenger RNA (tmRNA, also known as 10Sa RNA). The putative origin of replication was assigned to the same region of the genome as in *Tremblaya* PCIT based on a presence of oligonucleotide skew. Previously produced *Tremblaya* metabolic pathways were updated by hand using genome annotation results and EcoCyc (Keseler et al., 2005), MetaCyc (Caspi et al., 2006) and KEGG databases (Kanehisa and Goto, 2000) as guides. One possible homopolymer error was detected during the annotation process in β subunit of RNA polymerase (*rpoB*). PCR and Sanger sequencing of this region confirmed that the error was caused by 454 sequencing and the sequence was corrected accordingly. Circos v0.56 (Krzywinski et al., 2009) was used to generate graphical genome comparisons.

Planococcus citri RNA Preparation and Sequencing

Total RNA was extracted from 20 dissected mealybug bacteriomes and whole female bodies as reported previously (McCutcheon and von Dohlen, 2011). The samples were pooled submitted to eukaryotic (polyA) mRNA enrichment by TruSeq RNA Sample Preparation Kit and 99 bp paired-end libraries were sequenced by Illumina HiSeq 2000 at the Center for Genome Technology Sequencing Core, University of Miami. Illumina sequencing produced 131,944,592 and 85,597,850 paired-end reads for bacteriocytes only and whole female body samples respectively.

RNA-Seq and Differential Expression Analyses

De-novo transcriptome assemblies were carried out by the Trinity v_r2012-01-25 package (Grabherr et al., 2011) with default settings (fixed k-mer 25) from both RNA-seq samples (polyA enriched libraries from bacteriocytes and whole female bodies), and the resulting 96,981 and 82,968 contigs were preliminarily annotated by BLAST2Go (Conesa et al., 2005). The Perl script pipeline implemented in Trinity was followed to obtain FPKM expression values (fragments per kilobase of exon per million fragments mapped) and to identify differentially expressed transcripts. FastQ reads were mapped back to the transcripts by Bowtie 0.12.7 (Langmead et al., 2009), mapped reads were counted by RSEM v1.1.18 (Li and Dewey, 2011) and data normalization and identification of differentially expressed transcripts between the two samples was carried out in Bioconductor package edgeR v2.10 (Robinson et al., 2010). BAM alignment files were graphically visualized in IGV and Artemis browsers (Carver et al., 2012; Thorvaldsdottir et al., 2013). Coding regions for horizontally transferred transcripts were predicted either by the transcripts_to_best_scoring_ORFs.pl script provided in Trinity package or by NCBI ORF finder [http://www.ncbi.nlm.nih.gov/gorf/gorf.html] and checked by BLASTP searches against the nr database.

RT-qPCR Verification of ExHTG Enrichment in Bacteriome Tissue

Reverse-transcription quantitative PCR of 22 ExHGTs from whole insects and dissected bacteriomes was carried out to verify our bacteriocyte overexpression results determined by RNA-seq. Bacteriomes were dissected in 0.9% RNase free saline and immediately stabilized in TRI Reagent (Ambion). Total RNA was isolated from 20 to 30 bacteriomes and ten whole females (mealybug colony from Logan, Utah, USA) using Direct-zol RNA MiniPrep kit (Zymo Research). Extracted RNA was treated by RNase-free DNase I (Thermo Scientific) and first-strand cDNA synthesis was performed by Transcriptor Reverse Transcriptase (Roche) from 500 ng of RNA (using random hexamers in standard 20 μl reactions).

RT-qPCR primers were designed using Primer3Plus software for RT-qPCR (Untergasser et al., 2007) and checked for nonspecific products by MFEprimer-2.0 (Qu et al., 2012) against mealybug transcriptome and genome databases. Nonspecific products were also checked by melting curves and efficiencies of all primers were tested by standard curves in triplicates. Sequences and amplification efficiencies for used primers are listed in Table S2. The MIQE guidelines (Bustin et al., 2009) were followed to make the experiments as reproducible as possible.

Gene expression was normalized to 60S ribosomal protein L7 (rpl7) and relative quantification of gene expression was performed using $2^{-\Delta\Delta CT}$ methodology (Livak and Schmittgen, 2001). Rpl7 was selected based on previous work (R.P.D., unpublished data). Each experiment was performed in triplicate and included no template controls and no reverse transcription controls. Each 20 μl reaction comprised of 10 μl of LightCycler 480 SYBR green Master (Roche), 500 nM of forward and reverse primers and 5 μl of cDNA. PCR reactions were performed in white plates (Roche) on a LighCycler 480 (Roche) with thermal cycling conditions: 95°C of initial denaturation for 5 min, followed by 45 cycles at 95°C for 10 s, 60°C for 15 s, and 72°C for 15 s. The run was ended by a melting curve (95°C for 5 s, 65°C for 1 min and 97°C continuous acquisition). All analyses were carried out using LightCycler 480 software version 1.5 (Roche).

BLASTX-Based Screening for Functional Horizontal Gene Transfers of Bacterial Origin

We modified a previous pipeline (Nikoh et al., 2010) to detect genes of bacterial origin expressed in our RNA-seq data. First, the transcriptome assemblies from both RNA-seq samples were searched by BLASTX v2.2.25+ (-evalue 1×10^{-3} -outfmt 7) against the nr database (posted January 2012) and the blast results were visualized in Megan v4 (Huson et al., 2011) as metatranscriptomic data. By plotting number of sequences assigned to distinct taxonomic units, we identified several HGT candidates and contaminant bacteria.

Only those transcripts from the bacteriome transcriptome having top BLASTX hit to a bacterial sequence were filtered and used for further analyses. Transcripts with top hits to the *Tremblaya* and *Moranella* genomes were filtered based on sequence identity (>98%) and excluded for clarity. Transcripts with lower identity were checked manually. Since these transcripts did not contain recognizable transfers from the symbiont genomes and mostly represented short low-quality transcripts, they were excluded too. Importantly, we detected only a few individual transcripts from insect facultative symbionts and reproductive manipulators (such as *Wolbachia*, *Rickettsia*, *Cardinium*, *Arsenophonus, Hamiltonella*, *Regiella*, *Serratia* or *Spiroplasma*) and these transcripts were not associated with any housekeeping genes from the same taxa, which would be expected to be expressed in a facultative bacterium. The analysis thus showed that the RNA-seq data were free of facultative symbionts and confirmed previous metagenomic analysis showing that other bacteria were not present in the mealybug bacteriome at any significant level (McCutcheon and von Dohlen, 2011). Although the RNA-seq data were free of facultative symbionts, the analysis revealed contamination from common plant and soil-associated bacteria (particularly *Acidovorax* sp. and *Acinetobacter* sp.). Expression FPKM values obtained by differential expression analysis were added to the transcripts and the transcripts were filtered based on BLASTX e-value ($<1 \times 10^{-6}$), sequence identity (>40), FPKM values (>1), and sorted based on expression values. FPKM filtering (>1) allowed the filtering of low-quality transcripts and contaminants with low expression (*i.e., Acidovorax* and *Acinetobacter* spp.).

Finally, both the *P. citri* draft genome and transcriptome assemblies were divided into lengths of 1,000 nucleotides (nts), overlapping by 200 nts. This yielded 756,807 and 187,107 sequences from the genome and transcriptome assemblies respectively. These sequences were used as queries for BLASTX searches against the nr database (posted January 2012). As with the full-length transcript approach, BLASTX results were filtered to contain only contigs with top BLAST hit from the domain Bacteria and these results were processed similarly, except lower e-value cut-off was used ($<1 \times 10^{-8}$). Hits from genome scaffolds/contigs shorter than 1,000 bps and with average coverage higher than 15 were considered undetermined because our data did not allow to us to determine if these represented contamination or short duplicated HGTs.

Data from the divided transcriptome were used to look for HGT candidates cotranscribed with an insect gene, which could be missed by our search using full-length transcripts as queries. Data from the divided genome were used to detect possible unexpressed HGT candidates. BLASTN and TBLASTN searches (e-value 1×10^{-6}) of all HGT candidates against the *P. citri* genome assembly were used to check if the HGT candidates are present on a putative insect genome contig. All HGT candidates were checked by BLASTP search against the nr database.

Phylogenetic Analyses

HGT candidates were searched by PSI-BLAST against the nr database to detect approximate taxonomic position of individual transfers. Representatives for thorough taxon-sampling were then downloaded for individual HTG candidates according to their putative positions (Alphaproteobacteria: Rickettsiales, Gammaproteobacteria: Enterobacteriales and Bacteroidetes). As a taxon-sampling

guide for PSI-BLAST searches, available multi-gene phylogenies of these groups were used (Husník et al., 2011; McCutcheon and Moran, 2012; Williams et al., 2010; 2007; Wu et al., 2009). Protein sequences were aligned by the MAFFT v6 L-INS-i algorithm (Katoh and Toh, 2008). Ambiguously aligned positions were excluded by trimAL v1.2 (Capella-Gutiérrez et al., 2009) with the –automated1 flag set for likelihood-based phylogenetic methods. The resulting trimmed alignments were checked and manually corrected (if needed) in SeaView 4.3.4 (Gouy et al., 2010) or Geneious v5.6 (Kearse et al., 2012). Maximum likelihood (ML) and Bayesian inference (BI) phylogenetic methods were applied to the single-gene amino-acid alignments. ML trees were inferred using PhyML v3.0 (Guindon et al., 2010) under the LG+I+Γ model with subtree pruning and re-grafting tree search algorithm (SPR) and 100 bootstrap pseudo-replicates. BI analyses were conducted in MrBayes 3.2.1 (Ronquist et al., 2012) under WAG+I+Γ model with one to three million generations (prset aamodel = fixed(wag), lset rates = invgamma ngammacat = 4, mcmcp checkpoint = yes ngen = 1-3000000). For all ML and BI analyses, a proportion of invariable sites (I) was estimated from the data and heterogeneity of evolutionary rates was modeled by the four substitution rate categories of the gamma (Γ) distribution with the gamma shape parameter (alpha) estimated from the data. Exploration of MCMC convergence and burn-in determination was performed in AWTY (http://ceb.csit.fsu.edu/awty)and Tracer v1.5 (http://evolve.zoo.ox.ac.uk). Phylogenetic trees were rooted by outgroups and graphically visualized in FigTree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).

Planococcus citri DNA Preparation, Sequencing, and Genome Assembly

The *Pl. citri* line for genome sequencing was established from a long-term laboratory population at Wye College London (provided by Mike Copland). In May 2011, a single mated female was used to found an iso-female line. Three subsequent generations of this line were re-founded by a single female that was mated to her brother. After these three generations the line was kept as a mass culture. Genomic DNA was extracted from a single virgin adult female, and two short insert libraries of 200 and 800 bp were constructed and sequenced by the Beijing Genomics Institute (BGI). An addition 200 bp insert library was constructed in the McCutcheon lab and sequenced at the Vincent J. Coates Genomics Sequencing Laboratory at the University of California at Berkeley. A total of 81,628,073,600 nts of raw sequence was generated from these libraries (80 million 90 nt paired-end reads from the BGI 200 bp insert library, 78.8 million 90 nt paired-end reads from the BGI 800 bp insert library, and 337.2 million 100 nt paired-end reads from the Berkeley 200 bp insert library).

The raw sequencing reads were adaptor end-quality trimmed using the ea-utils tool fastq-mcf (http://code.google.com/p/ea-utils) using default parameters with the exception that the minimum remaining sequence length flag was set to 41. Overall sequence quality filtering was then performed using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit) program fastq_quality_filter using the flags -q 20 -p 80. Overlapping reads were combined using FLASH (Magoč and Salzberg, 2011). Any remaining singleton reads (i.e., those with a paired read that was thrown out during quality filtering) were removed. The combined quality filtered data set consisted of 206,570,756 reads, 19,602,678,710 nts in total, and was assembled using Velvet (Zerbino and Birney, 2008) with a k-mer size of 45 and the expected coverage set to "auto."

SUPPLEMENTAL REFERENCES

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., et al. (2009). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin. Chem. *55*, 611–622.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.

Carver, T., Harris, S.R., Berriman, M., Parkhill, J., and McQuillan, J.A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics *28*, 464–469.

Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S.Y., et al. (2006). MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res. *34*(Database issue), D511–D516.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics *21*, 3674–3676.

Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R., and Bateman, A. (2009). Rfam: updates to the RNA families database. Nucleic Acids Res. *37*(Database issue), D136–D140.

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. *27*, 221–224.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

Husník, F., Chrudimský, T., and Hypša, V. (2011). Multiple origins of endosymbiosis within the Enterobacteriaceae (γ-Proteobacteria): convergence of complex phylogenetic approaches. BMC Biol. *9*, 87.

Huson, D.H., Mitra, S., Ruscheweyh, H.J., Weber, N., and Schuster, S.C. (2011). Integrative analysis of environmental sequences using MEGAN4. Genome Res. *21*, 1552–1560.

Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Katoh, K., and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. *9*, 286–298.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647–1649.

Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M., and Karp, P.D. (2005). EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic Acids Res. *33*(Database issue), D334–D337.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Lagesen, K., Hallin, P., Rødland, E.A., Staerfeldt, H.H., Rognes, T., and Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. *35*, 3100–3108.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Livak, K.J., and Schmittgen, T.D. (2001). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods *25*, 402–408.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957–2963.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380.

McCutcheon, J.P., and Moran, N.A. (2007). Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proc. Natl. Acad. Sci. USA *104*, 19392–19397.

Qu, W., Zhou, Y., Zhang, Y., Lu, Y., Wang, X., Zhao, D., Yang, Y., and Zhang, C. (2012). MFEprimer-2.0: a fast thermodynamics-based program for checking PCR primer specificity. Nucleic Acids Res. *40*(WebServer issue), W205–W208.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. *61*, 539–542.

Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform.

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., and Leunissen, J.A.M. (2007). Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res. *35*(WebServer issue), W71–W74.

Williams, K.P., Sobral, B.W., and Dickerman, A.W. (2007). A robust species tree for the alphaproteobacteria. J. Bacteriol. *189*, 4578–4586.

Williams, K.P., Gillespie, J.J., Sobral, B.W., Nordberg, E.K., Snyder, E.E., Shallom, J.M., and Dickerman, A.W. (2010). Phylogeny of gammaproteobacteria. J. Bacteriol. *192*, 2305–2314.

Wu, D.Y., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N.N., Kunin, V., Goodwin, L., Wu, M., Tindall, B.J., et al. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature *462*, 1056–1060.

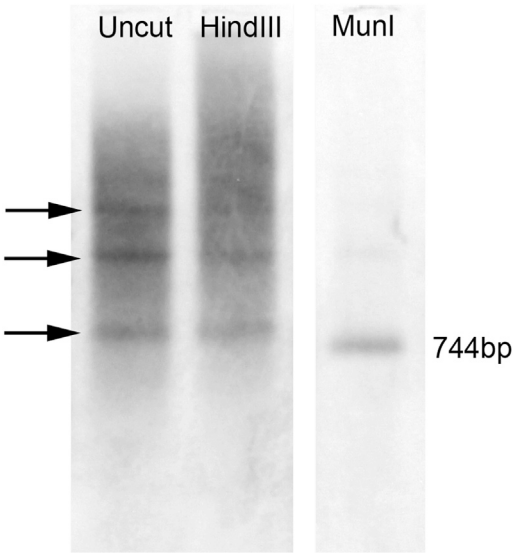

Figure S1. Southern Blot of *Tremblaya* PAVE Plasmid-Like Molecule, Related to Experimental Procedures