University of Montana

ScholarWorks at University of Montana

Graduate Student Theses, Dissertations, & Professional Papers

Graduate School

2015

The evolution of nutritional co-endosymbionts in cicadas

James Theodore Van Leuven

Follow this and additional works at: https://scholarworks.umt.edu/etd

Let us know how access to this document benefits you.

Recommended Citation

Van Leuven, James Theodore, "The evolution of nutritional co-endosymbionts in cicadas" (2015). *Graduate Student Theses, Dissertations, & Professional Papers*. 10794.
https://scholarworks.umt.edu/etd/10794

This Dissertation is brought to you for free and open access by the Graduate School at ScholarWorks at University of Montana. It has been accepted for inclusion in Graduate Student Theses, Dissertations, & Professional Papers by an authorized administrator of ScholarWorks at University of Montana. For more information, please contact scholarworks@mso.umt.edu.

THE EVOLUTION OF NUTRITIONAL CO-ENDOSYMBIONTS IN CICADAS

By

JAMES THEODORE VAN LEUVEN

B.A., Carroll College, Helena, Montana, 2008

Dissertation

presented in partial fulfillment of the requirements
for the degree of

Ph.D.
in Cellular, Molecular, and Microbial Biology

The University of Montana
Missoula, MT

December 2015

Approved by:

Sandy Ross, Dean of The Graduate School
Graduate School

Dr. John McCutcheon, Research Advisor
Division of Biological Sciences - CMMB

Dr. Scott Miller, Examination Chair
Division of Biological Sciences - CMMB

Dr. J. Stephen Lodmell
Division of Biological Sciences - CMMB

Dr. Frank Rosenzweig
Division of Biological Sciences - CMMB

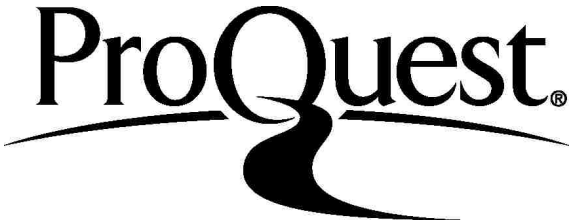
Dr. Jeffrey Good
Division of Biological Sciences - OBE

ProQuest Number: 10098631

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.



ProQuest 10098631

Published by ProQuest LLC (2016). Copyright of the Dissertation is held by the Author.

All rights reserved.
This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Van Leuven, James, Ph.D., December 2015 Cellular, Molecular, and Microbial Biology
Division of Biological Sciences

The evolution of nutritional co-endosymbionts in cicadas

Chairperson: Dr. John McCutcheon

Symbiosis occurs between organisms in all domains of life. The evolution of obligate symbionts from free-living bacteria typically results in the loss of genes involved in metabolic independence and an overall reduction in genome size. Outside the organelles, the most extreme examples of genome reduction come from the intracellular symbionts of sap-feeding insects. The genomes of these bacteria encode very few genes other than those involved in translation, replication, and amino acid synthesis. *Candidatus* Hodgkinia cicadicola (*Hodgkinia*) and *Candidatus* Sulcia muelleri (*Sulcia*) live in specialized insect cells (bacteriocytes) of the cicada *Diceroprocta semicincta,* and have undergone severe gene loss. *Hodgkinia* in particular retains one of the smallest gene sets of all bacteria, and even less than many organelles. As a result, the *Hodgkinia* genome is left with a seemingly incomplete set of genes that are required for cellular life, including core genes in the translational machinery. I analyzed a set of *Hodgkinia* genomes and performed several experiments to uncover the constraints guiding the evolution of *Hodgkinia*. What mutational and selective pressures are acting on the *Hodgkinia* genome? How do essential cellular enzymatic reactions occur in *Hodgkinia* cells? Does the cicada host complement *Hodgkinia's* limited genetic repertoire? How does the evolution of insect endosymbionts compare to the evolution of organelles? My work provides answers to many of these questions, and deepens our understanding of intracellular symbioses.

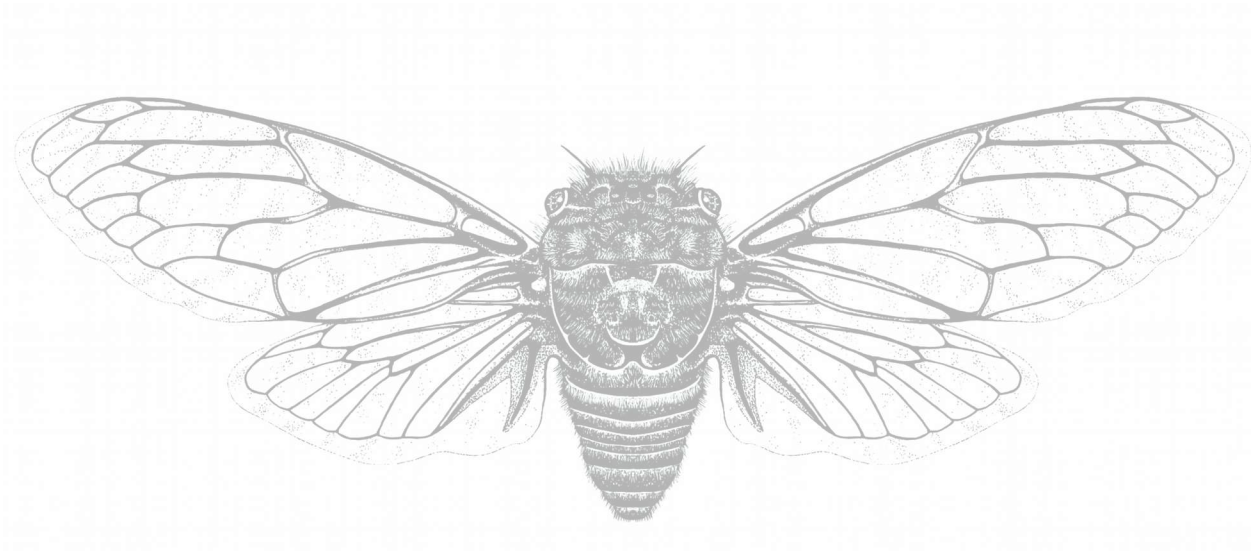


Illustration by David Tuss, appears on Sep. 11th 2014 cover, Cell

Acknowledgements

I am indebted to many people who contributed to the completion of this thesis. First, I thank my mentor, Dr. John McCutcheon for endless patience and unwavering support. Six years is a long time to work together, and over this time, John guided me in teaching others, developing collaborations, and writing with clarity. Without doubt, I am a far better scientist having learned from John. I thank my thesis chair, Dr. Scott Miller, for walking me through this enduring process; and my thesis committee, Dr. Frank Rosenzweig, Dr. J. Stephen Lodmell, and Dr. Jeffrey Good for scientific and professional advice. All the members of the McCutcheon lab have contributed to this work through either discussion or laboratory assistance; I thank you for your help and support. I extend extra gratitude to Dan Vanderpool, Filip Husník, Piotr Łukasik, Matt Campbell, and Amanda Brown for trudging through the conceptual challenges of the *Hodgkinia* story. Lastly, I thank my wife, Michelle Van Leuven, my parents, and the rest of my family for their encouragement and uplifting confidence. Despite how fast the past six years went by, these years were not without their trials and tribulations. If not for the support of my family I would have yielded to the challenges years ago. Thank you all for your contributions, I hope you consider this thesis as much yours as mine.

Chapter 1: Introduction

1.1 Sap-feeding insects: ecological importance and feeding habits

There are over one million documented insect species, with 2-8 million estimated to exist globally—far more than all other animals combined (Novotny et al. 2002; Engel 2015; Stork et al. 2015). About 85% of insects belong to the Holometabola, the monophyletic group that undergo complete metamorphosis (Truman and Riddiford 1999). Another 11% are comprised of the Paraneoptera, the superorder that contain lice, thrips, and hemipterans (Engel 2015). Only insects in the order Hemiptera feed solely on plant sap (Figure 1). Their defining characteristic is a specialized proboscis that is hardened to pierce the epidermis and cortex of plant stems and roots (Cobben 1978; Engel 2015). The transition to plant feeding (phytophagy) opened up a novel resource for insects, and was accompanied by increased rates of species diversification (Cobben 1978; Mitter et al. 1988). The adaptive radiation of phytophagous insects resulted in an

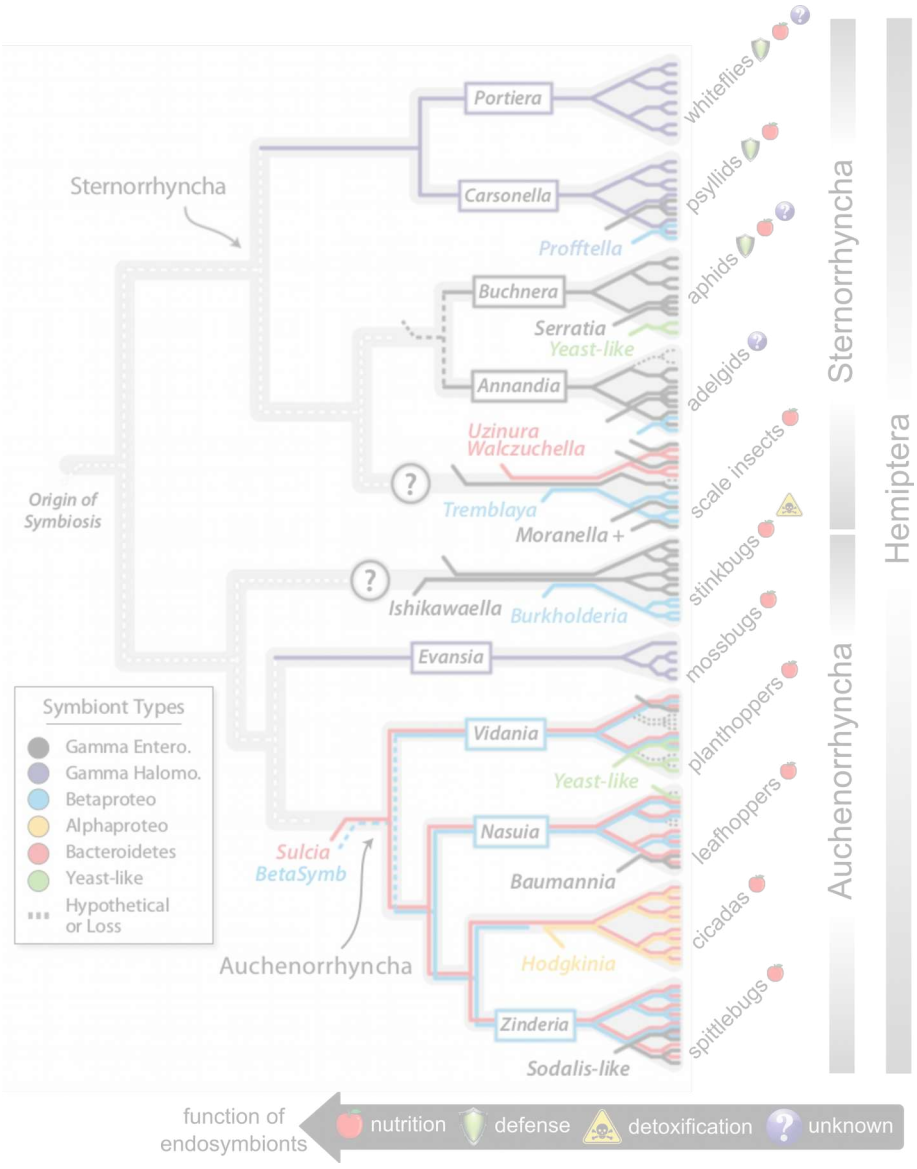


Figure 1. Cartoon cladogram adapted from Bennett and Moran 2015. Hemipteran insect phylogeny shown in grey with their microbial symbionts colored according to key. Lineages assumed to have obligate symbionts are represented with dotted lines. Common insect names for each group shown with symbionts' functional role. Citations in text.

incredibly specious, but phylogenetically narrow clade; half of all insect species that feed on plants belong to only 9 of 30 extant orders (Mitter et al. 1988; Bennett and O'Grady 2012).

Transitioning to exclusive sap feeding, however, came with several challenges to be overcome before hemipterans could successfully utilize this resource (Sandström and Moran 1999; Douglas 2006). Namely, insects that feed on sap must eat and concentrate large volumes of food that has low and unbalanced nutrient concentrations. The nutrient composition of phloem sap is generally rich in carbohydrates (mostly sucrose), and contains some proteins and amino acids (Hayashi and Chino 1986; Sandström and Moran 1999; Douglas 2006; Will et al. 2013; Hijaz and Killiny 2014). The composition of xylem sap on the other hand contains about 10-fold more dilute amino acids and proteins, and is largely devoid of carbohydrates (Jeschke et al. 1995; Sandström and Moran 1999; Kehr et al. 2005; Christensen and Fogel 2011; Krishnan et al. 2011; D'Mello 2015). In both, the amino acid composition is uneven and the nitrogen content low. Asparagine comprises 75% of the amino acid content, although glutamine and aspartic acid can rise to high levels during seasonal fluctuations (Sandström and Pettersson 1994; Grassi et al. 2002). To compensate for their nutrient-poor food, phloem feeding insect produce and expel honeydew—a carbohydrate-rich substance that also contains high proportions of non-essential amino acids (Douglas 2006). However, enriching nutrients cannot alleviate the insect from the complete lack of some essential compounds from plant sap. For this reason, sap-feeding insects have almost universally developed symbioses with microbes that can synthesize compounds missing from their diets (Figure 1).

Sap-feeders affect plant productivity by causing tissue damage via feeding, laying eggs (Meyer 1993; Zvereva et al. 2010; Stephens and Westoby 2015), and spreading microbial plant pathogens (Hill 1987; Dedryver et al. 2010). The damage caused by sap-feeders may not be as visually obvious as defoliating insects, but they cause a reduction in plant health as indicated by reduced seed production, slower growth rates, and higher exposure to other herbivorous insects (Crawley 1989; Zvereva et al. 2010). Range expansions into crop plants can be particularly devastating. For example, in the late 19th century the grape vine pest *Daktulosphaira vitifoliae* was introduced to Europe. These sap-feeding, gall-forming insects are closely related to adelgids (Figure 1) and are native to Northern America, where grape vines are partially resistant. British botanists brought *D. vitifoliae* to Europe, and they rapidly spread through vineyards. In France alone, two-thirds of the vineyards were completely destroyed (Powell et al. 2013). Microbial symbionts carried by the *D. vitifoliae* likely aid in the formation of damaging galls on grape vines (Vorwerk et al. 2007; Powell et al. 2013). In an age of global commerce and monoculture crops, particular attention should be paid to understanding if and how microbial symbionts facilitate range expansions of invasive insects (Brown et al. 2013).

The sheer number of sap-feeding insects suggests they play key roles in natural ecosystems. Cicadas in particular have been shown to provide substantial resource pulses that support insectivores and provide nutrient-rich detritus material (Yang 2004; Menninger et al. 2008). Sap-feeding insects link plants, microbial communities, and larger animals through trophic interactions (Hougen-Eitzman and Rausher 1994; Nowlin et al. 2007; Becerra 2015). Insects can also connect different ecosystems through plant-mediated interactions; insects feeding on above-ground plants affect insects feeding below ground by interspecific competition (Johnson et al. 2012). Moreover, the sheer abundance of insects makes them significant carriers of pathogens, which can infect both plants and animals (Elderd et al. 2013).

1.2 Functional role of endosymbionts in sap-feeding insects

Insect-bacterial symbioses were first documented by Robert Hooke in the late 17th century (Hooke 1665). He described the microbe-harboring mycetomes [bacteriomes] in the human body louse, although at the time he did not recognize that these organs housed bacteria, nor did he guess their function. It was not until the late 19th century when the mycetomes of plant-feeding insects were documented, probably most accurately by Leydig in 1850. Their discovery led to extensive microscopic studies of many plant-feeders including aphids (Leydig 1850; Huxley 1858), phyllids (Metschnikoff 1866), ants (Blochmann 1884), scale insects (Berlese 1893), cicadas (Heymons 1899), spittlebugs (Porta 1900), and weevils (Holmgren 1902). The discovery of morphologically similar organs across all of these insects was quite curious, and many functions were proposed (Buchner 1965). At the time, the idea that these organs carried symbiotic microorganisms was beyond conceptual reach, so these "albuminous bodies" were often described as having some nutrient storage function (Metschnikoff 1866). However, better histology combined with further description of their faithful transmission into eggs supported the idea of stable microbial symbioses. While studying the eggs and symbiont tissues of cockroaches, Blochmann wrote, "In the light of our present knowledge one can scarcely do otherwise than declare these rodlets to be bacteria" (Blochmann 1884).

In 1909 the true function of mycetomes were published simultaneously and independently by Umberto Pierantoni and Vytváření Karel Šulc. Šulc was the first to use the term mycetome during a lecture in Prague on November 5, 1909. Once their true function was revealed, the research on hemipteran symbioses exploded. In his landmark book, *Endosymbioses of Animals with Plant Microorganisms*, Paul Buchner described the contributions of Pierantoni and Šulc: "With the publication of these reports it seemed as though a blindfold had been removed from the eyes!" It was Buchner that likely contributed the most to symbiosis research in his careful microscopy studies of many blood- and sap-feeding insects. His 1965 book (translation from German) contains 371 figures, most hand drawn images of the symbiont-containing tissues of many blood- and sap-feeing insects. Buchner's work ushered in an era of insect symbiosis research. Through the 20th century, microscopic studies were used to investigate insects at all life stages, elucidating the unusual cellular biology of mycetomes, including multi-nucleated or syncytial cells, migration of bacterial cells into eggs during development, and the innervation of mycetomes with nutrient supplying trachea (Buchner 1965).

In the early 1970's, a series of explorative and manipulative studies uncovered the primary role of bacterial symbionts in aphids (Auclair 1965; Dadd et al. 1967; Dadd and Krieger 1968). In short, a disparity was noticed between the amino acid content of aphids and their phloem food source. When aphids were cured of their bacterial symbionts with antibiotics, their growth was severely stunted. However, normal growth was restored by adding essential amino acids to the aphid diet. The presence of the bacterium *Ca.* Buchnera aphidicola in most aphids suggested its importance and aposymbiotic aphids reared without the ten essential amino acids required by all metazoan have growth defects (Auclair 1965; Buchner 1965; Dadd et al. 1967; Dadd and Krieger 1968; Mittler 1971; Douglas 1989; Douglas and Prosser 1992; Lai et al. 1994). Incorporation of 14C, 15N, and 35S into essential amino acids from non-essential amino acids or elemental atoms shows that *Buchnera* is likely responsible for essential amino acid synthesis

(Douglas 1988; Febvay et al. 1995; Sasaki and Ishikawa 1995). In 2000 the complete genome of *Buchnera* from the pea aphid was published, showing a beautiful example of complementarity between the nutritional needs of the aphid and the amino acid biosynthesis pathways present in the *Buchnera* genome (Shigenobu et al. 2000). Of the 20 amino acids that bacteria can typically make from metabolites and sugars, the genes present in the *Buchnera* genome suggest that it can only make 10—precisely the ones needed by the aphid. In return, the aphid may provide *Buchnera* with a suitable environment rich in carbohydrates, fatty acids, and other metabolites.

Phylogenetic analyses suggest that *Buchnera* was acquired by aphids 160-280 million years ago (Moran et al. 1993). Over time, the host-restricted environment inhabited by *Buchnera* has allowed many genes that overlap with services provided by the aphid host to be lost. Depending on the aphid species, the *Buchnera* genome has shrunk to 0.42-0.67 Mb, compared to its free-living relatives that have genome sizes around 5 Mb (Shigenobu et al. 2000; Moran and Mira 2001; Pérez-Brocal et al. 2006). Many of the genes lost are thought to be dispensable when living in a restricted environment (e.g. pathways for anaerobic respiration), but some losses are uniquely characteristic of obligate, intracellular symbionts (e.g. genes involved in DNA repair, recombination and cell membrane synthesis). Sequencing the aphid genome revealed that while some genes have been transferred from the *Buchnera* genome to the aphid genome, these transfers have been pseudogenized and thus cannot offset gene loss in the *Buchnera* genome (Nikoh et al. 2010). However, genes transferred from other bacteria (e.g. *Wolbachia*) are upregulated in symbiont tissues, implicating a symbiotic role of these gene products (Nikoh et al. 2010). The protein product from one of these genes localizes to *Buchnera* cells, showing that the aphid host uses genes acquired by horizontal gene transfer (HGT) to support *Buchnera* (Nakabachi et al. 2014). Additionally, host-encoded amino acid transporters gene families are enriched in many endosymbiont containing insects, and amino acid transporters are found at the host-symbiont interface in aphids (Price et al. 2011; Duncan et al. 2014; Price et al. 2014). Despite sharing genetic resources, each organism retains its own signature of independence by encoding their own translational machinery (e.g. ribosomal RNA and protein, tRNAs, and aminoacyl tRNA

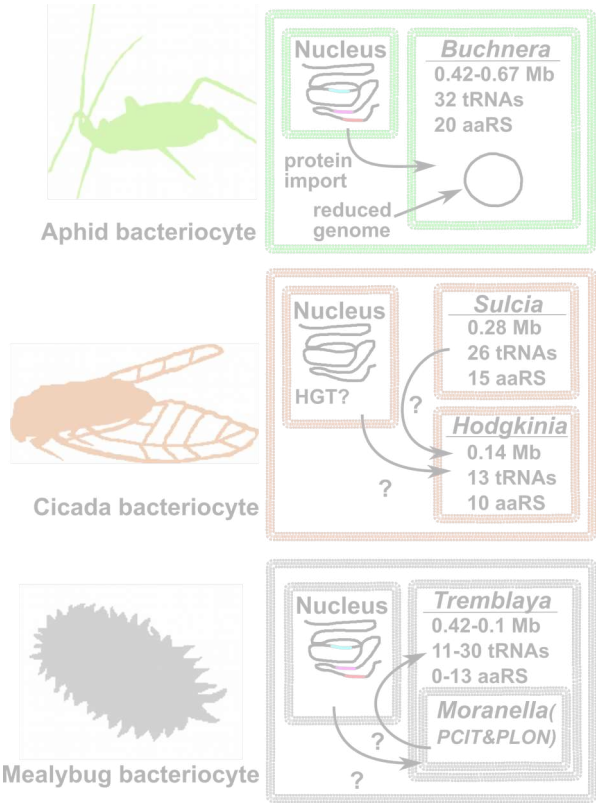


Figure 2. Summary of three nutritional endosymbiont systems. Each box represents a membrane-bound compartment, with the outermost being an insect cell that harbors intracellular symbionts (bacteriocyte). HGTs in the host genome are colored to reflect the diversity of donor species. Arrows indicate how the transport of gene products could support the bacterial symbionts.

sythetases (aaRSs)). Only the most extremely degenerate endosymbiont genomes lack these genes.

Mealybugs (sub-order Sternorrhyncha) are phloem-feeding insects that feed on wild and crop plants, posing a serious risk to several staple crops (Baumann 2005). The bacterial symbionts that they carry are noted for their unique arrangement. In mealybugs that have two bacterial symbionts, the more ancient one, *Ca.* Tremblaya princeps (hereafter *Tremblaya*), harbors a second symbiont within its cytoplasm (McCutcheon and von Dohlen 2011). The intrabacterial symbiont *Ca.* Moranella endobia (hereafter *Moranella*) has a much more gene-rich genome than *Tremblaya,* suggesting that it was more recently acquired. In basal mealybug lineages, only *Tremblaya* is present, providing an amazing glimpse into the genome evolution of *Tremblaya* during a critical point in mealybug evolutionary history (Husnik et al. 2013). Comparing the gene content of *Tremblaya* with and without *Moranella* clearly shows that the acquisition of *Moranella* resulted in genome degradation in *Tremblaya*. I will elaborate more on the particulars of genome structure and content evolution in this system in chapter 1.3, but for now, it should be mentioned that essential amino acid production in mealybugs with the *Tremblaya* and *Moranella* pair requires both bacteria and the mealybug host. The genomic and transcriptomic data from this system show that pathways required for essential amino acid production are partitioned between all three symbiotic partners. Some enzymatic steps are likely fulfilled by genes on the *Tremblay*a genome, some by genes on the *Moranella* genome, and a few by genes on the mealybug genome. Interestingly, several of the mealybug genes required are actually horizontal gene transfers from diverse bacterial donors. These genes are highly expressed in bacteriome tissue, strongly suggesting a functional role (Husnik et al. 2013).

While many endosymbionts in sap-feeding insects have a nutritional role, some can provide insect hosts with other functions (Figure 1) (Hosokawa et al. 2007; Hedges et al. 2008; Nakabachi et al. 2013; Kaltenpoth and Engl 2014). In aphids for example, alternative functions of beneficial symbionts include reduced rates of viral infections, protection from pathogenic fungi, resistance to parasitoid wasps, and higher heat tolerance (Oliver et al. 2010). The gammaproteobacterium *Regiella insecticola* is present in about 16% of aphid species and reduces the rate of infection by the entomopathic fungus *Pandora neoaphidis* by up to 5 fold (Ferrari et al. 2004; Scarborough et al. 2005). *Hamiltonella defensa* is found in about 14% of aphid species and reduceds the rate of parasitism by up to 100% (Oliver et al. 2005). Escape from pathogens is an adaptation that precedes range expansions and speciation (Hardin 1960; Connell 1972; Takiya et al. 2006; Bennett and O'Grady 2012). Understanding how microbial symbionts contribute to evolutionary and ecological changes in their host provides us with important insight into the potential benefits of symbioses. For example, aphids can be provided with instantaneous heat tolerance with a simple symbiont switch by exchanging one *Buchnera* strain for another in the laboratory (Moran and Yun 2015). No host adaptation is necessary to make this habitat shift.

1.3 Overview of endosymbiont genomics

By and large, bacterial genomes are single, circular molecules that contain on average 5 million basepairs (Mb) of DNA sequence (ranging from 0.112 to 17.5Mb). The genomes of bacteria can be dynamic, with genome rearrangements and horizontal gene transfer between bacteria occurring regularly (Smith et al. 1993; Joyce et al. 2002; Thomas and Nielsen 2005;

Touchon et al. 2009). Additionally, bacteria are wildly diverse in terms of metabolic and sequence diversity (Pace 1997). These characteristics are evident in the genomes of free living bacteria, where the need to adapt to changing environmental conditions and interact with other microorganisms requires a robust and ever-changing set of genes.

A clear transition in genome content occurs when bacteria become host associated (Mira et al. 2001; Moran 2002; Wernegreen 2015). The first genomic changes are evident in recently evolved facultative pathogens, which have slightly reduced genome sizes and a proportional increase in virulence factors. Obligate (often intracellular) pathogens have these changes plus may loose genes needed to sustain an extracellular lifestyle. Symbionts that become obligate and mutually beneficial lack virulence factors and undergo rapid genome reduction. Obligate mutualists that are very recently acquired can have average sized genomes, but often contain many pseudogenes that have not yet been completely removed from the genome (Dale et al. 2003; Clayton et al. 2012). A group of Enterobacteriaceae called the Sodalis-allied symbionts are particularly well known for frequently making the transition from free-living to host-associated (Toju et al. 2010; Koga et al. 2013; Koga and Moran 2014; Oakeson et al. 2014). These bacteria are clustered with species that live in soils, on trees, or other environmental substrates and give rise to more derived, obligate insect endosymbionts like *Baumannia*, *Blochmannia*, *Buchnera,* and *Moranella*. In grain weevils, Philaenine spittlebugs, and scale insects, Sodolis-allied bacteria have established obligate symbioses with insect hosts, replacing old endosymbionts (Koga et al. 2013; Bennett et al. 2014). Sodalis-allied symbionts have increased amino acid substitution rates, pseudogenization, and the proliferation of insertion sequence (IS) elements (Clayton et al. 2012). More derived Sodalis genomes are reduced in size and are lacking large sections of the genome that encode virulence factors, along with continued pseudogenization.

As selection purges non-functional DNA, endosymbiont genomes experience massive size reductions (Mira et al. 2001; McCutcheon and Moran 2012). Most endosymbionts that are ancient and stably associated with their hosts (like *Buchnera*) have genomes which are greatly reduced in size and gene content (Moran and Mira 2001). *Buchnera* for example, has a genome size of ~0.5 Mb and is lacking many genes that are conserved in all free-living bacteria. These genomes retain only the genes most critical for cellular life (energy production, metabolism, replication, ect.) and always retain the genes needed to support the symbiosis (amino acid production, defensive compound synthesis, ect.). An example of this is seen in the 0.46 Mb genome of the citrus psyllid symbiont *Ca.* Profftella armatura (Nakabachi et al. 2013). A full 15% of the *Profftella* genome is devoted to the biosynthesis of polyktides, while *Profftella* has completely lost the ability to synthesize any amino acids. In sharp contrast, the other citrus psyllid symbiont, *Carsonella-DC,* has an incredibly dense genome of only 0.17 Mb that encodes 30 genes involved in amino acid biosynthesis. These examples show how the nutritional roles of bacterial symbionts are clearly manifest in their genome sequences.

Endosymbiont pairs are common in Auchenorrhyncha (Figure 1), where the ancient symbiont *Ca.* Sulcia muelleri is joined or replaced by newer symbionts (Moran et al. 2005; Koga et al. 2013; Bennett and Moran 2015). In the many instances where a co-symbiosis is established, convergent loss of genes occurs in the newly established symbiont so that they retain only genes needed to complement *Sulcia* in their supplementation of the host insect (McCutcheon and Moran 2010). Across all of Auchenorrhyncha, *Sulcia* is highly conserved and has very low substitution rates (Moran et al. 2005; McCutcheon et al. 2009a; Bennett et al. 2014). Its genome

size varies from about 0.19 to 0.28 Mb, and its genomic GC content is consistently between 21 to 23%. This genome has remained essentially unchanged in gene content and genome synteny for at least 200 million years (McCutcheon and Moran 2010). The cosymbionts of *Sulcia* vary among host insects. In spittlebugs, *Sulcia* is accompanied by *Ca.* Zinderia insecticola, a betaproteobacterium with a 0.2 Mb, 13.5% GC genome which complements *Sulcia* in the biosynthesis of histidine, methionine, and tryptophan. In leafhoppers, *Sulcia* is joined by *Ca.* Nasuia deltocephalinicola, a close relative of *Zindera* (Bennett and Moran 2013). The *Nasuia* genome is only 0.112 Mb in length, 17% in GC content, and makes histidine and methionine, but not tryptophan (Bennett and Moran 2013). This is also true for the sharpshooter symbiont *Ca.* Baumannia cicadellinicola, except that this bacterium is a gammaproteobacterium with a genome size of 0.7 Mb and contains additional genes for vitamin biosythesis and amino acid membrane transport (Moran et al. 2003; Wu et al. 2006). Cicadas harbor *Sulcia* and *Ca.* Hodgkinia cicadicola (McCutcheon et al. 2009a; McCutcheon et al. 2009b). *Hodgkinia* is an alphaproteobacterium with a highly reduced genome that complements *Sulcia* in producing the 10 essential amino acids needed by their insect host. I will discuss this symbiont pair in much greater detail in the last section of the introduction.

Despite their frequency across diverse insect species, there is little evidence addressing why symbiotic mutualistic consortia are so common. Certainly, there must be evolutionary hurdles to overcome before mutualists become so intimate that they share most of the metabolic duties needed for cellular life. Many examples of microbe-microbe and microbe-animal symbioses are observed in nature and through symbiosis the ecological range and metabolic capabilities of the combined partners are often advantageous (Greenberg 2003; Tyson et al. 2004; Ueda et al. 2004; Woyke et al. 2006). Models suggest that the formation of symbiotic pairs is evolutionary favored in the right conditions (Estrela et al. 2015; Kiers and West 2015; Kümmerli et al. 2015; Pande et al. 2015).

With notable exceptions (like *Sulcia*), the typical evolutionary trajectory of endosymbionts is genome degradation to an unknown end-point (Bennett and Moran 2013; Moran and Bennett 2014; Bennett and Moran 2015). The *Tremblaya* genome from the mealybug species *Planococcus citri* (PCIT) is incredibly degenerate, containing only ~120 protein coding genes (McCutcheon and von Dohlen 2011). Amazingly, this genome is missing some of the most important genes known to cellular life, including all aminoacyl tRNA sythetases (aaRSs). *Hodgkinia* is also quite degenerate and contains only 10 of the required 20 aaRSs (McCutcheon et al. 2009b). This level of gene loss, combined with the frequency of symbiont replacement in hemipterans suggests a process whereby symbionts are acquired, used until their genomes become completely destroyed by mutation, then replaced with fresh symbionts (Bennett and Moran 2013; Bennett and Moran 2015). Why some symbionts like *Sulcia* and *Buchnera* persist over very long periods of time, remains a mystery.

One potential mechanism to buffer against fluctuations in symbiont consortia is for the host itself to facilitate these symbioses. The genomes of aphids, mealybugs, whiteflies, and psyllids have experienced massive gene family expansions of amino acid transport genes (Duncan et al. 2014). Presumably, these expansions have improved amino acid exchange at the symbiosomal membrane (Price et al. 2011; Duncan et al. 2014). Aphids have additional adaptations that have not yet been discovered in other symbiont-harboring insects. Through gene loss and transcriptional regulation, aphids have reduced immunological responses to bacterial

infections (Gerardo et al. 2010; Burke and Moran 2011a). Horizontal gene transfer from bacteria to the insect genome could also facilitate symbiosis. The *Planococcus citri* mealybug, for example, has acquired genes by HGT that are complementary to the metabolic pathways present in *Tremblaya* and *Moranella* (Husnik et al. 2013). Incredibly, these HGTs are not from *Tremblaya* or *Moranella*, but instead from a phylogenetically diverse set of bacterial donors that were presumably associated with mealybugs at some point in their evolutionary history. Pea aphids also have HGTs, although the source bacteria of their HGTs are phylogenetically distinct from the mealybug donors (Nikoh et al. 2010; Husnik et al. 2013). HGTs from bacteria are quite common in insect genomes however; entire Wolbochia genomes exist in the genome of *Drosophila ananassae* (Hotopp et al. 2007). Functional HGT is less frequently observed and stands as one of the last characteristics differentiating endosymbionts from organelles. This will be discussed later in the introduction. Host adaptation, obviously, does not require HGT. Surely, as more insects harboring degenerate microbial symbionts are studied, further adaptations will be discovered that inform our understanding on the evolutionary potential of symbiotic partnerships.

1.4 Molecular evolution of endosymbionts

Strictly intracellular mutualists have higher substitution rates than their free living relatives (Moran et al. 1993; Woolfit and Bromham 2003). The factors that impact substitution rate include mutation rate, DNA repair, recombination, purifying selection, and genetic drift. Although mutation rate has not been measured for any nutritional endosymbiont, it is hypothesized that elevated mutation rates in these bacteria contribute to their increased substitution rate (Itoh et al. 2002; Marais et al. 2007; Hershberg and Petrov 2010; Hildebrand et al. 2010; Van Leuven and McCutcheon 2011). Measurements made on cultivable organisms show little variation in mutation rates across divergent taxa, suggesting that selective processes or loss of DNA repair mechanisms are responsible for lineage specific increased substitution rates (Drake et al. 1998; Ochman et al. 1999). As many endosymbionts with reduced coding content are missing key enzymes involved in DNA repair (*Hodgkinia* is missing *mutS*, *mutL*, and *mutH*), this explanation seem likely, albeit unsupported by experimental evidence (McCutcheon 2010; McCutcheon and Moran 2012). The only genes involved in DNA repair and replication universally conserved in reduced endosymbiont genomes are the alpha (*dnaE*) and epsilon (*dnaQ*) subunits of DNA polymerase III, although even these genes are missing from some genomes in the *Hodgkinia* genome complexes of some cicada species, the *Portiera* genome, and the *Uzinura* genome (Sabree et al. 2013; Sloan and Moran 2013; Van Leuven et al. 2014; Campbell et al. 2015). Also missing from most endosymbiont genomes are enzymes involved in DNA recombination (*recA*, *recF* and the *uvr* operon), preventing DNA repair by homologous recombination, although some recombination does occur even in bacteria missing these genes, by some unknown mechanism (Dale et al. 2003; McCutcheon and von Dohlen 2011; Sloan and Moran 2013; Van Leuven et al. 2014). Adding insult to injury, nutritional endosymbionts only live in insect cell cytoplasm and are transmitted transovarially in small numbers, so their effective population sizes are much smaller than free-living bacteria (Mira and Moran 2002). Thus, the evolution of strict intracellular mutualists is characterized by relaxed purifying selection, rapid sequence evolution, and gene loss due to deletional biases (Moran 1996; Mira et al. 2001; Woolfit and Bromham 2003; Hershberg et al. 2007; Van Leuven and McCutcheon

2011). Indeed, genome-wide dN/dS is elevated in insect endosymbionts, even above values calculated for strictly clonal bacteria (Kuo and Ochman 2009; Hershberg and Petrov 2010; Van Leuven and McCutcheon 2011; Burke and Moran 2011b; Van Leuven et al. 2014).

1.5 Comparison of nutritional endosymbionts and organelles

Mitochondria are the evolutionary end-product of symbiosis between an intracellular alphaproteobacterium and a primitive eukaryote (Gray et al. 1999). Despite the differences between mitochondria and *Hodgkinia*, their translation systems are worth comparing because both have undergone severe genome reduction in the cytoplasm of eukaryotic cells (Figures 3 and 4), and like organelles, it is likely that degenerate endosymbiont genomes require coordination with host cells for function (Timmis et al. 2004; Gray 2012; Pett and Lavrov 2015). Mitochondria have lost all of their aaRS genes, but many retain a minimal number (~25) of tRNA genes (Suzuki et al. 2011; Burger et al. 2013; Salinas-Giegé et al. 2015). Only 13 non-redundant tRNA genes (16 total) and 10 aaRSs can be identified in the *Hodgkinia* genome (McCutcheon et al. 2009b). Even for insect nutritional endosymbionts, this is a very reduced gene set. In contrast, *Buchnera* strains have 31-32 tRNA genes and a full complement of 20

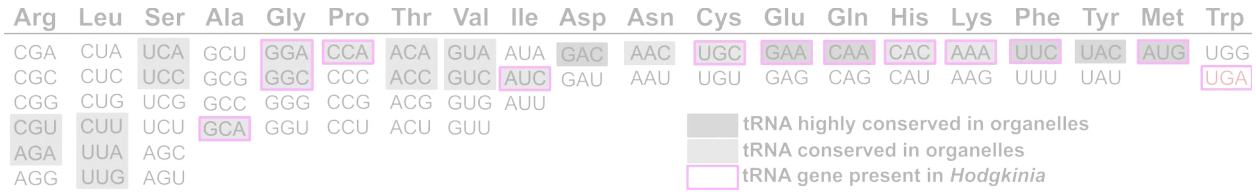


Figure 3. *Hodgkinia* and organelles conserve a similar set of tRNA genes. Among all codons, those shaded in gray have tRNA genes that are highly conserved in mitochondria, chloroplasts, and plastids genomes. Darker shades indicate higher conservation. *Hodgkinia* encodes tRNAs for codons with purple boxes around them. Red lettering indicates a codon reassignment in *Hodgkinia*.

aaRSs (Hansen and Moran 2012). However, the *Hodgkinia* genome encodes all 61 possible codons, and shotgun proteomics revealed that all 20 amino acids are used in *Hodgkinia* proteins (McCutcheon et al. 2009b). How *Hodgkinia* could carryout translation with so few tRNAs and aaRSs is unknown, but a few hints may be gained from reviewing how translation works in eukaryotic organelles, which encode similar sets of tRNA genes (Figure 3).

It is now clear that aaRS genes that were lost from mitochondrial genomes were transferred to the nuclear genome and subsequent import of aaRS proteins and tRNAs across the mitochondrial membrane occurs (Schneider 2011). However, the mitochondrial version of $aaRS^{Lys}$ and $aaRS^{Gly}$ have been completely lost and are replaced by splice variants of their nuclear equivalent (Schneider 2011). The mechanisms facilitating tRNA import and the extent to which it occurs are still not well understood, but membrane transport is known to occur through independent and co-import mechanisms (Rubio and Hopper 2011). Mitochondrial translation is now completely controlled by the host, as the host regulates the expression of mitochondrial aaRS genes and the membrane proteins responsible for transport of tRNAs and aaRS. This scenario provides the potential for conflict to occur between organelle and host, which face very

different evolutionary pressures. On one hand, organelles are semi-autonomous in that they divide by binary fission, and do not undergo meiosis like the genomes of their hosts. And despite accumulating evidence on the frequency and extent of mitochondrial recombination between distinct lineages (Eyre-Walker et al. 1999; Alverson et al. 2011; Rice et al. 2013; Sanchez-Puerta et al. 2015; B. Wu et al. 2015; Z. Wu et al. 2015), the general picture of organelle evolution is that of stability; most mitochondrial genomes encode the same set of genes (Gray et al. 1999) and are like tiny bacterial genomes (Burger et al. 2013). Eukaryotic genomes however, have complex genomic architectures, sexual recombination (except asexual eukaryotes), and different effective population sizes than their organelles (Cooper et al. 2015). These differences in

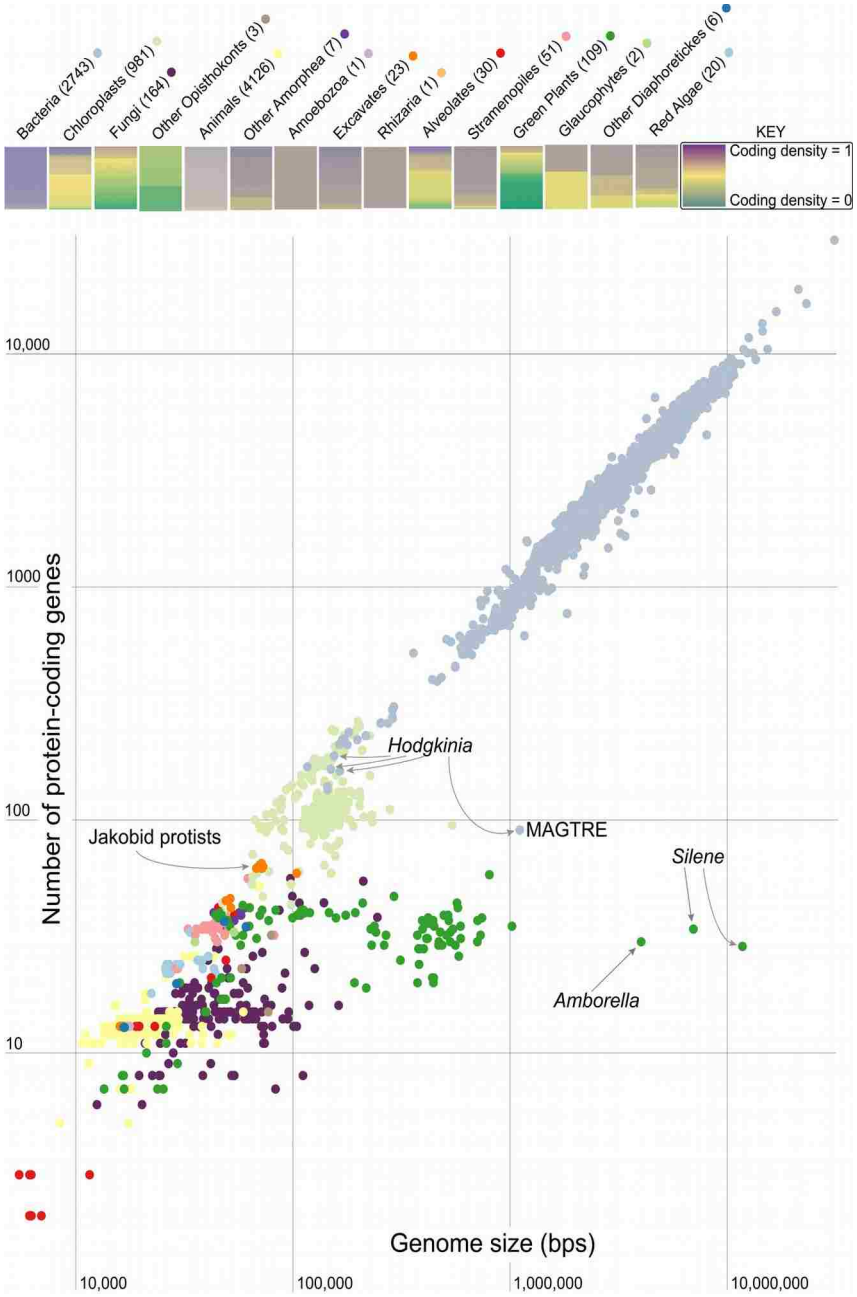


Figure 4. Genome sizes, gene numbers, and coding density for all sequenced bacterial and organelle genomes. Number of protein coding genes are plotted as a function of genome size. Genomes from organisms called out in the main text are noted. The color coded heat maps on the right show the coding density of every genome in each major group as defined by (Eme et al. 2014). A color coded key is shown at the upper right. The number of mitochondrial genomes in the major groups of eukaryotes is show to the right of each heatmap (Eme et al. 2014). Figure from (Campbell et al. 2015).

evolutionary pressures acting on host and organelle genome can cause conflict between cellular components that must interact for function of the organelle (Meiklejohn et al. 2013; Chou and Leu 2015). Nevertheless, most organellar proteins are encoded on the host genome. Do bacteria lacking conventionally essential genes—like *Sulcia* and *Hodgkinia*—import the missing cellular components like organelles? Have these bacteria cooped an entirely different strategy allowing loss of tRNA and aaRS genes? How does the interplay between interacting host-encoded proteins and bacterially-encoded proteins influence the evolutionary dynamic of the partners? My thesis seeks to address these questions, using the cicada symbiosis as a model.

1.6 The endosymbionts of the cicada *Diceroprocta semicincta*

Cicadas have a unique life history that is not shared by any other insect (Williams and Simon 1995). Like their close relatives, the spittlebugs, cicadas feed on plant xylemsap (Meyer 1993). However, cicadas feed almost exclusively on plant roots while in the nymphal stage of their life cycle. Depending on the cicada species, the nymphal stage can last between 1-17 years. At the appropriate time, entire broods synchronously emerge from underground to mate. After mating, females lay eggs in twigs and die. The eggs hatch a few months later and the nymphs drop to the ground to repeat the cycle.

Despite their unique lifestyle and global distribution, only one cicada symbiont metagenome was published before my thesis work (McCutcheon et al. 2009b). The *Sulcia-Hodgkinia* symbiont pair in this species is so far completely unique in its lack of combined aaRS genes, making this system particular interesting for learning about genome complementarity in mutualistic symbionts. The *Sulcia* genome is very similar to other *Sulcia* genomes in the Auchenorrhyncha, but *Hodgkinia* displays several unusual characteristics. Its genome is small (143,795 bp), its genomic GC content is very high (58.4%) for such a small genome, and it uses an alternative genetic code in which the base triplet UGA encodes for tryptophan instead of signaling for the termination of translation (McCutcheon et al. 2009b). While there are a few examples of smaller sized genomes, most other genomes of this size have genomic GC content of 15-20%. The only other exception to this rule is *Tremblaya PCIT*, which has a genomic GC content of 58.8% and a total genome size of 138,927 bp. A complete gene count reveals about 169 protein coding genes, of which about 140 can be assigned some hypothetical function, 16 tRNA genes, and 1 ribosomal operon. Conspicuously missing are 10 aminoacyl tRNA sythetase genes, many tRNA genes need to read all codons, RNase P, tmRNA, an ATP synthase, genes involved in cell membrane biosythesis, the majority of genes involved in DNA repair, and nearly all genes involved in metabolism. However, it is really only the apparent loss of genes involved in translation that is unusual for genomes smaller than 200,000 bp.

With dozens of bacterial genomes smaller than 0.75 Mb now available, it is largely recognized that a complete loss of almost all genes involved in metabolic processes is tolerable in the intracellular environment. Similarly, it is seemingly acceptable to loose most genes involved in DNA replication and translational control. A few subunits of the core DNA holoenzyme (*holA*, *dnaQ*, *dnaN*, *dnaX*) are present in most bacterial genomes that are smaller than 0.75 Mb, while DNA repair enzymes like *mutS* are almost always lost early on. Parts of the TCA cycle and electron transport pathways are retained in some sub-0.75 Mb genomes, but are mostly gone in sub-0.5 Mb genomes, although several cytochrome C oxidase and ATP synthase

genes remain in even the most degenerate bacterial genomes. Given the currently available sub-0.5 Mb genomes, three stand out in their extent of gene loss: *Hodgkinia, Tremblaya*, and *Nasuia*. All three have lost at least half of the 20 required aminoacyl tRNA sythetase (aaRS) genes, the ability to generate ATP, and the ability to make their own cellular membranes. The *Nasuia/Sulcia* pair is unique among the three listed above in having lost the most aaRS genes between the symbiont pair; together they retain only 9. The *Hodgkinia/Sulcia* pair is unique in having fewer than 20 combined aaRS genes, and in *Hodgkinia* encoding insufficient tRNA genes (*Nasuia* has 30). *Tremblaya* PCIT is unique in have the most degenerate genome, with no aaRS genes and only 7 tRNA genes. However, its intrabacterial endosymbiont, *Moranella*, contains a complete complement of both. How do these organisms survive? Are the hosts or co-symbionts supplying tRNAs and aaRSs? Are the degenerative processes occurring in these symbionts homologous to the process that occurred to organelles billions of years ago? How common is severe genome degeneration and what is its endpoint?

The primary focus of my thesis to better understand the evolutionary processes shaping endosymbiont genomes. My results will help us understand animal-microbe symbioses, bacterial genome evolution, and the formation of organelles. Each part of my thesis provides answers to these questions, but also raises many more as we discovered unusual biology in the cicada symbiosis. In chapter two, I investigate the mutational pressures acting on the *Hodgkinia* genome to test if mutation or selection is shaping the nucleotide content of *Hodgkinia*. I show that like most bacteria, *Hodgkinia* has a strong mutational bias and should have an AT-rich genome. This suggests that another process such as purifying selection is responsible for *Hodgkinia's* high GC content, which is perplexing because it is typically thought that selection is greatly relaxed on endosymbiont genomes. In chapter 3, I sequence and compare *Hodgkinia* genomes from three distantly related cicada species. I show that nucleotide content, gene content, and genome structure can vary drastically between cicada species. In some cicadas, unusual “speciation” events result in two or more cellularly distinct, but interdependent *Hodgkinia* lineages within a single cicada host. Chapter four addresses the conspicuously depleted gene set found in all the *Hodgkina* genomes that we have sequenced so far. I test if *Hodgkinia* and *Sulcia* tRNAs are processed despite missing the genes that encode for the enzymes that carry out these processing reactions. I also look for unconventional tRNAs in these genomes that might have been missed by traditional computational scans of the genome. Chapter five is the last chapter that presents data. Here I look for evidence that the cicada host is supporting *Hodgkinia* and *Sulcia*. I find upregulation of host genes involved in tRNA maturation, which is highly suggestive of host complementation. Although not conclusive, the apparent localization of some of these host-encoded proteins in *Hodgkinia* cells confirms this supportive role.

References

Alverson AJ, Rice DW, Dickinson S, Barry K, Palmer JD. 2011. Origins and Recombination of the Bacterial-Sized Multichromosomal Mitochondrial Genome of Cucumber. Plant Cell 23:2499–2513.

Auclair JL. 1965. Feeding and Nutrition of the Pea Aphid, Acyrthosiphon pisum (Homoptera: Aphidae), on Chemically Defined Diets of Various pH and Nutrient Levels. Ann. ent.

Soc. Am. 58:855–875.

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

Becerra JX. 2015. On the factors that promote the diversity of herbivorous insects and plants in tropical forests. PNAS 112:6098–6103.

Bennett GM, McCutcheon JP, MacDonald BR, Romanovicz D, Moran NA. 2014. Differential Genome Evolution Between Companion Symbionts in an Insect-Bacterial Symbiosis. mBio 5:e01697–14.

Bennett GM, Moran NA. 2013. Small, Smaller, Smallest: The Origins and Evolution of Ancient Dual Symbioses in a Phloem-Feeding Insect. Genome Biol Evol 5:1675–1688.

Bennett GM, Moran NA. 2015. Heritable symbiosis: The advantages and perils of an evolutionary rabbit hole. Proc Natl Acad Sci U S A:201421388.

Bennett GM, O'Grady PM. 2012. Host–plants shape insect diversity: Phylogeny, origin, and species diversity of native Hawaiian leafhoppers (Cicadellidae: Nesophrosyne). Molecular Phylogenetics and Evolution 65:705–717.

Berlese A. 1893. Le cocciniglie italiane, viventi sugli agrumi, 1: Dactylopius. Riv. patol. begetale 2.

Blochmann F. 1884. Uber die Metamorphose der Kerne in den Ovarialeiern und iiber dnt Btgilffl tkr Blastodermbildung bei den Ameisen. Verhandl. naturhist. med. Verein Heidelberg 3.

Brown AMV, Huynh LY, Bolender CM, Nelson KG, McCutcheon JP. 2013. Population genomics of a symbiont in the early stages of a pest invasion. Molecular Ecology:n/a – n/a.

Buchner P. 1965. Endosymbiosis of animals with plant microorganisms. Interscience Publishers

Burger G, Gray MW, Forget L, Lang BF. 2013. Strikingly Bacteria-Like and Gene-Rich Mitochondrial Genomes throughout Jakobid Protists. Genome Biol Evol 5:418–438.

Burke GR, Moran NA. 2011a. Responses of the pea aphid transcriptome to infection by facultative symbionts. Insect Mol Biol 20:357–365.

Burke GR, Moran NA. 2011b. Massive Genomic Decay in Serratia symbiotica, a Recently Evolved Symbiont of Aphids. Genome Biol. Evol. 3:195–208.

Campbell MA, Van Leuven JT, Meister RC, Carey KM, Simon C, McCutcheon JP. 2015. Genome expansion via lineage splitting and genome reduction in the cicada endosymbiont *Hodgkinia*. Proc Natl Acad Sci U S A 112:10192–10199.

Chou J-Y, Leu J-Y. 2015. The Red Queen in mitochondria: cyto-nuclear co-evolution, hybrid breakdown and human disease. Front Genet [Internet] 6. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4437034/

Christensen H, Fogel ML. 2011. Feeding ecology and evidence for amino acid synthesis in the periodical cicada (Magicicada). Journal of Insect Physiology 57:211–219.

Clayton AL, Oakeson KF, Gutin M, Pontes A, Dunn DM, von Niederhausern AC, Weiss RB, Fisher M, Dale C. 2012. A Novel Human-Infection-Derived Bacterium Provides Insights into the Evolutionary Origins of Mutualistic Insect–Bacterial Symbioses. PLoS Genet 8:e1002990.

Cobben RH. 1978. Evolutionary Trends in Heteroptera. Part II. Mouthpart-Structure and Feeding Strategie. Meded. Landbouwhogesch. Wageningen [Internet] 78. Available from: http://www.jstor.org/stable/2412577

Connell JH. 1972. Community interactions on marine rocky intertidal shores. Annual Review of Ecology and Systematics 3:169–192.

Cooper BS, Burrus CR, Ji C, Hahn MW, Montooth KL. 2015. Similar Efficacies of Selection Shape Mitochondrial and Nuclear Genes in Both Drosophila melanogaster and Homo sapiens. G3 5:2165–2176.

Crawley MJ. 1989. Insect Herbivores and Plant Population Dynamics. Annual Review of Entomology 34:531–562.

Dadd RH, Krieger DL. 1968. Dietary amino acid requirements of the aphid, Myzus persicae. Journal of Insect Physiology 14:741–764.

Dadd RH, Krieger DL, Mittler TE. 1967. Studies on the artificial feeding of the aphid Myzus persicae (Sulzer)—IV. Requirements for water-soluble vitamins and ascorbic acid. Journal of Insect Physiology 13:249–272.

Dale C, Wang B, Moran N, Ochman H. 2003. Loss of DNA Recombinational Repair Enzymes in the Initial Stages of Genome Degeneration. Mol Biol Evol 20:1188–1194.

Dedryver C-A, Le Ralec A, Fabre F. 2010. The conflicting relationships between aphids and men: A review of aphid damage and control strategies. Comptes Rendus Biologies 333:539–553.

Douglas A. 1989. Mycetocyte Symbiosis in Insects. Biol. Rev. Cambridge Philosophic. Soc. 64:409–434.

Douglas AE. 1988. Sulphate utilization in an aphid symbiosis. Insect Biochemistry 18:599–605.

Douglas AE. 2006. Phloem-sap feeding by animals: problems and solutions. J. Exp. Bot. 57:747–

754.

Douglas AE, Prosser WA. 1992. Synthesis of the essential amino acid tryptophan in the pea aphid (Acyrthosiphon pisum) symbiosis. Journal of Insect Physiology 38:565–568.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of Spontaneous Mutation. Genetics 148:1667–1686.

Duncan RP, Husnik F, Van Leuven JT, Gilbert DG, Dávalos LM, McCutcheon JP, Wilson ACC. 2014. Dynamic recruitment of amino acid transporters to the insect/symbiont interface. Mol Ecol 23:1608–1623.

Elderd BD, Rehill BJ, Haynes KJ, Dwyer G. 2013. Induced plant defenses, host–pathogen interactions, and forest insect outbreaks. PNAS 110:14978–14983.

Eme L, Sharpe SC, Brown MW, Roger AJ. 2014. On the Age of Eukaryotes: Evaluating Evidence from Fossils and Molecular Clocks. Cold Spring Harb Perspect Biol 6:a016139.

Engel MS. 2015. Insect evolution. Current Biology 25:R868–R872.

Estrela S, Morris JJ, Kerr B. 2015. Private benefits and metabolic conflicts shape the emergence of microbial interdependencies. Environ Microbiol [Internet]. Available from: http://onlinelibrary.wiley.com/doi/10.1111/1462-2920.13028/abstract

Eyre-Walker A, Smith NH, Smith JM. 1999. How clonal are human mitochondria? Proc Biol Sci 266:477–483.

Febvay G, Liadouze I, Guillaud J, Bonnot G. 1995. Analysis of energetic amino acid metabolism in Acyrthosiphon pisum: A multidimensional approach to amino acid metabolism in aphids. Arch. Insect Biochem. Physiol. 29:45–69.

Ferrari J, Darby AC, Daniell TJ, Godfray HCJ, Douglas AE. 2004. Linking the bacterial community in pea aphids with host-plant use and natural enemy resistance. Ecological Entomology 29:60–65.

Gerardo NM, Altincicek B, Anselme C, Atamian H, Barribeau SM, de Vos M, Duncan EJ, Evans JD, Gabaldón T, Ghanim M, et al. 2010. Immunity and other defenses in pea aphids, Acyrthosiphon pisum. Genome Biology 11:R21.

Grassi G, Millard P, Wendler R, Minotta G, Tagliavini M. 2002. Measurement of xylem sap amino acid concentrations in conjunction with whole tree transpiration estimates spring N remobilization by cherry (Prunus avium L.) trees. Plant, Cell & Environment 25:1689–1699.

Gray MW. 2012. Mitochondrial Evolution. Cold Spring Harb Perspect Biol 4:a011403.

Gray MW, Burger G, Lang BF. 1999. Mitochondrial Evolution. Science 283:1476–1481.

Greenberg EP. 2003. Bacterial communication: Tiny teamwork. Nature 424:134.

Hansen AK, Moran NA. 2012. Altered tRNA characteristics and 3′ maturation in bacterial symbionts with reduced genomes. Nucleic Acids Res. 40:7870–7884.

Hardin G. 1960. The Competitive Exclusion Principle. Science 131:1292–1297.

Hayashi H, Chino M. 1986. Collection of Pure Phloem Sap from Wheat and its Chemical Composition. Plant Cell Physiol 27:1387–1393.

Hedges LM, Brownlie JC, O'Neill SL, Johnson KN. 2008. Wolbachia and Virus Protection in Insects. Science 322:702–702.

Hershberg R, Petrov DA. 2010. Evidence That Mutation Is Universally Biased towards AT in Bacteria. PLoS Genet 6:e1001115.

Hershberg R, Tang H, Petrov D. 2007. Reduced selection leads to accelerated gene loss in Shigella. Genome Biology 8:R164.

Heymons R. 1899. Beitriige zur Morphologie und Entwicklungsgeschichte der Rynchoten. Nova Acta Leopoldina Carol. Akad. 74.

Hijaz F, Killiny N. 2014. Collection and Chemical Composition of Phloem Sap from Citrus sinensis L. Osbeck (Sweet Orange). PLoS ONE 9:e101830.

Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of Selection upon Genomic GC-Content in Bacteria. PLoS Genet 6:e1001107.

Hill DS. 1987. Agricultural Insect Pests of Temperate Regions and Their Control. CUP Archive

Holmgren N. 1902. Uber die Exkretionsorgane des Apionfiavipes und Dasytes niger. Anat. Anz. 22.

Hooke R. 1665. Micrographia. London

Hosokawa T, Kikuchi Y, Shimada M, Fukatsu T. 2007. Obligate symbiont involved in pest status of host insect. Proceedings of the Royal Society B: Biological Sciences 274:1979.

Hotopp JCD, Clark ME, Oliveira DCSG, Foster JM, Fischer P, Torres MCM, Giebel JD, Kumar N, Ishmael N, Wang S, et al. 2007. Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes. Science 317:1753–1756.

Hougen-Eitzman D, Rausher MD. 1994. Interactions between Herbivorous Insects and Plant-Insect Coevolution. The American Naturalist 143:677–697.

Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, Tanaka M, Satoh N, Bachtrog D, Wilson ACC, et al. 2013. Horizontal Gene Transfer from Diverse Bacteria to an Insect Genome Enables a Tripartite Nested Mealybug Symbiosis. Cell 153:1567–1578.

Huxley T. 1858. On the agamic reproduction and morphology of Aphids. Trns. Linn. Soc. London 22.

Itoh T, Martin W, Nei M. 2002. Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. PNAS 99:12944–12948.

Jeschke WD, Klagges S, Bhatti AS. 1995. Collection and composition of xylem sap and root structure in two halophytic species. Plant Soil 172:97–106.

Johnson SN, Clark KE, Hartley SE, Jones TH, McKenzie SW, Koricheva J. 2012. Aboveground–belowground herbivore interactions: a meta-analysis. Ecology 93:2208–2215.

Joyce EA, Chan K, Salama NR, Falkow S. 2002. Redefining bacterial populations: a post-genomic reformation. Nature Reviews Genetics 3:462–473.

Kaltenpoth M, Engl T. 2014. Defensive microbial symbionts in Hymenoptera.Clay K, editor. Functional Ecology 28:315–327.

Kehr J, Buhtz A, Giavalisco P. 2005. Analysis of xylem sap proteins from Brassica napus. BMC Plant Biology 5:11.

Kiers ET, West SA. 2015. Evolving new organisms via symbiosis. Science 348:392–394.

Koga R, Bennett GM, Cryan JR, Moran NA. 2013. Evolutionary replacement of obligate symbionts in an ancient and diverse insect lineage. Environmental Microbiology:n/a – n/a.

Koga R, Moran NA. 2014. Swapping symbionts in spittlebugs: evolutionary replacement of a reduced genome symbiont. ISME J 8:1237–1246.

Krishnan HB, Natarajan SS, Bennett JO, Sicher RC. 2011. Protein and metabolite composition of xylem sap from field-grown soybeans (Glycine max). Planta 233:921–931.

Kümmerli R, Santorelli LA, Granato ET, Dumas Z, Dobay A, Griffin AS, West SA. 2015. Co-evolutionary dynamics between public good producers and cheats in the bacterium Pseudomonas aeruginosa. J. Evol. Biol. [Internet]. Available from: http://onlinelibrary.wiley.com/doi/10.1111/jeb.12751/abstract

Kuo C-H, Ochman H. 2009. Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. Biology Direct 4:35.

Lai CY, Baumann L, Baumann P. 1994. Amplification of trpEG: adaptation of Buchnera

aphidicola to an endosymbiotic association with aphids. PNAS 91:3819–3823.

Van Leuven JT, Meister RC, Simon C, McCutcheon JP. 2014. Sympatric Speciation in a Bacterial Endosymbiont Results in Two Genomes with the Functionality of One. Cell 158:1270–1280.

Leydig F. 1850. Einige Bemerkungen iiber die Entwicklung der Blattliiuse. Zool. 2.

Marais GAB, Calteau A, Tenaillon O. 2007. Mutation rate and genome reduction in endosymbiotic and free-living bacteria. Genetica 134:205–210.

McCutcheon JP. 2010. The bacterial essence of tiny symbiont genomes. Curr Opin Microbiol 13:73.

McCutcheon JP, von Dohlen CD. 2011. An Interdependent Metabolic Patchwork in the Nested Symbiosis of Mealybugs. Curr Biol 21:1366–1372.

McCutcheon JP, McDonald BR, Moran NA. 2009a. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. Proc Natl Acad Sci U S A 106:15394–15399.

McCutcheon JP, McDonald BR, Moran NA. 2009b. Origin of an Alternative Genetic Code in the Extremely Small and GC–Rich Genome of a Bacterial Symbiont. PLoS Genet 5:e1000565.

McCutcheon JP, Moran NA. 2010. Functional Convergence in Reduced Genomes of Bacterial Symbionts Spanning 200 My of Evolution. Genome Biol Evol 2:708–718.

McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10:13–26.

Meiklejohn CD, Holmbeck MA, Siddiq MA, Abt DN, Rand DM, Montooth KL. 2013. An Incompatibility between a Mitochondrial tRNA and Its Nuclear-Encoded tRNA Synthetase Compromises Development and Fitness in Drosophila. PLoS Genet 9:e1003238.

D'Mello JPF. 2015. Amino Acids in Higher Plants. CABI

Menninger HL, Palmer MA, Craig LS, Richardson DC. 2008. Periodical Cicada Detritus Impacts Stream Ecosystem Metabolism. Ecosystems 11:1306–1317.

Metschnikoff H. 1866. Untersuchungen iiber die Embryologie der Hemipteren. Z. wiss. Zool. 16.

Meyer GA. 1993. A Comparison of the Impacts of Leaf- And Sap-Feeding Insects on Growth and Allocation of Goldenrod. Ecology 74:1101–1116.

Mira A, Moran NA. 2002. Estimating Population Size and Transmission Bottlenecks in

Maternally Transmitted Endosymbiotic Bacteria. Microbial Ecology 44:137–143.

Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends in Genetics 17:589–596.

Mitter C, Farrell B, Wiegmann B. 1988. The Phylogenetic Study of Adaptive Zones: Has Phytophagy Promoted Insect Diversification? The American Naturalist 132:107–128.

Mittler TE. 1971. Dietary Amino Acid Requirements of the Aphid Myzus persicae Affected by Antibiotic Uptake. J. Nutr. 101:1023–1028.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A 93:2873–2878.

Moran NA. 2002. Microbial minimalism: genome reduction in bacterial pathogens. Cell 108:583–586.

Moran NA, Bennett GM. 2014. The Tiniest Tiny Genomes. Annu Rev Microbiol 68:195–215.

Moran NA, Dale C, Dunbar H, Smith WA, Ochman H. 2003. Intracellular symbionts of sharpshooters (Insecta: Hemiptera: Cicadellinae) form a distinct clade with a small genome. Environmental Microbiology 5:116–126.

Moran NA, Mira A. 2001. The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. Genome Biol 2:1–12.

Moran NA, Munson MA, Baumann P, Ishikawa H. 1993. A Molecular Clock in Endosymbiotic Bacteria is Calibrated Using the Insect Hosts. Proc. R. Soc. Lond. B 253:167–171.

Moran NA, Tran P, Gerardo NM. 2005. Symbiosis and Insect Diversification: an Ancient Symbiont of Sap-Feeding Insects from the Bacterial Phylum Bacteroidetes. Appl. Environ. Microbiol. 71:8802–8810.

Moran NA, Yun Y. 2015. Experimental replacement of an obligate insect symbiont. PNAS 112:2093–2096.

Nakabachi A, Ishida K, Hongoh Y, Ohkuma M, Miyagishima S. 2014. Aphid gene of bacterial origin encodes a protein transported to an obligate endosymbiont. Curr Biol 24:R640–R641.

Nakabachi A, Ueoka R, Oshima K, Teta R, Mangoni A, Gurgui M, Oldham NJ, van Echten-Deckert G, Okamura K, Yamamoto K, et al. 2013. Defensive Bacteriome Symbiont with a Drastically Reduced Genome. Current Biology 23:1478–1484.

Nikoh N, McCutcheon JP, Kudo T, Miyagishima S, Moran NA, Nakabachi A. 2010. Bacterial Genes in the Aphid Genome: Absence of Functional Gene Transfer from Buchnera to Its

Host. PLoS Genet 6:e1000827.

Novotny V, Basset Y, Miller SE, Weiblen GD, Bremer B, Cizek L, Drozd P. 2002. Low host specificity of herbivorous insects in a tropical forest. Nature 416:841–844.

Nowlin WH, González MJ, Vanni MJ, Stevens MHH, Fields MW, Valente JJ. 2007. Allochthonous subsidy of periodical cicadas affects the dynamics and stability of pond communities. Ecology 88:2174–2186.

Oakeson KF, Gil R, Clayton AL, Dunn DM, Niederhausern AC von, Hamil C, Aoyagi A, Duval B, Baca A, Silva FJ, et al. 2014. Genome Degeneration and Adaptation in a Nascent Stage of Symbiosis. Genome Biol Evol 6:76–93.

Ochman H, Elwyn S, Moran NA. 1999. Calibrating bacterial evolution. PNAS 96:12638–12643.

Oliver KM, Degnan PH, Burke GR, Moran NA. 2010. Facultative Symbionts in Aphids and the Horizontal Transfer of Ecologically Important Traits. Annual Review of Entomology 55:247–266.

Oliver KM, Moran NA, Hunter MS. 2005. Variation in resistance to parasitism in aphids is due to symbionts not host genotype. PNAS 102:12795–12800.

Pace NR. 1997. A molecular view of microbial diversity and the biosphere. Science 276:734–740.

Pande S, Shitut S, Freund L, Westermann M, Bertels F, Colesie C, Bischofs IB, Kost C. 2015. Metabolic cross-feeding via intercellular nanotubes among bacteria. Nat Commun [Internet] 6. Available from: http://www.nature.com.offcampus.lib.washington.edu/ncomms/2015/150223/ncomms7238/full/ncomms7238.html

Pérez-Brocal V, Gil R, Ramos S, Lamelas A, Postigo M, Michelena JM, Silva FJ, Moya A, Latorre A. 2006. A Small Microbial Genome: The End of a Long Symbiotic Relationship? Science 314:312–313.

Pett W, Lavrov DV. 2015. Cytonuclear Interactions in the Evolution of Animal Mitochondrial tRNA Metabolism. Genome Biol Evol 7:2089–2101.

Porta A. 1900. Ricerche sull' Aphrophora spumaria. Rend. 1st lomb. sci. lett 2.

Powell KS, Cooper PD, Forneck A. 2013. The Biology, Physiology and Host-Plant Interactions of Grape Phylloxera Daktulosphaira vitifoliae. In: Johnson SN, Hiltpold I, Turlings TCJ, editors. Behaviour and Physiology of Root Herbivores. Vol. 45. San Diego: Elsevier Academic Press Inc. p. 159–218.

Price DRG, Duncan RP, Shigenobu S, Wilson ACC. 2011. Genome Expansion and Differential

Expression of Amino Acid Transporters at the Aphid/Buchnera Symbiotic Interface. Mol Biol Evol 28:3113–3126.

Price DRG, Feng H, Baker JD, Bavan S, Luetje CW, Wilson ACC. 2014. Aphid amino acid transporter regulates glutamine supply to intracellular bacterial symbionts. PNAS 111:320–325.

Rice DW, Alverson AJ, Richardson AO, Young GJ, Sanchez-Puerta MV, Munzinger J, Barry K, Boore JL, Zhang Y, dePamphilis CW, et al. 2013. Horizontal Transfer of Entire Genomes via Mitochondrial Fusion in the Angiosperm Amborella. Science 342:1468–1473.

Rubio MAT, Hopper AK. 2011. tRNA travels from the cytoplasm to organelles. Wiley Interdiscip Rev RNA 2:802–817.

Sabree ZL, Huang CY, Okusu A, Moran NA, Normark BB. 2013. The nutrient supplying capabilities of Uzinura, an endosymbiont of armoured scale insects. Environ Microbiol 15:1988–1999.

Salinas-Giegé T, Giegé R, Giegé P. 2015. tRNA Biology in Mitochondria. International Journal of Molecular Sciences 16:4518–4559.

Sanchez-Puerta MV, Zubko MK, Palmer JD. 2015. Homologous recombination and retention of a single form of most genes shape the highly chimeric mitochondrial genome of a cybrid plant. New Phytol 206:381–396.

Sandström J, Moran N. 1999. How nutritionally imbalanced is phloem sap for aphids? Entomologia Experimentalis et Applicata 91:203–210.

Sandström J, Pettersson J. 1994. Amino acid composition of phloem sap and the relation to intraspecific variation in pea aphid (Acyrthosiphon pisum) performance. Journal of Insect Physiology 40:947–955.

Sasaki T, Ishikawa H. 1995. Production of essential amino acids from glutamate by mycetocyte symbionts of the pea aphid, Acyrthosiphon pisum. Journal of Insect Physiology 41:41–46.

Scarborough CL, Ferrari J, Godfray HCJ. 2005. Aphid Protected from Pathogen by Endosymbiont. Science 310:1781–1781.

Schneider A. 2011. Mitochondrial tRNA Import and Its Consequences for Mitochondrial Translation. Annual Review of Biochemistry 80:1033–1053.

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

Sloan DB, Moran NA. 2013. The Evolution of Genomic Instability in the Obligate

Endosymbionts of Whiteflies. Genome Biol Evol 5:783–793.

Smith JM, Smith NH, O'Rourke M, Spratt BG. 1993. How clonal are bacteria? PNAS 90:4384–4388.

Stephens AEA, Westoby M. 2015. Effects of insect attack to stems on plant survival, growth, reproduction and photosynthesis. Oikos 124:266–273.

Stork NE, McBroom J, Gely C, Hamilton AJ. 2015. New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. PNAS 112:7519–7523.

Suzuki T, Nagao A, Suzuki T. 2011. Human Mitochondrial tRNAs: Biogenesis, Function, Structural Aspects, and Diseases. Annu Rev Genet 45:299–329.

Takiya DM, Tran PL, Dietrich CH, Moran NA. 2006. Co-cladogenesis spanning three phyla: leafhoppers (Insecta: Hemiptera: Cicadellidae) and their dual bacterial symbionts. Mol. Ecol. 15:4175–4191.

Thomas CM, Nielsen KM. 2005. Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. Nat Rev Micro 3:711–721.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nature Reviews Genetics 5:123–135.

Toju H, Hosokawa T, Koga R, Nikoh N, Meng XY, Kimura N, Fukatsu T. 2010. "Candidatus Curculioniphilus Buchneri," a Novel Clade of Bacterial Endocellular Symbionts from Weevils of the Genus Curculio. Appl. Environ. Microbiol. 76:275–282.

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, et al. 2009. Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. PLoS Genet 5:e1000344.

Truman JW, Riddiford LM. 1999. The origins of insect metamorphosis. Nature 401:447–452.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43.

Ueda K, Yamashita A, Ishikawa J, Shimada M, Watsuji T, Morimura K, Ikeda H, Hattori M, Beppu T. 2004. Genome sequence of Symbiobacterium thermophilum, an uncultivable bacterium that depends on microbial commensalism. Nucl. Acids Res. 32:4937–4944.

Van Leuven JT, McCutcheon JP. 2011. An AT mutational bias in the tiny GC-rich endosymbiont genome of *Hodgkinia*. Genome Biology and Evolution [Internet]. Available from: http://gbe.oxfordjournals.org/content/early/2011/11/23/gbe.evr125.abstract

Vorwerk S, Martinez-Torres D, Forneck A. 2007. Pantoea agglomerans-associated bacteria in grape phylloxera (Daktulosphaira vitifoliae, Fitch). Agricultural and Forest Entomology 9:57–64.

Wernegreen JJ. 2015. Endosymbiont evolution: predictions from theory and surprises from genomes. Ann N Y Acad Sci [Internet]. Available from: http://onlinelibrary.wiley.com/doi/10.1111/nyas.12740/abstract

Williams KS, Simon C. 1995. The Ecology, Behavior, and Evolution of Periodical Cicadas. Annual Review of Entomology 40:269–295.

Will T, Furch ACU, Zimmermann MR. 2013. How phloem-feeding insects face the challenge of phloem-located defenses. Front Plant Sci [Internet] 4. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3756233/

Woolfit M, Bromham L. 2003. Increased Rates of Sequence Evolution in Endosymbiotic Bacteria and Fungi with Small Effective Population Sizes. Mol Biol Evol 20:1545–1555.

Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, Boffelli D, Anderson IJ, Barry KW, Shapiro HJ, et al. 2006. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature 443:950.

Wu B, Buljic A, Hao W. 2015. Extensive Horizontal Transfer and Homologous Recombination Generate Highly Chimeric Mitochondrial Genomes in Yeast. Mol Biol Evol 32:2559–2570.

Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL, Khouri H, Tallon LJ, Zaborsky JM, Dunbar HE, Tran PL, et al. 2006. Metabolic Complementarity and Genomics of the Dual Bacterial Symbiosis of Sharpshooters. PLoS Biol 4:e188.

Wu Z, Cuthbert JM, Taylor DR, Sloan DB. 2015. The massive mitochondrial genome of the angiosperm Silene noctiflora is evolving by gain or loss of entire chromosomes. PNAS:201421397.

Yang LH. 2004. Periodical Cicadas as Resource Pulses in North American Forests. Science 306:1565–1567.

Zvereva EL, Lanta V, Kozlov MV. 2010. Effects of sap-feeding insect herbivores on growth and reproduction of woody plants: a meta-analysis of experimental studies. Oecologia 163:949–960.

Chapter 2: Nucleotide content diversity in *Hodgkina* genomes
Published as, Van Leuven JT, McCutcheon JP. 2012. An AT Mutational Bias in the Tiny GC-Rich Endosymbiont Genome of *Hodgkinia*. Genome Biol Evol 4:24–27

Summary
The fractional guanine-cytosine (GC) contents of sequenced bacterial genomes range from 13% to 75%. Despite several decades of research aimed at understanding this wide variation, the forces controlling GC content are not well understood. Recent work has suggested that a universal adenine-thymine (AT) mutational bias exists in all bacteria and that the elevated GC contents found in some bacterial genomes is due to genome-wide selection for increased GC content. These results are generally consistent with the low GC contents observed in most strict endosymbiotic bacterial genomes, where the loss of DNA repair mechanisms combined with the population genetic effects of small effective population sizes and decreased recombination should lower the efficacy of selection and shift the equilibrium GC content in the mutationally favored AT direction. Surprisingly, the two smallest bacterial genomes, *Candidatus* Hodgkinia cicadicola (144 kb) and *Candidatus* Tremblaya princeps (139 kb), have the unusual combination of highly reduced genomes and elevated GC contents, raising the possibility that these bacteria may be exceptions to the otherwise apparent universal bacterial AT mutational bias. Here, using population genomic data generated from the *Hodgkinia* genome project, we show that *Hodgkinia* has a clear AT mutational bias. These results provide further evidence that an AT mutational bias is universal in bacteria, even in strict endosymbionts with elevated genomic GC contents.

2.1 Introduction

The Smallest Bacterial Genomes Tend to Be Strongly AT Biased, with the exception of Hodgkinia and Tremblaya

Genome reduction in bacteria is usually associated with a genome-wide shift towards increased AT content (Moran 2002; Bentley and Parkhill 2004; McCutcheon et al. 2009). This pattern is especially pronounced in bacteria that live exclusively in the cytoplasm of host cells; for example, the two most extremely AT biased bacterial genomes yet reported are from the insect nutritional endosymbionts *Candidatus* Zinderia insecticola (13.5% GC) (McCutcheon and Moran 2012) and *Candidatus* Carsonella ruddii (16.5% GC) (Nakabachi et al. 2006). Two mechanisms are thought to explain the reduced GC content of endosymbiont genomes. First, endosymbionts tend to lose genes involved in DNA repair and recombination during genome reduction (Dale et al. 2003; Moran et al. 2008), which increases the load of unrepaired DNA damage. Second, endosymbionts have small effective population sizes and reduced rates of recombination, which reduces the efficacy of selection and allows more slightly deleterious mutations to be fixed by random genetic drift (Moran 1996; Woolfit and Bromham 2003). Combined with what seems to be an AT mutational bias in bacteria lacking DNA repair enzymes (Lind and Andersson 2008), these forces are thought to shift the GC-AT equilibrium towards AT in endosymbiont genomes. Until recently, empirical data from complete bacterial genomes universally supported this hypothesis. Remarkably, the only two known exceptions to this trend are from bacteria with the smallest reported genomes: *Candidatus* Hodgkinia cicadicola (hereby referred to as *Hodgkinia* for simplicity, 144 kb, 58.4% GC (McCutcheon et al. 2009)) and *Candidatus* Tremblaya princeps (*Tremblaya*, 138 kb, 58.8% GC (McCutcheon and von Dohlen 2011)). *Hodgkinia* is a member of the Alphaproteobacteria, a group in which most free-living members have GC-rich genomes, and most obligate intracellular members have reduced genomes that show the expected decrease in GC content (McCutcheon et al. 2009). These observations led to the hypothesis that the high GC content of *Hodgkinia* resulted from the retention of a GC mutational bias that was present its free-living alphaproteobacterial ancestor (McCutcheon et al. 2009). That the GC content at the 3rd position of 4-fold degenerate codons (GC4) in *Hodgkinia* is higher than the overall GC content in the genome (62.5% vs. 58.4%) seemed to support this hypothesis, as these positions are expected to be under little or no selection for protein-coding sequence, and were therefore thought to more clearly reflect the mutational biases inherent in *Hodgkinia's* replication machinery (McCutcheon et al. 2009).

Recent Work Suggests that all Bacteria Have an Inherent AT Mutational Bias

Two recent reports provide evidence that an AT mutational bias exists in all bacteria (Hershberg and Petrov 2010; Hildebrand et al. 2010). The authors of both papers conclude that selection for increased GC content, or a selection-like process such as biased gene conversion (BGC), is the most likely explanation for the diverging patterns of AT biased mutation and GC biased substitution observed in most bacterial genomes (Hershberg and Petrov 2010; Hildebrand et al. 2010). Both papers also single out *Hodgkinia* as an outlier and possible exception to this rule (Hershberg and Petrov 2010; Hildebrand et al. 2010). To help clarify the roles of mutational

biases and selection on the GC content of the *Hodgkinia* genome, we sought to determine the direction of *Hodgkinia's* mutational bias (if any) from existing population data generated during genome sequencing.

2.2 Measuring mutation in pooled DNA samples

Single Nucleotide Polymorphisms in the Hodgkinia Genome Reveal an AT Mutational Bias

The published *Hodgkinia* genome was generated by combining samples from 10 wild-caught individuals of the cicada *Diceroprocta semicincta* (McCutcheon et al. 2009). We reasoned that it might be possible to calculate mutational patterns from these population genomic data. We first reconfirmed that the pooled sample was from a single species of cicada by verifying a low level of sequence polymorphisms in the mitochondrial cytochrome c oxidase I (COI) sequence of the cicada (about 0.6% of 815 sites were polymorphic, well within the 1-2% divergence levels typically seen in conspecific pairs of animal COI sequences (Hebert et al. 2003)). We then calculated the number of single nucleotide polymorphisms (SNPs) in the *Hodgkinia* genome falling into all possible nucleotide change categories, and found that the majority of mutations (115 of 179, or 64%) were in the GC to AT direction. (The *Tremblaya* genome was generated from only 3 lab-reared insects, and no high-quality SNPs were observed in these data.)

2.3 Direction of mutation in *Hodgkinia* from *D. semicincta*

To unambiguously assign a mutational direction to the SNPs, we used a draft *Hodgkinia* genome assembly from a closely related but undescribed cicada species (referred to here as the cryptic species) as an outgroup to verify the ancestral state of each position where a SNP was identified (see Supplementary Materials for a complete description of the methods). The pairwise nucleotide divergence between partial mitochondrial CO1 sequences from *Diceroprocta semicincta* and the cryptic species was 3.5%. Of the 179 SNPs

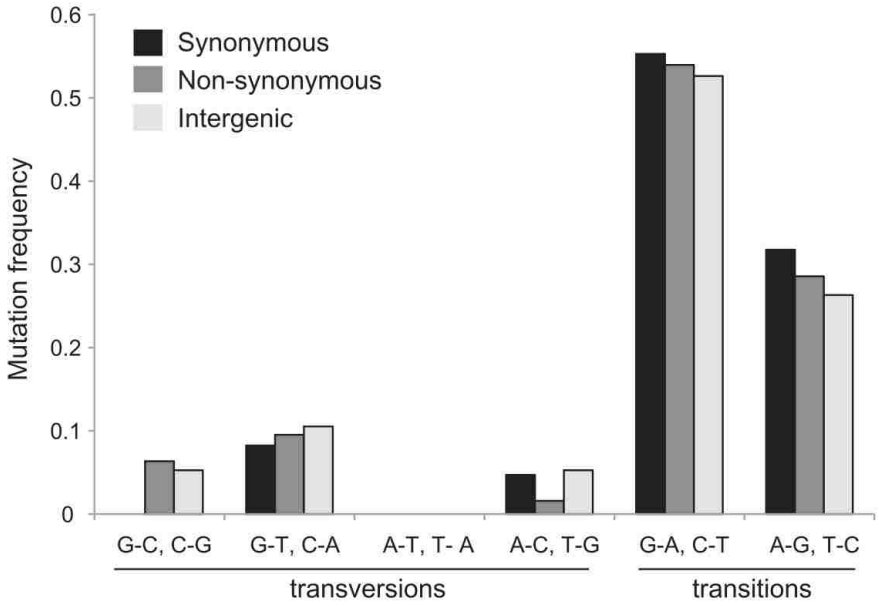


Figure 1. The majority of SNPs in the *Hodgkinia* genome are G to A or C to T transitions and collectively show a pronounced AT mutational bias. SNPs are shown as a percentage of the total number in each category (synonymous, nonsynonymous, and intergenic sites).

initially identified, 12 were not covered by contigs from the cryptic species. These 12 were removed from the dataset, resulting in 167 SNPs in which the direction of mutation could be confidently determined (Figure 1, Table 1, Table S1). The expected equilibrium GC content (GCeq) given the mutational patterns observed in the polarized data is 42%, significantly lower that the observed genomic value of 58% (Table 1).

Table 1. Raw SNP counts, mutation rates, and expected GC equilibrium values for synonymous (S), nonsynonymous (NS), and intergenic (IG) sites.

	Counts		Rates			
	Number of GC → AT	Number of AT → GC	$r_{GC \to AT}$	$r_{AT \to GC}$	Current GC	GC_{eq}
S	54 (43–67)	31 (22–40)	1.9×10^{-3}	1.9×10^{-3}	63	**50** (40–58)
NS	40 (29–51)	19 (12–26)	8.0×10^{-4}	4.9×10^{-4}	56	**38** (27–50)
IG	12 (7–19)	6 (2–10)	2.6×10^{-3}	1.7×10^{-3}	56	39 (17–56)
All	106 (89–122)	56 (44–68)	$\mathbf{1.3 \times 10^{-3}}$	$\mathbf{9.5 \times 10^{-4}}$	58	**42** (36–49)

NOTE.—The numbers in parentheses are 95% confidence intervals; significant values are bolded. The values do not sum to 167 because five SNPs did not alter the GC content (i.e., a G to C mutation).

2.4 Effects of purifying selection on segregating polymorphisms

To estimate the strength of selection acting on these SNPs, we calculated the ratio of non-synonymous and synonymous polymorphisms per non-synonymous and synonymous site (dN/dS), and found evidence for weak purifying selection (dN/dS = 0.37). This value is slightly lower but consistent with values reported previously for populations of clonal bacterial pathogens, which range from 0.45 to 0.64 (Hershberg and Petrov 2010). Differences in the magnitude of dN/dS need to be interpreted with caution in this situation, as this measure assumes that sequence polymorphisms are fixed substitutions between species, not intraspecific mutations segregating in a population (Kryazhimskiy and Plotkin 2008). Some SNPs in the pooled dataset include those at high frequencies, and we assume that these SNPs have been segregating in the population for some time and may have been exposed to significant levels of purifying selection. To assess whether we could measure differences in (1) the levels of purifying selection and (2) the magnitude of the AT mutational bias for SNPs partitioned into different frequency bins, we calculated dN/dS and GCeq values for SNPs binned at 0.1 frequency intervals (Figure 2). As ten individuals were pooled for sequencing, an ideal experiment would reveal SNPs clustering at frequencies of 0.1, 0.2, 0.3 and so on up to 0.9. We did not observe an increased number of SNPs near these expected frequencies, and attribute this non-ideal behavior to numerous potential experimental and computational artifacts (see Supplementary Methods for a full discussion). Nevertheless, these results confirm that SNPs present in the population at lower frequencies have been exposed to less purifying selection (indicated by a higher dN/dS value) and are more strongly AT biased than SNPs present at higher frequencies (Figure 2). For example, the GCeq content of the *Hodgkinia* genome is calculated to be 37% using only SNPs called at a frequency of 0.1 or less, lower than the 42% calculated when all SNPs are included. The true *Hodgkinia* GCeq is therefore probably closer to 37%, or perhaps even lower. From these data, we conclude that *Hodgkinia* has an AT mutational bias.

2.5 Genomic GC content of *Hodgkinia* genomes

Sequenced 16S PCR product from a number of *Hodgkinia* species suggest that the genome GC content of *Hodgkinia* many vary dramatically between cicada hosts (McCutcheon et al. 2009 and unpublished). To confirm this, we sequenced the metagenomes from the cicada species *Tettigades ulnaria*, *Tettigades undata*, *Tettigades chilensis*, and *Magicicada tredicium*, which have genomic GC contents of 46%, 47%, 45%, and 28% (Van Leuven et al. 2014; Campbell et al. 2015 and unpublished). To my knowledge, this dramatic range is unprecedented in any other set of monophyletic bacterial sub-species. Strangely, the genomic GC content of the mealybug symbiont *Tremblaya* is also quite variable, with the published genome varing from 42- 59%, and the unpublished genomes dropping well below 42% (McCutcheon and von Dohlen 2011; Husnik et al. 2013). It is unknown why these groups of bacteria have such a broad range of genomic GC content, although, we do propose that purifying selection is relaxed on the *Hodgkinia* genomes of some cicada species (see chapter 3).

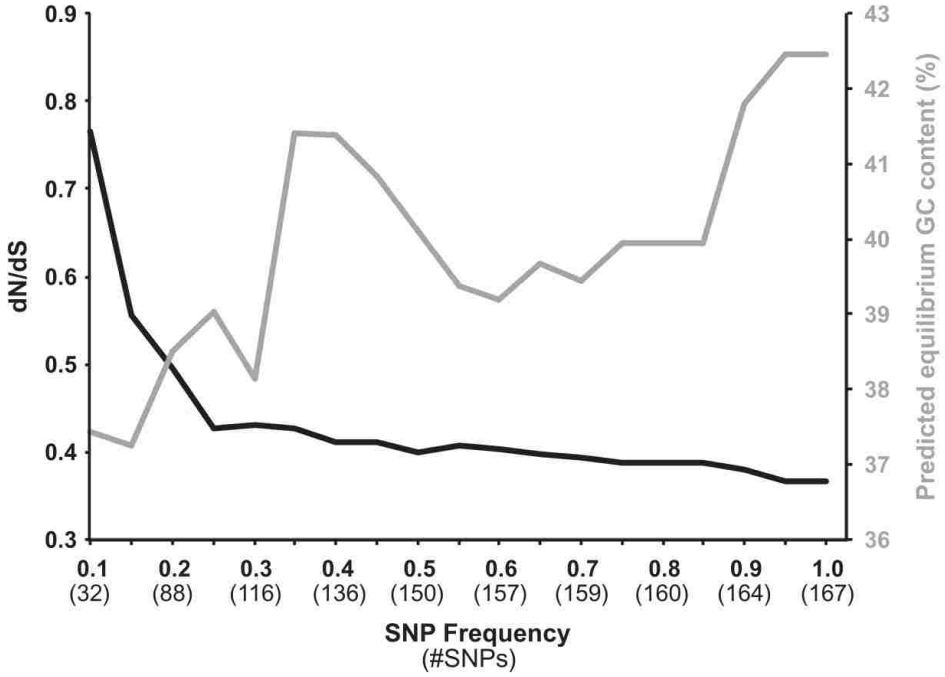


Figure 2. Plotting GCeq (gray line) and dN/dS (black) at different SNP frequency cutoffs shows that SNPs present at lower frequencies (which are likely more recent mutations) have been subjected to less selection and are more AT biased.

2.6 Discussion

Why Does Hodgkinia Have an Elevated Genomic GC Content?

While our data clearly show an AT mutational bias in *Hodgkinia*, they do not directly implicate the force(s) responsible for the disparity between the observed patterns of mutation and substitution. Hershberg, Hildebrand and co-workers suggest selection, or a selection-like process such as biased gene conversion, as the force driving the difference in bacteria (Hershberg and

Petrov 2010; Hildebrand et al. 2010). In bacteria, biased gene conversion involves horizontal gene transfer, recombination and DNA repair-based mechanisms (Rocha and Feil 2010). As *Hodgkinia* encodes no gene homologs capable of these processes (McCutcheon et al. 2009), biased gene conversion seems unlikely to be responsible for *Hodgkinia's* elevated GC content. Therefore, it appears that an unidentified selective force (or forces) is the most likely explanation for the GC bias in the *Hodgkinia* genome, although other explanations cannot be ruled out given the present data. For example, it is possible that GC content in *Hodgkinia* is mostly driven by mutational patterns, and that it recently underwent a shift from a GC to an AT mutational bias. Were this true, we would have had to have measured the mutational pattern soon after the change from a GC to an AT bias, but before this shift had the chance to alter the genome-wide nucleotide composition. This seems unlikely based simply on parsimony. Rather, given the results of Hershberg, Hildebrand and co-workers, we favor the explanation that *Hodgkinia* has, and has always had, an inherent AT mutational bias.

Our results seem to present a paradox in the way that the population genetics of endosymbionts are normally considered. The prevailing view that endosymbionts have less efficacious selection resulting from reduced effective population sizes (Moran 1996; Andersson and Kurland 1998; Woolfit and Bromham 2003) fits well with some features of the *Hodgkinia* genome, in particular with its tiny size and overall rapid rate of sequence evolution. The disparity between *Hodgkinia's* AT biased mutational pattern and GC biased genome does not fit easily into this framework, as these results seem to require either an atypically large effective population size for *Hodgkinia* or an unusually large selection coefficient for each individual AT-GC polymorphism in the population, or some combination of the two. It is possible that the population size of the host cicada is large and thus inflates the effective population size of *Hodgkinia*; theoretical work has shown that host population size can have large effects on mutation accumulation in *Buchnera aphidicola* in the context of its symbiosis with aphids (Rispe and Moran 2000). Why G or C nucleotides would be globally favored over A or T nucleotides is unclear, and is an interesting area of future study.

Hodgkinia is found as a symbiont throughout the cicada lineage (data not shown), and it will be of interest to examine the GC contents and mutational biases of *Hodgkinia* across the diversity of cicadas. If GC-poor lineages of *Hodgkinia* are found, then it may be possible to narrow the list of possible selective forces responsible for the elevated GC levels in *Hodgkinia* from *D. semicincta*, by considering factors such as the environmental conditions and population structures of the insect hosts. The mutational results reported here would predict that a lineage of *Hodgkinia* in which the selective restraints on elevated GC were severely reduced or eliminated would have a genomic GC content as low as, or possibly lower than, 37%.

2.7 Methods and supplementary materials

Identification of SNPs in the Hodgkinia genome from Diceroprocta semicincta. A total of 179 SNPs were identified in the Hodgkinia genome generated from 10 pooled *Diceroprocta semicincta* individuals by combining the output from the 454 GSmapper software (using default parameters and only considering the "high-quality" SNPs written to the HCDiffs.txt file) and SWAP454 (relevant parameters for MapNQSCoverage: MIN_QUAL=15 NQ=10; relevant parameters for CallPolymorphismsFromMap: MIN_RATIO=0 MIN_READS=2

NEED_RC=True). A total of 139 and 166 SNPs were identified using GSmapper and SWAP454, respectively; 126 were called by both programs. All SNPs were verified by manual inspection.

Polarization of SNPs using a draft *Hodgkinia* genome assembly from an undescribed but closely related cicada. During an unpublished, initial attempt at sequencing the *Hodgkinia* genome, DNA from the target species (*D. semicincta*) was unintentionally sequenced in combination with an unknown, but clearly distinct cryptic *Diceroprocta* species (hereby referred to as the "cryptic" species). The published *Hodgkinia* genome was generated in a completely separate subsequent experiment, and the SNPs were called from these data. In the initial mixed species assembly, several *Hodgkinia* contigs of equal length were present as duplicates, with pair-wise sequence identities of about 95% between homologous contigs. Some regions of the two *Hodgkinia* genomes were assembled together because of increased sequence identity (e.g., as in fig. S1C), but the majority of the genome (approximately 70%) fell out into two easily separable contig sets (e.g., as in fig. S1B). We verified the presence of two cicada species in this mixed dataset by identifying distinct insect mitochondrial COI sequences in the genome assembly data; this was verified by PCR and Sanger sequencing of a pinned individual of the cryptic species (the pairwise differences between COI sequences from *D. semicincta* and the cryptic species was 3.5%). *Hodgkinia* contigs from the cryptic species alone, as well as the mixed-species contigs, were used to polarize the direction of mutation in the pure sample of *D. semicincta* (see fig. S1 for a schematic overview of this process).

Issues related to determining SNP frequency bins in figure 2. In an ideal experiment, identical amounts of *Hodgkinia* DNA would be pooled from each of the 10 individuals dissected, and library creation and genome sequencing protocols would be immune to bias. In this ideal case, the assembled genome would be represented by an equal number of reads from each of the 10 individuals, resulting in SNPs frequencies very close to 0.1, 0.2, 0.3, and so on up to 0.9. The data in the present analysis does not conform to this ideal because of several sources of variability. First and foremost, different amounts of bacteriome tissue (and therefore different amounts of *Hodgkinia* DNA) were isolated from each insect and combined into a single sample. In some cases, nearly all of the bacteriome tissue was recovered from an animal, and in other cases only parts of the complete bacteriome could be recovered. Secondly, we explicitly required any called SNP to be supported by at least two polymorphic reads, and this effort to eliminate false positives should further exacerbate the unevenness of the data. In particular, this computational filtering has the effect of somewhat reducing the number of low frequency SNPs, even though the average sequencing coverage for a SNP in our analyses was 61X (that is, about 6X per individual). This has particular relevance for the identification of SNPs that fall into the [0, 0.1) bin in figure 2, as we do not expect a lower number of SNPs in the [0, 0.1) bin compared to the [0.1, 0.2) bin. It is likely that the SNPs in the [0, 0.1) bin are present in one insect, but the ratio has been artificially lowered from 0.1 by some of the experimental and computational idiosyncrasies described above. In summary, these confounding factors should diffuse the expected peaks at 0.1 frequency intervals into a much more complex pattern, and the precise boundaries for the bins shown in figure 2 should be interpreted with caution. The primary role of these bins was to allow broad trends to be inferred from calculations of dN/dS and GCeq on frequency binned data.

Calculation of dN/dS and GC_{eq}. To determine the total number of synonymous and non-synonymous sites in the published *Hodgkinia* genome (CP001226.1), the coding sequence was downloaded from NCBI and compiled into one sequence. DnaSP version 5.10.01 was used to create a codon usage table (Librado and Rozas 2009). As most of the SNPs were either GC→AT or AT→GC, all 2-box codons were considered synonymous. The calculation of genomic dN/dS was done as described (Hershberg and Petrov 2010), using the equations:

$$dN = \frac{3}{4}\ln\left(1 - \frac{4n}{3N}\right) \quad \text{and} \quad dS = \frac{3}{4}\ln\left(1 - \frac{4s}{3S}\right)$$

where n is the number of non-synonymous SNPs, s the number synonymous SNPs, N the number of non-synonymous sites, and S the number of synonymous sites.

The equilibrium GC content was calculated using the equation $GC_{eq} = r_{AT\rightarrow GC} / (r_{AT\rightarrow GC} + r_{GC\rightarrow AT})$, where $r_{AT\rightarrow GC} = AT\rightarrow GC/AT_{sites}$ and $r_{GC\rightarrow AT} = GC\rightarrow AT/GC_{sites}$. The COI sequence for the cryptic cicada species was amplified from DNA isolated from a small portion of tissue removed from the thorax of a pinned specimen. DNA was isolated using QIAGEN DNeasy Blood and Tissue kit. PCR was performed using the following conditions: initial denaturation at 94°C for 1 min, followed by 30 cycles of 94°C for 30 sec, 55°C for 1min 30sec, 68°C for 30sec, finished by 5min at 68°C. The primer sequences were: COI-F (5'-TCAGCCATCCCAATATGAAAAAGTGG-3') and COI-R (5'-CGACGAGGTATTCCTCTCAGTCCA-3').

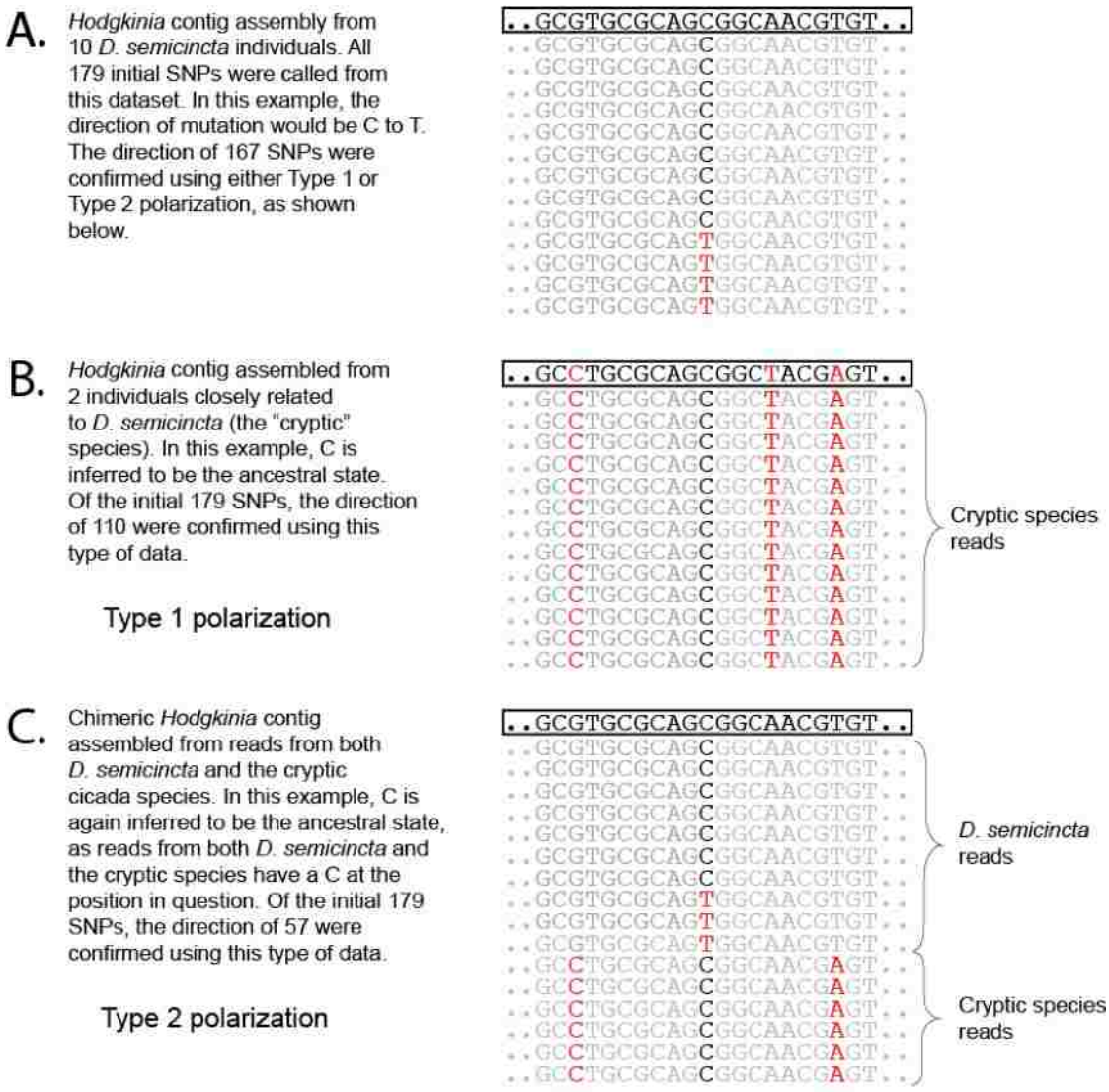


Fig. S1. Schematic of the process used to polarize SNPs in the *Hodgkinia* genome. Bold boxes represent the consensus sequence called in each *Hodgkinia* genome assembly. Nucleotides that differ from the published *Hodgkinia* consensus are shown in red. **A.** An example of a called C to T SNP in the published *Hodgkinia* genome. **B.** A total of 110 SNPs were polarized by what we are calling "Type 1 polarization," which involved mapping separately assembled *Hodgkinia* contigs from the cryptic cicada species onto the published *Hodgkinia* genome and verifying the state of the position in question in part A. **C.** A total of 57 SNPs were polarized by what we are calling "Type 2 polarization," which involved mapping *Hodgkinia* contigs that resulted from merged assemblies of *D. semicincta* and the cryptic species reads. In this case, C can be inferred to be the ancestral state because all reads from the cryptic species have C at the position in question.

Table S1. List of all polarized SNPs used for the calculations. Mutation type is abbreviated;

synonymous (S), non-synonymous (NS), and intergenic (IG). Direction is shown as one of six possibilities. Frequency indicates the decimal proportion of reads with the SNP at each location. In a few instances the polarization informed the direction and the frequency was corrected.

Pos.	Type	Freq.	Direction
507	S	0.92	A-G, T-C
1769	S	0.88	A-G, T-C
2492	S	0.19	G-A, C-T
2664	NS	0.23	G-A, C-T
2755	NS	0.05	G-A, C-T
4775	S	0.19	G-A, C-T
6777	S	0.15	G-A, C-T
7178	NS	0.05	G-A, C-T
7644	NS	0.04	A-G, T-C
7876	S	0.36	G-T, C-A
8888	NS	0.04	A-G, T-C
8935	S	0.55	G-A, C-T
9646	S	0.22	G-A, C-T
9680	NS	0.08	G-A, C-T
9753	S	0.49	G-A, C-T
9938	NS	0.37	G-A, C-T
10878	IG	0.16	G-A, C-T
10884	IG	0.11	A-G, T-C
11525	NS	0.06	G-A, C-T
11688	S	0.19	G-A, C-T
12260	NS	0.19	G-A, C-T
13006	IG	0.18	G-A, C-T
13841	IG	0.58	G-A, C-T
14736	S	0.31	A-G, T-C
15868	S	0.24	G-A, C-T
16378	S	0.02	G-A, C-T
17222	NS	0.18	G-A, C-T
17279	S	0.38	A-G, T-C
17618	S	0.10	G-T, C-A
19591	S	0.46	A-G, T-C
20081	NS	0.34	A-G, T-C
20187	NS	0.19	G-C, C-G
20414	S	0.11	G-A, C-T
20997	NS	0.17	G-A, C-T
21329	NS	0.04	G-A, C-T
21483	S	0.43	G-A, C-T
22364	NS	0.10	G-C, C-G
23862	S	0.05	G-A, C-T
25040	S	0.13	A-G, T-C
26451	IG	0.21	G-A, C-T
28348	S	0.10	G-A, C-T
28425	NS	0.04	A-G, T-C
29265	NS	0.57	A-G, T-C
31563	S	0.15	A-G, T-C
32188	S	0.37	A-G, T-C
32873	NS	0.11	G-A, C-T
33781	S	0.40	A-G, T-C
35389	IG	0.34	A-G, T-C
36152	NS	0.14	G-A, C-T
36430	NS	0.19	G-A, C-T
36750	S	0.07	A-G, T-C
39242	IG	0.04	A-G, T-C
39455	NS	0.46	G-A, C-T
39655	S	0.22	A-G, T-C
41334	NS	0.18	A-G, T-C

41534	NS	0.04	G-C, C-G
41800	NS	0.14	G-A, C-T
42021	IG	0.46	G-A, C-T
43033	IG	0.27	G-A, C-T
43165	IG	0.16	G-T, C-A
43230	S	0.91	A-G, T-C
43387	S	0.10	G-A, C-T
43548	NS	0.46	A-G, T-C
44645	S	0.28	A-G, T-C
44893	NS	0.38	G-A, C-T
45125	S	0.15	G-A, C-T
45890	S	0.14	A-G, T-C
45935	NS	0.12	A-C, T-G
48623	S	0.24	A-G, T-C
49013	S	0.47	A-G, T-C
49751	S	0.16	A-G, T-C
50038	NS	0.05	G-A, C-T
50369	S	0.11	G-A, C-T
50375	S	0.49	G-A, C-T
51809	NS	0.41	G-A, C-T
53076	S	0.35	G-A, C-T
53121	S	0.21	G-A, C-T
53379	S	0.14	G-A, C-T
53683	NS	0.47	G-T, C-A
58309	NS	0.49	G-A, C-T
58824	S	0.15	A-G, T-C
59182	NS	0.03	G-A, C-T
59338	NS	0.07	A-G, T-C
59484	S	0.27	G-A, C-T
60905	NS	0.27	G-A, C-T
62059	NS	0.54	G-A, C-T
62902	S	0.53	G-A, C-T
63016	S	0.46	G-A, C-T
63752	S	0.56	G-T, C-A
64580	S	0.33	G-A, C-T
65642	S	0.23	A-G, T-C
66374	S	0.46	G-A, C-T
67728	S	0.10	G-A, C-T
67737	S	0.09	G-A, C-T
68120	NS	0.50	G-A, C-T
68941	NS	0.20	G-T, C-A
69319	S	0.17	G-A, C-T
70215	NS	0.10	G-T, C-A
70307	S	0.13	G-T, C-A
70750	IG	0.05	G-A, C-T
72479	NS	0.16	G-T, C-A
73058	S	0.23	G-A, C-T
74914	NS	0.19	A-G, T-C
74921	NS	0.17	G-A, C-T
74926	NS	0.14	G-A, C-T
75447	S	0.21	G-A, C-T
77532	S	0.09	G-T, C-A
78799	S	0.26	G-A, C-T
78890	IG	0.09	G-C, C-G
81048	IG	0.10	A-G, T-C
82138	NS	0.07	G-A, C-T
83224	S	0.24	G-A, C-T
83527	S	0.04	G-A, C-T
84008	S	0.68	G-T, C-A
84155	S	0.13	G-A, C-T
85768	S	0.18	A-G, T-C
88626	IG	0.12	A-G, T-C
90506	S	0.35	G-A, C-T

90664	NS	0.06	G-A, C-T
91094	S	0.13	G-A, C-T
92515	S	0.08	A-G, T-C
98226	IG	0.18	A-C, T-G
98563	S	0.23	G-A, C-T
100661	S	0.21	G-A, C-T
101069	S	0.19	A-C, T-G
101631	S	0.31	A-C, T-G
102967	S	0.23	A-C, T-G
103698	NS	0.22	A-G, T-C
103915	NS	0.30	G-A, C-T
108013	S	0.47	G-A, C-T
108794	S	0.03	G-A, C-T
109142	S	0.61	A-C, T-G
110246	S	0.91	G-A, C-T
110385	NS	0.29	G-C, C-G
112516	S	0.15	G-A, C-T
114230	NS	0.28	G-A, C-T
115559	NS	0.21	G-A, C-T
115796	IG	0.31	G-A, C-T
115956	IG	0.14	G-A, C-T
119165	NS	0.02	A-G, T-C
120692	NS	0.34	A-G, T-C
122731	NS	0.21	G-A, C-T
122872	NS	0.18	A-G, T-C
123653	NS	0.20	A-G, T-C
124649	NS	0.12	A-G, T-C
125590	S	0.70	A-G, T-C
126897	NS	0.36	G-T, C-A
126932	NS	0.86	A-G, T-C
127006	S	0.13	G-A, C-T
127933	NS	0.12	G-A, C-T
129930	NS	0.06	G-A, C-T
130992	NS	0.11	A-G, T-C
131016	S	0.22	G-T, C-A
131487	S	0.31	G-A, C-T
133553	S	0.85	A-G, T-C
135687	S	0.13	A-G, T-C
137218	S	0.30	A-G, T-C
138364	NS	0.34	A-G, T-C
139305	S	0.16	A-G, T-C
139665	S	0.06	G-A, C-T
139871	NS	0.18	G-A, C-T
140058	NS	0.08	G-T, C-A
140751	S	0.16	G-A, C-T
142019	S	0.11	G-A, C-T
142943	IG	0.14	G-A, C-T
143001	IG	0.78	G-A, C-T
143059	S	0.87	A-G, T-C

2.8 References

Andersson SGE, Kurland CG. 1998. Reductive evolution of resident genomes. Trends Microbiol. 6:263–268.

Bentley SD, Parkhill J. 2004. COMPARATIVE GENOMIC STRUCTURE OF PROKARYOTES. Annu. Rev. Genet. 38:771–791.

Campbell MA, Van Leuven JT, Meister RC, Carey KM, Simon C, McCutcheon JP. 2015. Genome expansion via lineage splitting and genome reduction in the cicada endosymbiont Hodgkinia. Proc. Natl. Acad. Sci. 112:10192–10199.

Dale C, Wang B, Moran N, Ochman H. 2003. Loss of DNA Recombinational Repair Enzymes in the Initial Stages of Genome Degeneration. Mol. Biol. Evol. 20:1188–1194.

Hebert PDN, Ratnasingham S, deWaard JR. 2003. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc. R. Soc. B Biol. Sci. 270:S96–S99.

Hershberg R, Petrov DA. 2010. Evidence That Mutation Is Universally Biased towards AT in Bacteria. PLoS Genet 6:e1001115.

Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of Selection upon Genomic GC-Content in Bacteria. PLoS Genet 6:e1001107.

Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, Tanaka M, Satoh N, Bachtrog D, Wilson ACC, et al. 2013. Horizontal Gene Transfer from Diverse Bacteria to an Insect Genome Enables a Tripartite Nested Mealybug Symbiosis. Cell 153:1567–1578.

Kryazhimskiy S, Plotkin JB. 2008. The Population Genetics of dN/dS. PLoS Genet 4:e1000304.

Van Leuven JT, Meister RC, Simon C, McCutcheon JP. 2014. Sympatric Speciation in a Bacterial Endosymbiont Results in Two Genomes with the Functionality of One. Cell 158:1270–1280.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452.

Lind PA, Andersson DI. 2008. Whole-genome mutational biases in bacteria. Proc. Natl. Acad. Sci. 105:17878–17883.

McCutcheon JP, von Dohlen CD. 2011. An Interdependent Metabolic Patchwork in the Nested Symbiosis of Mealybugs. Curr. Biol. 21:1366–1372.

McCutcheon JP, McDonald BR, Moran NA. 2009. Origin of an Alternative Genetic Code in the Extremely Small and GC–Rich Genome of a Bacterial Symbiont. PLoS Genet 5:e1000565.

McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. 10:13–26.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. U. S. A. 93:2873–2878.

Moran NA. 2002. Microbial minimalism: genome reduction in bacterial pathogens. Cell 108:583–586.

Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. Annu. Rev. Genet. 42:165–190.

Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M. 2006. The 160-Kilobase Genome of the Bacterial Endosymbiont Carsonella. Science 314:267.

Rispe C, Moran NA. 2000. Accumulation of Deleterious Mutations in Endosymbionts: Muller's Ratchet with Two Levels of Selection. Am. Nat. 156:425–441.

Rocha EPC, Feil EJ. 2010. Mutational Patterns Cannot Explain Genome Composition: Are There Any Neutral Sites in the Genomes of Bacteria? PLoS Genet 6:e1001104.

Woolfit M, Bromham L. 2003. Increased Rates of Sequence Evolution in Endosymbiotic Bacteria and Fungi with Small Effective Population Sizes. Mol. Biol. Evol. 20:1545–1555.

Chapter 3: Comparative genomics of *Hodgkinia*

Published as, Van Leuven JT, Meister RC, Simon C, McCutcheon JP. 2014. Sympatric Speciation in a Bacterial Endosymbiont Results in Two Genomes with the Functionality of One. Cell 158:1270–1280.

and, Campbell MA*, Van Leuven JT*, Meister RC, Carey KM, Simon C, McCutcheon JP. 2015. Genome expansion via lineage splitting and genome reduction in the cicada endosymbiont *Hodgkinia*. Proc Natl Acad Sci U S A 112:10192–10199. *equally contributed to manuscript

Summary
Some insects have developed intracellular associations with communities of bacteria, where interdependencies are manifest in patterns of complementary gene loss and retention among members of the symbiosis. Gene loss events are most evident in the bacterial partners, where genome reduction is followed by genome structure stability. Here, using comparative genomics and microscopy, we show that a three-member symbiotic community has become a four-way assemblage through a novel bacterial lineage-splitting event. In some but not all cicada species of the genus *Tettigades*, the endosymbiont *Candidatus* Hodgkinia cicadicola has split into two new cytologically distinct but metabolically interdependent species. Although these new bacterial genomes are partitioned into discrete cell types, the inter-genome patterns of gene loss and retention are almost perfectly complementary. These results defy easy classification: they show genomic patterns consistent with those observed after both speciation and whole genome duplication. We suggest that our results highlight the potential power of non-adaptive forces in shaping organismal complexity. We test this non-adaptive hypothesis by sequencing the *Hodgkinia* genome from a very long-lived cicada, *Magicicada tredecim*, and compare the patterns of evolution observed in these endosymbionts to eukaryotic organelles—the most highly derived bacteria.

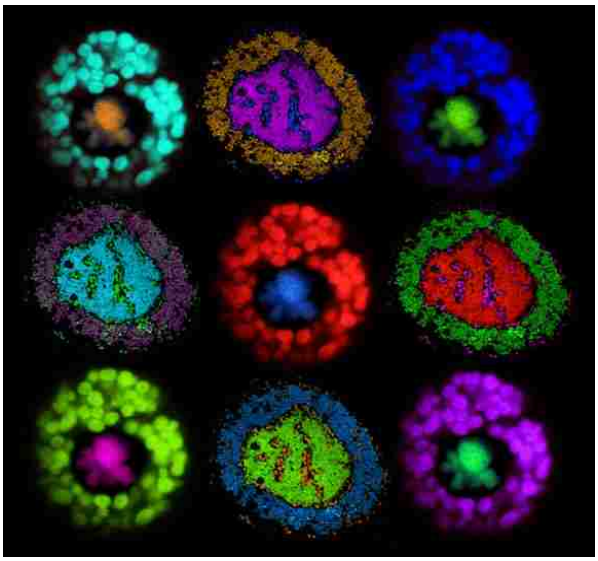


Illustration by Patrick Keeling and James Van Leuven
appears on Aug. 18th 2015 cover, PNAS

3.1 Introduction

An overview of endosymbiont genome size and structure.

The first published genome from a nutritional bacterial endosymbiont of an insect was *Buchnera aphidicola* from the pea aphid *Acyrthosiphon pisum* (AP) (Shigenobu et al. 2000). This landmark paper provided many key insights that would be repeatedly reinforced in different bacterial symbioses during the subsequent 15 years, including extreme gene loss and genome reduction, precise metabolic complementarity and interdependence with the host insect, highly biased nucleotide and amino acid compositions, and limited gene sets involved in DNA repair, gene regulation, and cell envelope biosynthesis (McCutcheon and Moran 2012). The second complete *Buchnera* genome, from the aphid *Schizaphis graminum* (SG), provided the next archetype for endosymbiont genomes: the *Buchnera* AP and SG genomes showed no rearrangements or gene acquisitions despite large amounts of sequence evolution and 50+ million years of divergence (Tamas et al. 2002). Unusual genomic structural stability has been repeatedly found in many other insect endosymbiont genera, including *Blochmannia* (Gil et al. 2003; Degnan et al. 2005), an ant endosymbiont; *Sulcia (McCutcheon et al. 2009a; McCutcheon and Moran 2010; Bennett and Moran 2013)*, which forms a widespread and ancient association with sap-feeding insects such as sharpshooters, spittlebugs, and cicadas (Moran et al. 2005); and *Carsonella*, an endosymbiont of psyllids (Nakabachi et al. 2006; Sloan and Moran 2012). A pattern thus emerged whereby the process of genome reduction in endosymbionts resulted in small and stable genomes. But several other examples, some recently published, have placed small cracks into the façade of genomic stability in endosymbionts. Sequencing of *Buchnera* from a third more diverged aphid genus showed two inversion rearrangements and two small translocations relative to the first two genomes (van Ham et al. 2003). Genomes from various endosymbiont genera found in cockroaches (Sabree et al. 2010), tsetse fly (Rio et al. 2012) mealybugs (McCutcheon and von Dohlen 2011), leafhoppers (Bennett and Moran 2013), and especially whiteflies (Sloan and Moran 2013) also showed some structural rearrangements in otherwise completely co-linear genomes (reviewed in (Sloan and Moran 2013)). While these results do not much change the general picture of genomic stability in highly reduced endosymbionts, they do suggest an alternative to unalterable co-linearity and stability given the right circumstances.

How do insect endosymbiont genomes become so degenerate?

Communities of independent organisms that develop stable, long-term associations can reciprocally lose traits that become redundant in the symbiotic context (Ellers et al. 2012). One of the clearest examples of this phenomenon occurs in symbioses involving insects and mutualistic endocellular bacteria. In these systems, symbionts provide nutrients that the host cannot make on its own and that are not found at high levels in the insect diet (Douglas 1998; Moran et al. 2003; Baumann 2005). The metabolic contributions of these bacteria are often clearly defined by their genomes, where patterns of gene loss and retention show precise inter-organism, and sometimes inter-pathway, genomic complementation (Shigenobu et al. 2000; Zientz et al. 2004; Wu et al. 2006; McCutcheon and Moran 2007; McCutcheon and Moran 2010;

Lamelas et al. 2011; McCutcheon and von Dohlen 2011; Sloan and Moran 2012). Over time, endosymbionts become deeply metabolically integrated with their hosts (Wilson et al. 2010; Macdonald et al. 2012; Husnik et al. 2013; Sloan et al. 2014), and evolve genomes encoding few genes outside of the core processes of replication, transcription, translation, and nutrient provisioning (McCutcheon and Moran 2012). In extreme cases, nutritional endosymbionts of sap-feeding insects rival organelles in their levels of genome reduction (McCutcheon and Moran 2012).

Similar to organelles, the evolutionary pressures faced by intracellular symbionts are driven primarily by their exclusive existence inside host cells, the need to continue making nutrients in the face of unrelenting genome reduction, and strong genetic drift (Moran 1996; Andersson and Kurland 1998). Thus, long-term endosymbiosis not only leads to massive genome reduction, but also to an overall degradation in symbiont function (Baumann et al. 1996; Moran 1996; Fares et al. 2002). Perhaps to compensate for this decrease in symbiont quality, a long-term single founding symbiont is often supplemented with additional unrelated bacteria (Moran et al. 2005; McCutcheon and Moran 2010; Lamelas et al. 2011; McCutcheon and von Dohlen 2011), or replaced altogether with a new symbiont (Koga et al. 2013). For example, some ancient lineages of sap-feeding insects possessed a single bacterial endosymbiont, *Sulcia muelleri* (Moran et al. 2005), which was repeatedly supplemented with additional bacterial partners several times as this ancestral symbiosis diversified (McCutcheon and Moran 2007; McCutcheon and Moran 2010). These transitions from the single- to double-symbiont state are followed by rapid genome degradation in both bacteria, the end result being clear inter-organism genomic complementarity (McCutcheon and Moran 2010; McCutcheon and von Dohlen 2011). Symbioses can therefore become more complex by adding new members: an insect with a single bacterial symbiont acquires a second, and a two-member assemblage becomes tripartite. If the secondary bacterium is established as a stable member of the symbiosis, the system evolves to a state dependent on all three organisms for survival of the whole (Wu et al. 2006; McCutcheon and Moran 2007; McCutcheon and Moran 2010; Lamelas et al. 2011; McCutcheon and von Dohlen 2011).

3.2 *Hodgkinia* genome structures and sizes

Genome sequencing recovers two symbiont genomes where one was expected.

Previous work in the cicada *Diceroprocta semicincta* (DICSEM) showed that some cicadas have two bacterial endosymbionts, *Sulcia* and *Hodgkinia* (McCutcheon et al. 2009a; McCutcheon et al. 2009b). While analyzing genomic data from the cicada *Tettigades undata*, we recovered the expected single circular *Sulcia* chromosome, co-linear with all other sequenced *Sulcia* genomes. Unexpectedly, we found that the *Hodgkinia* genome assembled into two distinct circular chromosomes. These chromosomes, which we call *Hodgkinia cicadicola* from *Tettigades undata* chromosome 1 (TETUND1) and TETUND2, showed different depths of sequencing coverage (405X and 640X, respectively) and were verified and closed into two separate circular molecules by PCR and Sanger sequencing (Table 1). Because many coding regions from these two chromosomes were alignable, we used average synonymous divergence (dS) and rRNA dissimilarity values calculated from across a diversity of bacteria, including symbionts (Kuo and Ochman 2009), to estimate a rough age of divergence. The average dS value

between protein-coding homologs in the two chromosomes is 0.168, and the small subunit (SSU) rRNA sequence dissimilarity is 0.6%, corresponding to roughly 5-25 million years of divergence.

Table 1. Cicada Species and Properties of Their Associated *Hodgkinia* Genomes

Cicada Species	Abbreviation	*Hodgkinia* Genome Size (bp)	All Predicted Coding Genes	Nonhypothetical Protein-Coding Genes	Pseudogenes
D. semicincta	DICSEM	143,795	170	134	1
T. ulnaria	TETULN	150,297	175	137	1
T. undata	TETUND1	133,698	121	92	38
	TETUND2	140,570	140	116	18
T. auropilosa	TETAUR1	not known	not known	not known	not known
	TETAUR2	not known	not known	not known	not known

The DICSEM and TETULN genomes differ in only four genes of predicted function (*glyA* is encoded on DICSEM, but not TETULN, and *rsmI*, *rsmA*, and *rpmE* are encoded on TETULN, but not DICSEM). The *Hodgkinia* TETAUR genome is in draft form, and so the genome statistics are not presently known.

Further screening identifies other duplicated Hodgkinia lineages, and a close non-duplicated. relative

Because the duplicated genomic structure of *Hodgknia* TETUND was highly unusual, we sought to confirm the generality of this result by screening *Hodgkinia* from related cicada species. We first verified the duplicated nature of *Hodgkinia* in another cicada species, *Tettidages auropilosa* (TETAUR), by draft genome sequencing (Fig. 1). Next, we identified a closely related cicada species, *Tettigades ulnaria* (TETULN), where the *Hodgkinia* genome was a single chromosome, completely co-linear and very similar in gene content to the first sequenced *Hodgkinia* genome from DICSEM (Fig. 1 and Table 1).

In addition to symbiont genomes, our sequencing effort also provided mitochondrial genomes. Phylogenetic reconstruction using complete cicada COI sequences shows that TETULN is sister to the group containing TETUND and TETAUR, verifying that the ancestral state of the *Hodgkinia* genome was a single highly reduced chromosome (Fig. 1 and Table 1). To test the 5-25 million year divergence times calculated from the duplicated TETULN sequences by another method, we estimated a model-corrected mitochondrial COI distance between TETULN and TETUND. The value was 0.104, which roughly corresponds to 3.0 to 4.5 My of divergence in insects (Brower 1994; Papadopoulou et al. 2010). Because this is consistent with but on the low end of estimate from TETUND comparisons, we estimate that the *Hodgkinia* lineage duplicated in some *Tettigades* genera approximately 5 My ago.

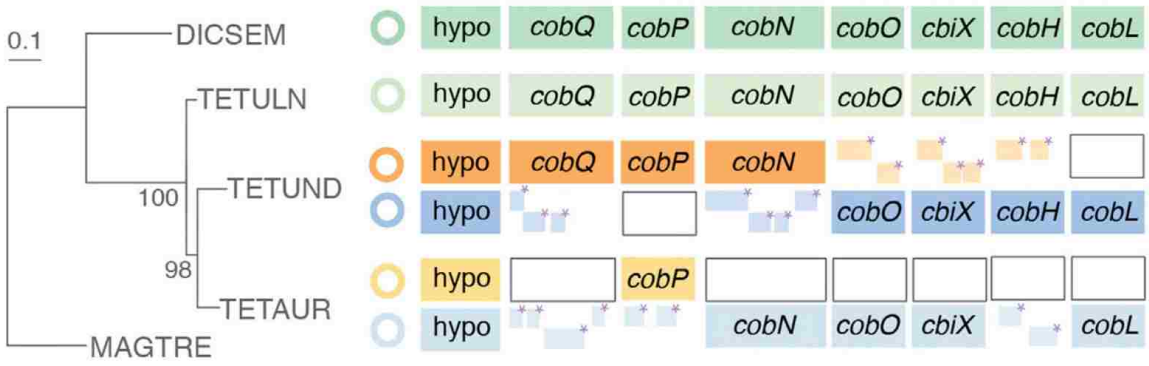


Figure 1. Origin of duplicated *Hodgkinia* genomes in the cicada genus *Tettigades*. At left, an unrooted maximum likelihood phylogeny based on cicada COI is shown with bootstrap support values (the scale bar is 0.1 expected substitutions per site; MAGTRE is *Magicicada tredecim*). The number of *Hodgkinia* genomes are indicated by colored circles to the right of the tree. The ancestral nature of the single *Hodgkinia* genome is evident from the sister group relationship between TETULN and the clade containing TETUND and TETAUR. The right side of the figure shows representative sections of genome, where intact genes are shown by large colored boxes, gene loss is indicated by empty boxes, and pseudogenes are shown as small open reading frames broken by frameshifts (small filled boxes) and stop codons (asterisks).

3.3 Gene contents of *Hodgkinia* genomes

The two Hodgkinia chromosomes have complementary patterns of gene loss and retention.

TETUND1 and TETUND2 are both co-linear with the single TETULN and DICSEM genomes with the exception of a 32 kb inversion present on TETUND1. This inversion inactivated the methionine synthase gene (*metH*), which is of interest as methionine is thought to be a critical nutrient supplied by *Hodgkinia* in the symbiosis (McCutcheon et al. 2009b). Because the *metH* homolog was intact and seemingly functional on TETUND2, we investigated patterns of gene loss and retention across the two *Hodgkinia* TETUND chromosomes and found a clear reciprocal pattern (Fig. 2). Of the 137 protein-coding genes on TETULN, 72 are present as apparently functional copies on both TETUND genomes, 20 were present and functional on TETUND1 but nonfunctional on TETUND2, 44 were present and functional on TETUND2 but nonfunctional on TETUND1, and 1 was nonfunctional on both TETUND chromosomes. In total, 136 of 137 TETULN protein-coding genes are retained and apparently functional in one or both TETUND genomes (Fig 2A). The complementary gene loss and retention patterns are found across gene functional categories (Table S1), including those involved in nutrient provisioning (McCutcheon et al. 2009a). Every gene present in the histidine, methionine, and vitamin B_{12} (cobalamin) pathways in TETULN and DICSEM is retained in one or both of the two TETUND chromosomes, but in no case can any single pathway be completed with predicted gene products from an individual TETUND chromosome (Fig. 2B). We note that the patterns of gene loss and

retention differ somewhat between duplicated regions of *Hodgkinia* from TETUND and TETAUR (Fig. 1).

3.4 Molecular evolution of *Hodgkinia* sister species in *T. undata*

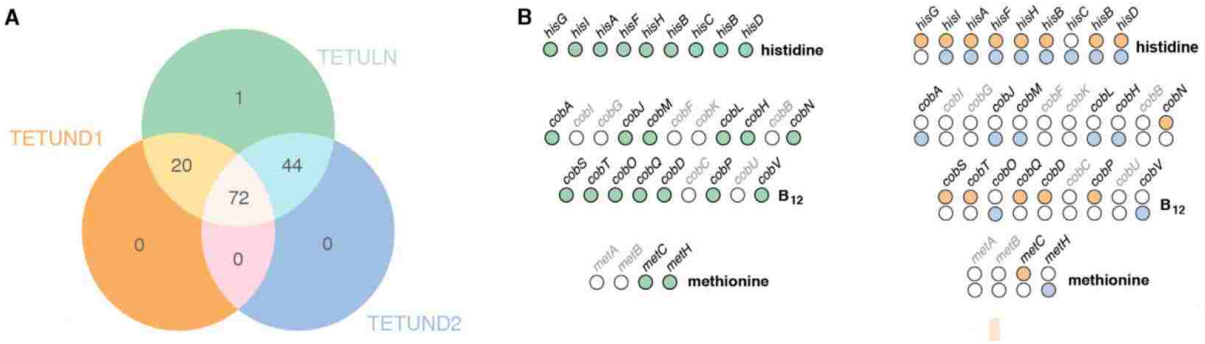


Figure 2. Reciprocal patterns of gene loss and retention in TETUND1 and TETUND2. (A) Venn diagram showing genes retained in the TETULN and TETUND genomes. (B) Nutrient provisioning genes encoded on the TETULN and DICSEM (green circles), TETUND1 (orange circles), TETUND2 (blue circles) genomes, or missing or pseudogenized (open circles).

Molecular evolutionary analyses reveal possible incipient pseudogenes and little evidence for positive selection in duplicates.

We next investigated the nature of sequence changes that have occurred between predicted homologs in TETUND1-TETUND2 comparisons. We observed some instances where both gene copies were apparently functional, and others where one copy was apparently functional but the other somehow inactivated. These inactivation events seemed of different ages: some were the result of single inactivating frameshift substitutions, some were regions that were barely recognizable as remnants of functional genes, and others were complete deletion events (Fig. 1, Fig. 3).

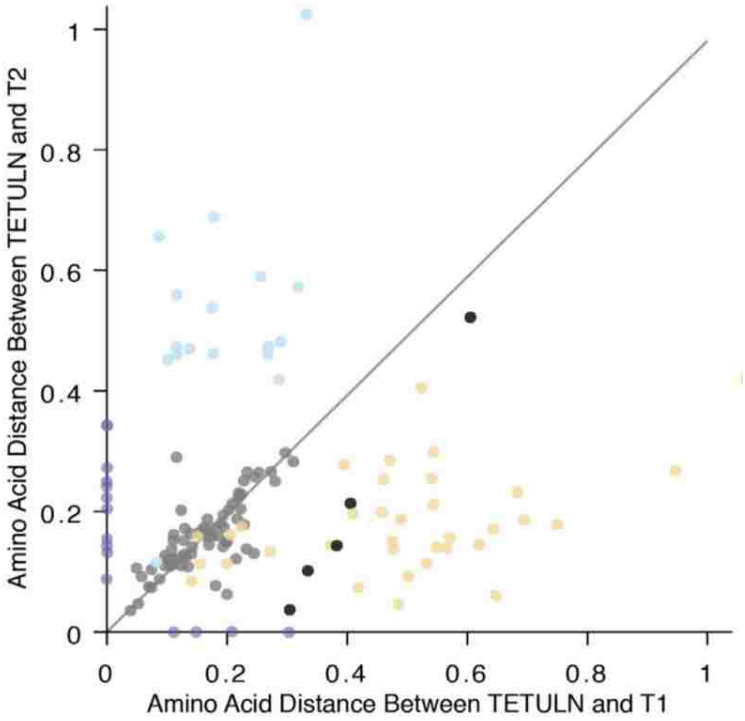
Given the large number of pseudogenized genes in different states of degradation we observed in the TETUND genomes, we hypothesized that some apparently functional genes may in fact be incipient pseudogenes that have not yet acquired an inactivating substitution. To test this idea, we compared pairs of TETUND homologs in which both were apparently functional and where one copy was a recent pseudogene to their TETULN counterpart. We estimated the ratio of nonsynonymous to synonymous substitution rates (d_N/d_S) for these comparisons, and as expected found evidence for relaxed selection in pseudogenes (functional gene—functional gene comparisons averaged 0.25 +/- 0.02, and functional gene—pseudogene comparisons averaged 0.57 +/- 0.05; $p=7.5e^{-24}$, t-test). Estimates of per-site amino acid substitution rates also show pronounced differences (Fig. 3, Fig. 4), with pairwise model-corrected distances higher for pseudogene—functional comparisons (0.52 ± 0.07) than for inferred protein sequences of apparently functional ORFs (0.19 ± 0.02; $p=9.9e^{-18}$, paired t-test). To find incipient pseudogene candidates, we looked for different rates of evolution between the 72 genes present and apparently functional on both TETUND chromosomes homologs and their TETULN homolog. Five of 72 genes show unequal rates of evolution when compared to TETULN by the likelihood

ratio test ($p<0.05$), with one copy evolving at a rate similar to bona fide pseudogenes (Fig. 4).

Figure 3. Patterns of TETUND gene retention, pseudogene formation, and rates of amino acid evolution mapped onto the TETULN genome. Annotated genes on the TETULN genome are shown as green boxes along the center of the image. Grey boxes are RNA genes. White boxes are genes that have been deleted in either TETUND1 or TETUND2, or both. If a gene is present and apparently functional on TETUND1, it is shown as a dark orange box, the height of which is proportional to the number of amino acid substitutions between the TETUND1 protein and the homolog in TETULN. If a gene is present as a pseudogene on TETUND1 it is shown as a light orange box. TETUND2 genes follow the same pattern as TETUND1 but are shown as blue bars below the TETULN genome. Rates of 0.5 and 1.0 amino acid changes per site are shown as horizontal black lines. Fig. 1 details the genomic region highlighted in light grey (the first eight genes in the genome). See also Table S1.

While it is possible that this signature is due to recent positive selection in one of the two gene copies, these results, together with the overall pattern of gene degradation we observe in the two TETUND chromosomes, suggest that these five rapidly evolving genes are incipient pseudogenes that have not yet acquired an inactivating substitution. Consistent with this interpretation, the average d_N/d_S for these five genes is 1.04.

Gene duplication is thought to sometimes enable the evolution of new function in one of the gene duplicates, with the ancestral function maintained in the other (Ohno 1970; Hughes 1994; Lynch and Conery 2000). We looked for evidence of this by testing for positive selection in pairs of TETUND homologs where both were retained and apparently functional. We found little evidence of positive selection using sensitive branch-site models, which are ideal for detecting positive selection on gene duplicates (Yang and Nielsen 2002). Only one gene, encoding the 50S ribosomal subunit protein L16 (*rplP*), showed weak evidence of positive selection on certain amino acids using branch-site models (likelihood ratio test, $p=0.045$; no genes show evidence of positive selection when d_N/d_S was averaged over the entire coding length). However, the use of branch-site models with only three taxa and rapidly evolving sequences may yield spurious results and should be interpreted with caution. Illustrating this problem, 25% of the pseudogenes we analyzed with branch-site models show evidence ($p<0.05$) of positive selection acting on at least one site.



\ **Figure 4.** Differential rates of amino acid sequence evolution identify possible incipient psuedogenes. Homologs present in TETULN, TETUND1, and TETUND2 are shown as grey dots. Homologs pseudogenized in TETUND1 or TETUND2 are shown as yellow and blue dots, respectively. Genes lost in either TETUND1 or TETUND2 are shown as purple dots along the axis. The five putative incipient pseudogenes (*rplU*, *rpsK*, *rplP*, *rpmJ*, and *hisB*) are shown as black dots. The graph is cropped at 1 expected amino acid change per site. See also Table S2.

3.5 The *Hodgkinia* genomes are cytologically distinct.

Because these complementary patterns of gene loss and retention show that the evolution of TETUND1 and TETUND2, and the *Hodgkinia* MAGTRE complex are intimately linked, we sought to test whether the *Hodgkinia* chromosomes were co-localized in the same *Hodgkinia* cells using fluorescence *in situ* hybridization (FISH) microscopy. Experiments using fluorescently labeled DNA probes targeting SSU rRNA showed that the *Sulcia* and *Hodgkinia* cells were distinct and isolated from each other in the bacteriome tissue of the cicada (Fig. 5). Data from other symbionts (including *Sulcia* (Woyke et al. 2010)) indicate that these bacteria are likely to be polyploid with hundreds of chromosomes per cell (Komaki and Ishikawa 1999). We therefore reasoned that both *Sulcia* and *Hodgkinia* cells could be visualized with FISH using probes targeting unique regions of their chromosomes. Experiments confirmed this, and revealed very little overlap in fluorescent signal (approximately 2-4%) between probes targeting the TETUND1 and TETUND2 chromosomes across all z-stack images (Pearson's coefficient=0.126, overlap coefficient=0.14, Fig. 5B). Thus, it seems that these chromosomes are not localized to the same *Hodgkinia* cells at any appreciable level, but rather are cytologically distinct genomes (Fig. 5B). Consistent with this interpretation, the fractional volume of space taken up by each TETUND probe set across a series of 60 Z-stack slices is similar to the proportion of total *Hodgkinia* reads assigned to each chromosome. Specifically, 43% of the total *Hodgkinia* fluorescence volume is estimated from TETUND1 (compared to 405/405+640 = 39% of the *Hodgkinia* sequencing coverage) and 57% of the fluorescence volume is estimated from

TETUND2 (compared to 61% of the sequencing coverage). We performed these microscopy experiments on a single insect out of necessity; we had no other individuals with which to work. But we did have another individual in our 2006 collections which appears to be a related species or subspecies that we designated *Tettigades near undata* (1.3% divergent at COI from *T. undata*). Genome-targeted microscopy on this cicada confirmed the distinct nature of the two TETUND genomes (Fig. S1).

3.6 *Hodgkina* genome complex in long-lived cicadas

Because of the novel arrangement of symbionts found in *Tettigades* which have a longer life cycle than DICSEM, we sought to investigate the genome structure of cicada symbionts from species with very long life cycles. The periodical cicada species *Magicicada tredecim* (MAGTRE) remains underground for 13 years, one of the longest insect life cycles documented (Williams and Simon 1995).

Extravagant complexity in Hodgkinia from Magicicada tredecim.

We first attempted to sequence the *Sulcia* and *Hodgkinia* genomes from MAGTRE using short-insert Illumina sequencing methods. The *Sulcia* MAGTRE assembly highlights the structural stability of some endosymbiont genomes: it cleanly assembled into one circular-mapping 268 kb molecule that was completely co-linear with all other *Sulcia* genomes. In contrast, the reads associated with the *Hodgkinia* genome assembled into an extremely complex mix of small contigs. We added sequencing reads from a 2.5 kb large-insert Illumina library with the aim joining these small contigs into larger scaffolds. We found 233 scaffolds assembled from these combined data that totaled 1.1 Mb and encoded recognizable *Hodgkinia* sequence. The assembled *Hodgkinia* scaffolds were present at different depths of coverage, consistent with what would be expected if the scaffolds did not arise from the

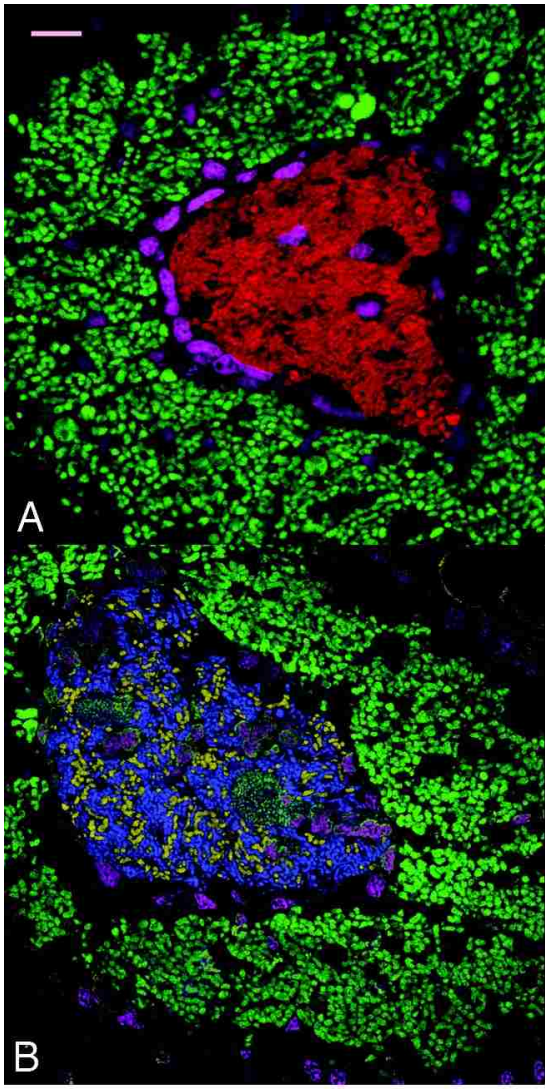


Figure 5. FISH microscopy. (A) rRNA targeted probes distinguish Sulcia (green) from *Hodgkinia* (red). (B) Genome-targeted probes distinguish *Sulcia* (green), TETUND1 (yellow), and TETUND2 (blue). Hoechst stained DNA is colored magenta, and primarily stains insect nuclei. Scale bar is 20μm. See also Fig. S1.

same physically linked DNA molecule.

We also found many cases where different versions of the same gene were present on several different scaffolds, consistent with gene duplication and/or lineage splitting. The variation in depth of sequencing coverage combined with the existence of related stretches of sequence at various levels of similarity made it difficult for us to finish the entire 1.1 Mb *Hodgkinia* assembly into distinct molecules. However, we identified 27 scaffolds totaling 739 kb of sequence where mate-pair information suggested the two scaffold ends were joined to each other (Fig. 6). Of these 27, we were able to verify that 17 scaffolds were circular-mapping molecules by PCR and Sanger sequencing, or by using approximately 421 Mb of PacBio long read data. These 17 verified circles totaled 512 kb of sequence. The remaining 10 circular scaffolds could not be closed by PacBio reads and did not provide clean PCR results because of stretches of sequence that was shared by many scaffolds. We therefore consider these putative circular-mapping molecules. The remaining 206 scaffolds contained 424 kb of sequence, ranged in size from 200 bp to 27 kb in length (166 of these were less than 2 kb in length), were frequently broken at stretches of sequence that were shared among several different scaffolds, and were left as a draft assembly.

The Hodgkina MAGTRE genome is fragmented and very degenerate

We next searched the entire 1.1 Mb *Hodgkinia* MAGTRE assembly for full-length open reading frames (ORFs). We found only 160 ORFs that were apparently functional. Ninety-six were unique (that is, 64 were duplicates of other genes), representing about 60% of the 155 ORFs we expected based on previous *Hodgkinia* genomes (McCutcheon et al. 2009b; Van Leuven and McCutcheon 2012). Seventy-six of the 160 ORFs were on the 17 closed circular molecules; 50 of these were unique (Fig. 6). Because we found no additional ORFs outside of these 160, we conclude that, like in TETUND, homologs of the ~150 genes present on the single-genome versions of *Hodgkinia* are the only genes present in the entire MAGTRE assembly. The *Hodgkinia* assembly also contained many pseudogenes, but we restricted the analysis in this paper to the 17 verified circles because of the difficulty in identifying non-functional duplicated genes in draft assemblies of rapidly evolving sequence (the average percent identity at the amino acid level was approximately 35% between MAGTRE and DICSEM orthologs, and 40% for MAGTRE-TETULN comparisons). The intergenic regions of these closed circular molecules contained mostly sequence that had no significant similarity to anything in sequence databases. The coding density of these 17 molecules was extremely low, the most gene dense circle being 45% coding DNA. It is worth noting that the assembled region of the 13 kb scaffold PUTATIVE006 (p6) contains three pseudogenes but no obviously functional genes (Fig. 6). It is possible that a functional gene exists in the unfinished gap; if not we would expect this chromosome to be under little selection to be maintained and likely to be lost in other cicada lineages.

Genes for the biosynthesis of methionine, histidine, and a vitamin B_{12}-like molecule have been found on all previous *Hodgkinia* genomes. This is thought to reflect the nutritional contribution of *Hodgkinia* to the symbiosis (McCutcheon et al. 2009a). We looked for evidence that these genes were conserved in the *Hodgkinia* MAGTRE assembly, and found that they were distributed on several scaffolds. For example, apparently functional copies of all genes in the

histidine biosynthesis pathway except *hisC* are present on at least one of the 27 circles shown in

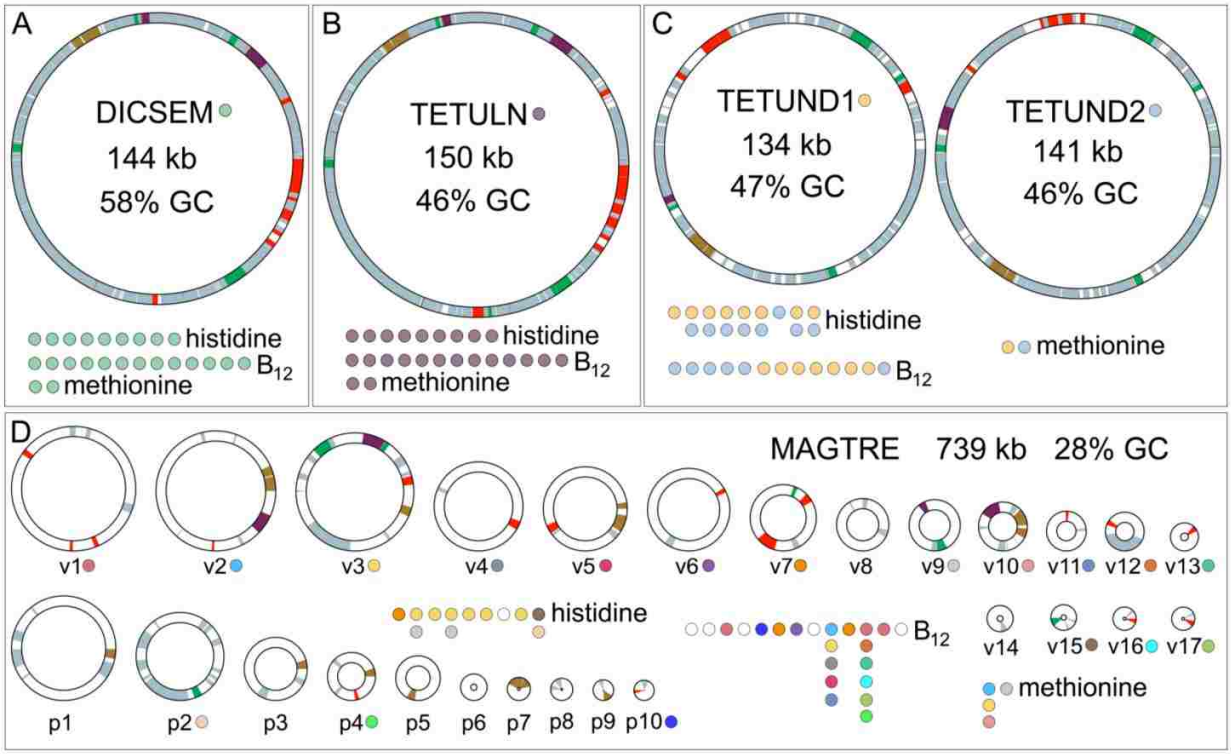


Figure 6. Schematic representations of *Hodgkinia* genomes from across cicadas. Schematic representations of all sequenced *Hodgkinia* genomes from (A) DICSEM, (B) TETULN, (C) TETUND, and (D) MAGTRE drawn to scale. On the genome diagrams, genes involved in methionine biosynthesis are shown in purple, vitamin B 12 biosynthesis in red, histidine biosynthesis in green, the 16S and 23S rRNA genes are shown in brown, and all other genes are shown in light blue. Regions of genomes encoding pseudogenes or other apparently nonfunctional DNA are shown in white. In each box, the gene homologs present on each genome from the methionine, B12 , and histidine pathways are shown as colored circles. The *Hodgkinia* genomes from DICSEM (green dots) and TETULN (purple dots) encode all of these genes on one genome, TETUND on two (blue and orange dots), and MAGTRE encode these gene distributed over several genomes (18 dot of different colors). In (D) v1-v17 are the verified circular genomes and p1-p10 are the putative circular genomes. Figure generated by Matt Campbell.

Fig. 6 (*hisC* is present as a pseudogene on a small scaffold outside of the 27 circles). It is presently unclear if functional copies of the genes missing in the histidine or B_{12} pathways are present but poorly assembled, or if like in mealybugs and psyllids, the insect host has taken over these functions.

Hodgkina MAGTRE circular molecules reside in different cells

We also looked for evidence that the 17 closed MAGTRE circular genomes arose through

the lineage-splitting and reductive process that we hypothesized for the duplicated TETUND genomes (Fig 6). While we could not exhaustively check all combinations of the 233 *Hodgkinia* MAGTRE scaffolds using genome-targeted fluorescence microscopy, we did find evidence that 4 of the 17 finished genomes were partitioned into separate cells (Fig. 7). None of the four tested genomes showed overlapping signal, suggesting that at least these four genomes (and perhaps many others) remain separated into discrete cells. We also find that the genome assembly coverage, which corresponds to the frequency with which the genome is present in the sample, correlates with the number of cells producing signal such that lower coverage scaffolds were

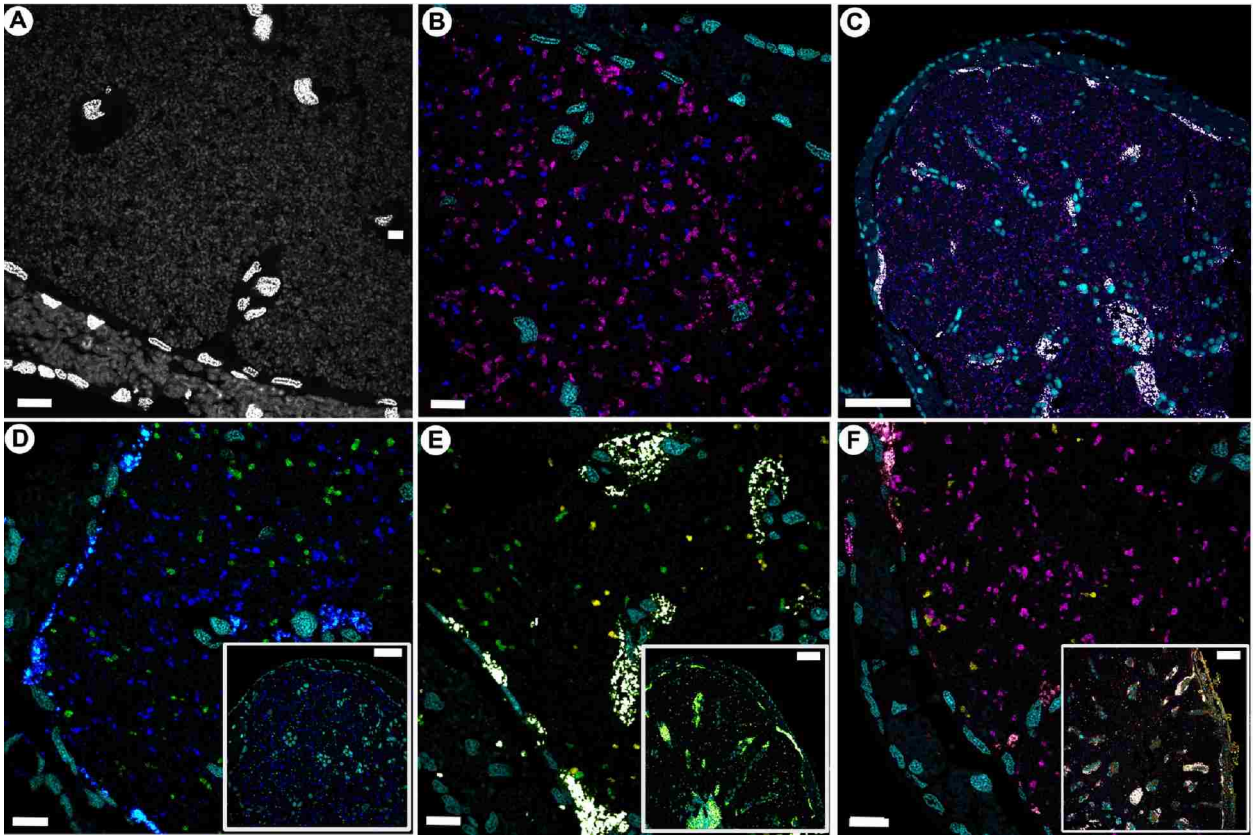


Figure 7. FISH microscopy shows some *Hodgkinia* genomes remain cytologically distinct throughout cicadas. Scale bars in the main panels of A, B, D-F are 20 µm; all others, including the insets, are 100 µm. (A) A section of bacteriome tissue from MAGTRE stained only with the general DNA Hoechst dye showing that all *Hodgkinia cells* (upper right 4/5 of image) and *Sulcia* cells (band across the lower left 1/5 of image) contain DNA. Insect nuclei are large bright punctate spots. In panels B-I, insect nuclei are teal. Panels B-F show sections of MAGTRE bacteriome tissue stained with DNA probes targeting specific genomes; in no case do the signals overlap. (B) Cells containing two high-coverage genomes MAGTREv1 (blue) and MAGTREv5 (purple) are both present a high numbers. (C) A lower resolution image of the tissue shown in (B). In D-F, lower resolution images of the tissue are shown in insets. (D) Cells containing the high-coverage genome MAGTREv1 (blue) are more abundant than cells containing the low-coverage genome MAGTREv6 (green). (E) Cells containing the two low-coverage genomes MAGTREv6 (green) and MAGTREv12 (orange) are both present at low numbers. (F) Cells containing the high-coverage genome MAGTREv5 (purple) are more abundant than cells containing the low-coverage genome MAGTREv12 (orange).

present in fewer cells that higher coverage scaffolds (Fig. 6, 7). These data are consistent with a process where new *Hodgkinia* genomic species are created when ancestral lineages split into new cytologically distinct lineages (Fig 8). It is presently unclear if all 17 of the circular genomes we have found are present in separate cells. Work from other endosymbionts shows that small plasmid-like subgenomic molecules can stably fracture from the main chromosome (Sloan and Moran 2013), so it is possible that part of what we are seeing in this complex mix of molecules is a combination of genomes that have split into new lineages combined with sub- genomic circles that have split off from larger chromosomes.

3.7 Discussion

Complex organismal interdependencies have been described for many symbioses involving intracellular bacteria. For example, the dual endosymbionts of mealybugs have adopted an unusual structure where one bacterium lives inside the other (von Dohlen et al. 2001). These bacteria have been shown to display high levels of inter-pathway dependency, where gene products from the insect and both symbionts seem to be required to produce compounds needed by the entire symbiosis (McCutcheon and von Dohlen 2011; Husnik et al. 2013). Like all other known endosymbioses involving more than one bacterium, one of the mealybug symbionts, the betaproteobacterium *Tremblaya princeps*, is older and longer established while the other, the gammaproteobacterium *Moranella endobia*, is a more recent addition (Thao et al. 2002; Gruwell et al. 2010). Thus, the increase in complexity of this symbiosis—going from a system involving one insect with one symbiont to one comprised of an insect with two symbionts—resulted from the acquisition of a second bacterium unrelated to the first. Here we described a similar increase in the complexity of a cicada symbiotic community, with the notable exception that the additional symbiont was derived from one of two existing bacterial lineages.

Cytologically distinct genomes that evolve like they aren't. The patterns of molecular evolution we describe in the *Hodgkinia* TETUND genomes look surprisingly like those that are observed after a whole genome duplication (WGD) event, where the entire genetic complement of an organism is doubled. When a WGD persists over evolutionary time, it imparts stereotypical signatures in the newly duplicated genes: (*i*) one or the other copy can be deleted (nonfunctionalization), (*ii*) one copy can retain its ancestral function, while the other acquires a novel function (neofunctionalization), (*iii*) the two new copies can reciprocally partition the ancestral functions of the non-duplicated gene leaving only the original function (subfunctionalization), or (*iv*) both copies can be retained in a functional state (Lynch and Conery 2000; Conant and Wolfe 2006; Otto 2007). In most cases, nonfunctionalization is the typical result, leading to genomes encoding nearly the same number of genes as the ancestral genome on twice the number of chromosomes (Wolfe 2001). This is precisely the pattern we see in the doubled *Hodgkinia* genomes (Fig. 3), with nonfunctionalization apparently dominating the evolution of gene duplicates (although at present it is difficult to rule out subfunctionalization).

But what we describe here is mechanistically unrelated to eukaryotic WGD. Our microscopy data show that the TETUND genomes are isolated into distinct cells (Fig. 5B, 7), whereas in WGD the new genetic material is co-localized in the same nucleus. It therefore appears that TETUND1 and TETUND2 are not chromosomes sharing the same cellular location,

but rather genomes that are faithfully partitioned in two discrete *Hodgkinia* cell types. Our molecular data suggest this cytological isolation has been stable for approximately 5 million years, long enough for two clearly different genomes to evolve. Nevertheless, despite these mechanistic differences, the evolutionary framework previously described for WGD (Lynch and Conery 2000; Wolfe 2001; Otto 2007) remains useful for understanding the genomic patterns we observe, and for predicting what we might expect to see in other lineages.

Two bacterial 'species' that evolve like one. If not WGD, then what? Viewed from the perspective of the *Hodgkinia* lineage, our data seem most simply described as a sympatric speciation event. Because *Hodgkinia* only exist in cicada cells, the new duplicated genomes we describe here emerged from the same environment; they evolved in sympatry. A single genomic species irreversibly split into two, and these new lineages differ in encoded genes, their predicted ecological function, and are cytologically distinct. After the split, the new genotypes evolved just as other endosymbiotic communities have after the acquisition of a new bacterial lineage—they lost genes through reciprocal nonfunctionalization and now perform divergent but interdependent functions. Interestingly, the highly interdependent gene loss and retention patterns we see in the TETUND genomes implies that this lineage-splitting event not only *happened* in sympatry, but also *required* sympatry. An interesting corollary to this hypothesis is that while WGD is thought to sometimes drive speciation in sexual eukaryotes (Otto 2007), here the opposite seems likely: a bacterial 'speciation' event has driven patterns of molecular evolution in two new genomes that mimic those occurring after a WGD.

The nature of the *Hodgkinia* cell envelope. How can data that look like WGD from one perspective, but speciation from another, be reconciled? We hypothesize that the answer lies in the nature of the *Hodgkinia* cell envelope. While our microscopy data show that the TETUND1 and TETUND2 genomes are cytologically distinct, the *Hodgkinia* cells seem to be intermixed in the same host insect cell (Fig. 5, 7). The widespread reciprocal gene loss and retention patterns we observe suggest a very low barrier to the sharing of gene products—but apparently not genomes—between the two *Hodgkinia* cell types. These gene products apparently include proteins, because the ε subunit of DNA polymerase III (*dnaQ*) and the α subunit of RNA polymerase (*rpoA*) are among the genes reciprocally nonfunctionalized in TETUND. This is important because the protein products of these genes act on the genome itself, and our data indicate that genomes are not shared among *Hodgkinia* cells. Because the *Hodgkinia* genomes encode no transporters or genes involved in cell envelope biosynthesis (McCutcheon et al. 2009b), we assume that transport occurs though host-encoded cell membranes and transporters. The nature of these transporters is unclear; we find no obvious bias in predicted protein size or charge of lost versus retained genes. We hypothesize that this free sharing of gene products produces nonfunctionalizaton patterns that look just like WGD, but that the membrane system of *Hodgkinia* restricts physical mixing of the genomes so that new bacterial species can form in sympatry.

A model for endosymbiotic lineage splitting. Although *Hodgkinia* does not encode much of the machinery necessary to perform recombination, TETUND1 does contain an inversion, and our comparative genomic and phylogenetic work (Fig. 1) shows that this inversion occurred after the loss of genes that encode recombinogenic enzymes. As in other symbionts that show some evidence of recombination without encoding genes for this activity (McCutcheon and von Dohlen 2011; Sloan and Moran 2013), we hypothesize that this inversion must have been catalyzed by host-encoded enzymes. However, aside from this inversion, we see no evidence of recombination within the *Hodgkinia* lineage. We therefore assume that *Hodgkinia* evolves primarily as an asexual organism.

In asexual organisms, lineage cohesion can be maintained by a combination of genetic drift and periodic selection favoring beneficial genotypes (Atwood et al. 1951; Cohan 2002). The paired *Hodgkinia* genomes we describe here appear to have resulted from an event or series of events that disrupted the cohesive force uniting the ancestral single *Hodgkinia* lineage. Given present data, it is not clear what the event(s) were that lead to this loss of cohesion, but we can propose a model based on our data and assumptions from previous work (Fig. 7). As both *Sulcia* and *Buchnera* have been shown to be highly polyploid (Komaki and Ishikawa 1999; Woyke et al. 2010), we expect that *Hodgkina* is also polyploid. Our genome-targeted FISH analyses support this assumption, as the fluorescence intensity is spread evenly throughout the *Sulcia* and *Hodgkinia* cells (Fig. 5B). Because *Hodgkinia* exists only in cicada cells, is likely bottlenecked at transovarial transmission (that is, they are passed vertically from mother to offspring though eggs) between insect generations, and is asexual, we assume that genetic drift plays a large role in its evolution.

In our model, the polyploid nature of *Hodgkinia* (Fig. 8A) masks deleterious mutations and allows them to rise to high frequency in the population through drift (Fig. 8B-8C). It is important to note that at least two separate and complementary gene-inactivating

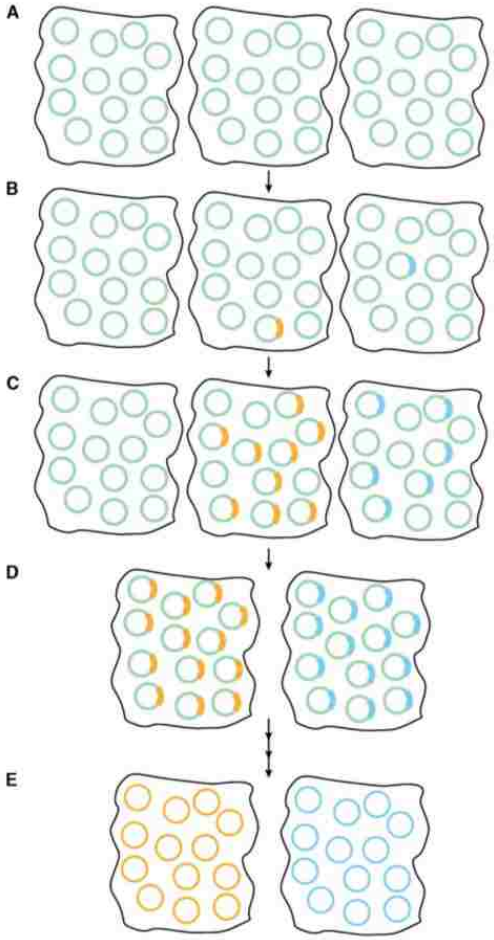


Figure 8. A model for the splitting of the ancestral *Hodgkinia* lineage. (A) We assume that *Hodgkinia* started as a population of cells with a single polyploid ancestral genotype (shown as green circles). (B) Mutations that inactive at least one gene occur in two different *Hodgkinia* cells (yellow and blue boxes) in the same insect. (C) These inactivating mutations rise to high levels, masked by the polyploid nature of *Hodgkinia*. (D) A bottleneck event purges the ancestral *Hodgkinia* genotype and fixes the reciprocal inactivating mutations in the population. (E) These new species lose genes in a reciprocal fashion to give rise to two discrete *Hodgkinia* genomes (yellow and blue circles).

mutations are required in distinct *Hodgkinia* cells, because single inactivating mutations that rise to high frequency in isolation would not 'lock in' complementary genotypes, and would eventually be purged by selection (this being the typical fate of a single deleterious mutation). A bottleneck event, perhaps at the level of the host population, eliminates the ancestral *Hodgkinia* genotype and fixes the derived genomes into a functionally obligate relationship (Fig. 8D). Finally, approximately 5 million years of evolution produces the two differentiated *Hodgkinia* species we see today (Fig. 8E).

We note that some of these steps, in particular the transition from 8B to 8D, may have been driven by selection in the symbiont or in the symbiont community, similar to what has been observed in experimental evolution studies of microbial populations (Treves et al. 1998; Friesen et al. 2004; Kinnersley et al. 2014; Plucain et al. 2014). For example, some *Hodgkinia* genotypes, freed through gene inactivation of the burden of making large proteins, may be driven to high frequency because of increased replication efficiency. If driven by selection, this event may be beneficial at the symbiont level, but would likely not be adaptive from the perspective of the entire symbiosis—here, what's good for the symbiont in the short term is not what's good for the entire symbiosis in the long term. One possible reason this phenomenon has not been observed in other insect symbiont systems may relate to the very long lifecycles of cicadas (Karban 1997). If host-level selection for symbiont quality is tested less frequently in long-lived insects, then this may allow slightly less fit symbiont genotypes to rise to high frequency without being purged by host-level selection. Further work targeting *Hodgkinia* from other cicada species with differing life cycles may help refine this model by establishing the frequency of these kind of lineage-splitting events.

Why does *Hodgkinia* fracture into many lineages while *Sulcia* remains cohesive? In all reported cicada species, *Sulcia* and *Hodgkinia* are contained within different insect cells but have been restricted to cicada tissues for tens of millions of years. Therefore, *Sulcia* and *Hodgkinia* should be subject to the same forces imposed by their extensive gene loss and living conditions —i.e., the effects of strict asexuality, intracellularity, host dependence, and transovarial transmission should be the same for both endosymbionts.

We suggest that the structural differences between the *Sulcia* and *Hodgkinia* genomes may be due to differences in their mutation rates. (While the mutation rate itself has not been measured in *Sulcia* or *Hodgkinia*, here we use the relative DNA substitution rates as a proxy.) *Sulcia* has been noted to have a very low DNA substitution rate in various insects, usually with its partner co-primary symbiont showing a more rapid rate of sequence evolution (Takiya et al. 2006; Powell 2009; McCutcheon and Moran 2012). For example, in sharpshooters, *Sulcia* has a 5X slower rate of DNA substitution than its partner symbiont *Baumannia cicadellinicola* (Powell 2009). Thus, symbiont pairs that are present in the same host can have different rates of sequence evolution, perhaps due to mechanical differences in their DNA replication machinery (Takiya et al. 2006).

The difference in DNA substitution rate between partner endosymbionts appears to be even more dramatic in the case of *Sulcia* and *Hodgkinia*. By comparing the average rates of synonymous site substitutions (d_S) in *Sulcia* and *Hodgkinia* homologs in different cicada species, we estimate

that the DNA substitution rate is between 17- to 137-fold higher in *Hodgkinia* than in *Sulcia* (Table S2). The model we propose for the lineage splitting events in TETUND and MAGTRE require at least two complementary and inactivating mutations to arise in different *Hodgkinia* cells (Fig 8). If the mutation rate is much higher in *Hodgkinia* compared to *Sulcia*, then the odds of acquiring two mutations in the *Hodgkinia* population for a given number of genome replication cycles is higher in *Hodgkinia*. *Sulcia* will still encounter inactivating mutations, and these may even rise to high frequency, but cell lineages that accumulate high levels of these deleterious genotypes will eventually be purged by selection (Fig. 8A). It is also possible that there are cell biological reasons why *Sulcia* and *Hodgkinia* are different. In particular, the patterns of genome evolution in *Hodgkinia* suggest that its cellular boundary is porous to most moleucles except genomes (McCutcheon et al. 2009b), while this may not be true in *Sulcia.* In this case, it would not be possible for inactivating mutations in two different *Sulcia* cells to interact, and thus cell lineages carrying inactivating mutations would not be masked from selection by other lineages with active gene copies and would eventually be purged by host-level purifying selection.

Why do symbionts fracture into many lineages in cicadas, but not in other insects? Aside from *Hodgkinia*, many other endosymbionts with tiny genomes have very high substitution rates (McCutcheon and Moran 2012). Why have the lineage-splitting events we observe in *Hodgkinia* not occurred to symbionts in other insects, and why are they found in only some lineages of cicadas? We suggest that it is related to the very long and variable life cycles of cicadas. While some exceptional insects have multi-year diapause stages that can last more than 25 years (e.g. (Denno and Roderick 1990)), the vast majority of sap-feeding insects have life cycles of one year or less (Heliövaara et al. 1994; Williams and Simon 1995; Nickel and Remane 2002). With known life cycles ranging from 2 to 19 years, cicadas are therefore among the longest-lived non-diapausing insects, (White and Strehl 1978; Karban 1997). Most cicada species for which we have data have life cycles of two to five years, with the synchronized thirteen- and seventeen-year life cycles of periodical cicadas in the genus *Magicicada* at the long end of the spectrum (Table S3).

We hypothesize that the number of splitting events experienced by a *Hodgkinia* lineage is proportional to the life cycle length of the cicada in which it resides. This could be the result of two factors. The first is the inferred high mutation rate in *Hodgkinia*—it could simply be that the longer an insect lives, the more genome replication cycles *Hodgkinia* undergoes and thus the likelihood of accumulating inactivating mutations is higher. The second factor relates the amount of time a cicada species exists in states of lowered metabolism such as winter diapause (Itô and Nagamine 1981; Logan et al. 2014), or the waiting period (Karban 1997) between when it has reached the critical 5th instar weight and when it emerges above ground. Because *Sulcia* and *Hodgkinia* provide essential amino acids to their host cicada (McCutcheon et al. 2009a), we assume that the host will test the quality of symbiont genotypes most intensley when protein synthesis is at its maximum; that is, when the insect is putting on mass during growth. Therefore, if there are periods during the cicada lifecycle where the symbionts are undergoing genome replication (in order to be maintained and passed to the next generation) but when the cicada is not putting on mass, then it may be possible for less fit symbiont genotypes to accumuate

because their 'symbiotic quality' would not be vigorously tested by host-level selection (McCutcheon et al. 2009b). The data from all cicada species surveyed so far support this hypothesis.

The benefits and costs to long-term endosymbiosis. The stable integration of mutualistic bacteria into host cells has profoundly altered the diversity and complexity of life. These 'key innovations' (Szathmáry and Smith 1995; Sachs et al. 2011) can promote rapid diversification by propelling the new symbiotic consortium into previously inaccessible ecological niches (Margulis 1981). While these events are often initially adaptive, endosymbionts that become stably associated inside host cells can undergo a long period of degenerative evolution as the partners become more intimately intertwined and genetic drift plays a larger role in their evolution (Moran 1996; Andersson and Kurland 1998).

The classic examples of this process are the mitochondria and plastids, the eukaryotic cellular organelles resulting from the endosymbiosis of an alphaproteobacterium and a cyanobacterium, respectively (Gray and Doolittle 1982). By allowing their hosts access to new forms of energy, these organelles are ultimately responsible for much of the macro-scale organismal diversity present on Earth today (Lane and Martin 2010). However, their exclusive presence in host cells and strict vertical transmission also limit their evolutionary potential. Mitochondrial genomes in particular are characterized by large amounts of gene loss relative to their bacterial ancestors, and by wild diversity in genomic architecture (Burger et al. 2003). It is now clear that this diversity is derived; the ancestral mitochondrial genome was probably a circular chromosome with a distinct bacterial nature (Lang et al. 1997). Flowering plants and some algae display dramatic organelle genome heterogeneity, with multi-circular chromosomes and enormous genome expansions resulting from both horizontal acquisition of foreign DNA and non-coding genome proliferation (Sloan et al. 2012; Rice et al. 2013; Smith et al. 2013). Because these increases in chromosome number and genome size do not increase the functional capacity of their hosts in any obvious way—most of the expanded DNA is non-coding—these events seem likely to reflect increases in genomic complexity that result from non-adaptive evolution (Lynch et al. 2006; Lynch 2007; Sloan et al. 2012; Rice et al. 2013; Smith et al. 2013).

Constructive neutral evolution as a mechanism for endosymbiont speciation. Although the mechanism of genome expansion is clearly different between *Hodgkinia* and organelle genomes, we suggest that these examples share the distinctive feature of having originated by chance rather than by necessity, i.e. by drift rather than by selection. In both cases, genome sizes are increased without adding any apparent functional capacity, and selection seems only to act to preserve ancestral gene function in the face of added genomic complexity. The splitting of the *Hodgkinia* lineage in some cicadas added an additional endosymbiont to the symbiosis, but differs importantly from other examples in insects, where the acquisition of an unrelated additional symbiont brought with it a large set of new bacterial genes upon which selection could act (McCutcheon and Moran 2010; Lamelas et al. 2011; McCutcheon and von Dohlen 2011; Sloan and Moran 2012). Here, the evolution of a new symbiont resulted from a speciation event that served only to partition existing *Hodgkinia* genes into two new lineages; this event apparently brought no new genetic capacity to the system.

We thus suggest that our results highlight the role chance can play in the evolution of biological complexity (Gould and Lewontin 1979; Doolittle and Sapienza 1980; Lynch 2007;

Gray et al. 2010; Finnigan et al. 2012). On one hand, the interdependence of the duplicated genomes—and they with their co-symbiont *Sulcia*, and all symbionts together with their cicada host—seem exquisitely engineered and fine-tuned. However, it is difficult to imagine that this symbiosis could have evolved to such extravagant complexity because the simpler way, with only *Sulcia* and a single *Hodgkinia* genotype, was less effective. Selection certainly maintains the complex organismal and genomic complementarity in this system, but we favor the idea that these patterns exist from adaptation born of necessity. Neutral, or even maladaptive, mutations can drift to high frequency in these populations because they are initially hidden from selection (Fig. 8B-8C), but once fixed have to be dealt with or the entire symbiosis collapses. This process, sometimes called "constructive neutral evolution" (CNE), has been argued to play a role in the generation of genomic and molecular complexity (Gray et al. 2010; Stoltzfus 2012). Here we suggest that a non-adaptive process similar to CNE has driven the evolution of new bacterial lineages. Our results reinforce the idea that, at least in some circumstances, neutral processes should be considered together with selection as a force driving the complexity of biological systems.

Differences and similarities in endosymbiont and organelle genome evolution. One important difference between mitochondria and *Hodgkinia* is the physical location of the genomes. The *Hodgkinia* genomes from TETUND (Van Leuven et al. 2014) and at least some of the circles from MAGTRE (Fig. 6, 7) appear to remain cytologically distinct, while this is likely not true in mitochondria because of the frequent fission and fusion events they undergo (Sheahan et al. 2005). Indeed, the frequency of mitochondrial fusion is the explanation proposed for the massive levels of foreign DNA acquisition seen in mitochondrial genomes from the plant genus *Amborella* (Rice et al. 2013) . Thus, even when mitochondrial genomes fragment into several chromosomes, those chromosomes stay distributed throughout a cell's mitochondria because of frequent organelle fusion. In contrast, when a *Hodgkinia* lineage fragments, each new genome seems to stay sequestered into discrete cells and mixing does not occur.

Despite these cell biological differences, decades of work on organelle and endosymbiont genomes has shown that genome reduction is a strong unifying theme of intracellular symbioses. While many organelle genomes remain small and gene dense, others have undergone secondary genome expansions through DNA proliferation or acquisition that make the genome larger but add little or no coding capacity (Rice et al. 2013; Smith and Keeling 2015; Wu et al. 2015). Similarly, most insect endosymbiont genomes are small and gene dense, but here we have shown that the '*Hodgkinia* genome complex' has grown in size by almost an order of magnitude and has drastically reduced its coding density, but through a different process involving lineage splitting and reciprocal gene inactivation. These examples of secondary genome expansion have three important similarities. The first is that they have all lead to the accumulation of large amounts of 'junk' DNA, inspiring arguments that these genome expansions are the result of nonadaptive evolution (Lynch et al. 2006; Boussau et al. 2011; Sloan et al. 2012; Rice et al. 2013; Van Leuven et al. 2014). The second is that mutation rate seems to an important correlate in the structure and stability of organelle (Sloan et al. 2012) and endosymbiont genomes. The third is that they both have evolved in the context of absolute co-dependency with their hosts. A eukaryotic cell is nothing without its mitochondria, just as an insect that only eats plant sap is nothing without its

endosymbiotic bacteria. It is likely that strong selection on the host to maintain the symbiosis provides a fertile ground for nonadaptive processes observed in organelles and endosymbionts. If conditions arise whereby an organelle acquires several genomes worth of foreign DNA, such as in *Amborella* (Rice et al. 2013); or if an insect host is not able to stop an endosymbiont splitting its genome into tens or hundreds of discrete cells, the host—and therefore the entire symbiosis—has no choice but to cope with the changes or die.

3.8 Methods and suppplementary materials

Supplemental Tables

Table S1. Distribution of genes by functional class in both *Hodgkinia* TETUND genomes. Raw gene counts and percent of total retained are shown. Gene functional classifications were obtained from (McCutcheon and Moran 2010).

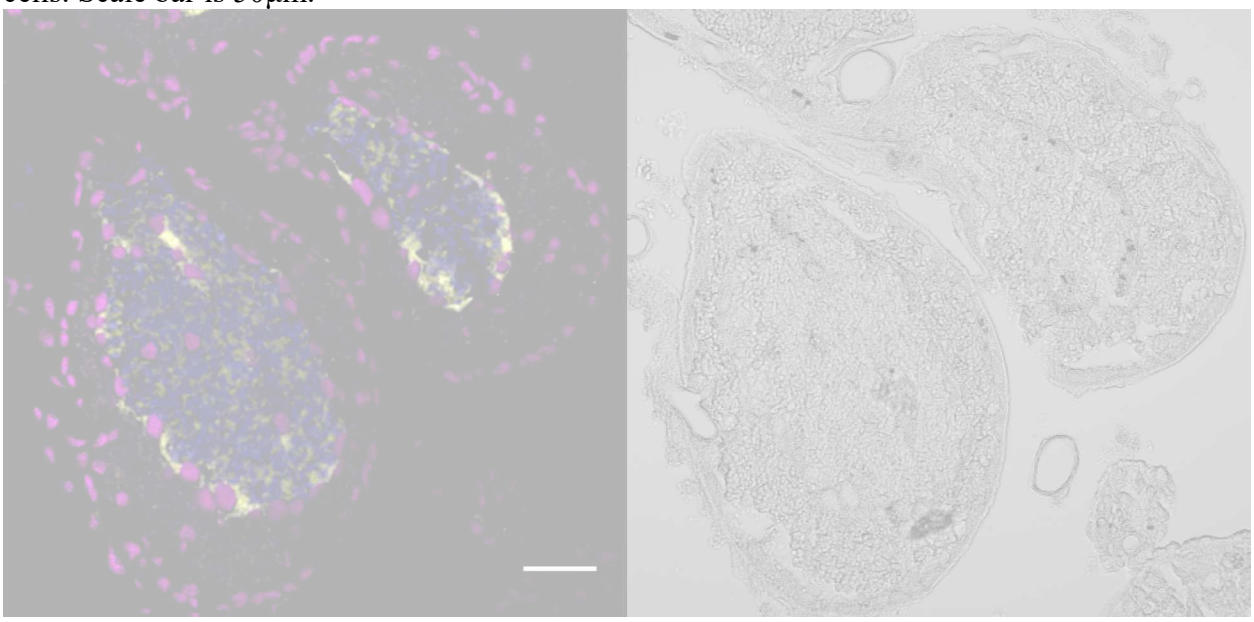
Functional class	# genes in category	# genes present in both TETUND genomes	% retained in both
Protein folding	4	4	100
Transcription	5	4	80
Aminoacyl tRNA formation	12	9	75
Ribosomal subunit	43	29	67
Amino acid biosynthesis	17	11	65
Unknown function	5	3	60
Replication	2	1	50
General metabolism	18	7	41
Vitamin biosynthesis	20	4	20
RNA processing	3	0	0
Translation	7	0	0

Table S2. Genome-wide nonsynonymous and synonymous nucleotide substitution estimations. Values are reported as the mean plus or minus two standard errors.

Ortholog set	TETULN vs. TETUND1	TETULN vs. TETUND2	TETUND1 vs. TETUND2
d*N* all	0.17±0.03	0.09±0.01	0.14±0.03
d*S* all	0.42±0.04	0.42±0.02	0.23±0.03
d*N*/d*S* all	0.35±0.04	0.29±0.03	0.54±0.06

d*N* functional in all three genomes	0.11±0.01	0.09±0.01	0.06±0.01
d*S* functional in all three genomes	0.42±0.03	0.41±0.02	0.17±0.01
d*N*/d*S* functional all in three genomes	0.27±0.03	0.24±0.03	0.35±0.06
d*N* functional vs. pseudogenes	0.32±0.06	0.32±0.05	0.26±0.05
d*S* functional vs. pseudogenes	0.55±0.10	0.58±0.01	0.31±0.07
d*N*/d*S* functional vs. pseudogenes	0.58±0.07	0.55±0.09	0.81±0.08
d*N* functional but pseudogenized in partner TETUND genome	0.11±0.02	0.11±0.01	
d*S* functional but pseudogenized in partner TETUND genome	0.48±0.05	0.42±0.03	
d*N*/d*S* functional but pseudogenized in partner TETUND genome	0.26±0.02	0.27±0.03	

Fig. S1. FISH microscopy using genome-targeted probes distinguish *Hodgkinia* in *Tettigades near undata*, related to Figure 4. (A) The two *Hodgkinia* genomes (blue and yellow) show from bacteriome tissue thin-sections as described in Fig. 4. Hoechst stained DNA is colored magenta, and primarily stains insect nuclei; no *Sulcia* probe was used in this experiment. DIC image (B) shows characteristic cell morphology, with *Hodgkinia* cells surrounded by *Sulcia* cells. Scale bar is 50μm.



Supplementary methods

Cicada provenance and identification. Genome sequencing and microscopy for *Tettigades undata* was performed from two ethanol preserved cicadas (a single female for the genomics, a single male for the microscopy), wild-caught in 2006 from the BioBio providence of Chile, approximately 2 km west of Cabrero (S37°4'0.5" W72°19'52.1"). Genome sequencing for *Tettigades ulnaria* was performed on a single ethanol preserved female cicada, wild-caught in 2013 near Pichilemu city, Chile (S34°28'51" W71°58'31"). Draft genome sequencing for *Tettigades auropolisa* was performed on a single ethanol preserved male cicada, wild-caught in 2012 from Cordillera providence of Chile (S33°48'55" W70°11'24"). Mitochondrial genome sequencing for *Magicicada tredecim* was performed on a single ethanol preserved female, wild-caught in 2011 from King William County, Virginia. Cicadas were identified by comparison to: 1) paratypes from the British Museum, 2) character descriptions in (Torres 1958), and 3) mtDNA COI-barcode ID and phylogenetic position in relation to other *Tettigades* species.

Total DNA was purified from dissected bacteriome tissue from twelve ethanol preserved cicadas, wild-caught in 2011 from King William County, Virginia, using the Qiagen DNeasy Blood and Tissue kit. DNA libraries from individual cicadas were separately barcoded for Illumina short-insert sequencing using NEXTflex adapters and protocols (Bioo Scientific). Pooled DNA from the same individuals was used to generate the Illumina Nextera large-insert and PacBio RS II DNA libraries using standard protocols from the manufacturer.

DNA sequencing and genome assembly. DNA was purified using the Qiagen DNeasy Blood and Tissue kit, and was prepared and sequenced using the Illumina GAII , HiSeq, or MiSeq platforms. Adapter sequences were trimmed with Trimmomatic v0.03 and reads were quality filtered using FASTX Toolkit v0.0.13 (fastq_quality_filter -q 30 -p 90 -Q 33). Meta-velvet v1.1.01 (Namiki et al. 2012) was used to assemble the bacterial and mitochondrial genomes using k-mers ranging from 81-161 and expected coverage equal to the approximate k-mer coverage of each bacterial genome. Genome scaffolds were connected and circularized by PCR and Sanger sequencing. The ribosomal operons (~7.5kb) of TETUND1 and TETUND2 were closed by long-range PCR and Sanger sequencing by primer-walking across the amplicons.

Adapter sequences were trimmed with trimmomatic (parameters: SLIDINGWINDOW:10:15 LEADING:3 TRAILING:3 MINLEN:60)(58) and quality filtered using FASTX Toolkit v0.0.13. High-quality, paired reads were assembled using SPAdes v3.1.1 (65), with kmer sizes of 91 and 95. Uncorrected PacBio reads were used to scaffold with SSPACE-LONGREADS v1.1 (66). Putatively circular scaffolds were confirmed with manual inspection of mate-pair read mapping and Sanger sequencing of PCR products. Internal gaps in the scaffolds were closed using PacBio reads and custom Python scripts.

For MAGTRE, the following data was generated for each sequencing technology: 136,081,956 pairs of 100x2 short insert Illumina HiSeq reads for about 27 Gb total; 50,884,070 pairs of 100x2 large insert HiSeq reads for about 10 Gb total; and 259,593 reads averaging 1600 nts for about 421 Mb total.

Molecular Evolution. Macse v0.9b1 (Ranwez et al. 2011) was used to align TETULN and TETUND amino acid sequences, which were back translated to produce alignments of the original nucleotide sequences. PAML v4.6 (Yang 2007) was used to estimate amino acid and d_N/d_S values using codeml (parameters: runmode=-2, seqtype=2, aaRatefile=wag.dat, model=2, cleandata=0 and yn00 parameters; icode=3, weighting=0, commonf3x4=0). Protest-3.2 (Darriba et al. 2011) was used on a subset of amino acid alignments to estimate the appropriate substitution matrix, and WAG (Whelan and Goldman 2001) consistently ranked in the top ten best fitting matrices. Likelihood scores were calculated using baseml (model=REV, fix_alpha=estimate, ncatG=5) for A) constrained clock in TETUND1 and TETUND2 lineages, or B) unconstrained clock. Likelihood scores used in determining positive selection on a branch and/or sites were calculated using codeml (parameters; runmode=0, seqtype=1, model=2(null) or 1, clock=0, NSsites=0 or 2, fix_kappa=0, kappa=2, fix_omega=0 or 1, omega=1, fix_alpha=1, alpha=0, ncatG=10). Omega was constrained on only one branch at a time (either TETUND1 or TETUND2).

Microscopy. Genome-targeted probes were generated from unique regions of the genomes by PCR, labeled by nick translation to incorporate fluorescently labeled dUTPs, and hybridized according to (Sarkar and Hopper 1998) except that no *E. coli* tRNAs were added to the hybridization or pre-hybridization buffer. Single, double, and triple probe hybridizations with and without Hoechst were done to check for channel bleed-through. Negative controls were used to assess insect tissue auto-fluorescence. Sixty slices, 0.146 uM spaced, imaged at 1024x1024 pixel resolution, were acquired on a Leica TCS SP5 inverted confocal microscope using a 63X 1.4 NA oil-immersion lens.

Supplemental Experimental Procedures

DNA sequencing and genome assembly. DNA was purified using the Qiagen DNeasy Blood and Tissue kit, and was prepared and sequenced using the Illumina GAII or MiSeq platforms. Adapter sequences were trimmed with Trimmomatic v0.03 and reads were quality filtered using FASTX Toolkit v0.0.13 (fastq_quality_filter -q 30 -p 90 -Q 33). Meta-velvet v1.1.01 (Namiki et al. 2012) was used to assemble the *Sulcia* TETUND and *Hodgkinia* TETUND genomes (-kmer 81 -ins_length 250 -ins_length_sd 50 -exp_covs 400_150_75 -max_divergence 0.03 -max_gap_count 2 or -kmer 161 -exp_covs 400_150_50 or -kmer 181 -exp_cov 20 -max_cov 200). Contigs containing blastx hits (E-value < $1e^{-5}$) to bacteria were retained and further binned into groups belonging to *Sulcia*, *Hodgkinia*, and the cicada mitochondrial genome. TETULN scaffolds were connected and circularized by PCR and Sanger sequencing. The ribosomal operons (~7.5kb) of TETUND1 and TETUND2 were closed by long-range PCR and Sanger sequencing by primer-walking across the amplicons.

Gene content and divergence comparisons. Prokaa v1.5.2 (http://vicbioinformatics.com) was used for initial protein-coding gene calls and was complemented by hand annotation. BLASTP v2.2.25 was used to compare all protein coding genes between the *Hodgkinia cicadicola* DICSEM, TETULN, TETUND1, and TETUND2 genomes (blastall -p blastp -m 8 -e 0.1 -b 1 -v 1). Reciprocal best hits (minimum e-value=0.1) were saved for each possible pair (genes shared among 2 genomes) and for genes shared in all three genomes (TETULN and TETUND1, and

TETUND2) using custom python and perl scripts. Pseudogenes and/or missed gene homologs were searched for by comparing genes only shared among 2 genomes to the other genome using TBLASTN v2.2.25 (-q 4 -F " " -e 0.1). Each hit was examined by eye. Errors in finding homologs caused by incorrect gene calling were fixed by hand for downstream analysis. For each set of homologs, each gene was designated at either 1) complete, 2) pseudogenized, or 3) deleted. Classification as a pseudogene required that ~50% of the gene length is retained with either stop and/or frameshift mutations causing disruption of the ORF. Classification as deleted required that less than ~50% of the gene remains in one of the genomes. ORFs missed or incorrectly called by Prokaa were corrected by hand in Artemis release 14.0.0 (Carver et al. 2012).

Of the 170 CDSs originally annotated in *Hodgkinia* DICSEM, 38 were given a "hypothetical protein" designation. Of these 38, we promoted one to an ORF with a 4-letter name (*cobL*) and one to a gene with putative function (16S rRNA m(4) methyltransferase). We excluded the remaining hypothetical ORFs from our analyses because we found that many were likely non-functional due to lack of conservation in TETUND or from subsequent work in our lab showing many hypothetical ORFs predicted in the original DICSEM annotation overlapped non-coding RNA (Van Leuven and McCutcheon, unpublished). This left 134 protein coding genes in *Hodgkinia* DICSEM which had some proposed function, and that we felt comfortable were likely to be functional genes, for comparative analysis. *Hodgkinia* TETULN contains three protein-coding genes than are not present in *Hodgkinia* DICSEM, but were included in the analysis. Macse v0.9b1 (Ranwez et al. 2011) was used to align TETULN and TETUND amino acid sequences, which were back translated to produce alignments of the original nucleotide sequences. PAML v4.6 (Yang 2007) was used to estimate amino acid and d*N*/d*S* values using codeml (parameters: runmode=-2, seqtype=2, aaRatefile=wag.dat, model=2, cleandata=0 and yn00 parameters; icode=3, weighting=0,commonf3x4=0). Protest-3.2 (Darriba et al. 2011) was used on a subset of amino acid alignments to estimate the appropriate substitution matrix, and WAG (Whelan and Goldman 2001) consistently ranked in the top ten best fitting matrices. Likelihood scores were calculated using baseml (model=REV, fix_alpha=estimate, ncatG=5) for A) constrained clock in TETUND1 and TETUND2 lineages, or B) unconstrained clock. Likelihood scores used in determining positive selection on a branch were calculated using codeml (parameters; runmode=0, seqtype=1, model=2(null) or 1, clock=0, NSsites=0, fix_kappa=0, kappa=2, fix_omega=0 or 1, omega=1, fix_alpha=1, alpha=0, ncatG=10). Omega was constrained on only one branch at a time (either TETUND1 or TETUND2). Custom perl scripts were used to run and parse the output of Macse and PAML. The TTEST, STEYX, and CHIDIST functions of LibreOffice Calc 3.5.72 were used to calculate statistical significance. The amino acid distance and dN/dS data were log transformed for statistical comparisons of means. The Shapiro-Wilk test was implemented in R to analyze normality. Processing v2.0.3 (http://processing.org) was used to generate the genome map and amino acid divergence boxplots in Fig. 2A and 2C. The mitochondrial gene sequences of 5 cicada species were aligned and back-translated by Macse. The optimal number of site-class partitions were determined using PartitionFinder v1.1.0 with parameters; models=all, model_selection=BIC, search=greedy (http://dx.doi.org/10.1093/molbev/mss020). Likelihood analyses were performed with GARLI v2.01 on partitioned alignments. Bootstrap values were overlaid on the most-likely topology with searchreps=10 with SumTrees v3.3.1. The bootstrap consensus, the most-likely tree, and a ML

tree generated from concatenated COI, COII, COIII, AP6 sequences all share identical topology.

Microscopy. A single ethanol-preserved male *Tettigades undata* cicada, collected from the same date and location as the female specimen used for whole genome sequencing (Simon Lab specimen number 06.CL.BI.CAB.02) and a *Tettidades near undata* male collected from Laguna del Laja National Park, Chile (Simon Lab specimen number 06.CL.BI.LLJ.01), were dissected in 70% ethanol. The bacteriomes were dehydrated through 1 hr incubations in 80%, 90% and 100% ethanol, then cleared in methylscylate for 2x1 hr. Paraffin embedding was done under vacuum for 2x1 hr. Paraffin blocks were thin sectioned to 5-10 uM. Thin sections were de-paraffinized in xylene for 2x5min, then hydrated though a 100%, 85%, 70% ethanol series.
SSU rRNA targeted FISH was done according to (Pernthaler et al. 2001) except that an Olympus FV 1000 IX inverted laser scanning confocal microscope was used for imaging with a 63X oil-immersion lens. The probe sequences were Cy3-CCAATGTGGGGGWACGC (*Sulcia*) and Cy5-CCAATGTGGCTGACCGT (*Hodgkinia*).

Fluorescently labeled DNA targeted probes were made to distinguish the three bacterial genomes present in the cicada bacteriome. A unique ~3kb sequence was chosen from *Hodgkinia cicadicola* TETUND1, *Hodgkinia cicadicola* TETUND2, and *Sulcia muelleri* TETUND corresponding to genome coordinates 85994-89330, 121433-124981, and 48013-51479, respectively and amplified with primers HC1F-AGTAGGCAACACGCCACAG, HC1R-ATAGCCACAAGCTGCCTTC, HC2F-AGTGTGCTAGCGTTAAGCTG, HC2R-AGCAAGGGCATCGCGCAATG, SM1F-GTTTCTCGCCATAATCTAGAAG, SM1R-AGATCTTGCAAAAGAGGCAG. The PCR mix was comprised of 1uL each of forward and reverse primers (10uM each), 1uL dNTPs (10 mM each), 10 ng template DNA, 10 uL OneTaq buffer, 0.25uL OneTaq (M0480), 35.75 water. Thermocycling conditions were 94° C for 1min, 94° C for 15 sec, 56° C for 30 sec, 68° C for 3 min, and 68° C for 5 min, with 35 cycles. Each amplicon was cloned into the PGEMT-Easy vector and a single insert positive clone was maintained in glycerol stocks of transformed Invitrogen Top10 *E. coli* cells. The transformed *E. coli* cells were used to test probe specificity. PCR was done on purified plasmid from each of the three clones using M13 primers. The PCR products were checked by Sanger sequencing, then subjected to nick-translation to incorporate fluorescently labeled dUTPs (Jena Biosciences: Aminoallyl-dUTP-Cy5, Aminoallyl-dUTP-Cy3, Aminoallyl-dUTP-ATTO488). The nick-translation mix contained ~200ng/uL PCR product, 1X nick-translation buffer, 0.25mM unlabeled dNTPs, 25uM labeled dNTPs, 2.5U/uL DNA polymerase I, and 10mU/uL Dnase. These reactions were incubated at 15° C for 2-4 hours. Probes in the size range of 100-500 nts in length were purified with AMPure XP beads (Agencourt A63880). Probes with at least 1.3 incorporated dNTPs per 1000 nucleotides (measured by UV spectrophotometry) were used for in-situ hybridization.
Genome-targeted probes were hybridized according to (Sarkar and Hopper 1998) except that no *E. coli* tRNAs were added to the hybridization or pre-hybridization buffer. Briefly, once hydrated, tissues were incubated in prehybridization solution (12.5% dextran sulfate, 2.5X SCC, 10ng/uL ssDNA, 0.25% BSA, 1.25U/uL RNaseOut) at 37° C for 1 hour in a humidity chamber. Slides were then briefly washed with warm 2XSCC and incubated overnight at 37° C with hybridization solution (prehybridization solution, 10ng/uL probe, 1.5ug/uL Hoechst 33258) in a humidity chamber. Slides were then incubated in 2XSCC at 37° C for 1 hour, briefly rinsed with

diH2O and preserved with FluorSave (CalbioChem). Single, double, and triple probe hybridizations with and without Hoechst were done to check for channel bleed-through. Negative controls were used to assess insect tissue auto-fluorescence. Sixty slices, 0.146 uM spaced, imaged at 1024x1024 pixel resolution, were acquired on a Leica TCS SP5 inverted confocal microscope using a 63X 1.4 NA oil-immersion lens. Fluoresence was collected by sequentially scanning using the following excitation and emission parameters: Hoechst, 405 nm laser for excitation, fluorescence emission collected from 420-465 nm; Alexa-488, 488 nm excitation, collected at 500-569 nm; CY3, 561 nm excitation, collected at 570-646 nm; and CY5 633 nm excitation, collected at 650-793 nm. Spectral separation was then performed on the z-stack using the on-board Leica channel separation software, and fluorescence spectra collected on singly stained samples.

Post-acquisition processing was done in ImageJ version 1.46a (Schneider et al. 2012). Background signal in the Cy3 and Cy5 lines was estimated by defining a region of interest (ROI) where insect auto-fluorescence is expected, but no Cy3 and Cy5 signal is expected. The average pixel intensity of this ROI was calculated for each slice individually. Custom perl scripts and an ImageJ macro were used to generate a "background" z-stack, where each slice has an even intensity equal to 4 times the background signal calculated previously. The "background" stacks were then subtracted from the real data to generate a background subtracted z-stack for TETUND1 (Cy3) and TETUND2 (Cy5). JACoP (Bolte and Cordelières 2006) was used to calculate colocalization on a ROI that avoids areas of insect auto-fluorescence and any bleed-through from the 405 line. Non-thresholded Manders' Coefficients were M1=0.306 and M2=0.239. The JACoP threshold values were 105 and 103 for Cy3 and Cy5, respectively. Manders' Coefficients using these threshold values were M1=0.04 and M2=0.024, Pearson's r=0.126, Overlap r=0.14. The volume of TETUND1 and TETUND2 was determined by taking the average volume across the entire 60-image z-stack after background subtraction. Fig. 3 is a maximum intensity projection of slices 38-41. A single image of *T. near undata* was taken with similar parameters and hybridization conditions (no *Sulcia* probe), except that the imaging was done on an Olympus FV 1000 IX inverted laser scanning confocal microscope with a 20X lens. The only post-processing done was level adjustment.

The same procedure was used in MAGTRE, except that these primers were used; MAGTRE001: AGGAGAAACTTAAAGTTCATTGATCC and ATTACAATCCTAGATGTCTACCC, MAGTRE0012: AGAAACAACAACATAATAAACAAAGC and AATTATCGAAACATTAACAACACAGC, MAGTRE005: ACACCTAAGCATAGCGTTCC and ATTTATCCAAGTTCATGTAAACCC, and MAGTRE006: AGTGGGTTTTGAATTTAATGTAGG and ATCCGAACTTAACCTTTGAAAACC.

3.9 References

Andersson SGE, Kurland CG. 1998. Reductive evolution of resident genomes. Trends Microbiol. 6:263–268.

Atwood KC, Schneider LK, Ryan FJ. 1951. Periodic Selection in Escherichia Coli. Proc. Natl. Acad. Sci. U. S. A. 37:146–155.

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

Baumann P, Baumann L, Clark MA. 1996. Levels of Buchnera aphidicola Chaperonin GroEL During Growth of the Aphid Schizaphis graminum. Curr. Microbiol. 32:279–285.

Bennett GM, Moran NA. 2013. Small, Smaller, Smallest: The Origins and Evolution of Ancient Dual Symbioses in a Phloem-Feeding Insect. Genome Biol. Evol. 5:1675–1688.

Bolte S, Cordelières FP. 2006. A guided tour into subcellular colocalization analysis in light microscopy. J. Microsc. 224:213–232.

Boussau B, Brown JM, Fujita MK. 2011. Nonadaptive Evolution of Mitochondrial Genome Size. Evolution 65:2706–2711.

Brower AV. 1994. Rapid morphological radiation and convergence among races of the butterfly Heliconius erato inferred from patterns of mitochondrial DNA evolution. Proc. Natl. Acad. Sci. U. S. A. 91:6491–6495.

Burger G, Gray MW, Franz Lang B. 2003. Mitochondrial genomes: anything goes. Trends Genet. 19:709–716.

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

Cohan FM. 2002. Sexual Isolation and Speciation in Bacteria. Genetica 116:359–370.

Conant GC, Wolfe KH. 2006. Functional Partitioning of Yeast Co-Expression Networks after Genome Duplication. PLoS Biol 4:e109.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics:btr088.

Degnan PH, Lazarus AB, Wernegreen JJ. 2005. Genome sequence of Blochmannia pennsylvanicus indicates parallel evolutionary trends among bacterial mutualists of insects. Genome Res. 15:1023–1033.

Denno RF, Roderick GK. 1990. Population Biology of Planthoppers. Annu. Rev. Entomol. 35:489–520.

von Dohlen CD, Kohler S, Alsop ST, McManus WR. 2001. Mealybug β-proteobacterial endosymbionts contain γ-proteobacterial symbionts. Nature 412:433–436.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284:601–603.

Douglas AE. 1998. Nutritional Interactions in Insect-Microbial Symbioses: Aphids and Their Symbiotic Bacteria Buchnera. Annu. Rev. Entomol. 43:17–37.

Ellers J, Toby Kiers E, Currie CR, McDonald BR, Visser B. 2012. Ecological interactions drive evolutionary loss of traits. Ecol. Lett. 15:1071–1082.

Fares MA, Ruiz-González MX, Moya A, Elena SF, Barrio E. 2002. Endosymbiotic bacteria: GroEL buffers against deleterious mutations. Nature 417:398–398.

Finnigan GC, Hanson-Smith V, Stevens TH, Thornton JW. 2012. Evolution of increased complexity in a molecular machine. Nature [Internet]. Available from: http://www.nature.com/nature/journal/vaop/ncurrent/full/nature10724.html

Friesen ML, Saxer G, Travisano M, Doebeli M. 2004. Experimental Evidence for Sympatric Ecological Diversification 2 to Frequency-Dependent Competition in Escherichia coli. Evolution 58:245–260.

Gil R, Silva FJ, 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. 100:9388–9393.

Gould S, Lewontin R. 1979. The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme.

Gray MW, Doolittle WF. 1982. Has the endosymbiont hypothesis been proven? Microbiol. Rev. 46:1.

Gray MW, Lukeš J, Archibald JM, Keeling PJ, Doolittle WF. 2010. Irremediable Complexity? Science 330:920–921.

Gruwell ME, Hardy NB, Gullan PJ, Dittmar K. 2010. Evolutionary Relationships among Primary Endosymbionts of the Mealybug Subfamily Phenacoccinae (Hemiptera: Coccoidea: Pseudococcidae). Appl. Environ. Microbiol. 76:7521–7525.

van Ham RCHJ, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, Fernández JM, Jiménez L, Postigo M, Silva FJ, et al. 2003. Reductive genome evolution in Buchnera aphidicola. Proc. Natl. Acad. Sci. U. S. A. 100:581–586.

Heliövaara K, Väisänen R, Simon C. 1994. Evolutionary ecology of periodical insects. Trends Ecol. Evol. 9:475–480.

Hughes AL. 1994. The Evolution of Functionally Novel Proteins after Gene Duplication. Proc. Biol. Sci. 256:119–124.

Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, Tanaka M, Satoh N, Bachtrog D, Wilson ACC, et al. 2013. Horizontal Gene Transfer from Diverse Bacteria to an Insect

Genome Enables a Tripartite Nested Mealybug Symbiosis. Cell 153:1567–1578.

Itô Y, Nagamine M. 1981. Why a cicada, Mogannia minuta Matsumura, became a pest of sugarcane: an hypothesis based on the theory of “escape.” Ecol. Entomol. 6:273–283.

Karban R. 1997. Evolution of prolonged development: a life table analysis for periodical cicadas. Am. Nat. 150:446–461.

Kinnersley M, Wenger J, Kroll E, Adams J, Sherlock G, Rosenzweig F. 2014. Ex Uno Plures: Clonal Reinforcement Drives Evolution of a Simple Microbial Community. PLoS Genet 10:e1004430.

Koga R, Bennett GM, Cryan JR, Moran NA. 2013. Evolutionary replacement of obligate symbionts in an ancient and diverse insect lineage. Environ. Microbiol.:n/a – n/a.

Komaki K, Ishikawa H. 1999. Intracellular bacterial symbionts of aphids possess many genomic copies per bacterium. J. Mol. Evol. 48:717–722.

Kuo C-H, Ochman H. 2009. Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. Biol. Direct 4:35.

Lamelas A, Gosalbes MJ, Manzano-Marín A, Peretó J, Moya A, Latorre A. 2011. Serratia symbiotica from the Aphid Cinara cedri: A Missing Link from Facultative to Obligate Insect Endosymbiont. PLoS Genet 7:e1002357.

Lane N, Martin W. 2010. The energetics of genome complexity. Nature 467:929–934.

Lang BF, Burger G, O’Kelly CJ, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Gray MW. 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature 387:493–497.

Logan DP, Rowe CA, Maher BJ. 2014. Life history of chorus cicada, an endemic pest of kiwifruit (Cicadidae: Homoptera). N. Z. Entomol. 37:96–106.

Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc. Natl. Acad. Sci. 104:8597–8604.

Lynch M, Conery JS. 2000. The Evolutionary Fate and Consequences of Duplicate Genes. Science 290:1151–1155.

Lynch M, Koskella B, Schaack S. 2006. Mutation Pressure and the Evolution of Organelle Genomic Architecture. Science 311:1727–1730.

Macdonald SJ, Lin GG, Russell CW, Thomas GH, Douglas AE. 2012. The central role of the host cell in symbiotic nitrogen metabolism. Proc. R. Soc. B Biol. Sci. 279:2965–2973.

Margulis L. 1981. Symbiosis in Cell Evolution. San Francisco: W. H. Freeman

McCutcheon JP, McDonald BR, Moran NA. 2009a. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. Proc. Natl. Acad. Sci. 106:15394–15399.

McCutcheon JP, McDonald BR, Moran NA. 2009b. Origin of an Alternative Genetic Code in the Extremely Small and GC–Rich Genome of a Bacterial Symbiont. PLoS Genet 5:e1000565.

McCutcheon JP, Moran NA. 2007. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proc. Natl. Acad. Sci. U. S. A. 104:19392–19397.

McCutcheon JP, Moran NA. 2010. Functional Convergence in Reduced Genomes of Bacterial Symbionts Spanning 200 My of Evolution. Genome Biol. Evol. 2:708–718.

McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. 10:13–26.

McCutcheon JP, von Dohlen CD. 2011. An Interdependent Metabolic Patchwork in the Nested Symbiosis of Mealybugs. Curr. Biol. 21:1366–1372.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. U. S. A. 93:2873–2878.

Moran NA, Dale C, Dunbar H, Smith WA, Ochman H. 2003. Intracellular symbionts of sharpshooters (Insecta: Hemiptera: Cicadellinae) form a distinct clade with a small genome. Environ. Microbiol. 5:116–126.

Moran NA, Tran P, Gerardo NM. 2005. Symbiosis and Insect Diversification: an Ancient Symbiont of Sap-Feeding Insects from the Bacterial Phylum Bacteroidetes. Appl. Environ. Microbiol. 71:8802–8810.

Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M. 2006. The 160-Kilobase Genome of the Bacterial Endosymbiont Carsonella. Science 314:267.

Namiki T, Hachiya T, Tanaka H, Sakakibara Y. 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res.:gks678.

Nickel H, Remane R. 2002. Check list of the planthoppers and leafhoppers of Germany, with notes on food plants, diet width, life cycles, geographic range and conservation status. Beitr. Zur Zikadenkunde 5:27–64.

Ohno S. 1970. Evolution by gene duplication. Springer-Verlag

Otto SP. 2007. The Evolutionary Consequences of Polyploidy. Cell 131:452–462.

Papadopoulou A, Anastasiou I, Vogler AP. 2010. Revisiting the Insect Mitochondrial Molecular Clock: The Mid-Aegean Trench Calibration. Mol. Biol. Evol. 27:1659–1672.

Pernthaler J, Glockner FO, Schonhuber W, Amann R. 2001. Fluorescence in situ hybridization (FISH) with rRNA-targeted oligonucleotide probes. Methods Microbiol. Vol 30 30:207–226.

Plucain J, Hindré T, Gac ML, Tenaillon O, Cruveiller S, Médigue C, Leiby N, Harcombe WR, Marx CJ, Lenski RE, et al. 2014. Epistasis and Allele Specificity in the Emergence of a Stable Polymorphism in Escherichia coli. Science 343:1366–1369.

Powell JA. 2009. Longest Insect Dormancy: Yucca Moth Larvae (Lepidoptera: Prodoxidae) Metamorphose After 20, 25, 30 Years in Diapause. Ann Entomol Soc Am 94:677–680.

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. PLoS ONE 6:e22594.

Rice DW, Alverson AJ, Richardson AO, Young GJ, Sanchez-Puerta MV, Munzinger J, Barry K, Boore JL, Zhang Y, dePamphilis CW, et al. 2013. Horizontal Transfer of Entire Genomes via Mitochondrial Fusion in the Angiosperm Amborella. Science 342:1468–1473.

Rio RVM, Symula RE, Wang J, Lohs C, Wu Y, Snyder AK, Bjornson RD, Oshima K, Biehl BS, Perna NT, et al. 2012. Insight into the Transmission Biology and Species-Specific Functional Capabilities of Tsetse (Diptera: Glossinidae) Obligate Symbiont Wigglesworthia. mBio 3:e00240–11.

Sabree ZL, Degnan PH, Moran NA. 2010. Chromosome Stability and Gene Loss in Cockroach Endosymbionts. Appl. Environ. Microbiol. 76:4076–4079.

Sachs JL, Skophammer RG, Regus JU. 2011. Evolutionary transitions in bacterial symbiosis. Proc. Natl. Acad. Sci. U. S. A. 108:10800–10807.

Sarkar S, Hopper AK. 1998. tRNA Nuclear Export in Saccharomyces cerevisiae: In Situ Hybridization Analysis. Mol. Biol. Cell 9:3041–3055.

Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. Nat. Methods 9:671–675.

Sheahan MB, McCurdy DW, Rose RJ. 2005. Mitochondria as a connected population: ensuring continuity of the mitochondrial genome during plant cell dedifferentiation through massive mitochondrial fusion. Plant J. 44:744–755.

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

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR. 2012.

Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates. PLoS Biol 10:e1001241.

Sloan DB, Moran NA. 2012. Genome Reduction and Co-evolution between the Primary and Secondary Bacterial Symbionts of Psyllids. Mol. Biol. Evol. 29:3781–3792.

Sloan DB, Moran NA. 2013. The Evolution of Genomic Instability in the Obligate Endosymbionts of Whiteflies. Genome Biol. Evol. 5:783–793.

Sloan DB, Nakabachi A, Richards S, Qu J, Murali SC, Gibbs RA, Moran NA. 2014. Parallel Histories of Horizontal Gene Transfer Facilitated Extreme Reduction of Endosymbiont Genomes in Sap-Feeding Insects. Mol. Biol. Evol. 31:857–871.

Smith DR, Hamaji T, Olson BJSC, Durand PM, Ferris P, Michod RE, Featherston J, Nozaki H, Keeling PJ. 2013. Organelle Genome Complexity Scales Positively with Organism Size in Volvocine Green Algae. Mol. Biol. Evol. 30:793–797.

Smith DR, Keeling PJ. 2015. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. Proc. Natl. Acad. Sci. 112:10177–10184.

Stoltzfus A. 2012. Constructive neutral evolution: exploring evolutionary theory's curious disconnect. Biol. Direct 7:35.

Szathmáry E, Smith JM. 1995. The major evolutionary transitions. Nature 374:227–232.

Takiya DM, Tran PL, Dietrich CH, Moran NA. 2006. Co-cladogenesis spanning three phyla: leafhoppers (Insecta: Hemiptera: Cicadellidae) and their dual bacterial symbionts. Mol. Ecol. 15:4175–4191.

Tamas I, Klasson L, Canbäck B, Näslund AK, Eriksson A-S, Wernegreen JJ, Sandström JP, Moran NA, Andersson SGE. 2002. 50 Million Years of Genomic Stasis in Endosymbiotic Bacteria. Science 296:2376–2379.

Thao ML, Gullan PJ, Baumann P. 2002. Secondary (γ-Proteobacteria) Endosymbionts Infect the Primary (β-Proteobacteria) Endosymbionts of Mealybugs Multiple Times and Coevolve with Their Hosts. Appl. Environ. Microbiol. 68:3190–3197.

Torres BA. 1958. Revision Del Genero Tettigades Amy. Y Serv. (Homo- ptera-Cicadidae). Revista del Museo de La Plata. Secc. Zool. VII:51–106.

Treves DS, Manning S, Adams J. 1998. Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of Escherichia coli. Mol. Biol. Evol. 15:789–797.

Van Leuven JT, McCutcheon JP. 2012. An AT Mutational Bias in the Tiny GC-Rich

Endosymbiont Genome of Hodgkinia. Genome Biol. Evol. 4:24–27.

Van Leuven JT, Meister RC, Simon C, McCutcheon JP. 2014. Sympatric Speciation in a Bacterial Endosymbiont Results in Two Genomes with the Functionality of One. Cell 158:1270–1280.

Whelan S, Goldman N. 2001. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Mol. Biol. Evol. 18:691–699.

White J, Strehl CE. 1978. Xylem feeding by periodical cicada nymphs on tree roots. Ecol. Entomol. 3:323–327.

Williams KS, Simon C. 1995. The Ecology, Behavior, and Evolution of Periodical Cicadas. Annu. Rev. Entomol. 40:269–295.

Wilson ACC, Ashton PD, Calevro F, Charles H, Colella S, Febvay G, Jander G, Kushlan PF, Macdonald SJ, Schwartz JF, 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. 19:249–258.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nat. Rev. Genet. 2:333–341.

Woyke T, Tighe D, Mavromatis K, Clum A, Copeland A, Schackwitz W, Lapidus A, Wu D, McCutcheon JP, McDonald BR, et al. 2010. One Bacterial Cell, One Complete Genome. PLoS ONE 5:e10314.

Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL, Khouri H, Tallon LJ, Zaborsky JM, Dunbar HE, Tran PL, et al. 2006. Metabolic Complementarity and Genomics of the Dual Bacterial Symbiosis of Sharpshooters. PLoS Biol 4:e188.

Wu Z, Cuthbert JM, Taylor DR, Sloan DB. 2015. The massive mitochondrial genome of the angiosperm Silene noctiflora is evolving by gain or loss of entire chromosomes. Proc. Natl. Acad. Sci.:201421397.

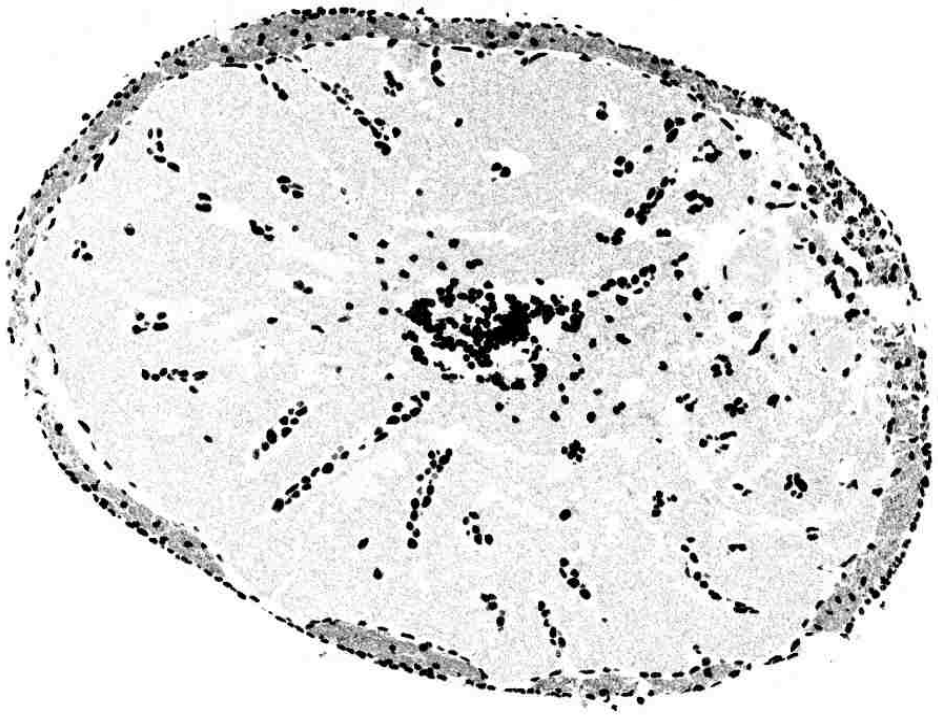
Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19:908–917.

Zientz E, Dandekar T, Gross R. 2004. Metabolic Interdependence of Obligate Intracellular Bacteria and Their Insect Hosts. Microbiol. Mol. Biol. Rev. 68:745–770.

Chapter 4: Transfer RNA presence and processing in the cicada *Diceroprocta* semicincta
Unpublished

Summary
Gene loss and genome reduction are defining characteristics of nutritional endosymbiotic bacteria. In extreme cases, even 'essential' genes related to core processes such as replication, transcription, translation are deleted from the endosymbiont genome. The bacterial symbionts of the cicada *Diceroprocta semicincta*, *Ca.* Hodgkinia cicadicola and *Ca.* Sulcia muelleri, encode only 26 and 16 tRNA, and 15 and 10 aminoacyl tRNA synthetase genes, respectively. Furthermore, the existing *Ca.* Hodgkinia is missing several essential genes involved in tRNA processing, such as RNase P and CCA tRNA nucleotidyltransferase, as well as several RNA editing enzymes required for tRNA maturation. How *Ca.* Sulcia and *Ca.* Hodgkinia preform basic cellular processes without these genes remains unknown, but could be explained by some combination of horizontal gene transfer (HGT) to the host genome, functional complementation from genes from the host lineage, incorrect or incomplete genome annotation, or other unknown compensatory mechanisms enabling the loss of certain functions. Here, we show that the limited *Ca.* Sulcia and *Ca.* Hodgkinia tRNA set predicted by computational annotation was correct. We show that despite the absence of genes encoding tRNA processing activities on the symbiont genomes, symbiont tRNAs have correctly processed 5' and 3' ends, and seem to undergo nucleotide modification at some positions. We conclude that these essential translation-related functions are most likely performed by host-encoded enzymes.



DNA stain of cicada bacteriome containing many multinucleated bacteriocytes.
Image by James Van Leuven

4.1 Introduction

The sequencing of endosymbiont genomes over the past two decades has revealed a series of genetic changes that occur in bacterial genomes during the transition from a free-living to a strictly intracelluar lifestyle (Toft and Andersson 2010; McCutcheon and Moran 2012). At the onset of symbiosis, endosymbiont genomes undergo genome rearrangement, mobile element proliferation, and pseudogenization of non-essential genes (Burke and Moran 2011; Manzano-Marín and Latorre 2014; Oakeson et al. 2014). Following this period of genomic turmoil, endosymbionts evolve towards structural stability, while continuing to lose non-coding DNA and genes that are not critical for symbiont function (Wernegreen 2015). The resulting small, gene-dense genomes are often stable in gene order and orientation, but experience rapid sequence evolution that is likely caused by the loss of recombination and DNA-repair machinery and sustained reductions in effective population size (Tamas et al. 2002; Woolfit and Bromham 2003; Sabree et al. 2010; McCutcheon and Moran 2012; Sloan and Moran 2012). The predicted destabilizing effects of accelerated substitution rates may be dampened by the high expression of protein chaperones like Hsp70 (Fares et al. 2002; McCutcheon et al. 2009a; Tokuriki and Tawfik 2009; Poliakov et al. 2011). The most gene-poor endosymbiont genomes have lost even seemingly essential genes, like those involved in genome replication and protein translation (Moran and Bennett 2014). In terms of genome size and coding capacity, these genomes span the gap between their less degenerate endosymbiotic cousins, which retain seemingly minimal sets of genes, and the bacterially derived organelles, which have lost most genes involved in replication, transcription, and translation (McCutcheon and Moran 2012; Moran and Bennett 2014). These extremely gene-poor endosymbiont genomes thus provide an opportunity to learn more about key adaptations enabling codependent symbioses, but in associations that are younger and still undergoing the integration process that the classic cellular organelles encountered billions of years ago.

Normal mitochondrial and plastid fuction requires extensive coordination between organelle and host genome (Timmis et al. 2004; Gray 2014). Most of the proteins present in organelles are encoded on the host genome, the products of which are imported into the organelle (Benz et al. 2009). Interestingly, the functioning of some bacterial endosymbionts in both amoeba (Nowack and Grossman 2012) and insects (Nikoh et al. 2010; Husnik et al. 2013; Sloan et al. 2014; Luan et al. 2015) also seem to be supported by horizontal gene transfer (HGT) to the host genomes, although protein import has been established in only two cases (Nowack and Grossman 2012; Nakabachi et al. 2014). In the insect examples, most transferred genes do not originate from the symbionts themselves, but from other unrelated bacteria (Nikoh et al. 2010; Husnik et al. 2013; Sloan et al. 2014; Luan et al. 2015). The taxonomic origin of these genes may not matter, however: just as in organelles, it is hypothesized that these HGT events enable the loss of complementary genes in the endosymbiont (Husnik et al. 2013; Gray 2014; Sloan et al. 2014; Bennett and Moran 2015).

Candidatus Hodgkinia cicadacola and *Candidatus* Sulcia muelleri (hereafter *Hodgkinia* and *Sulcia*) have two of the smallest bacterial genomes published (143kb and 277kb respectively) and are obligate nutritional endosymbionts of the cicada *Diceroprocta semicincta* (hereafter DICSEM) (McCutcheon et al. 2009a). Together these genomes encode complementary gene pathways to make the ten essential amino acids required by their cicada host (McCutcheon and Moran 2010). The *Hodgkinia* genome encodes only 10 of the 20 required aminoacyl tRNA synthetases (aaRSs), and 16 tRNA genes. Further, it is predicted to encode only three genes involved in tRNA maturation (*trmE*, *mmnA*, and *gidA*), all of which modify tRNAs at position 34. These gene loss patterns suggest that even if expressed, *Hodgkinia* tRNAs may lack the features necessary to be functional, such as correctly processed 5' and 3' ends. This pattern of dramatic tRNA and aaRS gene loss is extremely rare: only two bacterial species lack both tRNAs and aaRS on their genomes (Figure 1, Table 1). Because the genes encoding aaRSs and tRNA processing enzymes are large and typically highly conserved over evolutionary time, it is unlikely that these proteins were missed in the original annotation of *Sulcia* and *Hodgkinia*. In

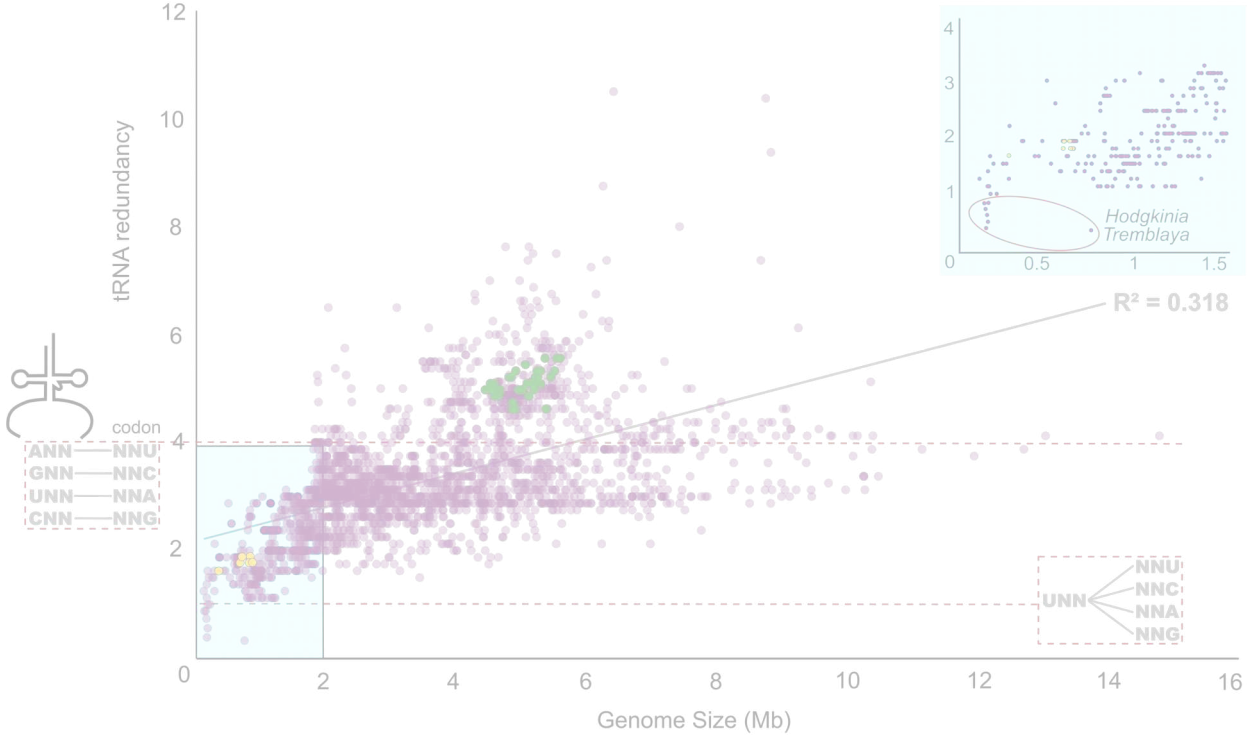


Figure 1. Genome size and tRNA redundancy are positively correlated. Each fully sequenced bacterial genome is shown as a dot (n=2761). tRNA redundancy represents the number of total 4-box tRNA genes in a genome over the number of 4-box families. The red-dashed line at y=1 shows a limit where only one tRNA is found from each of the eight 4-box families. Below this limit, it is unclear if the organism has enough tRNAs for translation. The red-dashed line at y=4 shows one tRNA gene for each 4-box codon. *Buchnera aphidicola* (*Buchnera*) and *Escherichia coli* (*E.coli*) are shown as yellow and green dots, respectively. Theoretically, all bacteria could function with redundancy value of 1.

contrast, the detection of tRNA genes in highly degraded genomes—in particular in mitochondrial genomes—is known to be difficult (Wolstenholme et al. 1987; Soma et al. 2007). Many mitochondrial tRNAs have unusual structures, in some cases missing entire D-loops, and can be missed by computational gene finders unless they are specifically trained to find them (Bruijn et al. 1980). Similarly, in the degenerate archaeal genome of *Nanoarchaeum*, tRNA prediction software initially missed its split and permuted tRNA genes (Randau and Söll 2008; Watanabe et al. 2014). It seemed quite possible therefore that the computational annotation of Hodgkinia's tRNAs may be incomplete.

4.2 tRNA gene content and codon usage in *Hodgkinia* and *Sulcia*

The number of tRNA genes encoded in bacterial genomes is variable, ranging between 30 and 167 with an average of 58 (Chan and Lowe 2009). Thirty tRNA genes are sufficient for translating all 61 possible codons, allowing for some tRNAs to pair with up to four different codon triplets (Andachi et al. 1989). Theoretical limits place the minimal number of tRNAs near 20 (Osawa et al. 1992; van der Gulik and Hoff 2011), and while several small bacterial genomes approach this limit, two bacterial species exceed it (Figure 1, Table 1). *Hodgkinia* is missing tRNA genes needed to decode leucine, valine, arginine, serine, threonine, aspartic acid, asparagine, and tyrosine codons (Figure 2). The mealybug endosymbiont *Candidatus* Tremblaya princeps (hereafter *Tremblaya*) also falls below the theoretical limit of 20, encoding only 8-12 tRNAs genes and 0 or 1 aaRSs, depending on strain (Table 1) (López-Madrigal et al. 2011; McCutcheon and von Dohlen 2011; Husnik et al. 2013). However, *Tremblaya* is unusual in hosting its own intrabacterial endosymbiont, *Ca.* Moranella endobia, which may provide the

Table 1. tRNA and aaRS genes in the smallest bacterial genomes. Shaded cells indicate that the aaRS gene in the left column is missing. [1]aaRS is heteromeric. [2]probable pseudogenes.

	DICSEM	TETULN	TETUND1	TETUND2	MAGTRE	TPPAVE	TPPLON	TPPCIT
valS		UAC		UAC		GAC	GAC(2)	[2]CAC
ileS	GAU	GAU		GAU	GAU	CAU,GAU	CAU	GAU
proS	UGG	UGG	UGG	UGG	UGG	UGG		
hisS	GUG	GUG		GUG		GUG	GUG	[2]GUG
trpS	UCA	UCA	UCA	UCA	UCA	CCA		CCA
metG	CAU(3)	CAU(3)	CAU(2)	CAU(2)	CAU	CAU(2)	CAU	CAU
gltX	UUC	UUC		UUG		UUC	UUC	
pheS[1]	GAA	GAA	GAA	GAA		GAA	GAA	GAA
pheT[1]								
alaS	UGC	UGC	[2]GGC,UGC(2)	UGC		UGC	ACG,CCU	UGC
glyS[1]	UCC,GCC(2)	UCC,GCC	UCC	UCC,GCC	UCC,GCC	GCC,UCC	GCC,UCC	
glyO[1]								
serS						CGA,GCU,GGA, UGA	ACU	UGA
asnS						GUU	GUU	
trvrS						GUA		
glnS	UUG	UUG	UUG	UUG	UUG	UUG		UUC
lysS	UUU	UUU		UUU	UUU	CUU,UUU	CUU(2)	CUU(3)
leuS						CAA,UAG	CAG	
argS						ACG,UCU		ACG
aspS		GUC	GUC	GUC		GUC	GUC	
thrS		UGU,GGU	UGU	UGU,GGU		CGU,GGU,UGU		
cysS	GCA	GCA	GCA	GCA	GCA	GCA	GCA	

missing tRNAs and aaRSs (Husnik et al. 2013). There is no such explanation for *Hodgkinia's* apparent lack of tRNA, aaRS, and tRNA processing genes.

However, the presence of a tRNA gene on a genome does not imply a functional tRNA molecule. Functional tRNAs are generated by a complex multistep process that can require trimming off transcribed nucleotides that precede (5' leader) and follow (3' trailer) the predicted tRNA gene, postranscriptional nucleotide editing at numerous positions, adding terminal CCA sequences when not encoded on the genome, and aminoacylation of the mature tRNA to produce a molecule that is active on the ribosome. After transcription, 5' leaders are trimmed by the near-universal ribozyme RNase P (Evans et al. 2006; Randau et al. 2008). The 3' trailer is cleaved off by a combination of endonucleases and/or exonucleases (Condon 2007) and if a terminal CCA is not encoded in the genome, one is added by a nucleotidyl transferase (Zhu and Deutscher 1987; Deutscher 1990). Finally, tRNA nucleosides are modified by a variety of enzymes

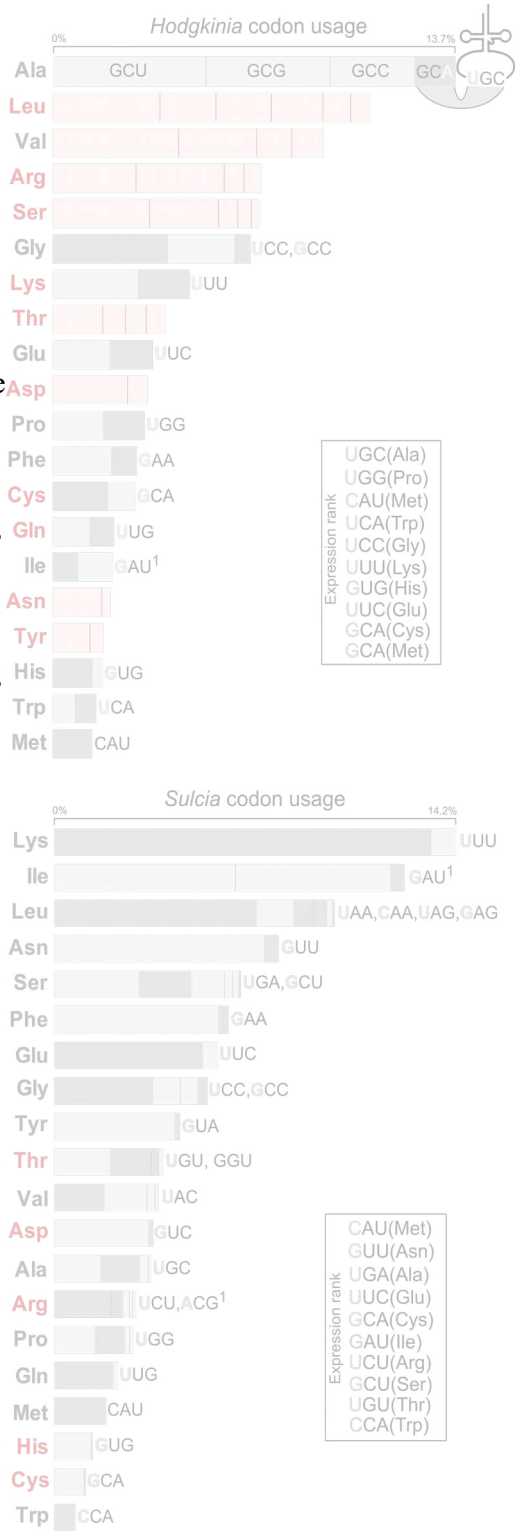


Figure 2. Codon usage and RNA expression in *Hodgkinia* and *Sulcia* genomes. Box size indicates codon frequency of all protein coding genes in *Hodgkinia* and *Sulcia*. Codons are grouped by amino acid; e.g. of the four alanine codons found in *Hodgkinia* protein coding genes, GCU is used most frequently. The nucleotide sequences for *Hodgkinia* alanine codons (which make up 13.7% of the genome) are shown as an example, all others are omitted for simplicity of display. The presence of a perfectly paired tDNA is indicated by a dark grey box. Light grey fill indicates that a tRNA could possibly be used to translate the codon by N34 wobble. The anticodon sequence of each tRNA is shown to the right of its cognate codons and is written 5' to 3'. N34 modifications that are likely needed for tRNA-codon pairing are indicated 1 . A red colored three letter amino acid abbreviation indicates that the genome does not encode that aaRS. tRNA abundance is shown in the “Expression rank” box.

at various conserved positions (Limbach et al. 1994; Söll and RajBhandary 1995; Jackman and Alfonzo 2013). These modifications greatly influence tRNA tertiary structure and how interactions with cellular enzymes and proteins (Jackman and Alfonzo 2013).

The annotated genome of *Hodgkinia* from *D. semicincta* lacks genes related to tRNA processing (McCutcheon et al. 2009a). It is missing the RNA (*rnpB*) and protein (*rnpA*) subunits of RNase P, and the nucleases implicated in 3' trimming. Despite only one tRNA possessing a genome-encoded terminal CCA, *Hodgkinia* does not encode a CCAing enzyme. The *Hodgkinia* DICSEM genome contains only three genes involved in tRNA editing (*trmE*, *mnmA*, and *gidA),* all of which are likely to be involved in the conversion of uridine to 5-methylaminomethyl-2-thiouridine at U34 (Dunin-Horkawicz et al. 2006; McCutcheon et al. 2009a). These patterns raise some obvious questions: Have some *Hodgkinia* tRNAs been missed by computational prediction software? For tRNAs present on the genomes, are their 5' and 3' ends correctly processed? Are *Hodgkinia* tRNAs modified only at the U34 wobble position? Do host aaRSs fill in for the missing symbiont genes?

Here we address these questions by performing RNA-seq on mRNAs and small RNAs from DICSEM bacteriome tissue. Our data confirm the expression of all annotated *Sulcia* tRNAs, most *Hodgkinia* tRNAs, and some mitochondrial tRNAs, but fail to identify any tRNA genes not previously annotated by computational methods. We find a highly expressed, but previously unannotated RNase P RNA in *Hodgkinia*, but the majority of the enzymes used to perform tRNA processing remain missing. Despite lacking these processing-related genes, *Hodgkinia* and *Sulcia* tRNAs undergo 3' trailer trimming, RNA modification, and CCA addition. Our data suggest that cicadas have not experienced successful HGT from bacteria, but reveal host expression patterns that might compensate for many of the missing symbiont activities.

4.3 tRNA expression in cicada bacteriomes

Small RNA-seq identifies only one unannotated symbiont tRNA.

We sequenced small RNAs in the cicada bacteriome, then searched for novel tRNAs that might complete *Hodgkinia's* set of tRNAs (Figure 1 and 2). It was immediately clear that our data were messy and complex. In principle, our library size-selection efforts would have produced sequencing reads entirely comprised of full-length tRNAs (~75 nts) that unambiguously belong to either *Sulcia*, *Hodgkinia*, or the cicada host. Instead, we found at least low-level expression from across the entire genome (Figure 3). We mapped quality trimmed reads to the *Hodgkinia* and *Sulcia* genomes then manually inspected regions with coverage greater than the lowest expressed, annotated tRNA gene and called polymorphic sites in tRNA genes. Defining a basal expression level removed background expression noise and enabled a more detailed review of highly expressed genes. In the case of *Hodgkina*, where some tRNAs are not expressed, we manually scanned the genome for coverage spikes that reached a maximum depth of at least 10X. While this approach allowed us to characterize tRNA processing and uncover expression of unannotated genes, it was not well suited for identifying spliced or otherwise unconventional RNAs, such as intron containing tRNAs. Therefore, we collapsed

identical reads of length 48-90nts and searched highly abundant transcripts for sequences ending in CCA, having predicted tRNA genes, or that partially align to *Sulcia* and *Hodgkinia* with BLAST. This approach allowed us to identify all RNA transcripts in the cell, including any unusual tRNAs. This approach also necessitated a coverage cutoff, as the number of unique, or nearly unique reads was very high. Therefore, generated a histogram of transcript coverage, and choose an arbitrary cutoff of 100X, because the distribution is nearly flat above this value.

In *Sulcia*, we find that the majority (>99%) of reads (map to tRNAs, tmRNA, RNase P, and ribosomal RNAs (Figure 3, supplementary table S1). However, even with so many reads mapping to RNA genes, the average read depth across protein coding genes (CDSs) was 380X. There were regions not in RNA genes that showed pronounced spikes in read coverage, including the 5' end of most CDSs, the 3' end of *menA*, *groL*, and *ilvC*, the middle of *sucB*, and an intergenic region from 203779-203842. None of the reads from these regions have a terminal CCA, nor an RNAfold structure that resembles a tRNA (Gruber et al. 2008). We found one de novo assembled transcript that encodes a predicted Thr^{GGT} tRNA that was unannotated in the current GenBank *Sulcia* file. This transcript contained 13nt missing from the published *Sulcia* genomic sequence. This gap was confirmed to be a missassembly of the original genome sequence by Sanger sequencing (the NCBI Reference Sequence NC_012123.1 was updated; see supplementary table S2 for primer sequences).

In *Hodgkinia*, we find high expression from predicted tRNAs, ribosomal RNAs, and the 5' ends of protein coding genes. We also found high expression from genome regions encoding the non-coding RNAs RNase P and tmRNA (discussed in the "Expression of unusual RNase P

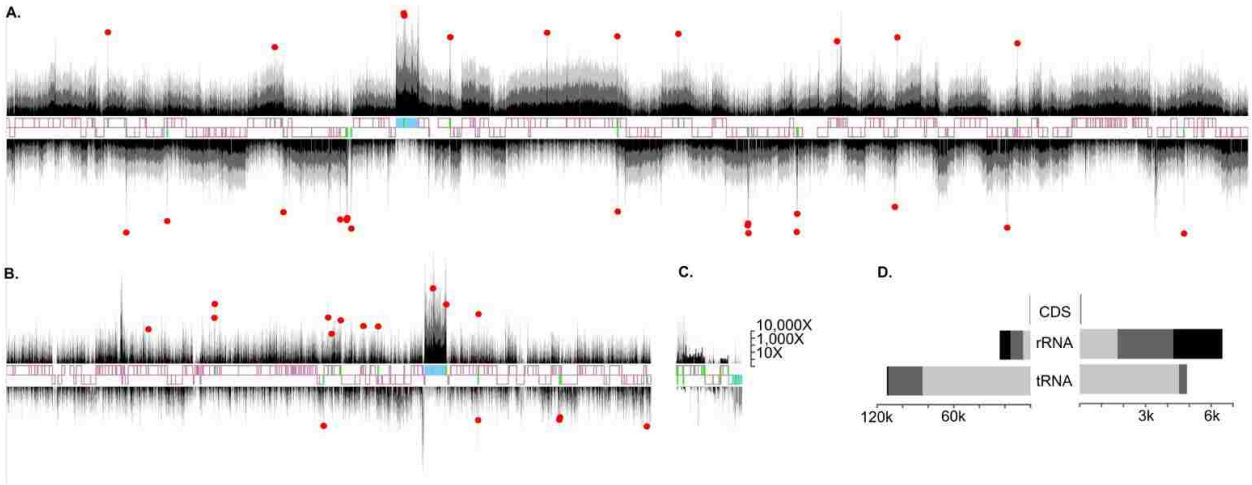


Figure 3. RNA expression patterns from the *Sulcia*, *Hodgkinia*, and mitochondrial genomes show relatively low expression of *Hodgkinia* tRNAs. Read depth plotted across the *Sulcia* (A), *Hodgkinia* (B) and mitochondrial (C) genomes. Protein coding, ribosomal RNA, and tRNA genes on the sense and anti-sense strands are shown in pink, blue, and green, respectively. Red dots show the highest read depth for each tDNA. Coverage depth for reads of length 18-47, 48-89, and 90-100 are shown in light grey, grey, and black, respectively and each are drawn on a log scale, then summed. (D) shows median coverage depths for *Sulcia* (left) and *Hodgkinia* (right) are shown for each gene category and read length in. The bars are colored as in A-C.

and permuted tmRNA" section below). There are no de novo assembled transcripts (min 10X identical reads) that BLAST to the *Hodgkinia* genome with an e-value less than 1E-25 aside from those of predicted RNA genes. Given these data, we conclude that *Hodgkinia* does not encode any tRNAs other than those previously annotated. Additionally, of *Hodgkinia's* sixteen total tRNA genes, many are not expressed at high levels (Figure 4, Supplementary table S3). The tRNA genes Gly061 and Gly108 each have no full-length reads aligning to them, even when allowing for 5-8 mismatches (Supplementary table S3). However, many shorter-than-full-length reads map to these genes, allowing us to predict modification sites.

Most tRNAs are found as tRNA halves.

The vast majority of reads mapping to tRNA genes were shorter than the gene itself (Figures 4-6, Supplementary table S3). These transcripts could be due to RNA degradation, PCR bias towards short amplicons during library creation, or from bona fide stable tRNA halves (Haiser et al. 2008; Thompson and Parker 2009; Jackowiak et al. 2011). Because reverse transcription occurs after RNA adapter ligation, these short reads are not likely due to reverse transcriptase failing to proceed through modified nucleotides. The presence of high levels of tRNA halves was corroborated by randomly selecting and Sanger sequencing a small RNA

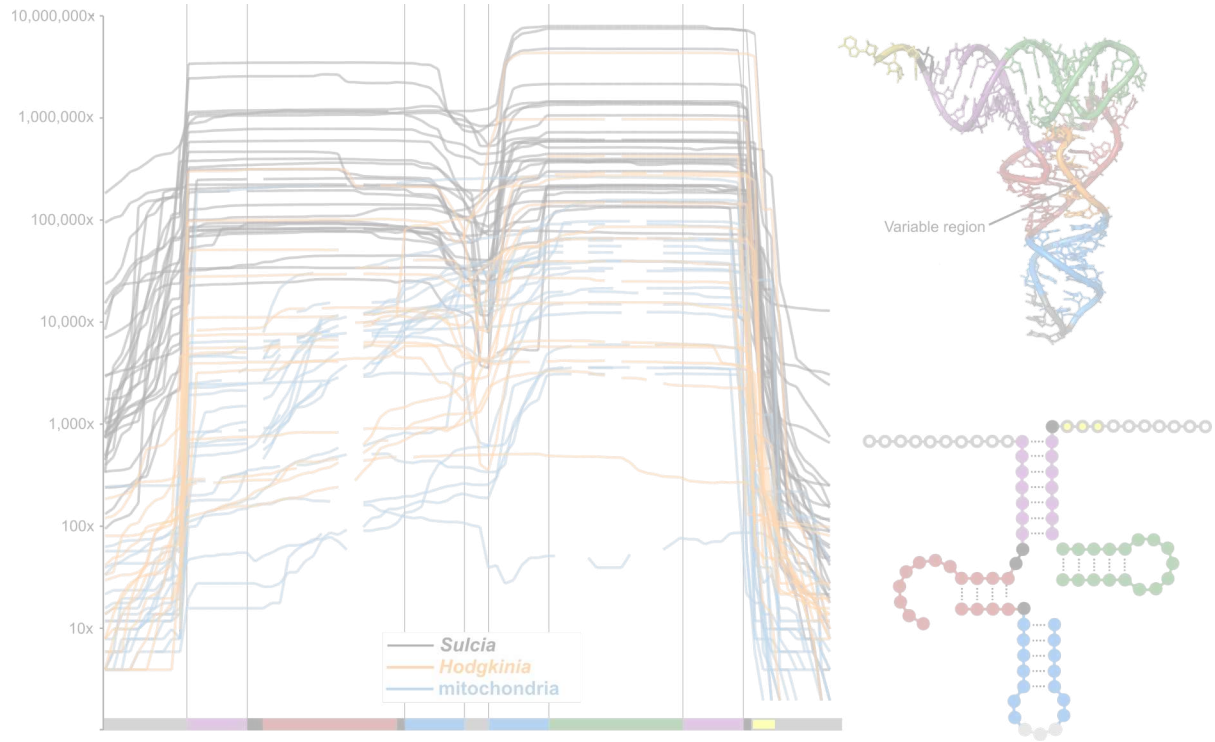


Figure 4. Dynamic range of tDNA half expression in *Sulcia* and *Hodgkinia*. Line graphs show read depth across each tDNA in *Hodgkinia* (orange), *Sulcia* (black), and the cicada mitochondria (blue). Data for only tRNA regions conserved in *Hodgkinia* and *Sulcia* are shown, mitochondrial tRNAs are often missing these regions, as indicated by gaps in the line graph. 18-100 nucleotide reads were mapped for this figure.

library prior to PCR amplification (n=9). Thus, it seems that a large proportion of either tRNA degradation products or stable tRNA halves are present in the cicada bacteriome. The majority of halves correspond to the 3' end of tRNA genes, where the break point is most often just 3' to the anticodon sequence (Figure 4). Figure 5 shows that there are exceptions to this pattern, especially in reads mapping to *Sulcia*, where many 5' halves are present.

Of 145,176,847 million quality filtered reads greater than 18nts in length, only 0.05% (74,651), and 1.7% (2,520,749) map to *Hodgkinia* and *Sulcia* tRNAs, respectively, and are long enough to be functional (Supplementary table S1). Sequencing coverage at *Hodgkinia* tRNAs is much lower than that of *Sulcia* and less even across tRNA genes. The range between highly expressed and lowly expressed *Sulcia* tRNAs is 100 fold less than the range between *Hodgkinia* tRNA genes (Supplementary table S3). This suggests that the lack of coverage for many *Hodgkinia* tRNAs is not due to under sequencing, but rather to a fundamental difference in transcriptional regulation between *Hodgkinia* and *Sulcia* (Figure 3). Consistent with these expression differences, endpoint RT-PCR on total bacteriome RNA using *Hodgkinia* tRNA specific primers shows a clear difference between a highly expressed and a lowly expressed tRNAs (Supplementary figure S1). Nearly equal numbers of small RNA reads map to the whole genome of *Sulcia* and *Hodgkinia* (224 reads/bp and 164 reads/bp, respectively), suggesting that a fundamental difference in expression patterns between *Hodgkinia* and *Sulcia* might explain the disparity in tRNA expression levels between the two. The equality in whole genome coverage is due to a large number of reads mapping to *Hodgkinia* 23S, 16S, and 5S genes (Figure 3).
In most bacteria, tRNA abundance corresponds well to codon usage (Novoa et al. 2012). In contrast, we find that highly expressed *Hodgkinia* and *Sulcia* tRNAs are rarely those corresponding to abundant codons (Figure 2). A lack of correlation between tRNA and codon abundance was also observed int the aphid endosymbiont *Buchnera* (Hansen and Moran 2012). Similarly, we find no pattern linking abundant tRNAs to those that have cognate aaRSs encoded in the genome (Figure 2).

4.4 Processing of *Hodgkinia* and *Sulcia* tRNAs

tRNA modification and maturation occurs in Hodgkinia and Sulcia.

The *Hodgkinia* genome encodes only three genes known to be involved in tRNA modification, all of which act on U34: *mnmA, gidA, and trmE* (McCutcheon et al. 2009a). MnmA catalyzes the 2-thiolation of U to s2U; GidA and TrmE form a dimer that catalyzes the conversion of s2U to nm5s2U (Dunin-Horkawicz et al. 2006). The *Sulcia* genome encodes these three genes, along with *truA* and *tilS* (McCutcheon et al. 2009a). TruA modifies U38-U40 to pseudouridine and TilS converts C34 to I34, enabling the specific recognition of Met versus Ile anticodons (Dunin-Horkawicz et al. 2006). However, we find sequence polymorphisms—which we interpret as base modifications (Iida et al. 2009; Findeiß et al. 2011; Hansen and Moran 2012)—at several sites other than the expected position 34 in *Hodgkinia* (1-4, 6, 7, 9, 15, 16, 18, 20, 23, 26, 27, 37, 43, 46, 49, 57, 58, 62, and 68) and the expected positions 34 and 38-40 in *Sulcia* (7, 26, 34, 37, and 58) (Figure 5, supplementary table S4). For a position to be called polymorphic, we required at least 10X read depth and greater than 2% polymorphism at the modified site. Interestingly, *Hodgkinia* tRNAs are more highly modified than *Sulcia* tRNAs in both the diversity of modification and in the total number of tRNAs modified. Of *Hodgkinia's* 16

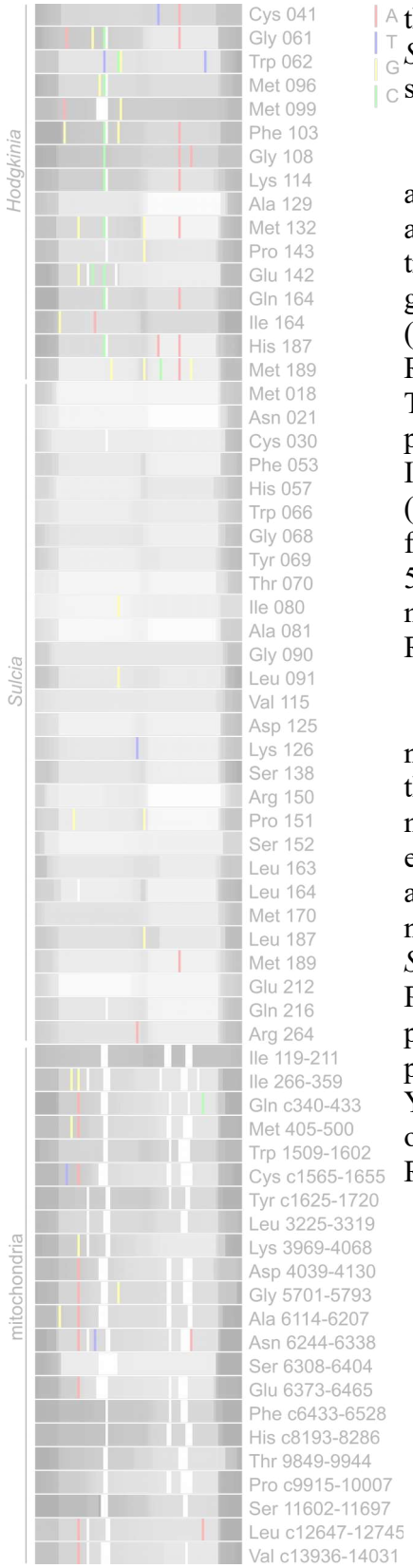


Figure 5. Expression level of individual tDNAs shown with polymorphic sites that have frequencies of greater than 2%. The per-base read depth was log transformed and is shown on a 0-255 color scale, making even large expression level differences difficult to distinguish by eye. Low expression is shown in black, high expression in white. Polymorphic sites are colored according to their genomic sequence. Ten bases of leader and trailer are shown as in Figure 5. Gaps are shown in white and are most apparent in mttDNAs. See supplementary table S2 for gene name descriptions.

tRNAs, 15 have at least one modified site, versus 8 of 28 in *Sulcia* and 10 of 22 in the mitochondrial genome (Figure 5, supplementary table S4).

In preparing an Illumina HiSeq compatible library, RNA adapter sequences were ligated directly to the small RNA pools at the 5' and 3' ends. Adapter ligation can be blocked by either a tri- or diphosphorylated 5' end, but a functional RNase P will generate 5' monophosphate ends which are active for ligation (Kazantsev and Pace 2006). By splitting the pool of small RNAs into two groups, one untreated, and one treated with Tobacco Acid Pyrophosphatase (TAP), we tested the 5' processed state of bacteriome tRNAs (Efstratiadis et al. 1977). In both *Hodgkinia* and *Sulcia*, we found no difference (Spearman's rank correlation, $P<0.005$) between the tRNA sets from each library (Supplementary table S5), suggesting that the 5' ends of tRNAs in the cicada bacteriome are monophosphorylated, consistent with the presence of an active RNase P enzyme.

Unprocessed tRNA transcripts typically include extra nucleotides on the 5' and 3' ends that are trimmed off during tRNA maturation (Söll and RajBhandary 1995). We find that many reads aligning to *Hodgkinia* and *Sulcia* tRNA genes extend past the predicted gene boundaries, suggesting that they are transcribed with 5' leaders and 3' trailers and that these extra nucleotides are quickly trimmed off (Figure 6). The 5' end of *Sulcia* tRNAs could be processed by the RNA moiety of RNase P that is present in the *Sulcia* genome. *Sulcia* also contains a putative ribonuclease (ACU52822.1) that could potentially process the 3' end, although the gene is most similar to RNase Y, which is involved in mRNA decay (Chen et al. 2013). The original *Hodgkina* genome annotation did not include any RNase P subunits or any nucleases, however, a putative *rnpB*

was found using a modified Infernal (Nawrocki et al. 2009) search (personal communication

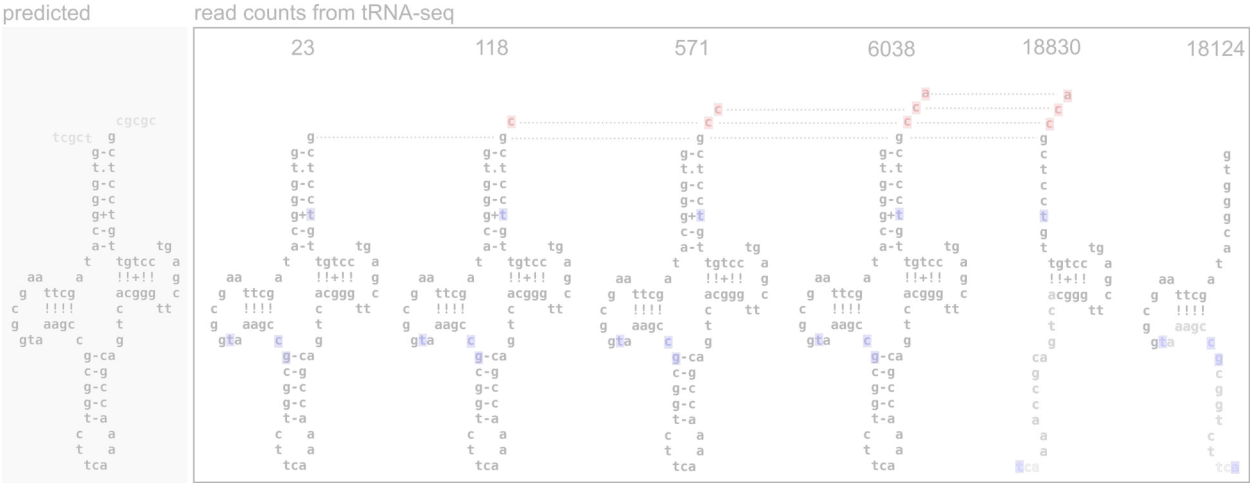


Figure 6. tRNA processing occurs in a stepwise manner, but full-length tRNAs comprise a small minority of the total reads. The majority of reads mapping to the *Hodgkinia* tDNA Trp062 gene (51,683) map to one of the secondary structures shown. Polymorphic sites (>2%) are shown in blue (RNA modifications) or red (CCA addition). tRNA halves are colored to indicate common sites of RNA degradation, where black letters indicate the highest read depth.

with Patricia Chan and Todd Lowe). We also observe reads ending in C, CC, and CCA that map to *Sulcia* and *Hodgkinia* tRNAs genes, indicating that each nucleotide of the terminal CCA is added one at a time to the 3' end of transcripts lacking 3' trailers (Figure 6). *Sulcia* contains a tRNA CCA nucleotidyl transferase, but *Hodgkinia* does not. Our mRNAseq data show upregulation of a cicada tRNA CCA nucleotidyl transferase (Chapter 5), however, we do not know if this enzyme is active on *Hodgkinia* tRNAs. In plants, mammals, and yeast, isoforms of this protein are localized to both the cytoplasm and organelle, and it can function in tRNA nuclear export and cytoplasmic tRNA quality control (Nagaike et al. 2001; Feng and Hopper 2002; Braun et al. 2007).

4.5 *Hodgkinia* RNase P and tmRNA

Discovery of unannotated RNase P and tmRNA genes in Hodgkinia.

By aligning small RNA reads to the *Hodgkinia* genome we found expression of previously unannotated RNase P (*rnpB*) and tmRNA (*ssrA*). Many reads map to the *Hodgkinia* genome (NC_012960.1) between 25448-25794 and 92713-93140, which correspond to *rnpB* and *ssrA*, respectively. Given that the 5' end of *Hodgkinia* tRNAs are processed and that we cannot find any other RNA nucleases in *Hodgkinia*, it seems likely that this RNase P is responsible for the observed tRNA processing. The permuted tmRNA is coded for in the reverse direction, on the anti-sense strand (Supplementary figure S2). All components typically conserved in tmRNA structure can be found in the proposed tmRNA gene, however the peptide tag does not end in the conserved YALAA sequence. The coding RNA and acceptor RNAs are separated by a 129nt

intervening sequence containing complementary sequences needed for folding, yet there are very few reads mapping to this region. There are mismatches that indicate CCA addition at the 3' end of both the coding and acceptor RNAs, thus the nucleotidyl transferase may not be highly specific to tRNAs. Also, we observe reads of varying length at the ends of the tmRNA gene, indicating that end-trimming probably occurs. While we might expect to see only the first 100nts of long transcripts in the data, the adapters can ligate to any RNA that has either a 2',3'-OH or a 2'-O-methyl,3'-OH at the 3' end and a monophosphate at the 5' end. Thus, the extent that breakdown products contribute to the sequencing data is unknown and may explain the presence of reads mapping to the entire 346bp of RNase P and the whole tmRNA gene.

4.6 Discussion

The effects of genome reduction on transcription and translation

Massive gene loss shapes the genomes of nutritional endosymbionts, and in most cases, results in gene complements that are intact enough for cellular, but not metabolic, automomy (McCutcheon and Moran 2012; Moran and Bennett 2014; Sloan et al. 2014). Endosymbiont genomes are able to loose genes that facilitate a free-living lifestyle because their hosts and/or co-symbionts support their newfound metabolic dependency (McCutcheon et al. 2009b; Hansen and Moran 2011; McCutcheon and von Dohlen 2011; Macdonald et al. 2012; Sloan and Moran 2012; Husnik et al. 2013; Nakabachi et al. 2014). The loss of genes essential for transcription, translation, and replication is rarer, but occurs in a few of the most gene-poor bacterial genomes. How these organisms compensate for the loss of these genes is unknown.

In this paper, we focus on the information processing systems of the *Hodgkinia* genome because so few bacterial genomes are missing genes in this category. However, the transcription, translation, and replication systems begin to show signs of disruption long before endosymbiont genomes become as reduced as *Hodgkinia*. During the initial period of genomic turmoil and subsequent settling, normal transcription is affected by the disruption of operons by rearrangement, altered codon usage patterns across the genome, and the loss of genes involved in regulating transcription (e.g. sigma factors). Although few studies investigate the impact that these changes have on transcription, evidence from the aphid symbiont *Buchnera aphidicola* (hereafter *Buchnera*) suggests that there is little apparent compensation for this disruption. Few *Buchnera* protein coding genes are differentially expressed across aphid life stages and in response to stress treatments (Wilcox et al. 2003; Wilson et al. 2006; Viñuelas et al. 2011).Also, mutations that disrupt tRNA basepair complementarity, combined with a reduction in box-family isoacceptors likely reduces the efficiency of translation in *Buchnera* (Hansen and Moran 2012). It is clear that compared to other bacteria, endosymbionts that have reduced genomes must manage with crippled protein expression systems. Alternative mechanisms for regulating gene expression, possibly with small antisense RNAs, may be important in endosymbionts (Hansen and Degnan 2014).

Some organisms have adapted to the loss of genes important for translation, and we

expected to find similar adaptations in *Hodgkinia*. For example, *Nanoarchaeum equitans* (an archea that has a reduced genome) does not contain RNase P and has eliminated its need by using equidistant promoters 5' to each tRNA gene (Randau et al. 2008). Several *Pyrobaculum* species (hyperthermophilic crenarchaeons with moderately reduced genomes) contain functional, but dramatically reduced RNase P genes (Lai et al. 2010). RNase P could not initially be found in *Hodgkinia*, but here we show expression of a putative *rnpB*. While still lacking the protein component (*rnpA*), trimming of the 5' leader can likely occur with only the RNA component (Guerrier-Takada et al. 1983). Cleavage of the 3' trailer from pre-tRNAs can be accomplished by a variety of redundant exo- and/or endonucleases (Condon 2007), none of which are encoded in *Sulcia, Hodgkinia,* or *Tremblaya* PCIT, but some combination of which are present in most organisms. In *E. coli,* RNase PH, RNase T, RNase D and RNase II can all trim back the 3' end of pre-tRNAs (Condon 2007). In *Sulcia*, a nuclease with similarity to RNase Y can be identified. Although RNase Y is typically though to initiate mRNA decay, it is implicated in multiple RNA processing tasks (Chen et al. 2013). After removal of excess nucleotides from the 3' end of a pre-tRNA, a terminal CCA must be added by a CCAing enzyme. All cellular genomes sequenced to date have this gene, except *Hodgkinia* and *Tremblaya*. However, in organisms with hard-coded CCAs, the gene can be deleted without major impacts to cell growth (Reuven et al. 1997). The CCAing enzyme gene can even be knocked out in organisms without hard-coded CCAs. These mutants often have growth defects, but are viable and tRNAs still get CCA'd by poly(A) polymerase 1 (Reuven et al. 1997). A few bacteria even require two enzymes for CCA addition, one adding the CC and one adding the terminal A (Tomita and Weiner 2001). Both genes share homology to the single CCAing enzyme that is present in most bacteria, however each contain mutations the presumably abolish their dual functionality (Neuenfeldt et al. 2008). We clearly see transcripts belonging to *Hodgkinia* and *Sulcia* that have a terminal C, CC, CCA, CCAC, CCACC, and CCACCA. The presence of these variants indicates an active CCAing enzyme and suggests that tRNA turnover occurs in *Hodgkinia* and *Sulcia*. Turnover is an important quality control mechanism that ensures correct folding structure of tRNAs (Wilusz et al. 2011). This function could potentially be performed by the mitochondrial CCAing enzyme that is upregulated in cicada bacteriome (Chapter 5). The mtCCAing enzyme is know to have broad specificity and functionality (Braun et al. 2007; Phizicky and Hopper 2010).

Base modifications are essential for tRNA aminoacylation and codon recognition, and have been well described in previous works (Söll and RajBhandary 1995). It is surprising that modifications are detected in *Hodgkinia* and *Sulcia* when the genes for these modification enzymes are not. On the other hand, we find polymorphism at sites that are commonly edited in many bacterial species. G37 edits are known to alter the specificity of codon-anticodon interaction and C20 edits affect tRNA secondary structure (Söll and RajBhandary 1995). Both *Hodgkinia* and *Sulcia* tRNAs are likely edited at G37. We find it very interesting that *Hodgkinia* and mitochondrial tRNAs share more tRNA modifications than do *Hodgkinia* and *Sulcia*. In general, however, reads mapped to *Sulcia* tRNAs have few polymorphic sites, despite *Sulcia* and *Hodgkinia* have very similar percent total mismatch across the entire genome (2.0% and 2.1%). The conservation of tRNA editing, even in intracellular symbionts, is evidence for their ubiquitous importance. It will be intriguing to determine the mechanism by which *Hodgkinia* tRNAs undergo base modifications.

The sets of retained tRNAs in *Hodgkinia* and *Tremblaya* overlap, but there are considerable differences (Table 1). Of 22 tRNA anticodon species in *Hodgkinia* DICSEM and *Tremblaya* PCIT, only trnA^{UGC}, trnI^{GAU}, trnM^{CAU}, and trnF^{UGC} are present in both genomes (Table 1). However, isoacceptor conservation between *Hodgkinia* DICSEM, mitochondria, and plastids is quite similar (Lohan and Wolfe 1998; Delannoy et al. 2011). Every *Hodgkinia* tRNA is highly conserved among organelles (Van Leuven et al. 2014; Campbell et al. 2015). *Tremblaya* PCIT, however, retains trnQ^{CUG} and trnK^{CUU} while *Hodgkinia* and organelles do not. A similar pattern of conserving core genes in the translational apparatus is seen with the ribosomal protein genes of mitochondrial and plastid genomes (Maier et al. 2013). Comparing the conservation of genes among reduced genomes (both bacterial and organellear) is a elegant way to learn about the functional importance of cellular processes. For example, endosymbionts tend to keep tRNA genes with the broadest codon recognition (5'-UNN) (Hansen and Moran 2012). Both *Sulcia* and *Hodgkinia* adhere to this pattern, with most tRNA isoacceptors belonging to 4-codon families having a uridine at the N34 position. Strangely, *Hodgkinia* seems to be entirely missing isoacceptors for five out of eight 4-codon families. Moreover, these codons are highly represented in the genome (Figure 2) and would be expected, if selection prevailed, to have highly expressed corresponding tRNA genes.

How to surpass the minimal microbial genome

Translation is a complex cellular process, requiring the coordination of ribosomal RNAs and proteins, tRNAs, aaRSs, initiation factors 1-3 (infA-C), elongation factors G and Ts (fusA, tsf), release factors 1 and 2 (prfA/B), ribosome recycling factor (frr), tmRNA (ssrA), RNase P (rnpA/B), and a handful of RNA editing enzymes. These genes are present in almost all organisms. However, like *Hodgkinia* and *Tremblaya*, many of these genes are missing from the genomes of organelles. The most gene-rich mitochondrial genomes of the Jakobid protists look very much like endosymbiont genomes, and contain a full set of about 30 tRNA genes (Burger et al. 2013). In contrast, the most gene-poor genomes of some trypanosomatides and alveolates contain no tRNA genes (Hancock and Hajduk 1990). The range is similar in plastids, from 1-30 tRNA genes (Barbrook et al. 2006; Bock 2007). Unlike in insect endosymbionts, organellar aaRSs have been completely transferred to the nuclear genome (Salinas et al. 2008; Alfonzo and Söll 2009; Gray 2012). Endosymbiont gene transfer (EGT) occurs at surprising frequency and magnitude (Martin 2003; Stegemann et al. 2003; Hotopp et al. 2007; Rice et al. 2013) and it is clear that the import of components (either host or symbiont derived) into the organelle is required for organelle function (Gray 2012). Even though the import of host components into symbiont cells has only been shown twice (Nowack and Grossman 2012; Nakabachi et al. 2014), our results suggest that something similar is happening in *Hodgkinia.*

The processes involved in tRNA import into organelles are complex (reviewed in Alfonzo and Söll 2009; Duchêne et al. 2009; Salinas-Giegé et al. 2015). For translation to occur in the organelle, fully processed and charged tRNAs must be localized to the organellar ribosome, and in no case are all the necessary components entirely encoded for in the organelle genome. The simplest hypothetical way to accomplish EGT-facilitated translation is for each essential gene to

be transferred to and expressed from the nuclear genome (1 copy of mitochondrial origin and 1 copy of chloroplast origin for plants). The components can then be targeted and imported into their respective organelle for translation (Duchêne et al. 2009). In all cases studied so far, the proteins/RNAs inside of organelles are mosaics of gene products of both eukaryotic and bacterial origin (Keeling and Palmer 2008; van Wijk and Baginsky 2011; Gray 2012). In human mitochondria, for example, almost all of the aaRSs are bacterially-derived; only two (Gly and Lys) are dually purposed, and only one tRNA (Gln) is known to be imported (Rubio et al. 2008; Suzuki et al. 2011). In contrast, of the 45 aaRS genes expressed from the *A. thaliana* genome (Iida et al. 2009), about half are localized to a single compartment; 21 are found only in the cytoplasm, 21 are dually-targeted, 2 are chloroplast specific, and 1 is targeted to all three cellular compartments. However, of the ~600 tRNA genes in the *A. thaliana* genome, only a couple have been shown to localize to the mitochondria (Duchêne and Maréchal-Drouard 2001). It is worth noting that the mitochondrial and plastid genomes of *A. thaliana* contain 22 and 30 tRNA genes, yet cytosolic tRNAs are still imported (Duchêne and Maréchal-Drouard 2001). The import of seemingly unnecessary tRNAs occurs quite frequenty, and in most cases, the role of redundant tRNAs in organelles is unknown (Salinas-Giegé et al. 2015). It is however, difficult in many cases to separate functional tRNA genes from tRNA pseudogenes, potentially overestimating the extent of redundancy. The low expression of some *Hodgkinia* tRNA genes and the high abundance of tRNA halves suggest the potential for ongoing psudogenization of *Hodgkinia* tRNA geness.

The evolution of promiscuous enzymes—or other multifuctional cellular components—could allow for genome reduction without the need for HGT or direct host supplementation. Endosymbionts with reduced genomes are often missing genes metabolic pathways, yet completely functional pathways are required for the symbiosis (Zientz et al. 2004; Macdonald et al. 2012; Husnik et al. 2013). In *Buchnera,* IlvC likely performs the function of successive genes in the vitamin B5 biosysnthesis pathway(Price and Wilson 2014). The expression of *Buchnera's* ilvC in *E. coli* can rescue *E. coli* panE- and ilvC- kncokout strains. It is hypothesized that this single enzyme is dually functioning in *Buchnera* cells, whereas two separate enzymes (IlvC and PanE) are required in *E. coli* (Price and Wilson 2014). Some aaRSs also act on multiple substrates. In the majority of prokaryotes, the noncognate aa-tRNA species Asp-tRNA^{Asn} and Glu-tRNA^{Gln} are formed by nondiscriminating aaRSs (Ibba and Söll 2004). The non-standard amino acids selenocysteine and pyrolysine are also incorporated by misaminoacylation (Ibba and Söll 2004). In these cases, it is unlikely that nondiscriminating aaRSs would aid in genome reduction because the noncognate aa-tRNAs are subsequently repaired by aminotransferases. These two lines of evidence perhaps suggest that *Hodgkinia* and *Sulcia* aaRSs could be broadly functioning, however, it is difficult to imagine how sloppy aminoacylation could happen, given the importance of maintaining fidelity in the translational system.

Despite the ancient nature and massive genetic integration of organelle with host (Maier et al. 2013; Ku et al. 2015), most mitochondria and plastids are partially autonomous. This suggests that there are challenges associated with complete host-symbiont integration. Gene retention patterns in the genomes of highly reduced bacterial symbionts also show a reluctancy to give up independence, especially for the processes of transcription, translation, and replication

(McCutcheon and Moran 2012; Moran and Bennett 2014). In organelles, these challenges are obviously overcome. Studying the cell and evolutionary biology of endosymbionts like *Tremblaya* and *Hodgkinia* provides insight on how host and symbiont become dependent and integrated with one another. In particular, the pattern of evolution that we observe suggest that the *Hodgkinia*-*Sulcia*-cicada symbiosis has slipped into a costly and irreversible path towards symbiont degradation (Bennett and Moran 2015). While symbioses initially promote adaptive resource utilization, the co-evolutionary dynamics between symbiont and host initiates a degeneration process that causes host-symbiont conflict and ends in either extinction or symbiont replacement (Bennett and Moran 2015). Similar conflicts occur between organelle and host. A clear example is seen in the drosophila *simw*^{501}*-OreR* hybrid, where single point mutations in the nuclearly-encoded, mitochondrially-derived tyrosyl-aaRS and its cognate, mitochondrially-encoded tRNA^{Tyr} interact to decrease the activity of the OXPHOS complexes I, III, and IV and cause growth defects (Meiklejohn et al. 2013). One interesting observation is that the most degenerate endosymbiont genomes are always in co-symbiosis with other bacteria (Moran and Bennett 2014). Perhaps one symbiont primes the system to enable massive genome reduction in the other symbiont. Very similar pre-adaptation processes may have enabled the establishment of mitochondria and the serial endosymbiosis of plastids (Larkum et al. 2007; Dorrell and Howe 2012; Gray 2014). Our data provide a little more insight into the process of host-symbiont integration. We suggest that the translational system in *Hodgkinia* is irrevocably broken, yet *Hodgkinia* proteins are still somehow made. Additionally, we show that the remaining parts seem to be functional, since processed tRNAs are present in the bacteriome. This work further supports the idea that obligate symbioses may undergo major transitions to become a single functional and co-evolving unit (Kiers and West 2015).

4.7 Methods and supplementary material

Method caveats.

We found several unexpected results while analyzing our data. First, we found highly abundant small RNAs containing predicted tRNAs that did not belong to *Hodgkinia*, *Sulcia*, or mtDNA tRNA genes. In these cases each half of the transcript aligned to separate genomic locations, or even the genomes of separate organisms (½ to *Sulcia* and ½ to *Hodgkinia*). In all cases, these were tRNA-like sequences that were joined near the anticodon. We could not amplify these RNAs from total RNA using gene-specific RT-PCR and thus concluded that they are a byproduct of the RNA ligation steps of the library preparation. This serves as a cautionary result of this method. In all cases, true *Hodgknina* and *Sulcia* tRNAs were also amplified, cloned, and sequenced as positive controls (see supplementary table S2 for primer sequences). Second, we found tRNA modification patterns that do not correlate with the functional capabilities of *Hodgkinia* and *Sulcia*, and reasoned that modified nucleosides could disrupt reverse transcriptase during library preparation **(Zheng et al. 2015)**. These cDNAs will not contain both primer binding sites (adapters) and will not be amplified during the PCR step of the library preparation, thereby selectively enriching for non-modified tRNAs. Since we find abundant tRNA sequences with polymorphism at conventionally modified sites, it seems likely that reverse-transcriptase can proceed over some modifications, consistent with previous findings (Ebhardt et al. 2009; Iida

et al. 2009; Findeiß et al. 2011; Hansen and Moran 2012; Cozen et al. 2015; Zheng et al. 2015). If this impacts our data significantly, we expect that i) tRNA abundance ranking is incorrect, ii) any tRNA with extremely low coverage (ie: *Hodgkinia* tRNAs) have modifications besides those that we describe here (Zheng et al. 2015). Regardless of any issues caused by library preparation, we can think of no alternative way to simultaneously i) assay the total tRNA pool, ii) determine if tRNA end processing occurs, iii) evaluate RNA editing, especially when starting with single bacteriome quanities of RNA

Sequencing small RNAs

The bacteriomes of three wild caught female *Diceroprocta semicincta* collected around Tucson, Arizona in July, 2010 and July, 2012 were dissected and stored in RNA-Later (Ambion). Total RNA was later purified using the Roche High Pure miRNA Isolation kit following the total RNA protocol. Small RNAs were isolated with the same kit, but following the 2-column protocol for <100nt RNAs. RNA-specific adapters were ligated to the 5' and 3' ends of the small RNAs using the Scriptminer™ Small RNA-Seq Library Preparation Kit from Epicenter. One index was treated with the supplied TAP enzyme to reduce the 5' end to a monophosphate. Reverse transcription was done with an adapter specific primer and each library was subjected to 15 rounds of PCR using FailSafe PCR Enzyme Mix (Epicenter) and the supplied primers (94°C for 15sec, 55°C for 5 sec, 65°C for 10sec). PCR bands of approximate size 50-300nt (including 113nt adapters) were cut from an 8% polyacrylamide gel after staining with SYBR® Safe (Invitrogen), and visualized on a standard UV transilluminator. The gel was shredded using a 0.5 mL tube with needle holes in the bottom, and eluted with 300 uL 0.5 M ammonium acetate for 3.5 hours at 37°C. The liquid was separated from gel particles using a 0.22 micron sterile filter and DNA was purified by standard isopropanol precipitation. Bioanalyzer traces of both libraries show DNA of about 100-275bp at sufficient concentration for Illumina sequencing. 226,712,931, 100nt single-end reads were generated on three HiSeq lanes at the UC Berkeley Vincent J. Coates Genomics Sequencing Laboratory.

Read processing for small RNAseq

Adapter sequences were trimmed using Cutadapt version 1.0 with options -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -g AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGACGATC -O 7 (Martin 2011). Then, reads less than 18nt in length were removed using a custom Perl-5.10.0 script. Reads were quality filtered using FASTX-Toolkit version 0.0.12 (http://hannonlab.cshl.edu/fastx_toolkit/) so that reads with a quality score less than 20 over more than 10% of the read were discarded (fastq_quality_filter -q 20 -p 90). Datasets with reads of length 18-90, 48-90, and 70-100nt were generated using a custom Perl script. The size of 18nt was chosen because the identical matches up to 16nt in length can be found between different symbiont tRNA genes. The size 48nt was chosen because the shortest tRNAs are about that length (Klimov and O'Connor 2009). At this point, each of these datasets were used for mapping to *Hodgkinia* and *Sulcia* genomes and tRNA genes using either bowtie-1.0.0, with settings –best –maqerr 150 –seedlen 18 or bwa-0.7.5 aln, with settings -n 0.08 -i 2 (Langmead et al. 2009; Li and Durbin 2009).

De novo RNA discovery

Identical reads from the 48-90nt dataset were compressed using FASTX-Toolkit (fastx-collapser). The majority of collapsed sequences were comprised of only one read, so a cutoff value was determined arbitrarily using a histogram of sequence coverage. The distribution of sequence coverage between 100X and 2E6X was quite even. The number of sequences with coverage from 100X-1X increases dramatically, so that there were 15,115 collapsed sequences with coverage higher than 100X and 6,478,420 collapsed sequences with coverage less than 100X. Therefore, all sequences comprised of less than 100 identical reads were discarded (6,463,305 sequences). The remaining 15,115 sequences were split into two sets: reads with BLASTN hits to *Hodgkinia* and *Sulcia* tRNA genes, and reads without hits (blastall 2.2.25, blastn -e 1E-25). Sequences that did not align with known, bacterial tRNAs were then aligned to the *Hodgkinia* and *Sulcia* full genome sequences (blastn -e 1E-10). The remaining sequences that did not align to the bacterial genomes were considered cicada sequences, and tRNAs were predicted using tRNAscan-SE 1.21 and ARAGORN 1.2.34 (Lowe and Eddy 1997; Laslett and Canback 2004). Nearly identical sequences were grouped into contigs using CAP3 (Huang and Madan 1999). Collapsed sequences with different anticodons, 5' leaders or 3' trailers that assembled together in CAP3 were separated into their own contigs for bowtie-0.12.7 and BWA-0.5.9 alignments using custom Perl scripts.

Comparing TAP treated to untreated libraries

Differential expression between libraries was compared using by expression rank changes and edgeR differential expression analysis. Reads from the 20-100nt and 70-100nt datasets were mapped to a multi-fasta file containing *Hodgkinia, Sulcia,* and mitochondrial tRNA genes plus 15bp of genome sequence flanking the gene using bowtie-0.12.7 with the -f, -S, and -n 3 options. tRNA abundance rankings were generated from the *.sam files by simply counting the number of reads that mapped to each tRNA sequence listed in Figure 3. The order of tRNA coverage was compared between indexes using Spearman-rank correlation (Supplementary table S5). Trinity v20140717 packages align_and_estimate_abundance.pl, abundance_estimates_to_matrix.pl, run_DE_analysis.pl, and analyze_diff_expr.pl scripts were used to compare differential transcription with parameters (--SS_lib_type F –est_method RSEM –aln_method bowtie –seedlen 18 –maqerr 150 –best). Using the de novo approach separately for library 1 and 2, with 48-100nt reads, we normalized tRNA coverage (number of reads per tRNA/total number of reads mapping to all tRNAs). A ratio of difference between library 1 and 2 coverage was calculated for each tRNA (library 1 tRNA normalized coverage/ library 2 tRNA normalized coverage). For all values less than zero, the inverse was taken and multiplied by -1. In this way, we tried to capture the relative difference in expression for all tRNAs from all organisms. These data were tabulated so that source organism, paired amino acid type, anticodon sequence, and relative expression change for every tRNA were in one row. In R, all non-numeric factors were changed using as.numeric(), a linear model was generated using lm(), and ANOVA was run using anova(). The results of this analysis are shown in supplementary table S5.

Cloning and sequencing of prepared libraries and tRNAs

Cloning was done using Invitrogen's TOPO TA cloning kit with OneShot TOP10 chemically competent cells using standard procedures. Primers designed to be specific to the tRNA of interest were used to prime reverse transcription using Invitrogen's SuperScript III First-Strand Synthesis kit. NEB OneTaq was used in end-point PCR prior to cloning (standard reaction with 2uL RT product and 40 cycles). Promega PCR ladder and NEB 6X loading dye was used to visualize PCR products prior to cloning. Plasmids were purified using Omega's Plasmid Mini Kit and sequencing was done with the standard M13F primer.

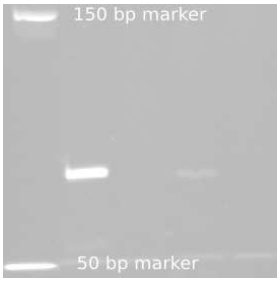
Bioinformatics

Complete bacterial genome sequences were downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz. Chromosomal sequences were searched for tRNA genes using tRNAscan-SE 1.21 using the bacterial model (Lowe and Eddy 1997). Genomic GC contents and 4-box family tRNA gene counts were calculated with custom PERL scripts. 6-box families were included in the analysis. tRNA redundancy is simply calculated by dividing the number of 4-box family tRNA genes by the number of 4-box families.

Acknowledgements

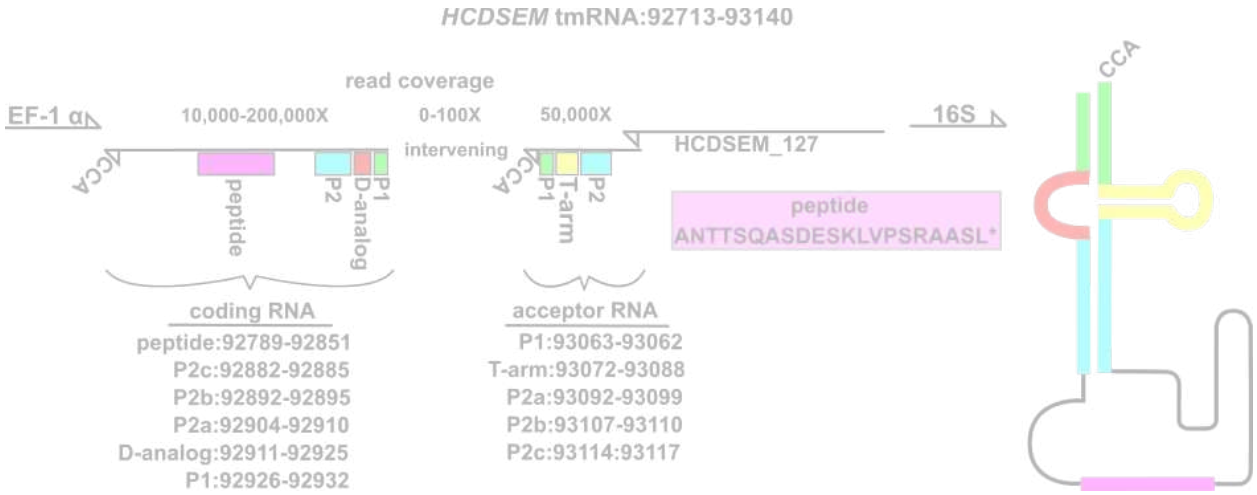
The authors thank the McCutcheon lab members, Juan Alfanzo, Todd Lowe, and Patricia Chan for technical assistance and conceptual feedback, Filip Husník for contributions to Table 1, and Bodil Cass for cicada collection. This work was supported by the National Science Foundation (NSF IOS-1256680). Raw Illumina reads are available in NCBI's SRA database for the small RNA (XXXXXXX) and transcriptome (SRR952383) datasets. Assembled transcript sequences are available from the NCBI TSA database (XXXXXX).

Supplementary figure 1. To show a difference between highly expressed and lowly expressed transcripts, 40 cycles of RT-PCR was done on total bacteriome RNA using primers specific for *Hodgkinia* $tRNA^{Ala}$ and *Hodgkinia* $tRNA^{Cys}$. Lanes 1-5: DNA marker, $tRNA^{Ala}$ primers, $tRNA^{Ala}$ primers no RT control, $tRNA^{Cys}$ primers, $tRNA^{Cys}$ primers no RT control.



Supplementary figure 2. Proposed tmRNA gene in *Hodgkinia* lies between genes for EF-1 alpha and 16S rRNA. The direction of transcription is indicated by an arrow. EF-1 alpha and 16S are encoded on the sense strand. The tmRNA and *Hodgkinia*_127 are encoded on the anti-sense

strand. Two small RNA transcripts with high coverage were identified as shown by separate arrows. Read depth across the tmRNA gene varies from 0-50,000X. Coordinates of tmRNA features are shown for the coding and acceptor RNAs.



Supplementary table 1. Number of reads in the dataset.

	Index1 (TAP)	Index2 (untreated)	Index3 (untreated)	Index4 (untreated)
Raw	77,189,680	19,096,461	47,564,122	82,862,668
Quality/length filtered	60,627,486	12,657,156	38,406,203	33,486,002
18-90nts	48,521,525	10,914,315	28,597,483	24,618,283
Mapped to *Hodgkinia*	7,484,021	1,565,097	7,677,511	5,903,119
Mapped to *Sulcia*	21,424,135	5,094,798	9,313,149	5,360,814
Mapped to mitochondria	261,693	83,062	57,994	77,055
Mapped to *Hodgkinia* tRNAs	2,545,941	635,453	101,749	386,447
Mapped to *Sulcia* tRNAs	15,582,990	3,489,686	1,732,283	2,646,973
Mapped to mitochondrial tRNAs	338,590	104,637	60,532	120,499
48-90nts	13,855,233	3,713,578	21,300,447	13,163,499
Mapped to *Hodgkinia*	3,706,879	24,261	7,151,011	4,812,614
Mapped to *Sulcia*	4,127,670	1,067,020	6,835,654	2,087,571
Mapped to mitochondria	81,003	24,261	47,331	38,985
Mapped to *Hodgkinia* tRNAs	13,254	4,277	17,712	39,408
Mapped to *Sulcia* tRNAs	885,324	229,957	691,271	714,197
Mapped to mitochondrial tRNAs	57,593	18,308	47,960	51,929
70-100nts	17,917,568	3,862,380	23,160,979	14,622,761
Mapped to *Hodgkinia*	3,263,929	554,659	6,775,227	4,046,344
Mapped to *Sulcia*	9,572,979	1,591,156	10,075,795	6,042,490
Mapped to mitochondria	26,122	4,996	6,320	3,718

Mapped to *Hodgkinia* tRNAs	5,644	1,767	8,209	8,723
Mapped to *Sulcia* tRNAs	520,386	144,229	617,538	659,736
Mapped to mitochondrial tRNAs	674	128	321	558

Supplementary table 2. Primer sequences to amplify tRNAs from total RNA, genomic DNA, and finished library preparations.

	Forward primer 5' to 3'	**Reverse primer 5' to 3'**
Ala_129_*Hodgkinia*	GGGGCTGTAGCTCAATTGG	TGGAGCTAAGCGGACTCG
Cys_041_*Hodgkinia*	GGCTTCGTGGTATAGGGGT	GGCTTCGCTCAGACTCG
Thr_*Sulcia*_flanking	CCTGGACAATCTACATGAGCA	GGTAGAGCATCAGCCTTCCA
Split_tRNA_1	AGAGTTGCCGGAGGGGTTAAC	TGGAGAATATCGGATTTGAACCG
Split_tRNA_2	TATGGCAATAACCAAG	TGGAGAATATCGGATTTGAACCG
Split_tRNA_3	GGTGGAGCAGTTGGTAGC	AGCTAAGCGGACTCGAACCGC
Split_tRNA_4	GGTGAACGTAGCTCAATTGG	TGGAGCTAAGCGGACTCG
Split_tRNA_5	GGATGTAGCGTAGGTTGG	CGGTACCGGGAATCGAACC
Split_tRNA_6	CGCGGGGTGGAGCAGTTGG	CAACGGGGGCAGGAGTCG

Supplementary table 3. Number of reads mapping to each tRNA gene (plus 15bp flanking sequence) using bowtie. The 18-90 SAM file was parsed for reads that map to the tRNA with nearly the perfect length and ending in CCA.

	18-90	48-90	70-100	tRNA count
SMDSEM_264_Arg	171822	5010	4216	3522
SMDSEM_216_Gln	569993	13458	12207	291
SMDSEM_212_Glu	2345816	170866	163305	120854
SMDSEM_189_Met	390387	33319	29288	12978
SMDSEM_187_Leu	80814	1747	1746	542
SMDSEM_170_Met	110296	16461	15727	8919
SMDSEM_164_Leu	499915	54916	679	219
SMDSEM_163_Leu	142900	14619	14306	2513
SMDSEM_152_Ser	602352	96386	156066	35534
SMDSEM_151_Pro	1192993	2460	1995	1459
SMDSEM_150_Arg	4004740	56468	54731	36894
SMDSEM_138_Ser	232144	122460	4854	2366
SMDSEM_126_Lys	220987	16320	14855	6793
SMDSEM_125_Asp	557210	56272	51479	40892
SMDSEM_115_Val	92904	52027	50574	30918

SMDSEM_091_Leu	127084	28779	22692	17190
SMDSEM_090_Gly	50342	32083	29976	4065
SMDSEM_081_Ala	3605265	263269	122850	91939
SMDSEM_080_Ile	645895	159599	147330	95551
SMDSEM_070_Thr	1204147	175998	132236	68762
SMDSEM_069_Tyr	297423	21774	4172	1938
SMDSEM_068_Gly	249994	24795	22514	17837
SMDSEM_066_Trp	226724	38543	36839	30700
SMDSEM_057_His	272166	11285	9114	7518
SMDSEM_053_Phe	374054	2042	1844	1203
SMDSEM_030_Cys	219277	97275	96060	70828
SMDSEM_021_Asn	4102498	540031	335843	178660
SMDSEM_018_Met	861790	412487	404391	305158
HCDSEM_189_Met	306977	2522	1118	831
HCDSEM_187_His	220501	1654	373	250
HCDSEM_164_Ile	10009	65	8	3
HCDSEM_163_Gln	15973	103	48	30
HCDSEM_143_Pro	52512	6800	5805	9
HCDSEM_142_Glu	141333	37552	981	946
HCDSEM_132_Met	498517	330	108	78
HCDSEM_129_Ala	2254846	11053	7459	5591
HCDSEM_114_Lys	20565	112	58	13
HCDSEM_108_Gly	1859	60	3	0
HCDSEM_103_Phe	3561	156	4	3
HCDSEM_099_Gly	379	171	138	8
HCDSEM_096_Met	11795	1107	83	74
HCDSEM_062_Trp	51683	11390	7899	6754
HCDSEM_061_Gly	75091	388	1	0
HCDSEM_041_Cys	3989	1188	257	168
DICSEMmt_Val_c(13936..14031)	51028	6575	2	295
DICSEMmt_Tyr_c(1625..1720)	3553	107	10	5
DICSEMmt_Trp_1509..1602	6858	75	4	28
DICSEMmt_Thr_9849..9944	17606	692	51	9
DICSEMmt_Ser_6308..6404	142889	114076	108	75887
DICSEMmt_Ser_11602..11697	6822	23	41	1
DICSEMmt_Pro_c(9915..10007)	37229	6479	117	55
DICSEMmt_Phe_c(6433..6528)	2035	8	0	1
DICSEMmt_Met_405..500	29118	1469	3	160
DICSEMmt_Lys_3969..4068	80942	3324	129	83
DICSEMmt_Leu_c(12647..12745)	23661	4271	788	645
DICSEMmt_Leu_3225..3319	15019	421	8	75
DICSEMmt_Ile_266..359	36577	2388	4	162

DICSEMmt_Ile_119..211	476	14	0	0
DICSEMmt_His_c(8193..8286)	1948	13	0	4
DICSEMmt_Gly_5701..5793	8489	3501	85	351
DICSEMmt_Glu_6373..6465	50668	17779	7	61
DICSEMmt_Gln_c(340..433)	34924	2491	323	479
DICSEMmt_Cys_c(1565..1655)	25463	3374	0	59
DICSEMmt_Asp_4039..4130	20488	1068	0	32
DICSEMmt_Asn_6244..6338	17953	3333	0	1370
DICSEMmt_Ala_6114..6207	10512	4309	1	396

Supplementary table 4. tRNA modifications sorted by site. Those shown are at 2% or greater in frequency. The number of reads matching each of the four nucletides is shown. The genome sequence at that position is greyed. [a]Edit occurs on mismatched base-pair in stem region, [b]tRNA secondary structure suggests that the gene is pseudogenized, [c]tDNA with high nucleotide similarity exists in nuclear genome. 48-90 nucleotide reads used in mapping.

		A	**T**	**G**	**C**
N1	*Hodgkinia*_164	18[a]	0	10	0
	Mito_Ala_6114-6207	137	2	174	0
N2	*Hodgkinia*_103	0	0	0	8[a]
N3	*Hodgkinia*_099	0	15	0	112
N4	*Hodgkinia*_061	19	0	2	0
	Mito_Cys_1565-1655	0	67	0	48[a]
N6	Mito_Ile_266-359	281	0	90	2
	Mito_Met_405-500	39	0	279	1
N7	*Sulcia*_151	64	4	2072	5
N9	*Hodgkinia*_132	3	16	105	3
	*Hodgkinia*_142	7	12	1023	15
	Mito_Ile_266-359	7	12	571	16
	Mito_Gln_340-433	210	77	38	0
	Mito_Met_405-500	323	11	3	1
	Mito_Cys_1565-1655	96	23	14	0
	Mito_Lys_3969-4068	4	26	108	20
	Mito_Asp_4039-4130	55	12	18	0
	Mito_Gly_5701-5793	656	190	144	1
	Mito_Ala_6114-6207	515	147	29	0

	Mito_Asn_6244-6338	618	622	1560	28
	Mito_Glu_6373_6465[c]	310	60	72	0
	Mito_Leu_12647-12745	610	268	113	4
	Mito_Val_13936-14031[c]	314	179	263	0
N15	*Hodgkinia*_061	22	0	6	0
	*Hodgkinia*_142	139	0	1122	0
N16	*Hodgkinia*_164	20	0	27	0
	Mito_Asn_6244-6338	0	1698	1445	0
N18	*Hodgkinia*_096	3	0	104	0
N20	*Hodgkinia*_061	0	10	0	28
	*Hodgkinia*_062	108	7813	4	43
	*Hodgkinia*_096	11	67	0	177
	*Hodgkinia*_103	0	3	0	51
	*Hodgkinia*_108	0	18	0	4
	*Hodgkinia*_114	1	7	0	50
	*Hodgkinia*_132	7	85	1	57
	*Hodgkinia*_142	97	1258	3	70
	*Hodgkinia*_163	2	51	0	18
	*Hodgkinia*_187	18	254	1	514
N23	*Hodgkinia*_189	478[a]	8	1824	0
N26	*Hodgkinia*_062	190	3	1	8394
	*Hodgkinia*_103	86	0	38	0
	*Sulcia*_080	102	2873	147868	29
	*Sulcia*_091	22	471	17805	6
	*Sulcia*_138	2	125	4770	1
	Mito_Gly_5701-5793	1257	0	2172	0
N27	*Hodgkinia*_062	12	224	7915	3
	*Hodgkinia*_099	0	27	0	116
N34	*Sulcia*_126	4	15802	7	407
	*Sulcia*_264	4546	2	380	0
N37	*Hodgkinia*_189	8	14	1403	981
	*Hodgkinia*_132	2	21	287	5
	*Hodgkinia*_143	63	243	6143	122
	*Sulcia*_151	38	89	2137	91
	Sulcia_187	4	15	1290	12
N43	*Hodgkinia*_041	0	442	0	371[a]

	*Hodgkinia*_187	28	1	1569	1
N45	*Hodgkinia*_061	91	1	264	0
	*Hodgkinia*_189	470	12	1902	1
	Mito_Asp_4039-4130	237	0	644	0
N46	*Hodgkinia*_187	27	1	1571	0
N49	*Hodgkinia*_189	12	468	0	1899
N57	*Hodgkinia*_132	2563	3	3922	0
	Mito_Gln_340-433	1488	0	246	0
N58	*Hodgkinia*_041	424	148	179	0
	*Hodgkinia*_061	188	26	123	1
	*Hodgkinia*_103	103	8	19	0
	*Hodgkinia*_108	33	1	5	0
	*Hodgkinia*_114	77	1	10	0
	*Hodgkinia*_163	40	26	30	0
	*Hodgkinia*_187	1501	22	29	0
	*Hodgkinia*_189	2158	42	17	0
	*Sulcia*_189	29461	671	17	2
N61	Mito_Asn_6244-6338	1811	1	1362	0
N62	*Hodgkinia*_189[b]	21[a]	0	115	0
	*Hodgkinia*_108	36	0	2	0
N67	Mito_Gln_340-433	0	222	0	1479
	Mito_Leu_12647-12745	1772	2085	2	2
N68	*Hodgkinia*_062	2	842	1	7800[a]
T-loop	Mito_Met_405-500	580	570	0	0
	Mito_Cys_1565-1655	0	267	0	1305
	Mito_Cys_1565-1655	1303	266	2	0
	Mito_Asp_4039-4130	643	0	0	242
D-loop	Mito_Val_13936-14031[c]	0	5108	1	588

Supplementary table 5. No difference found between TAP treated and untreated libraries by Spearman's rank correlation and ANOVA, indicating that the 5' of *Hodgkinia* and *Sulcia* tRNAs are properly processed. Spearman's rank shows significant correlation between tRNA expression of TAP treated and untreated samples. ANOVA shows no significant difference between tRNA relative abundance between treated and untreated samples. Relative abundance is the number of reads corresponding to each tRNA over the total number of reads assigned to all tRNAs in the sample. Categories for “Organism” include *Sulcia*, *Hodgkinia*, mitochondrial (DSEM), and other (unidentified). Significant F values for the ANOVA are F(0.01) = 3.14, F(0.05) = 4.95.

Spearman's Rank

	Sulcia n (p=0.005 critical value)	*Hodgkinia* n (p=0.005 critical value)	Rho value *Sulcia*	Rho value *Hodgkinia*
20-100nts	28 (0.496)	16 (0.666)	**0.932**	**0.988**
70-100nts	28 (0.496)	16 (0.666)	**0.943**	**0.962**

ANOVA

	Df	Sum of squares	Mean of squares	F value	Pr (>F)
Organism	4	8.17	2.04	0.4586	0.7658
Amino acid	45	193.69	4.30	0.9665	0.5423
Anticodon	16	59.13	3.70	0.8298	0.6482
Organism:Amino acid	3	4.51	1.50	0.3374	0.7984
Organism:Anticodon	35	208.40	5.95	1.3371	0.1529
Anticodon:Amino acid	3	5.61	1.87	0.4197	0.7394

Supplementary table 6. Number of differentially expressed tRNA genes encoded on the *Hodgkinia*, *Sulcia,* and the cicada mitochondria genomes by Edger analysis (66 total genes). The analysis was performed for all four small RNA samples and for three read size ranges. Index1: TAP treated 2010 sample, index2: 2010 sample, index3: 2012 sample, index4: 2012 sample.

	70-100 nt			
index1	0			
index2	0	0		
index3	9	10	0	
index4	8	5	3	0

	48-90 nt			
index1	0			
index2	0	0		
index3	18	16	0	
index4	11	9	4	0

	18-90 nt			
index1	0			
index2	0	0		
index3	20	20	0	
index4	11	11	3	0

4.8 References

Alfonzo JD, Söll D. 2009. Mitochondrial tRNA import – the challenge to understand has just

begun. Biol Chem 390:717–722.

Andachi Y, Yamao F, Muto A, Osawa S. 1989. Codon recognition patterns as deduced from sequences of the complete set of transfer RNA species in Mycoplasma capricolum: Resemblance to mitochondria. J Mol Biol 209:37–54.

Barbrook AC, Santucci N, Plenderleith LJ, Hiller RG, Howe CJ. 2006. Comparative analysis of dinoflagellate chloroplast genomes reveals rRNA and tRNA genes. BMC Genomics 7:297.

Bennett GM, Moran NA. 2015. Heritable symbiosis: The advantages and perils of an evolutionary rabbit hole. Proc Natl Acad Sci U S A:201421388.

Benz JP, Soll J, Bölter B. 2009. Protein transport in organelles: The composition, function and regulation of the Tic complex in chloroplast protein import: Translocation across the outer chloroplast membrane. FEBS Journal 276:1166–1176.

Bock R. 2007. Structure, function, and inheritance of plastid genomes. In: Bock R, editor. Cell and Molecular Biology of Plastids. Topics in Current Genetics. Springer Berlin Heidelberg. p. 29–63.

Braun SS von, Sabetti A, Hanic-Joyce PJ, Gu J, Schleiff E, Joyce PBM. 2007. Dual targeting of the tRNA nucleotidyltransferase in plants: not just the signal. J Exp Bot 58:4083–4093.

Bruijn MHL de, Schreier PH, Eperon IC, Barrell BG, Chen EY, Armstrong PW, Wong JFH, Roe BA. 1980. A mammalian mitochondrial serine transfer RNA lacking the “dihydrouridine” loop and stem. Nucl Acids Res 8:5213–5222.

Burger G, Gray MW, Forget L, Lang BF. 2013. Strikingly Bacteria-Like and Gene-Rich Mitochondrial Genomes throughout Jakobid Protists. Genome Biol Evol 5:418–438.

Burke, Moran NA. 2011. Responses of the pea aphid transcriptome to infection by facultative symbionts. Insect Mol Biol 20:357–365.

Campbell MA, Van Leuven JT, Meister RC, Carey KM, Simon C, McCutcheon JP. 2015. Genome expansion via lineage splitting and genome reduction in the cicada endosymbiont Hodgkinia. Proc Natl Acad Sci U S A 112:10192–10199.

Chan PP, Lowe TM. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res 37:D93–D97.

Chen Z, Itzek A, Malke H, Ferretti JJ, Kreth J. 2013. Multiple Roles of RNase Y in Streptococcus pyogenes mRNA Processing and Degradation. J Bacteriol 195:2585–2594.

Condon C. 2007. Maturation and degradation of RNA in bacteria. Curr Opin Microbiol 10:271–278.

Cozen AE, Quartley E, Holmes AD, Hrabeta-Robinson E, Phizicky EM, Lowe TM. 2015. ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. Nat Meth 12:879–884.

Delannoy E, Fujii S, Francs-Small CC des, Brundrett M, Small I. 2011. Rampant Gene Loss in the Underground Orchid Rhizanthella gardneri Highlights Evolutionary Constraints on Plastid Genomes. Mol Biol Evol 28:2077–2086.

Deutscher MP. 1990. Ribonucleases, tRNA nucleotidyltransferase, and the 3' processing of tRNA. Prog Nucleic Acid Res Mol Biol 39:209–240.

Dorrell RG, Howe CJ. 2012. Functional remodeling of RNA processing in replacement chloroplasts by pathways retained from their predecessors. Proc Natl Acad Sci U S A 109:18879–18884.

Duchêne A-M, Maréchal-Drouard L. 2001. The Chloroplast-Derived trnW and trnM-e Genes Are Not Expressed in Arabidopsis Mitochondria. Biochem Biophysl Res Commun 285:1213–1216.

Duchêne A-M, Pujol C, Maréchal-Drouard L. 2009. Import of tRNAs and aminoacyl-tRNA synthetases into mitochondria. Curr Genet 55:1–18.

Dunin-Horkawicz S, Czerwoniec A, Gajda MJ, Feder M, Grosjean H, Bujnicki JM. 2006. MODOMICS: a database of RNA modification pathways. Nucleic Acids Res. 34:D145–D149.

Ebhardt HA, Tsang HH, Dai DC, Liu Y, Bostan B, Fahlman RP. 2009. Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. Nucl. Acids Res. 37:2461–2470.

Efstratiadis A, Vournakis JN, Donis-Keller H, Chaconas G, Dougall DK, Kafatos FC. 1977. End labeling of enzymatically decapped mRNA. Nucleic Acids Res 4:4165–4174.

Evans D, Marquez SM, Pace NR. 2006. RNase P: interface of the RNA and protein worlds. Trends Biochem. Sci. 31:333–341.

Fares MA, Ruiz-González MX, Moya A, Elena SF, Barrio E. 2002. Endosymbiotic bacteria: GroEL buffers against deleterious mutations. Nature 417:398–398.

Feng W, Hopper AK. 2002. A Los1p-independent pathway for nuclear export of intronless tRNAs in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 99:5412–5417.

Findeiß S, Langenberger D, Stadler PF, Hoffmann S. 2011. Traces of post-transcriptional RNA modifications in deep sequencing data. Biol. Chem. 392:305–313.

Gray MW. 2012. Mitochondrial Evolution. Cold Spring Harb Perspect Biol 4:a011403.

Gray MW. 2014. The Pre-Endosymbiont Hypothesis: A New Perspective on the Origin and Evolution of Mitochondria. Cold Spring Harb Perspect Biol 6:a016097.

Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL. 2008. The Vienna RNA Websuite. Nuceic Acids Res. 36:W70–W74.

Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. 1983. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell 35:849–857.

van der Gulik P, Hoff W. 2011. Unassigned Codons, Nonsense Suppression, and Anticodon Modifications in the Evolution of the Genetic Code. J. Mol. Evol. 73:59–69.

Haiser HJ, Karginov FV, Hannon GJ, Elliot MA. 2008. Developmentally regulated cleavage of tRNAs in the bacterium Streptomyces coelicolor. Nucleic Acids Res 36:732–741.

Hancock K, Hajduk SL. 1990. The mitochondrial tRNAs of Trypanosoma brucei are nuclear encoded. J Biol Chem 265:19208–19215.

Hansen AK, Degnan PH. 2014. Widespread expression of conserved small RNAs in small symbiont genomes. ISME J 8:2490–2502.

Hansen AK, Moran NA. 2011. Aphid genome expression reveals host–symbiont cooperation in the production of amino acids. Proc Natl Acad Sci U S A 108:2849–2854.

Hansen AK, Moran NA. 2012. Altered tRNA characteristics and 3′ maturation in bacterial symbionts with reduced genomes. Nucleic Acids Res. 40:7870–7884.

Hotopp JCD, Clark ME, Oliveira DCSG, Foster JM, Fischer P, Torres MCM, Giebel JD, Kumar N, Ishmael N, Wang S, et al. 2007. Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes. Science 317:1753–1756.

Huang X, Madan A. 1999. CAP3: A DNA Sequence Assembly Program. Genome Res. 9:868–877.

Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, Tanaka M, Satoh N, Bachtrog D, Wilson ACC, et al. 2013. Horizontal Gene Transfer from Diverse Bacteria to an Insect Genome Enables a Tripartite Nested Mealybug Symbiosis. Cell 153:1567–1578.

Ibba M, Söll D. 2004. Aminoacyl-tRNAs: setting the limits of the genetic code. Genes Dev. 18:731–738.

Iida K, Jin H, Zhu J-K. 2009. Bioinformatics analysis suggests base modifications of tRNAs and miRNAs in Arabidopsis thaliana. BMC Genomics 10:155.

Jackman JE, Alfonzo JD. 2013. Transfer RNA modifications: nature's combinatorial chemistry playground. Wiley Interdiscip Rev RNA 4:35–48.

Jackowiak P, Nowacka M, Strozycki PM, Figlerowicz M. 2011. RNA degradome—its biogenesis and functions. Nucleic Acids Research 39:7361–7370.

Kazantsev AV, Pace NR. 2006. Bacterial RNase P: a new view of an ancient enzyme. Nat Rev Micro 4:729–740.

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9:605–618.

Kiers ET, West SA. 2015. Evolving new organisms via symbiosis. Science 348:392–394.

Klimov PB, O'Connor BM. 2009. Improved tRNA prediction in the American house dust mite reveals widespread occurrence of extremely short minimal tRNAs in acariform mites. BMC Genomics 10:598.

Ku C, Nelson-Sathi S, Roettger M, Garg S, Hazkani-Covo E, Martin WF. 2015. Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes. Proc Natl Acad Sci U S A:201421385.

Lai LB, Chan PP, Cozen AE, Bernick DL, Brown JW, Gopalan V, Lowe TM. 2010. Discovery of a Minimal Form of RNase P in Pyrobaculum. Proc Natl Acad Sci U S A 107:22493–22498.

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

Larkum AWD, Lockhart PJ, Howe CJ. 2007. Shopping for plastids. Trends in Plant Science 12:189–195.

Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 32:11–16.

Van Leuven JT, Meister RC, Simon C, McCutcheon JP. 2014. Sympatric Speciation in a Bacterial Endosymbiont Results in Two Genomes with the Functionality of One. Cell 158:1270–1280.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.

Limbach PA, Crain PF, McCloskey JA. 1994. Summary: the modified nucleosides of RNA. Nucleic Acids Res 22:2183–2196.

Lohan AJ, Wolfe KH. 1998. A subset of conserved tRNA genes in plastid DNA of nongreen plants. Genetics 150:425–433.

López-Madrigal S, Latorre A, Porcar M, Moya A, 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.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. Nucleic Acids Res. 25:0955–0964.

Luan J-B, Chen W, Hasegawa DK, Simmons AM, Wintermantel WM, Ling K-S, Fei Z, Liu S-S, Douglas AE. 2015. Metabolic Coevolution in the Bacterial Symbiosis of Whiteflies and Related Plant Sap-Feeding Insects. Genome Biol Evol 7:2635–2647.

Macdonald SJ, Lin GG, Russell CW, Thomas GH, Douglas AE. 2012. The central role of the host cell in symbiotic nitrogen metabolism. Proc Biol Sci 279:2965–2973.

Maier U-G, Zauner S, Woehle C, Bolte K, Hempel F, Allen JF, Martin WF. 2013. Massively Convergent Evolution for Ribosomal Protein Gene Content in Plastid and Mitochondrial Genomes. Genome Biol Evol 5:2318–2329.

Manzano-Marín A, Latorre A. 2014. Settling Down: The Genome of Serratia symbiotica from the Aphid Cinara tujafilina Zooms in on the Process of Accommodation to a Cooperative Intracellular Life. Genome Biol Evol 6:1683–1698.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12.

Martin W. 2003. Gene transfer from organelles to the nucleus: Frequent and in big chunks. Proc Natl Acad Sci U S A 100:8612–8614.

McCutcheon JP, von Dohlen CD. 2011. An Interdependent Metabolic Patchwork in the Nested Symbiosis of Mealybugs. Curr Biol 21:1366–1372.

McCutcheon JP, McDonald BR, Moran NA. 2009a. Origin of an Alternative Genetic Code in the Extremely Small and GC–Rich Genome of a Bacterial Symbiont. PLoS Genet 5:e1000565.

McCutcheon JP, McDonald BR, Moran NA. 2009b. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. Proc Natl Acad Sci U S A 106:15394–15399.

McCutcheon JP, Moran NA. 2010. Functional Convergence in Reduced Genomes of Bacterial Symbionts Spanning 200 My of Evolution. Genome Biol Evol 2:708–718.

McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10:13–26.

Meiklejohn CD, Holmbeck MA, Siddiq MA, Abt DN, Rand DM, Montooth KL. 2013. An Incompatibility between a Mitochondrial tRNA and Its Nuclear-Encoded tRNA Synthetase Compromises Development and Fitness in Drosophila. PLoS Genet

9:e1003238.

Moran NA, Bennett GM. 2014. The Tiniest Tiny Genomes. Annu Rev Microbiol 68:195–215.

Nagaike T, Suzuki T, Tomari Y, Takemoto-Hori C, Negayama F, Watanabe K, Ueda T. 2001. Identification and Characterization of Mammalian Mitochondrial tRNA nucleotidyltransferases. J Biol Chem 276:40041–40049.

Nakabachi A, Ishida K, Hongoh Y, Ohkuma M, Miyagishima S. 2014. Aphid gene of bacterial origin encodes a protein transported to an obligate endosymbiont. Curr Biol 24:R640–R641.

Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics 25:1335–1337.

Neuenfeldt A, Just A, Betat H, Moerl M. 2008. Evolution of tRNA nucleotidyltransferase: A small deletion generated CC-adding enzymes. Proc Natl Acad Sci U S A 105:7953–7958.

Nikoh N, McCutcheon JP, Kudo T, Miyagishima S, Moran NA, Nakabachi A. 2010. Bacterial Genes in the Aphid Genome: Absence of Functional Gene Transfer from Buchnera to Its Host. PLoS Genet 6:e1000827.

Novoa EM, Pavon-Eternod M, Pan T, Ribas de Pouplana L. 2012. A Role for tRNA Modifications in Genome Structure and Codon Usage. Cell 149:202–213.

Nowack ECM, Grossman AR. 2012. Trafficking of protein into the recently established photosynthetic organelles of Paulinella chromatophora. Proc Natl Acad Sci U S A 109:5340–5345.

Oakeson KF, Gil R, Clayton AL, Dunn DM, Niederhausern AC von, Hamil C, Aoyagi A, Duval B, Baca A, Silva FJ, et al. 2014. Genome Degeneration and Adaptation in a Nascent Stage of Symbiosis. Genome Biol Evol 6:76–93.

Osawa S, Jukes TH, Watanabe K, Muto A. 1992. Recent evidence for evolution of the genetic code. Microbiol Rev 56:229–264.

Phizicky EM, Hopper AK. 2010. tRNA Biology Charges to the Front. Genes Dev. 24:1832–1860.

Poliakov A, Russell CW, Ponnala L, Hoops HJ, Sun Q, Douglas AE, van Wijk KJ. 2011. Large-Scale Label-Free Quantitative Proteomics of the Pea aphid-Buchnera Symbiosis. Mol Cell Proteomics 10:M110.007039.

Price DR, Wilson AC. 2014. A substrate ambiguous enzyme facilitates genome reduction in an intracellular symbiont. BMC Biology 12:110.

Randau L, Schröder I, Söll D. 2008. Life without RNase P. Nature 453:120–123.

Randau L, Söll D. 2008. Transfer RNA genes in pieces. EMBO reports 9:623–628.

Reuven NB, Zhou Z, Deutscher MP. 1997. Functional Overlap of tRNA Nucleotidyltransferase, Poly(A) Polymerase I, and Polynucleotide Phosphorylase. J. Biol. Chem. 272:33255–33259.

Rice DW, Alverson AJ, Richardson AO, Young GJ, Sanchez-Puerta MV, Munzinger J, Barry K, Boore JL, Zhang Y, dePamphilis CW, et al. 2013. Horizontal Transfer of Entire Genomes via Mitochondrial Fusion in the Angiosperm Amborella. Science 342:1468–1473.

Rubio MAT, Rinehart JJ, Krett B, Duvezin-Caubet S, Reichert AS, Söll D, Alfonzo JD. 2008. Mammalian mitochondria have the innate ability to import tRNAs by a mechanism distinct from protein import. Proc Natl Acad Sci U S A 105:9186–9191.

Sabree ZL, Degnan PH, Moran NA. 2010. Chromosome Stability and Gene Loss in Cockroach Endosymbionts. Appl Environ Microbiol 76:4076–4079.

Salinas-Giegé T, Giegé R, Giegé P. 2015. tRNA Biology in Mitochondria. International Journal of Molecular Sciences 16:4518–4559.

Salinas T, Duchêne A-M, Maréchal-Drouard L. 2008. Recent advances in tRNA mitochondrial import. Trends in Biochemical Sciences 33:320–329.

Sloan DB, Moran NA. 2012. Genome Reduction and Co-evolution between the Primary and Secondary Bacterial Symbionts of Psyllids. Mol Biol Evol 29:3781–3792.

Sloan DB, Nakabachi A, Richards S, Qu J, Murali SC, Gibbs RA, Moran NA. 2014. Parallel Histories of Horizontal Gene Transfer Facilitated Extreme Reduction of Endosymbiont Genomes in Sap-Feeding Insects. Mol Biol Evol 31:857–871.

Söll D, RajBhandary UL. 1995. tRNA: structure, biosynthesis, and function. ASM Press

Soma A, Onodera A, Sugahara J, Kanai A, Yachie N, Tomita M, Kawamura F, Sekine Y. 2007. Permuted tRNA Genes Expressed via a Circular RNA Intermediate in Cyanidioschyzon merolae. Science 318:450–453.

Stegemann S, Hartmann S, Ruf S, Bock R. 2003. High-frequency gene transfer from the chloroplast genome to the nucleus. Proc Natl Acad Sci U S A 100:8828–8833.

Suzuki T, Nagao A, Suzuki T. 2011. Human Mitochondrial tRNAs: Biogenesis, Function, Structural Aspects, and Diseases. Annu Rev Genet 45:299–329.

Tamas I, Klasson L, Canbäck B, Näslund AK, Eriksson A-S, Wernegreen JJ, Sandström JP, Moran NA, Andersson SGE. 2002. 50 Million Years of Genomic Stasis in Endosymbiotic Bacteria. Science 296:2376–2379.

Thompson DM, Parker R. 2009. Stressing Out over tRNA Cleavage. Cell 138:215–219.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nature Reviews Genetics 5:123–135.

Toft C, Andersson SGE. 2010. Evolutionary microbial genomics: insights into bacterial host adaptation. Nat Rev Genet 11:465–475.

Tokuriki N, Tawfik DS. 2009. Chaperonin overexpression promotes genetic variation and enzyme evolution. Nature 459:668–673.

Tomita K, Weiner AM. 2001. Collaboration Between CC- and A-Adding Enzymes to Build and Repair the 3'-Terminal CCA of tRNA in Aquifex aeolicus. Science 294:1334–1336.

Viñuelas J, Febvay G, Duport G, Colella S, Fayard J-M, Charles H, Rahbé Y, Calevro F. 2011. Multimodal dynamic response of the Buchnera aphidicola pLeu plasmid to variations in leucine demand of its host, the pea aphid Acyrthosiphon pisum. Mol Microbiol 81:1271–1285.

Watanabe Y, Suematsu T, Ohtsuki T. 2014. Losing the stem-loop structure from metazoan mitochondrial tRNAs and co-evolution of interacting factors. Front Genet [Internet] 5. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4013460/

Wernegreen JJ. 2015. Endosymbiont evolution: predictions from theory and surprises from genomes. Ann N Y Acad Sci [Internet]. Available from: http://onlinelibrary.wiley.com/doi/10.1111/nyas.12740/abstract

van Wijk KJ, Baginsky S. 2011. Plastid Proteomics in Higher Plants: Current State and Future Goals. Plant Physiol. 155:1578–1588.

Wilcox JL, Dunbar HE, Wolfinger RD, Moran NA. 2003. Consequences of reductive evolution for gene expression in an obligate endosymbiont. Molecular Microbiology 48:1491–1500.

Wilson AC, Dunbar HE, Davis GK, Hunter WB, Stern DL, Moran NA. 2006. A dual-genome microarray for the pea aphid, Acyrthosiphon pisum, and its obligate bacterial symbiont, Buchnera aphidicola. BMC Genomics 7:50.

Wilusz JE, Whipple JM, Phizicky EM, Sharp PA. 2011. tRNAs Marked with CCACCA Are Targeted for Degradation. Science 334:817–821.

Wolstenholme DR, Macfarlane JL, Okimoto R, Clary DO, Wahleithner JA. 1987. Bizarre tRNAs inferred from DNA sequences of mitochondrial genomes of nematode worms. Proc Natl Acad Sci U S A 84:1324–1328.

Woolfit M, Bromham L. 2003. Increased Rates of Sequence Evolution in Endosymbiotic

Bacteria and Fungi with Small Effective Population Sizes. Mol Biol Evol 20:1545–1555.

Zheng G, Qin Y, Clark WC, Dai Q, Yi C, He C, Lambowitz AM, Pan T. 2015. Efficient and quantitative high-throughput tRNA sequencing. Nat Meth 12:835–837.

Zhu L, Deutscher MP. 1987. tRNA nucleotidyltransferase is not essential for Escherichia coli viability. EMBO J 6:2473–2477.

Zientz E, Dandekar T, Gross R. 2004. Metabolic Interdependence of Obligate Intracellular Bacteria and Their Insect Hosts. Microbiol Mol Biol Rev 68:745–770.

Chapter 5: Host complementation in cicada bacteriocytes
Unpublished

Summary
Sap-feeding insects occupy a nutrient-poor niche through obligate symbiosis with intracellular bacteria. While the bacterial endosymbionts across sap-feeding insects are phylogenetically diverse, they converge towards similar functionality; to metabolically complement their insect host. Adaptation to an intracellular lifestyle is manifested in a number of characteristic traits: endosymbiont genomes are typically smaller, more rapidly evolving, enriched in amino-acid and vitamin biosynthesis genes, and lacking in genes involved in basic cellular processes, when compared to free-living bacterial genomes. In the most reduced endosymbiont genomes, there are not enough genes to perform some of the most basic cellular processes, like translation. *Hodgkinia cicadicola, Tremblaya princeps*, *Nasuia deltocephalinicola*, and *Zindera insecticola* all have bacterial genomes that encode fewer than 150 genes and are missing components of the translational system. Here, we test for host complementation of *Hodgkinia* by looking for bacterial HGTs and overexpression of host-encoded genes that may function in the symbiosis. Unlike in other insect endosymbiontic partnerships, we find no evidence for HGT. We did, however, find several insect cytoplasmic and mitochondrial genes that are involved in tRNA processing that were significantly upregulated in bacteriome tissue. Interestingly, many of these overexpressed genes complement those missing from the *Sulcia* and *Hodgkinia* genomes, consistent with a possible supportive or compensatory role of the cicada host in symbiont translation. We also explore potential mechanisms for aaRS transport to *Hodgkinia* through confocal and electron microscopy.

Photograph taken by Piotr Łukasik

5.1 Introduction

Highly reduced endosymbiont genomes are missing critical genes

The smallest bacterial genomes are all insect nutritional endosymbionts (Moran et al. 2008; McCutcheon and Moran 2012). The most highly reduced genomes currently published are *Hodgkinia cicadicola, Tremblaya princeps*, *Carsonella ruddii*, and *Nasuia deltocephalinicola*, all of which appear to be missing genes that are thought to be essential for life (Moran and Bennett 2014). *Hodgkinia,* for example, which lives inside of the cicada species *Diceroprocta semicincta,* encodes only ten of the twenty amino acid tRNA synthetase (aaRS) genes (McCutcheon et al. 2009). Of all sequenced bacteria, only five have genomes containing fewer than 15 aaRSs. *Sulcia*, *Portiera, Zindera*, *Uzinura*, and *Blattabacterium* have 15-20 aaRS genes, while all other bacteria have 20. *Tremblaya PCIT* actually contains no functional aaRS homologs, but it has its own endosymbiont called *Moranella endobia,* which encodes all 20 aaRS genes. It is presumed that *Tremblaya* somehow has access to the aaRS proteins produced by *Moranella* cells (McCutcheon and von Dohlen 2011; Husnik et al. 2013). *Hodgkinia* also lives symbiotically with another bacterium, but they inhabit distinctly separate insect cells (McCutcheon et al. 2009; Campbell et al. 2015). However, even if *Hodgkinia* and its co-symbiont, *Sulcia*, were able to share aaRSs, together they endode insufficient aaRSs genes since in combination their genomes only encode 16. Together, they are missing the arginine, asparagine, threonine, and cystine aaRS genes (Table 1). There are several possible hypotheses that might explain how *Sulcia* and *Hodgkinia* survive: (1) host-derived aaRSs aminoacylate bacterial tRNAs; (2) horizontal gene transfer (HGT) from *Sulcia* and/or *Hodgkinia* to the cicada host has occurred and heterologous complementation results in full functionality; (3) similarly, heterologous complementation restores function, but HGT genes originate from other bacterial sources; (4) *Hodgkinia* and *Sulcia* import aminoacylated host tRNAs; and (5) these bacteria have found an alternative mechanism to aminoacylate tRNAs.

Table 1. Distribution of aaRS genes in the most degenerate bacterial genomes, plus *Sulcia* DICSEM for comparison.

	alaS	asnS	aspS	argS	cysS	glnS	gltX	glyS	hisS	ileS	leuS	lysS	metG	pheS	proS	serS	thrS	trpS	tyrS	valS
Carsonella		X		X	X			X						X	X		X			X
Nasuia			X			X					X					X			X	X
Tremblaya					ψ															
Hodgkinia	X						X	X	X	X			X	X	X			X		X
Sulcia	X	X				X	X	X		X	X	X	X	X	X	X		X	X	X

Evidence of HGT in insect-bacterial symbioses

Hypothesis (2) and (3) can be directly tested by sequencing the cicada transcriptome. If HGT has occurred, the transferred genes would need to be expressed from the cicada genome for functionality and should be detectable by RNA-seq. While HGT of aaRS genes have not been

found in any insect genome (except from organelles), HGT of other genes have (reviewed in Sloan et al. 2014). The Sternorrhyncha are the best studied, with HGTs having been discovered in mealybugs, whiteflys, psyllids, and aphids (22, 10, 4, and 2 HGTs respectively). In only one possible case are the HGTs ancestral to all four insects; independent gene acquisition has occurred in the insect lineages to complement endosymbiont genome degradation. By and large, these HGTs seem to complement critical steps in amino acid or vitamin synthesis that are missing from the bacterial endosymbiont genomes. However, many of the unique HGTs in the mealybug genome are involved in peptidoglycan biosynthesis and recycling (Husnik et al. 2013). It is hypothesized that these genes may be important in regulating the supply of cellular components from the intrabacterial symbiont, *Moranella*, to is host bacterium, *Tremblaya*, by controlling the cell wall stability of *Moranella*.

Host support of endosymbionts through transcriptional upregulation

As an alternative to HGT, host genes could heterologously support gene loss in endosymbiotic bacteria. Comparing eukaryotic gene expression between bacteriome tissue and other insect tissues has shown overexpression of host genes that are conspicuously complementary to genes missing from the genomes of *Buchnera*, *Tremblaya*, *Moranella*, and *Carsonella* that function in essential amino acid synthesis and nitrogen recycling (Hansen and Moran 2011; Poliakov et al. 2011; Macdonald et al. 2012; Sloan et al. 2014), including amino acid transporters to facilitate the transfer of amino acids between symbiont and host (Price et al. 2011; Duncan et al. 2014; Price et al. 2014). Upregulated genes with important proposed functions in nitrogen acquisition and recycling in mealybugs, psyllids, and aphids include glutamine sythetase, glutamine oxoglutarate aminotransferase, asparaginase, aspartate aminotransferase, and 1-pyrroline-5-carboxylate synthase (Sloan et al. 2014). These genes likely aid in making nitrogen available to the endosybionts from ammonia and the non-essential amino acids glutamate and glutamine. Since plant sap has a very low C:N ratio, these pathways are needed by the endosymbionts for making high levels of essential amino acids (Macdonald et al. 2012). Other genes universally upregulated in these insects' bacteriomes include genes involved in nonessential amino acid biosysthesis, presumably because most of the pathways for nonessential amino acid biosynthesis are missing from the genomes of these endosymbionts.

5.2 Identifying potential HGTs from bacteria to cicadas

As in previous studies (Husnik et al. 2013; Nakabachi et al. 2014), we used an RNA-Seq approach to find bacterial genes that may have been transferred to the host insect genome. In contrast to other related sap-feeding insects (Nikoh et al. 2010; Husnik et al. 2013; Sloan et al. 2014; Luan et al. 2015), we find no evidence for expression of important horizontally transferred bacterial genes in DICSEM other than genes of mitochondrial origin (Table 2, supplementary table S1-S2, supplementary fig. 2-11). We assembled 140,308 transcripts from 96,199,327 quality-filtered, adapter-trimmed reads. We removed 393 transcripts belonging to *Hodgkinia* and *Sulcia*, leaving 139,915 for HGT and differential expression (DE) analysis. The largest transcript was 18,931 bp in length, with 25 transcripts over 15 kb, suggesting that the sequencing coverage was sufficient to obtain a good assembly.

Our initial high-evalue BLASTP search of Trinity components to nr identified 13 amino acid sequences of putative bacterial origin (Table 2, supplementary table 1), after removal of *Sulcia* and *Hodgkinia* sequences. These components were further classified by more stringent BLASTP searches, Pfam domain searches, and phylogenetics, resulting in the removal of five components. Of the eight remaining components, several had high sequence identities to each other (Supplementary table 2) and were combined for downstream phylogenetic analyses.

Table 2. Taxonomic classification of HGT candidates closest BLAST hits to the orthoMCL database.

	Domain	**Phylum**	**Class**	**Order**	**Family**	**Genus**
m.1	No hits					
m.2	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Brevundimonas
m.3	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus
m.4	Bacteria	Bacteroidetes	Bacteroidetes	Cytophagia	Cytophagales	Cytophagaceae
m.5	Not assigned					
m.6	Not assigned					
m.7	Not assigned					
m.8	No hits					
m.9	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Rickettsiaceae	Rickettsieae
m.10	Eukaryota	Euglenozoa	Kinetoplastida	Trypanosomatidae	Trypanosoma	Trypanozoon
m.11	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Rickettsiaceae	Rickettsieae
m.12	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Clostridium
m.13	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus

Candidate m.2

The top blast hits of m.2 are ornithine carbamoyltransferases from a-proteobacteria. I sampled about 15 sequences from a-proteobacteria lineages, and a few from g-proteobacteria, b-proteobacteria, firmicutes, mitochondria and nuclear origins. The WAG+I+G substitution model gives the lowest likelihood value in ProtTest. A maximum likelihood tree was make with the following parameters: WAG+G+I, 100 bootstraps, #discrete G categories=5, initial tree=NJ,BioNJ. The ML tree is shown in Figure 1 and does not give good support for m.2 being monophyletic with any bacterial phylum, but instead is monophyletic with *Danaus plexippus* (monarch butterfly). The ornithine carbamoyltransferase gene family has previously been identified as a horizontally transferred gene in aphids, mealybugs, and psyllids (Nikoh et al. 2010; Macdonald et al. 2012; Husnik et al. 2013; Sloan et al. 2014).

Candidate m.3

The top blast hits of m.3 are hypothetical proteins from firmicutes. I sampled about 15 sequences from firmicute lineages, and a few from a-proteobacteria, fusobacteria, and eukaryotic taxa. The number of blast hits was low, making taxon sampling difficult. The JTT+G substitution model gives the lowest likelihood value is ProtTest. A maximum likelihood tree was make with the following parameters: JTT+G, 100 bootstraps, #discrete G categories=5, initial tree=NJ,BioNJ. The ML tree is shown in Figure 1 and does not give good support for m.2 being

monophyletic with any particular phylum.

Candidate m.4

The top blast hits of m.4 are aldehyde dehydrogenases from alpha and beta-proteobacteria. I sampled about 15 sequences from these lineages, and a few from actinobacteria, and eukaryotic taxa. The number of blast hits was low, making taxon sampling difficult. The LG+G substitution model gives the lowest likelihood value is ProtTest. An unweighted parsimony tree was made in PAUP*, but m.4 is found to be sister to the eukaryote *Saccharomyces cerevisiae*. ML and Baysien trees show similar relationships among the taxa, where m.4 falls out by itself between bacteria and eukaryotic taxa. The maximum likelihood tree was make with the following parameters: LG+G, 100 bootstraps, #discrete G categories=5, initial tree=NJ,BioNJ. The parameters used for the Bayesian were: LG+G, 500000 generations, 100000 generations burnin, samplefreq=10 generations.

Candidate m.9/m.11

The top blast hits of m.9 and m.11 are hypothetical proteins (possibly transcriptional regulators) from a-proteobacteria. I sampled mostly taxa from a-proteobacteria, with *E. coli* and *Volvox* sequences as outgroups. The LG+I+G substitution model gives the lowest likelihood value is ProtTest. Parsimony, ML, and Baysien trees all show that m.9 and m.11 are monophyletic with *Rikettsia*. The maximum likelihood tree was make with the following parameters: LG+I+G, 100 bootstraps, #discrete G categories=5, initial tree=NJ,BioNJ. The parameters used for the Bayesian were: LG+I+G, 500000 generations, 100000 generations burnin, samplefreq=10 generations.

Candidate m.12

The top blast hits of m.12 are hypothetical proteins from firmicutes. I sampled about 15 sequences from firmicute lineages, and a few from a-proteobacteria, b-proteobacteria, spirochete, and eukaryotic taxa. The JTT+G substitution model gives the lowest likelihood value is ProtTest. ML and Baysien trees for m.12 are not in very good agreement. The ML tree groups m.12 monophyletically with firmicutes, albeit with low support. The Bayesian analysis shows m.12 as being paraphyletic with firmicutes, but the analysis does not seem to group genes from closely related organisms together and does not have a similar topology as the ML tree. The ML tree was make with the following parameters: JTT+G, 100 bootstraps, #discrete G categories=5, initial tree=NJ,BioNJ. The parameters used for the Bayesian were: JTT+G, 500000 generations, 100000 generations burnin, samplefreq=10 generations.

Candidate m.13

The top blast hits of m.13 are hypothetical proteins (possible AAA-ATPases) from firmicutes. I sampled about 15 sequences from firmicute lineages, and a few from a-proteobacteria, fusobacteria, and eukaryotic taxa. m.13 also had a weak blastp (standalone blastp search to cicada symbionts only) hit to gene YP003108565.1 from *Sulica muelleri SMDSEM*, so this gene was included in the dataset. The WAG+I+G substitution model gives the lowest likelihood value is ProtTest. Despite most blastp hits being Firmicute lineages, the ML tree groups m.12 monophyletically with *Sulcia* and the outgroup (*Tribolium castaneum*, or the flour

beetle). The gene annotation for the *T. castaneum* gene is poor, leading me to believe that the gene could potentially be an HGT itself. Grouping with *S. muelleri* is an exciting finding that will require further investigation. More taxa should be sampled before any definite conclusions can be made. The Bayesian analysis also shows m.13 being sister to *S. muelleri.* The ML tree was make with the following parameters: WAG+I+G, 100 bootstraps, #discrete G categories=5, initial tree=NJ,BioNJ. The parameters used for the Bayesian were: WAG+I+G, 500000 generations, 100000 generations burnin, samplefreq=10 generations.

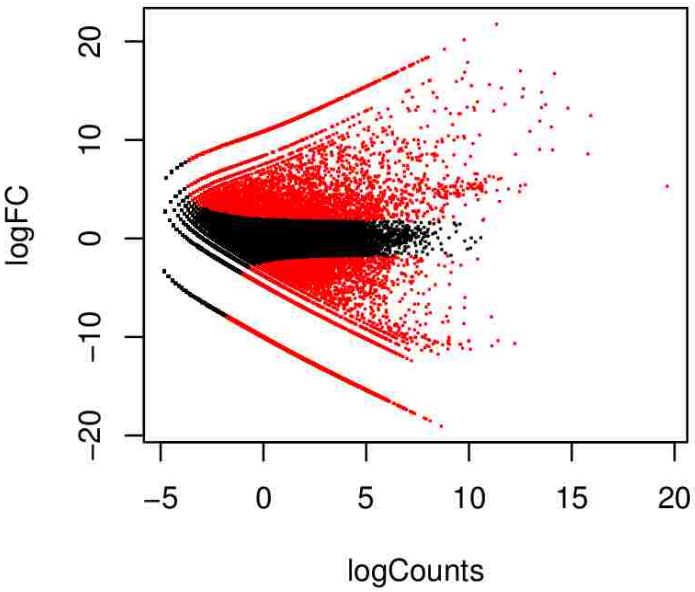


Figure 1. Differential gene expression between cicada bacteriome and body tissues. Each point represents a Trinity subcomponent, with the x axis indicating overall gene expression and the y axis indicating differential expression between tissue types. Genes identified by edgeR as being significantly upregulated or downregulated in the bacteriome are in red.

5.3 Differential expression analysis

Of 140,308 transcripts assembled by Trinity, 11987 were differentially expressed, with 8418 being upregulated in bacteriocytes and 3569 being upregulated in insect tissues (Figure 1). We found several insect cytoplasmic and mitochondrial aaRSs that were significantly upregulated in bacteriome tissue (edgeR, $p<0.01$). Interestingly, many of these overexpressed aaRSs are those missing from the *Sulcia* and *Hodgkinia* genomes (Tables 2,3), consistent with a possible supportive or compensatory role of the cicada host in symbiont translation. In addition to aaRSs, many other genes involved in tRNA processing were upregulated, including a mitochondrial CCA transferase (Tables 4) that could potentially perform the CCAing activity in *Hodgkinia* cells, however, *Sulcia* also encodes a CCA transferase that could hypothetically add the CCAs to *Hodgkinia* tRNAs, as we described in chapter 4.

Table 2. Upregulated cicada aaRS genes in complementing *Sulcia* and *Hodgkinia*.

	alaS	asnS	aspS	argS	cysS	glnS	gltX	glyS	hisS	ileS	leuS	lysS	metG	pheS	proS	serS	thrS	trpS	tyrS	valS
Hodgkinia	X						X	X	X	X			X	X	X			X		X
Sulcia	X	X				X	X	X		X	X	X	X	X	X	X		X	X	X
Cicada mitochon.					▲								▲							
Cicada cytoplasm.			▲	▲	▲				▲							▲	▲	▲		

Table 3. List of all aaRS transcripts identified in the cicada transcriptome by Trinotate.

Contig name	Trinotate identification
comp5977	Alanine--tRNA ligase, cytoplasmic
comp30585	Alanine--tRNA ligase, mitochondrial
comp988	Arginine--tRNA ligase, cytoplasmic
comp13625	Arginine--tRNA ligase, mitochondrial
comp2749	Asparagine--tRNA ligase, cytoplasmic
comp8449	Probable asparagine--tRNA ligase, mitochondrial
comp5101	Aspartate--tRNA ligase, cytoplasmic
comp15524	Aspartate--tRNA ligase, mitochondrial
comp12190	Cysteine--tRNA ligase, cytoplasmic
comp2006	Cysteine--tRNA ligase, mitochondrial
comp3009	Glycine--tRNA ligase
comp7912	Probable glutamate--tRNA ligase, mitochondrial
comp3592	Bifunctional glutamate/proline--tRNA ligase
comp2936	Probable glutamine--tRNA ligase
comp2122	Histidine--tRNA ligase, cytoplasmic
comp5369	Isoleucine--tRNA ligase, cytoplasmic
comp6001	Isoleucine--tRNA ligase, mitochondrial
comp5125	Leucine--tRNA ligase, cytoplasmic
comp3756	Probable leucine--tRNA ligase, mitochondrial
comp3095	Lysine--tRNA ligase
comp11367	Methionine--tRNA ligase, cytoplasmic
comp28836	Methionine--tRNA ligase, mitochondrial
comp17444	Phenylalanine--tRNA ligase alpha subunit
comp8874	Phenylalanine--tRNA ligase beta subunit
comp5410	Phenylalanine--tRNA ligase, mitochondrial
comp7714	Probable proline--tRNA ligase, mitochondrial
comp1617	Serine--tRNA ligase, cytoplasmic
comp99338	Serine--tRNA ligase, cytoplasmic - partial
comp17326	Serine--tRNA ligase, mitochondrial
comp2373	Serine--tRNA ligase, mitochondrial
comp798	Threonine--tRNA ligase, cytoplasmic
comp2783	Tryptophan--tRNA ligase, cytoplasmic
comp3642	Tryptophan--tRNA ligase, mitochondrial
comp1213	Tyrosine--tRNA ligase, cytoplasmic
comp9373	Tyrosine--tRNA ligase, mitochondrial
comp19834	Valine--tRNA ligase
comp22431	Valine--tRNA ligase

Table 4. List of transcripts that are involved in tRNA maturation and are up regulated in cicada bacteriocytes.

id	logFC	logCPM	PValue	FDR	edger_bac	edger_ins	RecName
comp988_c0_seq1	-4.10	5.94	3.95E-008	8.56E-007	79.26	2.32	Arginine--tRNA ligase, cytoplasmic
comp686_c0_seq2	-5.67	5.12	4.28E-012	1.93E-010	75.43	0.74	Aminoacyl tRNA synthase complex-interacting multifunctional protein 1
comp609_c0_seq1	-8.35	6.12	4.47E-019	7.11E-017	58.94	0.10	tRNA (uracil(54)-C(5))-methyltransferase homolog-B
comp1069_c0_seq1	-4.24	5.31	1.91E-008	4.47E-007	49.87	1.31	tRNA (uracil(54)-C(5))-methyltransferase homolog
comp423_c0_seq3	-7.38	5.33	2.53E-016	2.40E-014	46.43	0.13	Aminoacyl tRNA synthase complex-interacting multifunctional protein 1
comp1225_c0_seq6	-13.55	3.21	9.03E-016	7.73E-014	40.17	0.00	D-tyrosyl-tRNA(Tyr) deacylase 1
comp953_c0_seq1	-14.17	3.81	1.39E-017	1.63E-015	35.41	0.00	D-tyrosyl-tRNA(Tyr) deacylase
comp844_c0_seq2	-13.00	2.68	3.75E-014	2.37E-012	33.02	0.00	Eukaryotic translation initiation factor 3 subunit F
comp1195_c3_seq6	-7.87	2.75	4.28E-013	2.26E-011	29.97	0.06	D-tyrosyl-tRNA(Tyr) deacylase 1
comp1497_c0_seq1	-14.07	3.71	2.64E-017	2.97E-015	27.69	0.00	D-tyrosyl-tRNA(Tyr) deacylase 1
comp1645_c0_seq2	-3.04	5.52	1.88E-005	2.44E-004	25.50	1.54	Speckle targeted PIP5K1A-regulated poly(A) polymerase
comp1727_c2_seq6	-3.57	4.25	1.29E-006	2.10E-005	23.53	0.99	Queuine tRNA-ribosyltransferase subunit QTRTD1 homolog
comp798_c1_seq6	-5.47	4.67	1.92E-011	7.62E-010	23.27	0.27	Threonine--tRNA ligase, cytoplasmic
comp1373_c0_seq5	-13.16	2.84	1.21E-014	8.35E-013	19.02	0.00	Putative tRNA pseudouridine synthase Pus10
comp2626_c0_seq2	-4.21	4.12	3.68E-008	8.02E-007	18.95	0.51	Pseudouridylate synthase 7 homolog
comp2861_c0_seq3	-13.23	2.90	7.87E-015	5.61E-013	16.17	0.00	Threonylcarbamoyladenosine tRNA methylthiotransferase
comp2314_c0_seq2	-12.82	2.52	1.21E-013	7.00E-012	13.20	0.00	Peptidyl-tRNA hydrolase 2, mitochondrial
comp2122_c0_seq1	-14.07	3.71	2.72E-017	3.05E-015	12.59	0.00	Histidine--tRNA ligase, cytoplasmic
comp3759_c0_seq3	-12.45	2.17	1.39E-012	6.80E-011	10.27	0.00	tRNA methyltransferase 10 homolog A
comp1617_c0_seq4	-14.68	4.31	4.12E-019	6.63E-017	9.28	0.00	Serine--tRNA ligase, cytoplasmic
comp3307_c0_seq4	-7.32	2.24	1.56E-011	6.28E-010	9.02	0.02	Aminoacyl tRNA synthase complex-interacting multifunctional protein 1
comp2006_c0_seq2	-13.41	3.07	2.39E-015	1.89E-013	8.59	0.00	Cysteine--tRNA ligase, mitochondrial
comp1142_c0_seq4	-13.36	3.03	3.26E-015	2.50E-013	8.49	0.00	Fatty-acid amide hydrolase 2
comp3282_c2_seq1	-5.78	2.97	8.60E-011	3.05E-009	7.30	0.08	Methionyl-tRNA formyltransferase, mitochondrial
comp2712_c0_seq2	-12.43	2.15	1.62E-012	7.81E-011	5.98	0.00	tRNA (guanine(10)-N2)-methyltransferase homolog
comp5101_c0_seq3	-12.38	2.11	2.16E-012	1.02E-010	5.83	0.00	Aspartate--tRNA ligase, cytoplasmic
comp2783_c0_seq1	-12.80	2.50	1.39E-013	7.92E-012	4.58	0.00	Tryptophan--tRNA ligase, cytoplasmic
comp2470_c0_seq1	-12.35	2.08	2.59E-012	1.20E-010	4.43	0.00	tRNA (guanine(37)-N1)-methyltransferase
comp1117_c0_seq3	-4.82	1.78	8.15E-008	1.66E-006	4.33	0.08	CCA tRNA nucleotidyltransferase 1, mitochondrial
comp1645_c0_seq4	-11.51	1.32	5.79E-010	1.78E-008	2.87	0.00	tRNA 2'-phosphotransferase 1
comp5600_c0_seq2	-9.54	-0.40	1.71E-005	2.23E-004	2.82	0.00	Probable queuine tRNA-ribosyltransferase
comp2464_c0_seq1	-12.16	1.91	8.94E-012	3.78E-010	2.57	0.00	L-seryl-tRNA(Sec) kinase
comp9666_c0_seq1	-3.58	0.99	4.58E-005	5.36E-004	2.56	0.11	Telomerase reverse transcriptase
comp6810_c0_seq1	-10.97	0.83	1.68E-008	4.00E-007	1.72	0.00	Mitochondrial ribonuclease P protein 3
comp1551_c1_seq21	-9.68	-0.28	7.46E-006	1.05E-004	1.57	0.00	Eukaryotic translation initiation factor 5B
comp2075_c1_seq5	-5.10	0.27	9.47E-006	1.30E-004	1.38	0.02	tRNA-specific adenosine deaminase 2
comp17100_c0_seq1	-3.27	1.17	7.89E-005	8.57E-004	1.28	0.06	Selenocysteine-specific elongation factor
comp5512_c0_seq3	-10.70	0.60	2.53E-008	5.72E-007	1.18	0.00	D-aspartate oxidase
comp12864_c0_seq1	-10.99	0.85	1.43E-008	3.45E-007	1.00	0.00	Probable tRNA (guanine(26)-N(2))-dimethyltransferase
comp12190_c0_seq1	-10.36	0.30	1.87E-007	3.57E-006	0.75	0.00	Cysteine--tRNA ligase, cytoplasmic
comp6712_c0_seq1	-9.06	-0.81	5.78E-005	6.52E-004	0.61	0.00	Elongation factor Tu, mitochondrial
comp2373_c0_seq1	-9.73	-0.24	5.95E-006	8.52E-005	0.48	0.00	Zinc finger protein 593 homolog

comp18314_c0_seq3	-9.35	-0.56	4.23E-005	5.00E-004	0.25	0.00	Helicase sen1

5.4 Localization of *Sulcia* and cicada aminoacyl tRNA synthetases

Peptide antibodies were generated against mitochondrial and cytoplasmic cystine and aspartate aaRS proteins that were identified by mRNAseq. These candidates were chosen because these genes are missing in both *Sulcia* and *Hodgkinia*, but the *Sulcia* genome contains tRNA genes to decode these amino acids. The *Hodgkinia* genome encodes a $tRNA^{cys}$, but no $tRNA^{asp}$. This pattern of tRNA gene retention, but aaRS gene loss suggests that the proteins might be transported to the bacteria for the aminoacylation of the bacterial tRNAs. An antibody was also generated against *Hodgkinia* dnaQ to be used as a control for localizing *Hodgkinia* cells. When tested by western blot on total protein isolated from cicada bacteriomes, all antibodies display some activity, although the dnaQ antibody is poorly reactive, and the aaRS antibodies display some cross-reactivity with proteins of unexpected sizes (Supplementary figure S12). Since de novo assembled transcripts can sometimes give unreliable splice variants, we interpreted the observed antigenicity as a positive test of the antibodies.

In both laser scanning confocal microscopy (CLSM) and transmission electron microscopy (TEM), cicada mitochondrial CysRS and AspRS appear to be localized to small punctate spheres, primarily in *Hodgkinia*-containing bacteriocytes (Figures 2-4). Since the DICSEM mitochondrial genome contains $tRNA^{cys}$ and $tRNA^{asp}$, we expect the cognate aaRSs to be localized to the mitochondria, where they need to aminoacylate mitochondrial tRNAs. This expectation seems supported by CLSM (Figure 2-3). The dnaQ antibody did not have signal when tested on paraffin embedded tissue sections, even at high concentrations (1:10). Thus, we co-labeled tissue sections with *Hodgkinia* 16S rRNA probes and the aaRS antibodies. We find weak mt-CysRS signal in *Hodgkinia* cells, and strong signal near the periphery of *Hodgkinia* cells (Figure 5). Sections labeled with gold-conjugated secondary antibodies and visualized by TEM confirm this result, but conflict the CLSM in that mt-CysRS and mt-AspRS do not appear to be mainly localized in mitochondria, but rather in the nucleus. We do, however, see good labeling of what appears to be cytoplasmic compartments adjacent to *Hodgkina* cells, and diffuse labeling within *Hodgkinia* cells (Figure 6-7). The reason for this disparity is not clear, however, it is possible that the strongly florescent Hoechst dye in the confocal images swamp out the Cy3-labeled CysRS antibody in CLSM. Polyploid nulcei with holocentric chromosomes (found in several hemipterans (Wigglesworth 1967; Braendle et al. 2003; Gagnon et al. 2014)) can bind antibodies and cause non-specific signal, this explanation seems likely, but remains untested. The cytosolic antibodies seemed to label all tissues equally in CLSM (Figure 3), but is not seem in *Hodgkinia* cells by immuno-TEM (Figure 6-7).

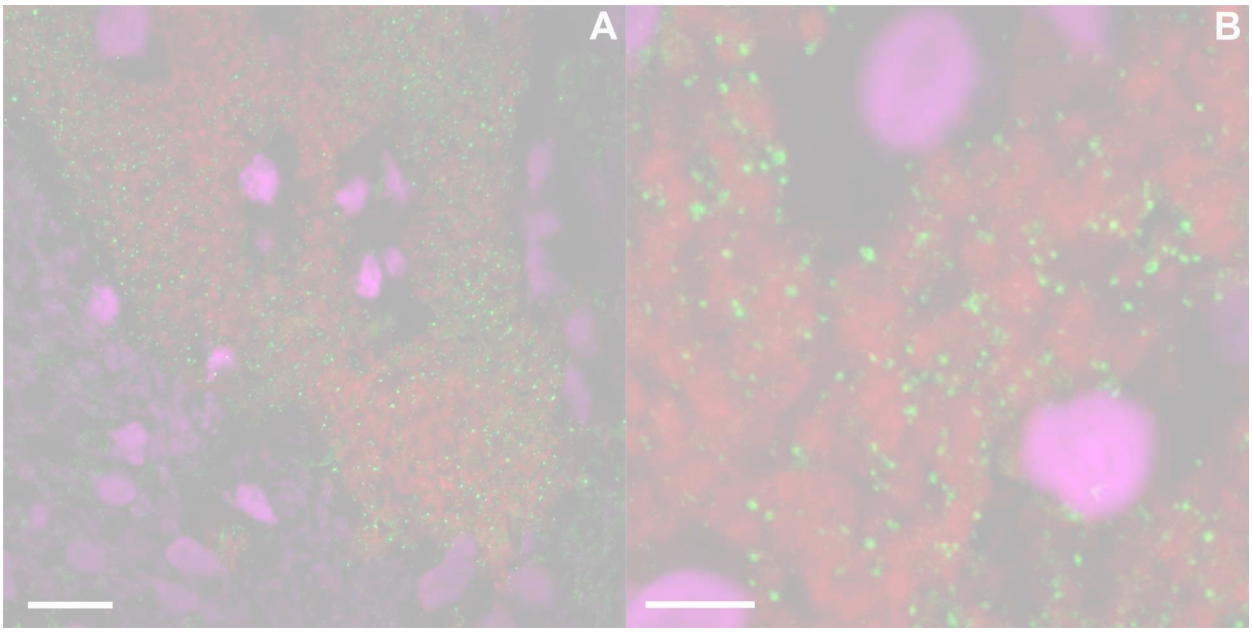


Figure 2. Mitochondrial CysRS (green) and *Hodgkinia* rRNA (red) labeled tissue sections show strong aaRS signal around *Hodgkinia* cells, with weak, punctate signal from within bacterial cells. *Sulcia* cells are not specifically labeled, but can be visualized at the bottom left of panel (A) due to the Hoechst DNA stain (magenta) which primarily labels insect cell nuclei. Scale bars are 20 μm and 5μm in panels (A) and (B), respectively.

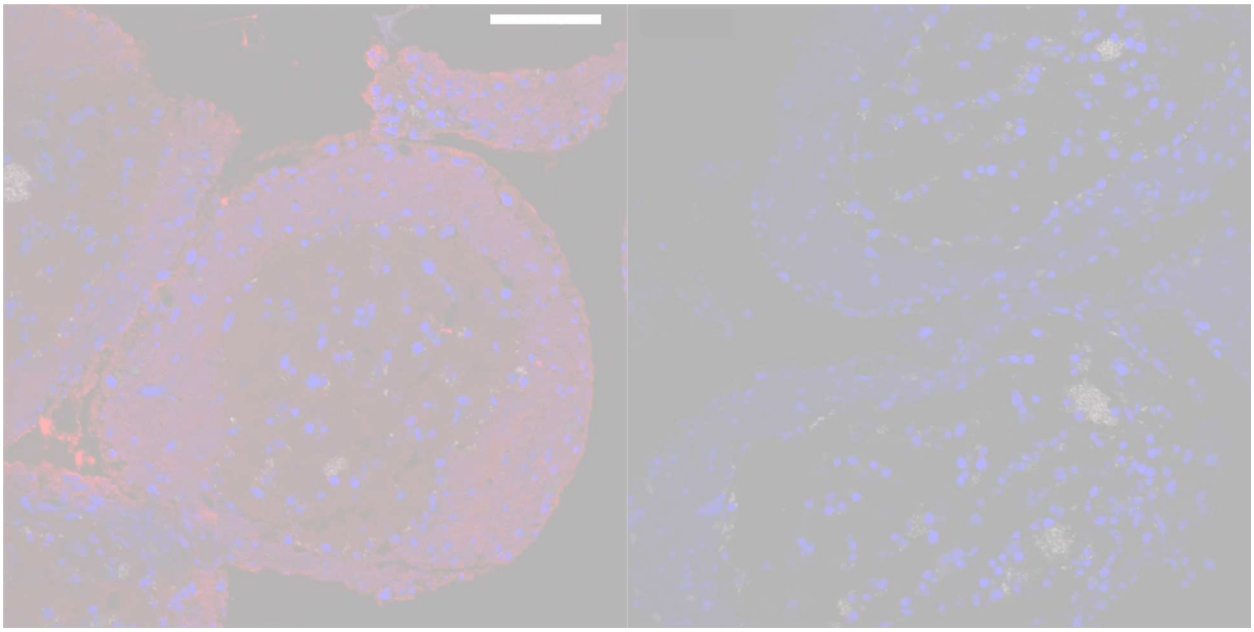


Figure 3. Cytosolic CysRS shows diffuse, weak signal by CLSM. No primary control on right.

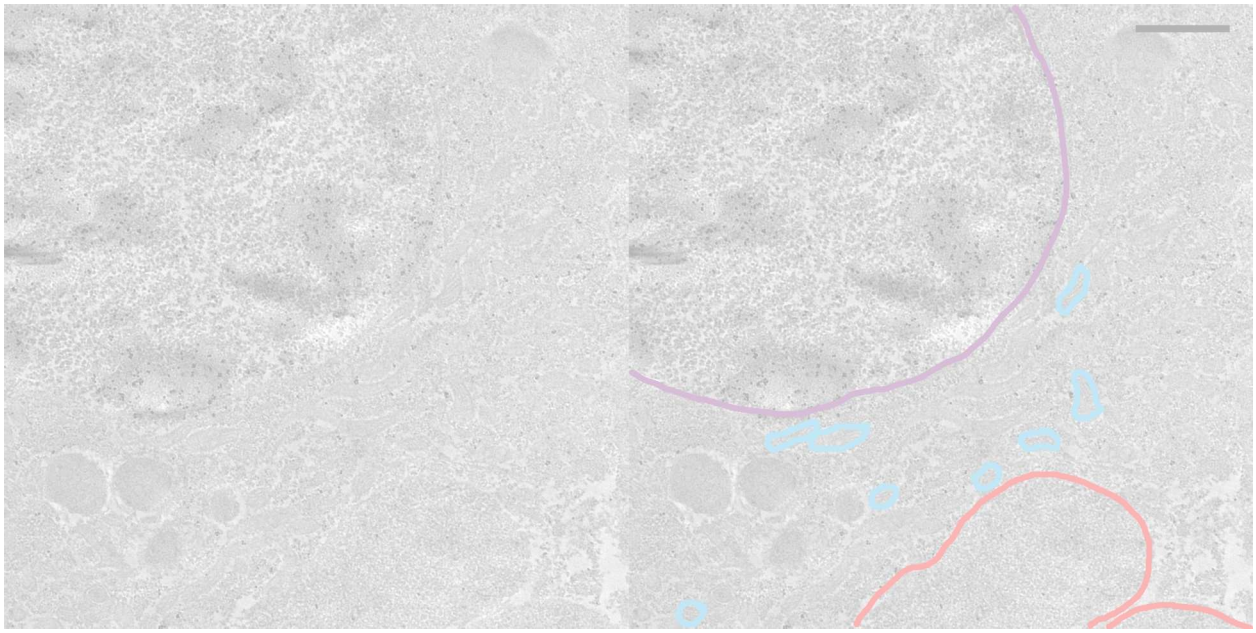


Figure 4. Mitochondrial CysRS labeled tissue sections imaged by TEM show one insect cell nulceus (magenta), two partial *Hodgkinia* cells (red), and many mitochondria (some pseudo-colored blue). The gold-beads are 10nM, the scale bar is 1μm.

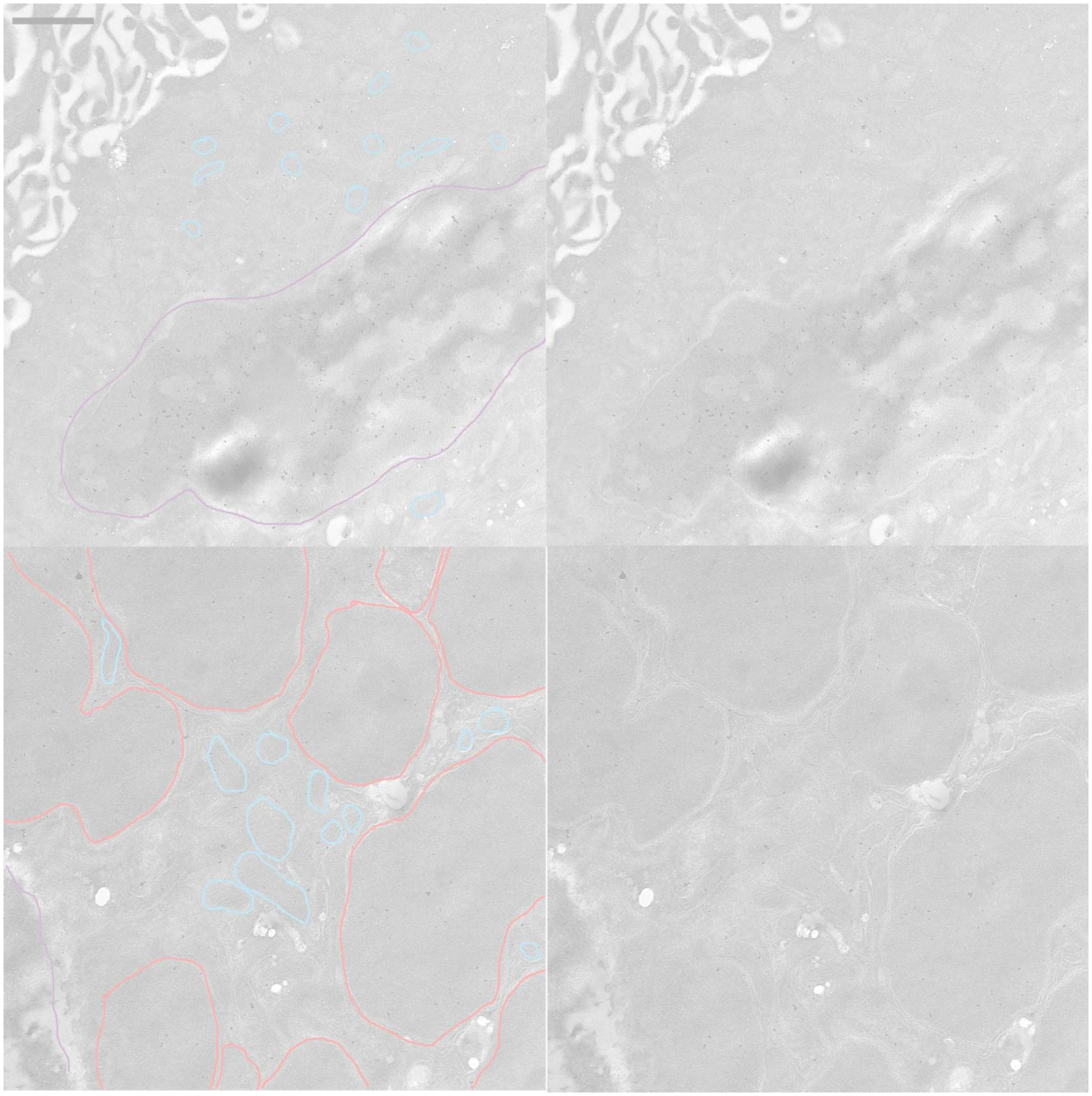


Figure 5. Immuno-TEM of cicada tissue section labeled with anti-cytoplasmic AspRS with *Hodgkinia* (red), insect cell nuclei (magenta), and mitochondria (blue). Scale bar is 1 μm.

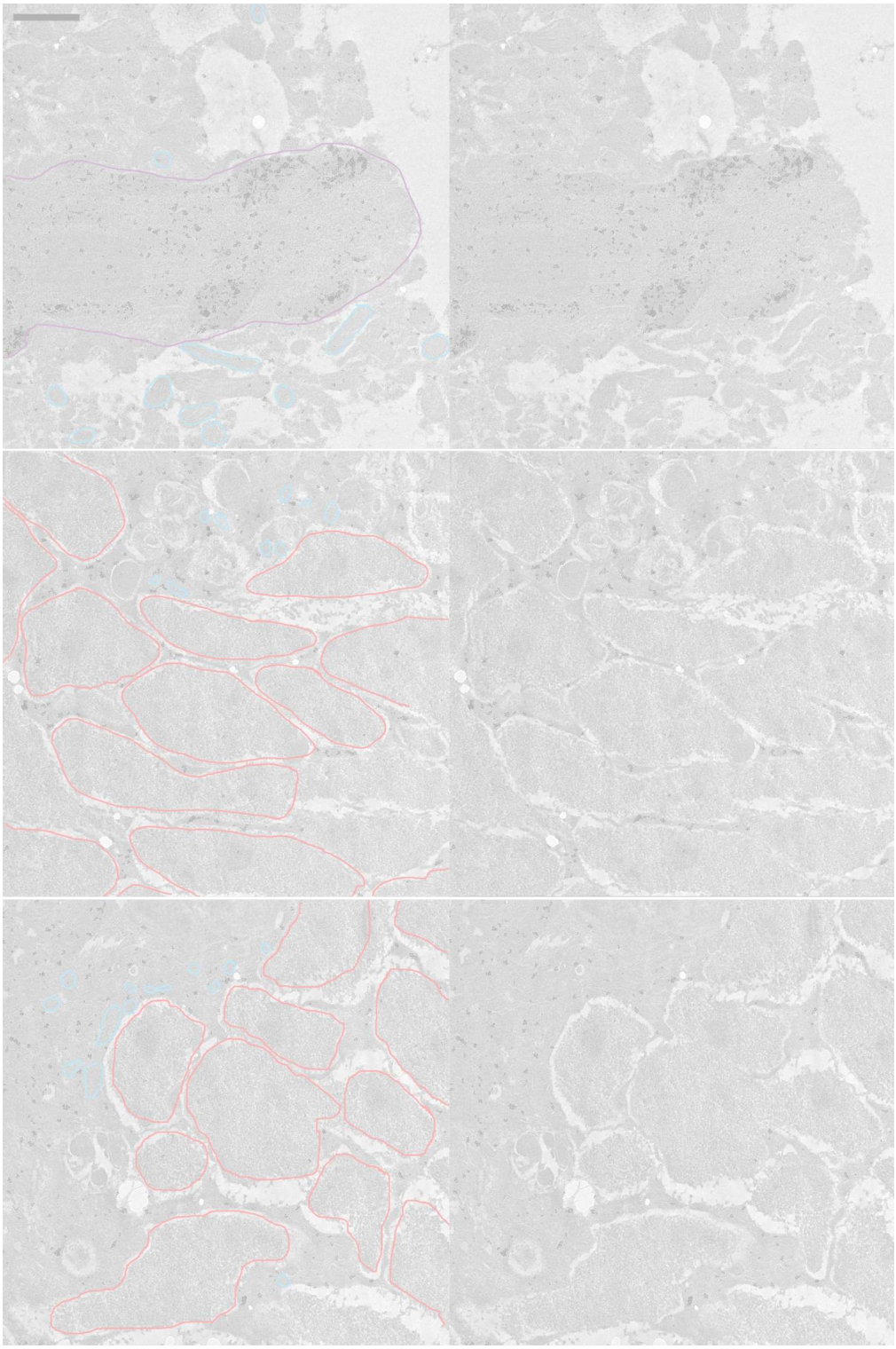


Figure 6. Immuno-TEM of cicada tissue section labeled with anti-cytoplasmic CysRS with *Hodgkinia* (red), insect cell nuclei (magenta), and mitochondria (blue). Scale bar is 1 μm.

5.5 Electron microscopy of cicada bacteriocytes

In an effort to gain an understanding on the cellular organization of cicada bacteriocytes, we performed TEM on DICSEM bacteriomes. We observe a complex arrangement of densely packed intracellular membranes, *Hodgkinia* cells, and mitochondria (Figure 7, 8). As with many insect endosymbionts, the shape of *Hodgkinia* and *Sulcia* cells is amorphous, with cross-sectional diameters of 1-3 µm. Confocal microscopy from previous work shows that these cells can be up to about 10 µm in length (McCutcheon et al. 2009; Van Leuven et al. 2014). They are surrounded by three membranes (presumably two bacterial and one symbiosomal), with no

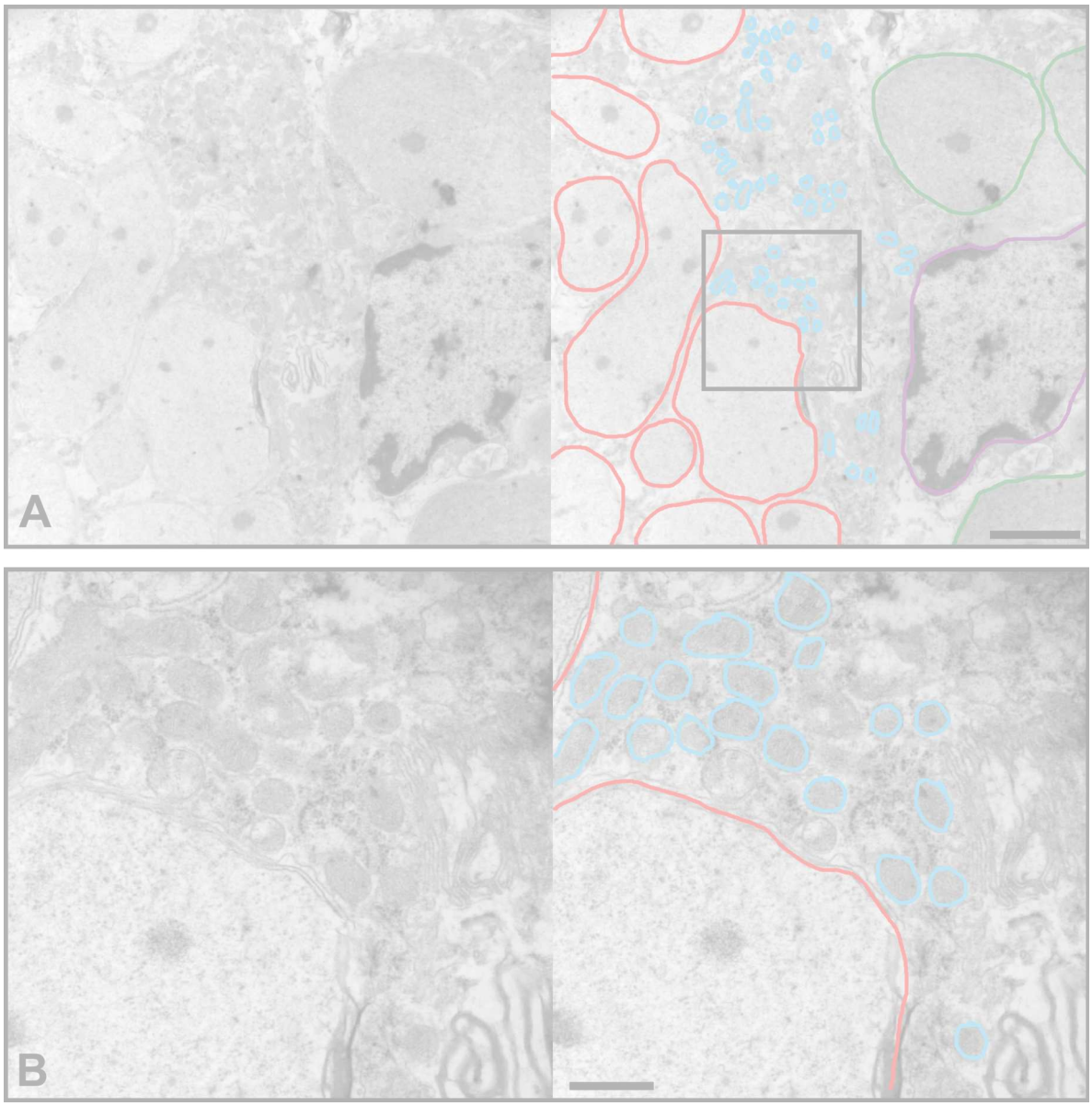


Figure 7. TEM of cicada tissue section showing *Sulcia* (psudo-colored green), *Hodgkinia* (red), insect cell nucleus (magenta), and mitochondria (blue). Panel (B) is an inset of panel (A). Scale bars are 2µm in (A) and 0.5µm in (B).

visible peptidioglycan layer. *Sulcia* and *Hodgkinia* are partitioned into different areas of the bacteriome, that are separated by a section of host cells and is about 5-20 µm thick (Figure 7A). This region, as well as the *Hodgkinia* containing bacteriocytes are densely packed with mitochondria. As with bacteriocytes of other insects, cicada bacteriocytes are probably multinucleated and contain many unidentifiable membrane compartments, making it hard to define a single cicada cell. Some of these membrane compartments seem to interact with the membranes of the *Sulcia* and *Hodgkinia* (Figure 7B, 8), although it is difficult to say if this is an artifact of the preservation methods.

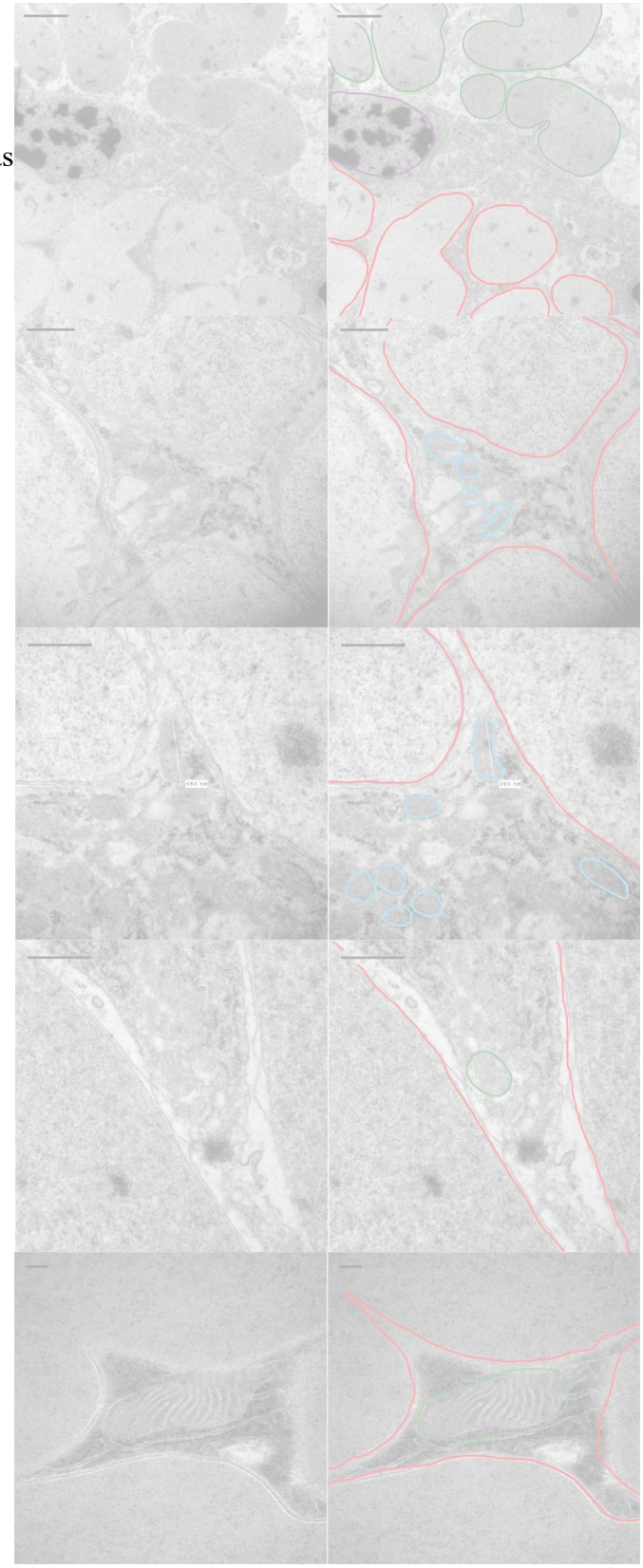


Figure 8. TEM of cicada tissue section showing *Sulcia* (psudo-colored green), *Hodgkinia* (red), insect cell nuclei (magenta), and mitochondria (blue). Scale bars are 2µm in (A), 0.5µm in (B-D), and 0.1µm in (E).

5.6 Discussion

HGT influences the ecology and evolution of organisms (Keeling and Palmer 2008). In bacterial-eukaryote symbioses, HGT likely enables genome reduction in the bacterial partner (Sloan et al. 2014). However, I find no predicted aaRS genes of bacterial origin being expressed from the cicada genome. Additionally, the genes identified in RNA-seq originate from a diverse set of bacteria including α-proteobacteria, Firmicutes, and possibly Bacteroidetes. The taxonomy of the most closely related bacteria for each gene candidate is shown in Table 2 (based off blast searches to the orthoMCL database). However, the phylogenetic analysis (by parsimony, ML, and Bayesian methods) show that categorizing the HGT candidates to a particular bacterial clade may be more difficult than the orthoMCL results suggest.

My results corroborate the results of others; HGT from the current symbiont to the host genome seems to be rare. HGT, followed by import of gene products back into their originating organism (the organelle) is a defining property distinguishing bacterial symbionts from organelles. To date, only two examples of protein transport of HGT gene products into bacterial symbiont have been shown (Nowack and Grossman 2012; Nakabachi et al. 2014). However, many cases of HGT alone have been shown, suggesting that genomic information is often transferred from symbiont to host, but infrequently incorporated into the host's functional genomic repertoire. The evolutionary implications of this observation are interesting, because it implies a tenancy for acquiring DNA, but an innate reluctance for maintaining foreign DNA. The rate of acquiring and maintaining DNA is probably dynamically variable, depending on exposure frequency and environmental conditions (stress, for example).

Further evaluation of candidate m.13 will be interesting. The other HGT candidates found in this study seem to belong to gene families (AAA-ATPases, hypothetical proteins, aldehyde dehydrogenases, and ornithine carbamoyltransferases) that are commonly transferred by HGT. m.9 and m.11 deserve further functional evaluation since they are implicated to be involved in transcriptional regulation. Transcription in these highly reduced genomes is thought to be more-or-less constitutive since they have lost most regulatory mechanisms (eg: both have only one specificity factor, sigma-70).

The pattern of aaRS overexpression that we describe is intriguing, especially since some of these proteins seems to be localized in *Hodgkinia* and *Sulcia* cells. This result has not been observed in other endosymbiont systems and it is worth a second look at the transcriptome data of mealybugs and psyllids to check the expression levels of host aaRS genes with those missing in *Carsonella* and *Tremblaya* PAVE. The data that we present in this chapter suggest that the cicada host is contributing cellular components to fill in core processes of translation missing from the endosymbiont genomes. If true, this further breaks down the barriers distinguishing organelles from endosymbionts and, surprisingly, suggests that aaRS proteins from very distantly related organisms can likely charge the tRNAs of *Sulcia* and *Hodgkinia*.

5.7 Methods and supplementary materials

Wild cicadas were caught on palo verde trees in the Tuscon, AZ area and were

decapitated and immediately placed in RNAlater (Ambion) and stored at -20°C until dissection. RNA was purified from bacteria-harboring tissues and pooled head, leg, and wing muscle tissues according to kit instructions (MO-BIO: Biofilim RNA Isolation kit). RNA was prepared and sequenced on a Illumina HiSeq sequencing machine at HudsonAlpha Sequencing Center in Huntsvillle, AL. Raw reads from one HiSeq lane were quality filtered to Q=20 over 90% of the read and 5bp were trimmed from the end of each read. 96,199,327 reads were assembled in Trinity (January 25, 2012 release) using kmer_length=25 and min_contig_length=48. All assembled transcripts were classified by domain (Bacteria, Eukarya, Archea) using a blastx search against the NCBI protein database followed by filtering using custom Perl scripts. All transcripts that blast to bacterial sequences were extracted from the complete assembly. This subset was further filtered by removing sequences that have high identity to the *H. cicadicola* and *S. muelleri* genomes (blastn 97% identity). Of the remaining 41 transcripts, the best ORFs were picked using Trinity's "transcripts_to_best_scoring_ORFs.pl" program, which uses a Markov model to choose the most likely full-length transcripts. The resulting sequences were used in phylogenetic analysis.

Sequence alignments

Sequences related to each potential HGT protein sequence were obtained using blastp. The resulting taxonomy profile was parsed to heavily sample closely related genes and broadly sample divergent genes from other phyla. A minimum of 30 gene sequences were downloaded and aligned using MAFFT (v7.027b) L-INS-i followed by manual correction in SeaView (v4.3.1). The appropriate substitution model was chosen separately for each dataset using ProtTest (v3.2).

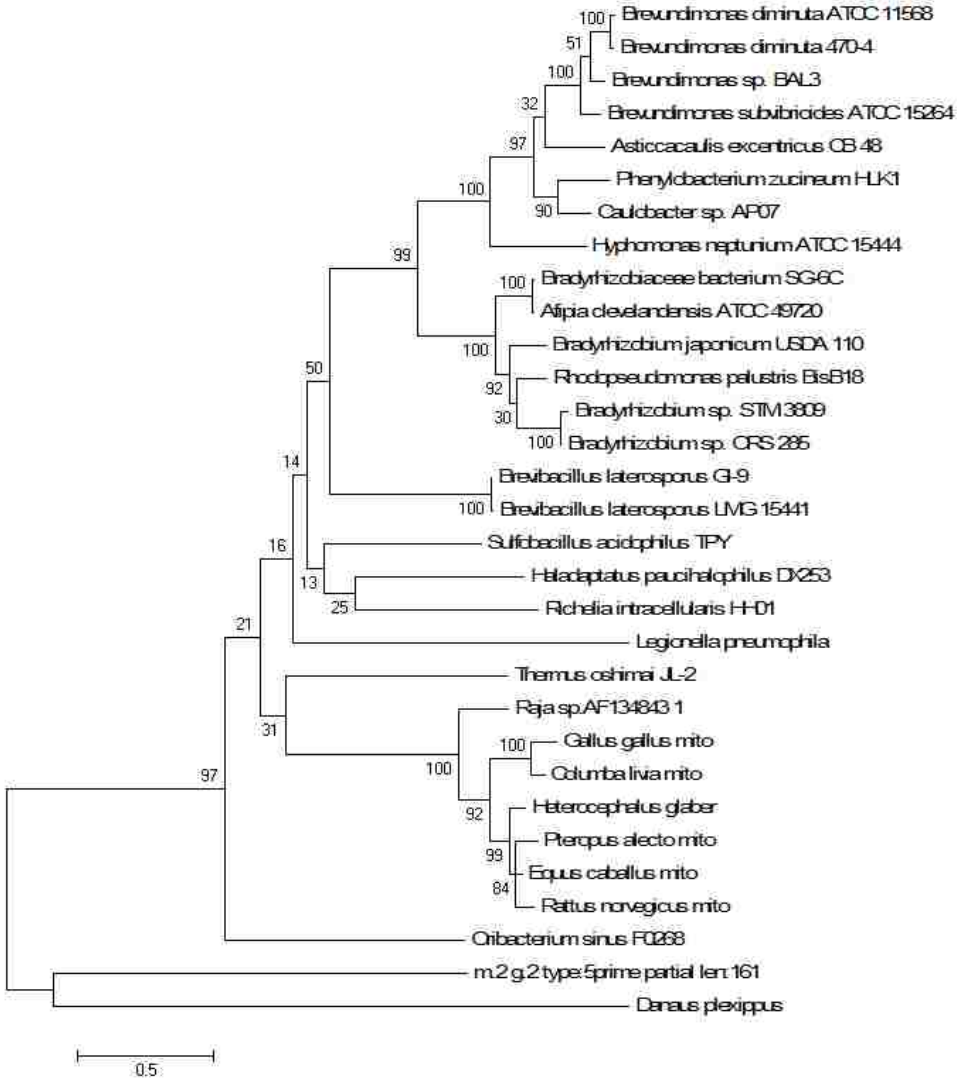
Tree building and visualization

Unweighted parsimony trees were created using PAUP*. Maximum likelihood and Bayesian inference phylogenetic methods were applied to each set of amino acid alignments using MEGA5 and MrBayes, respectively. Bootstrap values, credibility intervals, and burn-in generations are indicated for each alignment set. Trees were visualized and edited in TreeView.

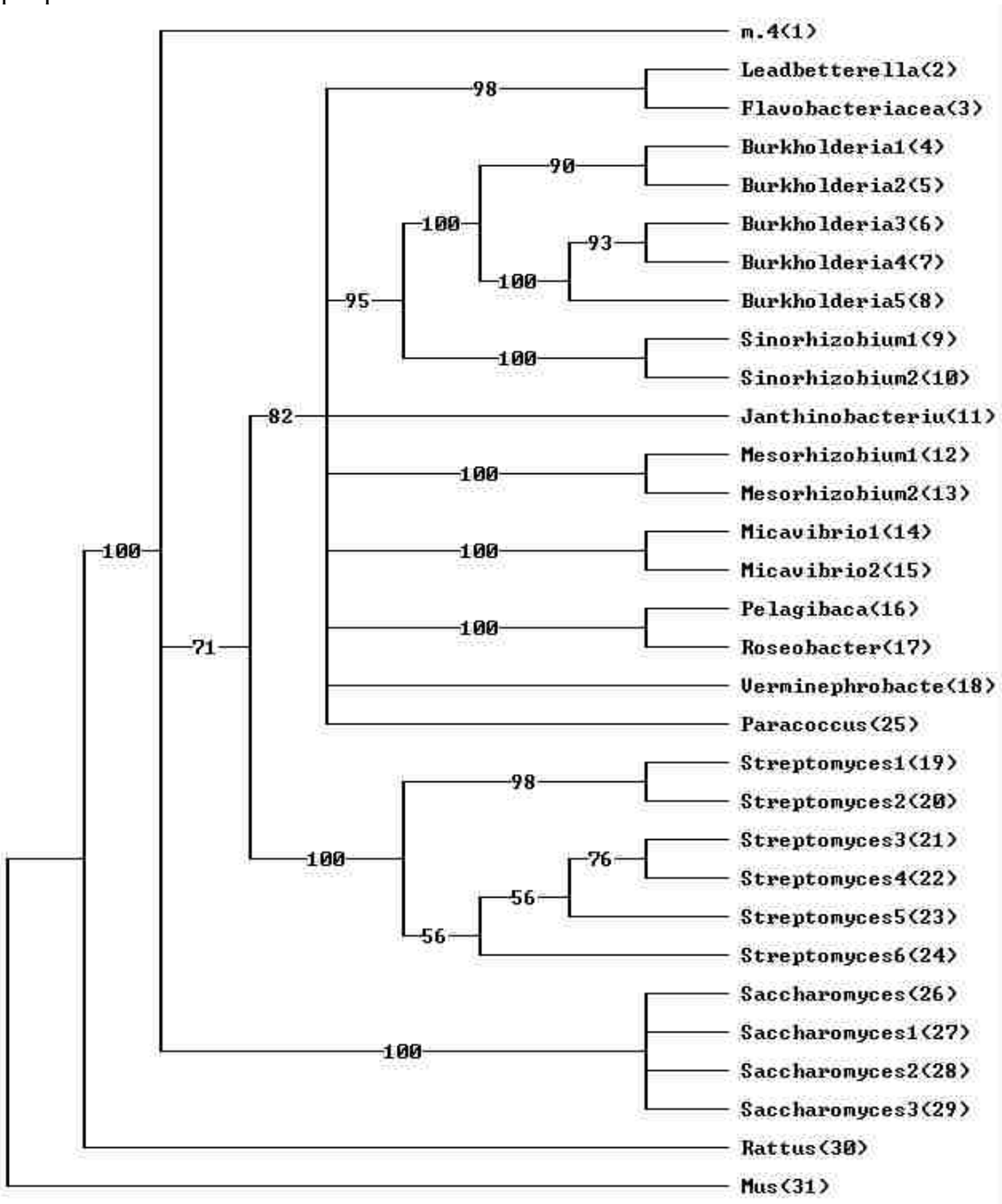
mRNA-seq analyses

Illumina reads from cicada bacteriome and non-bacteriome tissues (SRR952383) were pooled and assembled using TRINITY (25January2012 release) using kmer_ length = 25 and min_contig_length = 48 (Grabherr et al. 2011). The edgeR package was used to analyze differential expression with RSEM quatification and bowtie alignments (Robinson et al. 2010). Assembled transcripts belonging to *Sulcia* and *Hodgkinia* were removed by mapping with bwa-mem v07.5a-2. Resulting sam files were visualized in Tablet v1.14.04.10 to ensure correct mapping. The remaining transcripts were annotated using Trinotate (10November2013 release) and linked to differentially expressed genes with custom Perl scripts. De novo assembled transcripts were also searched for tRNA genes using tFIND v1.4 (Hudson and Williams 2015).

Supplementary figure 1. Maximum-likelihood tree for HGT candidate m.2. *Danaus plexippus* is the American monarch butterfly, gene GI:357606220.

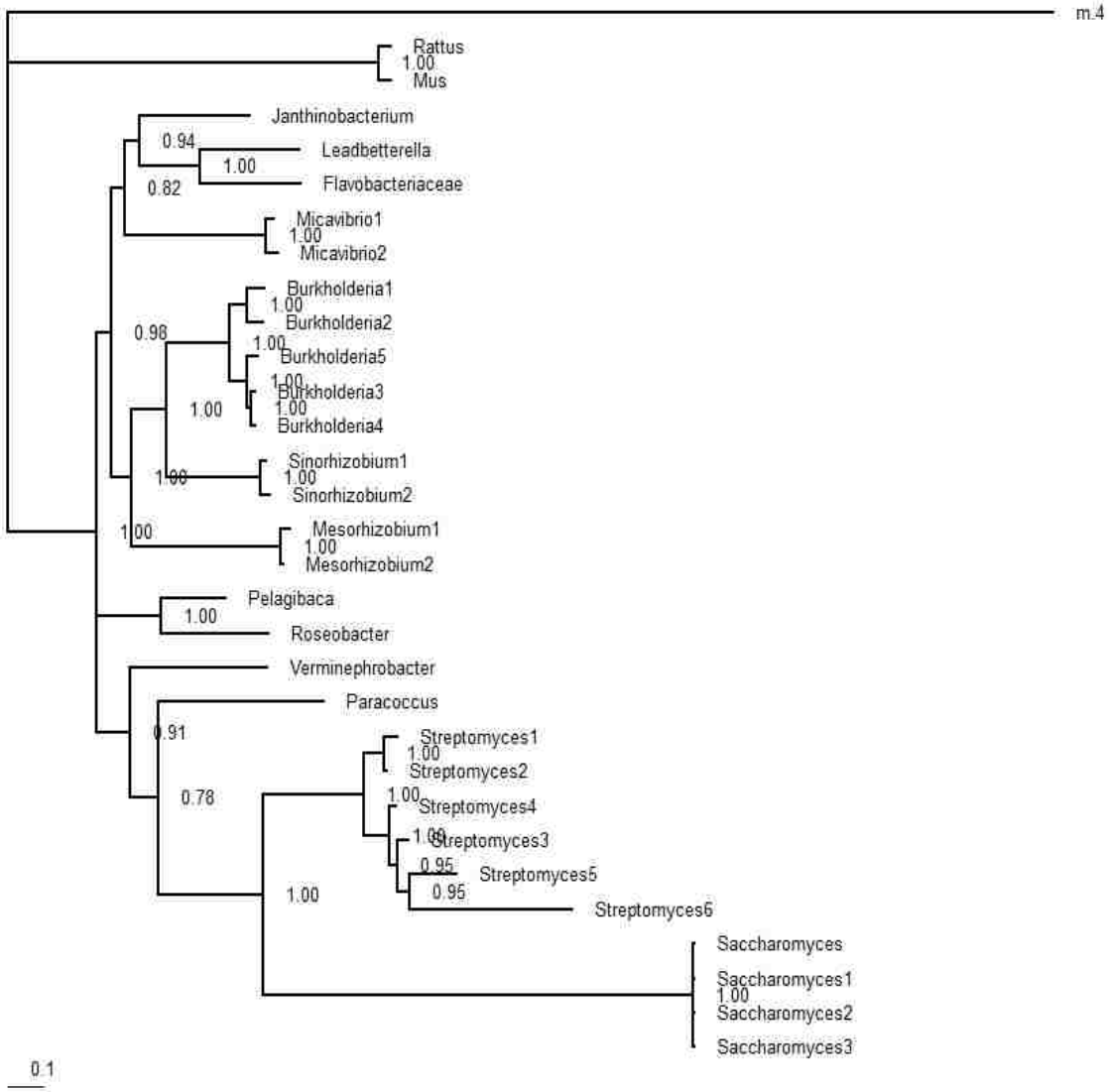


Supplementary figure 2. Unweighted parsimony tree for HGT candidate m.4, with 1000 bootstrap replicates.

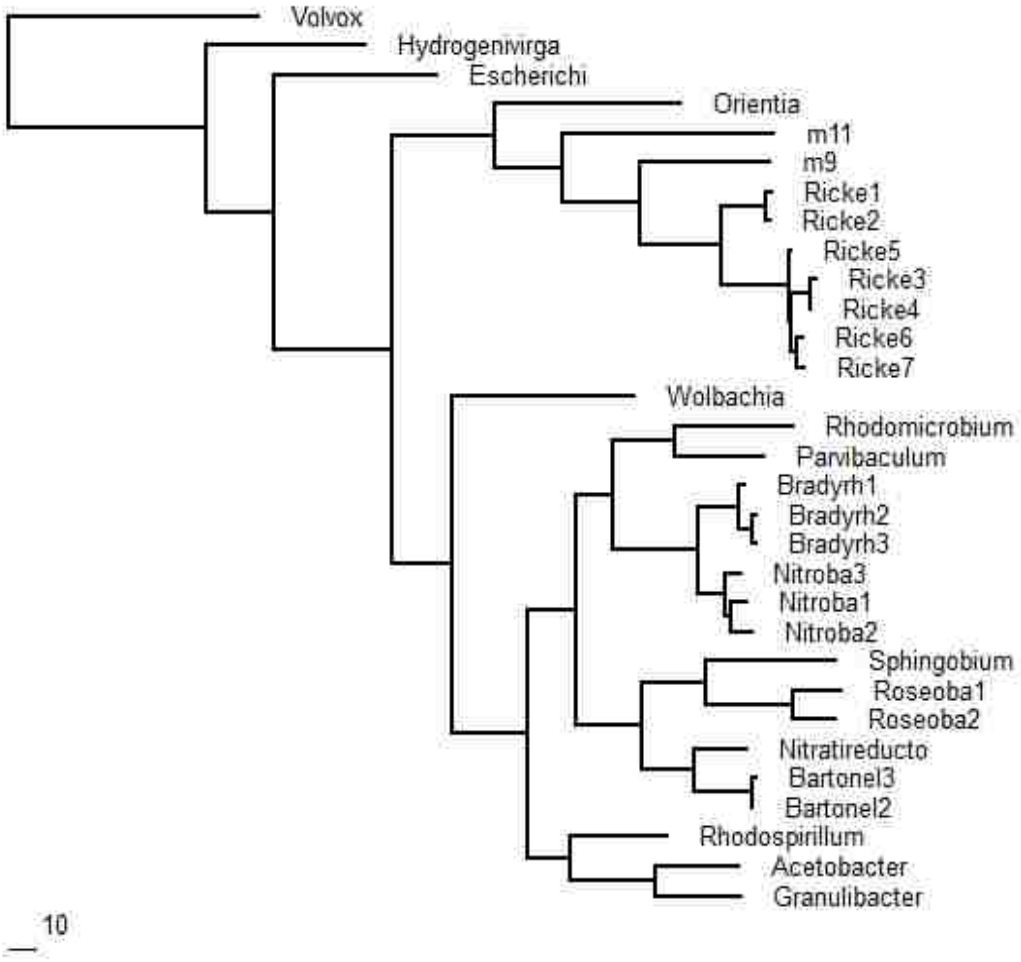


Supplementary figure 3. ML tree for HGT candidate m.4, with 100 bootstrap replicates.

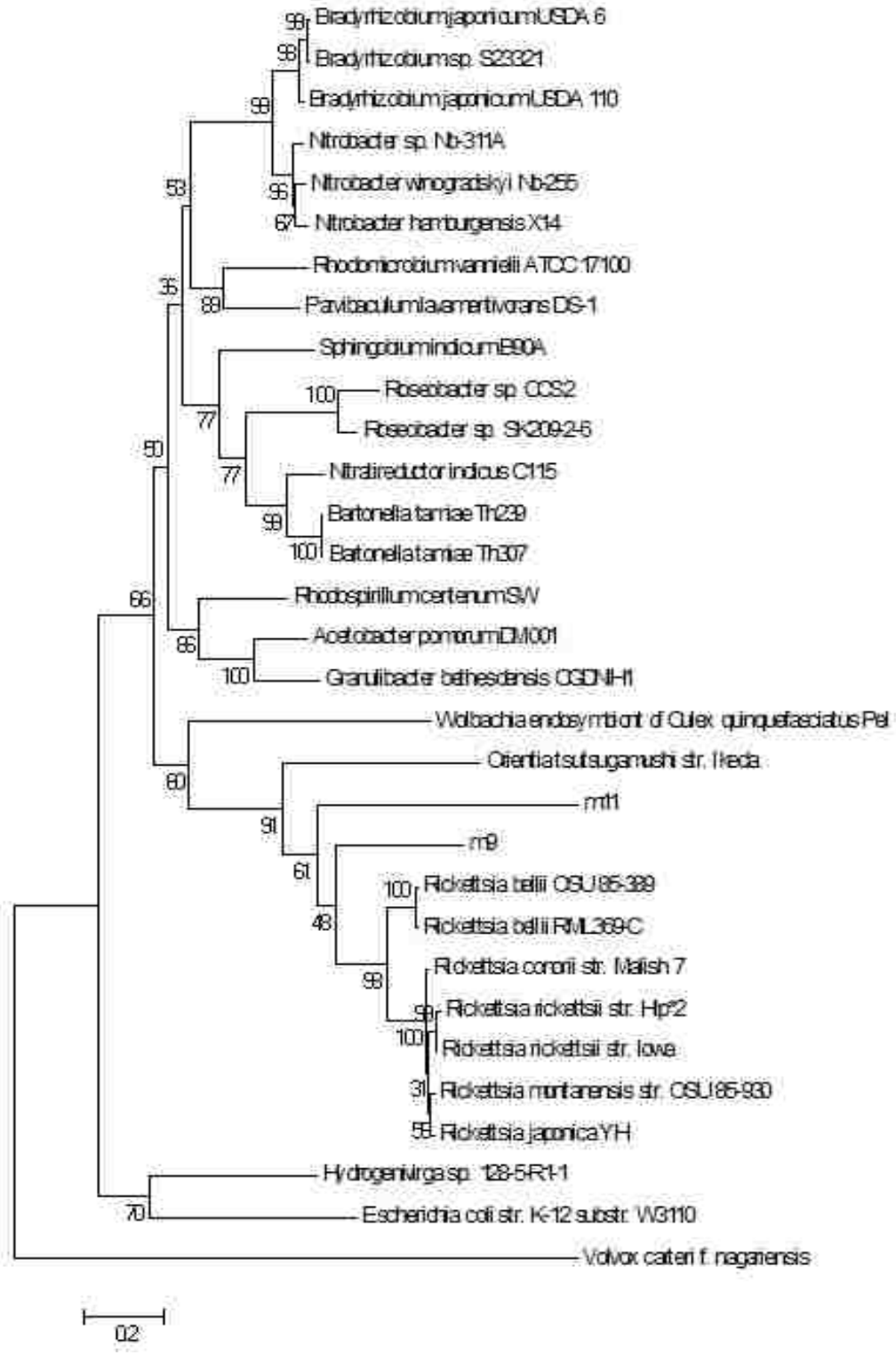
Supplementary figure 4. Baysien tree for HGT candidate m.4.



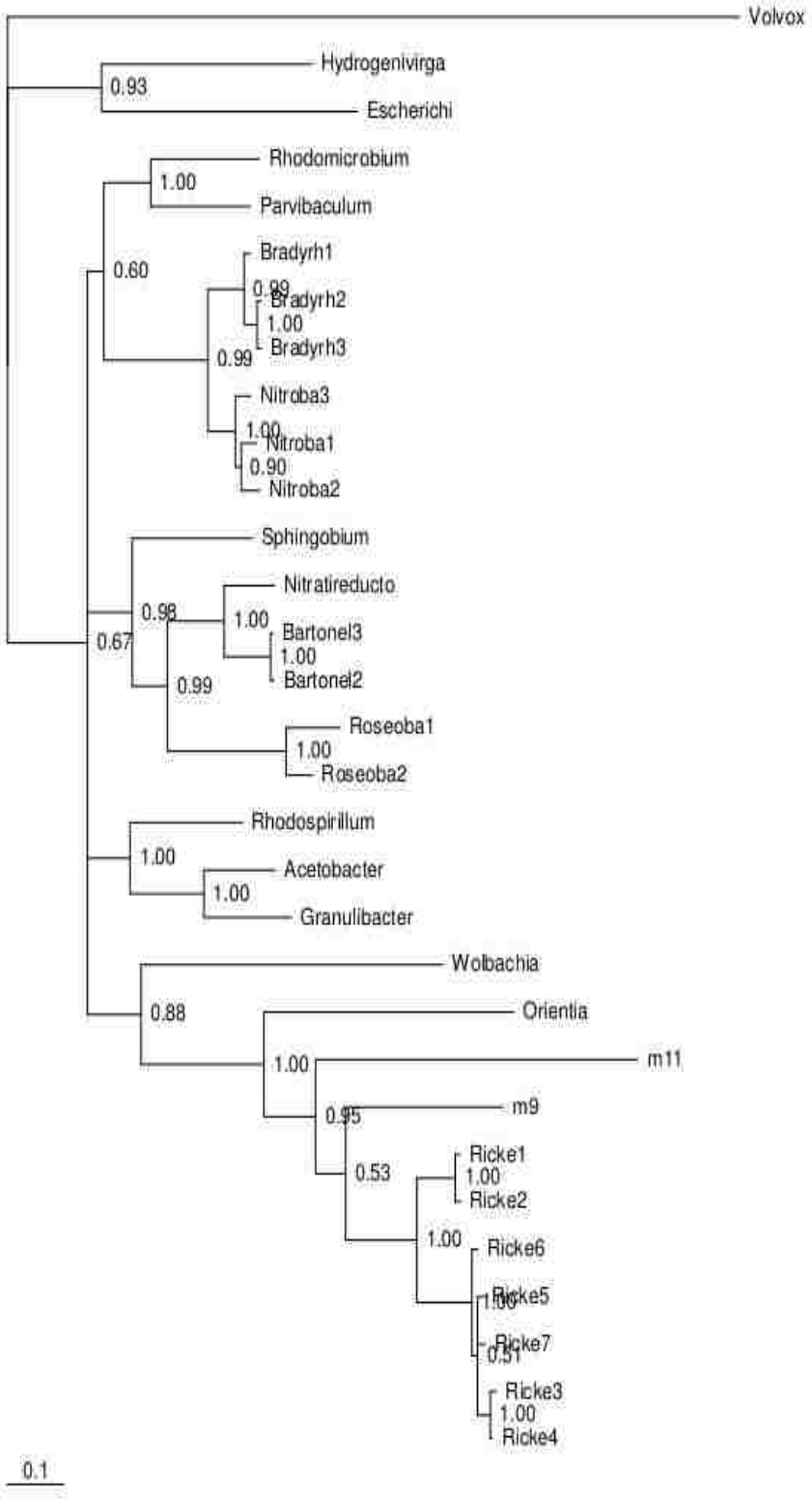
Supplementary figure 5. Unweighted parsimony tree for HGT candidate m.9/m.11.



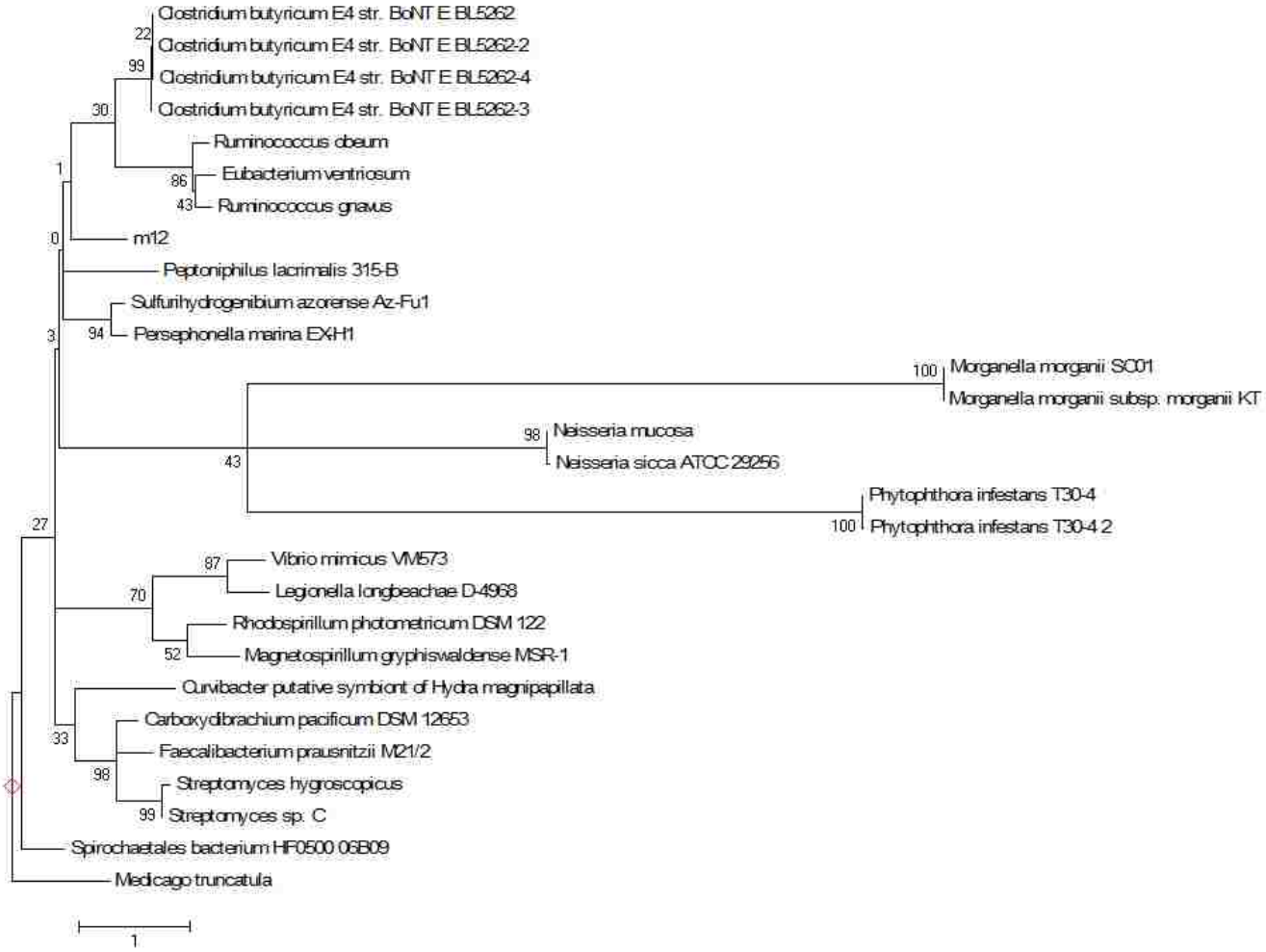
Supplementary figure 6. ML tree for HGT candidate m.9/m.11, with 100 bootstrap replicates.



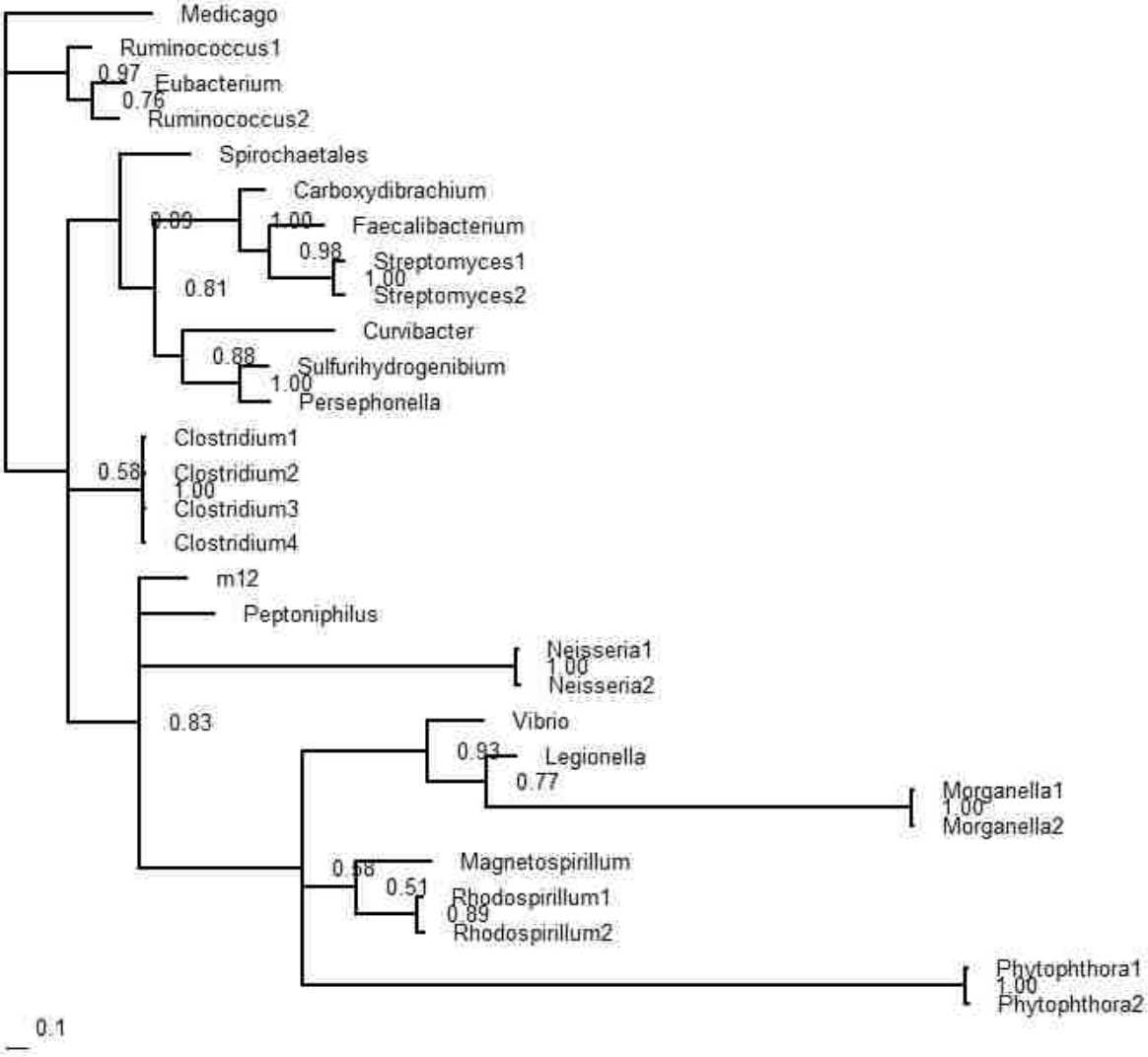
Supplementary figure 7. Baysien tree for HGT candidate m.9/m.11.



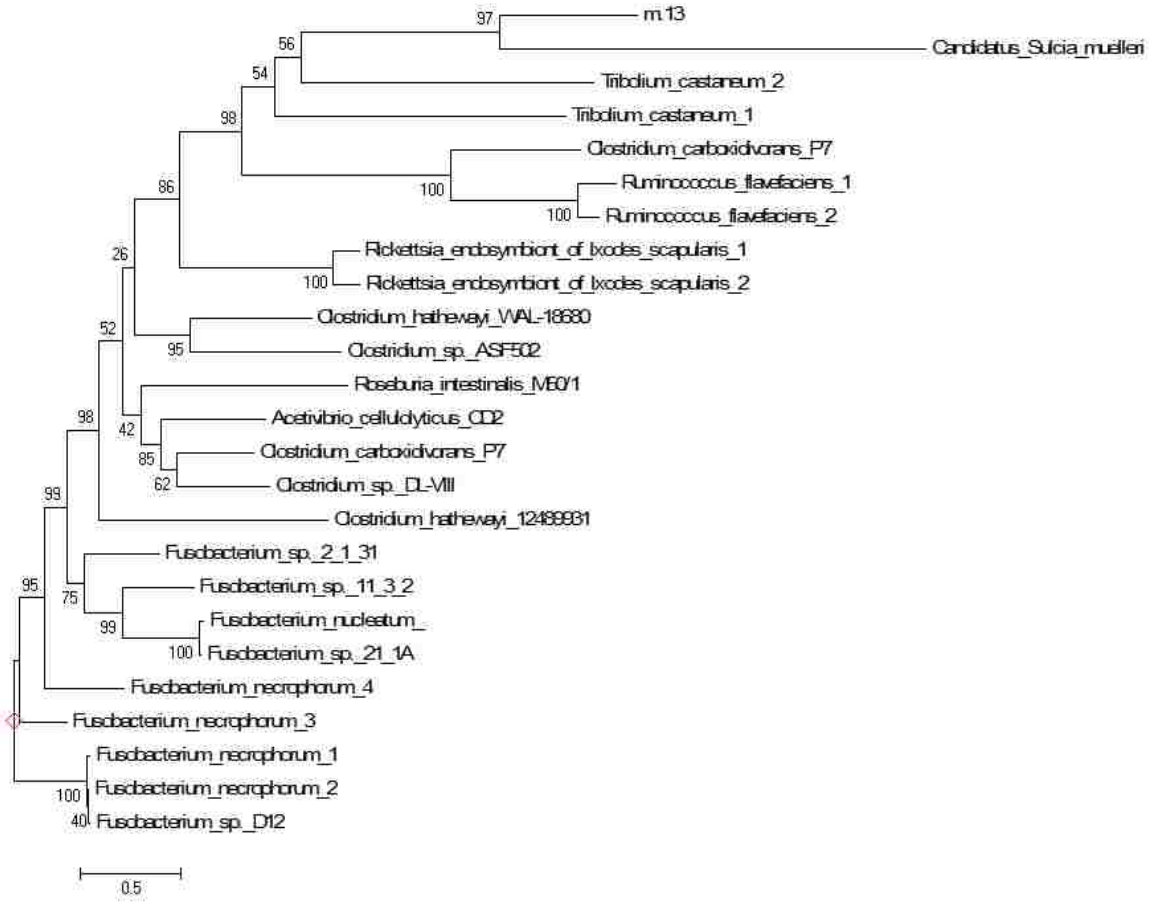
Supplementary figure 8. ML tree for HGT candidate m.12, with 100 bootstrap replicates.



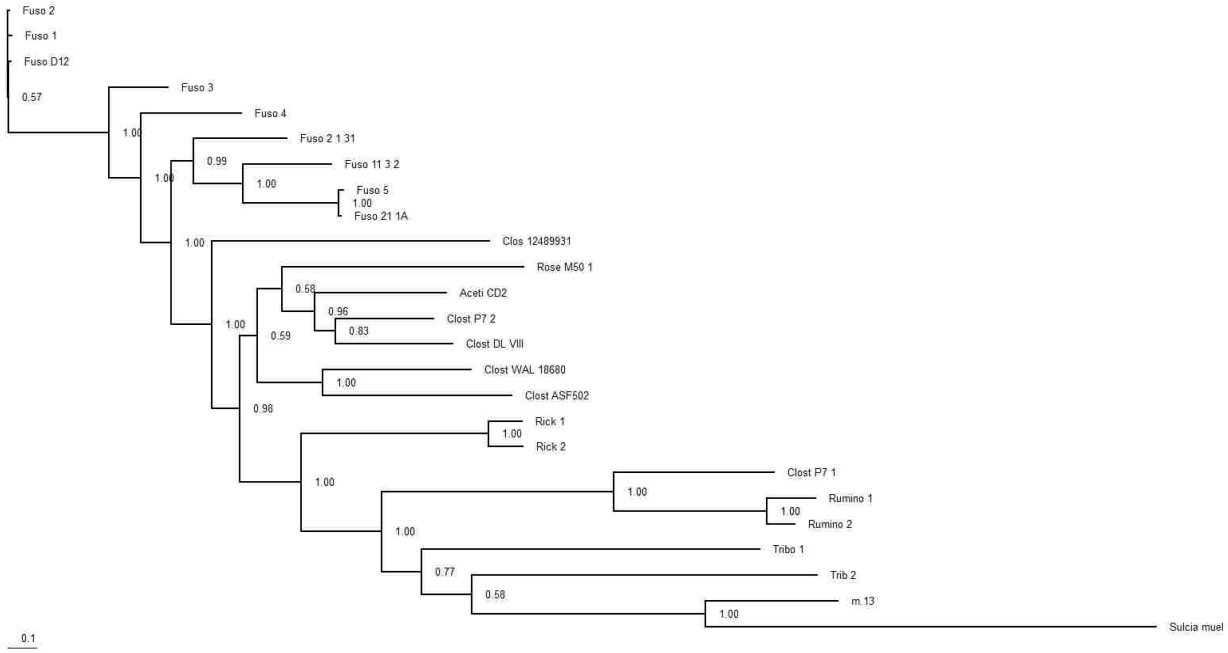
Supplementary figure 9. Baysien tree for HGT candidate m.12.



Supplementary figure 10. ML tree for HGT candidate m.13, with 100 bootstrap replicates.



Supplementary figure 11. Baysien tree for HGT candidate m.13.



Supplementary figure 12. Western blot on total protein from cicada bacteriomes. Antibodies are listed with expected target protein size.

Supplementary table 1. Amino acid sequences for all HGT candidates.

>m.1 g.1 type:internal len:216
DRTLDESRDDVQLLNNSTSDFKDNSLNKSADEIGEEVGNRSESNSSFSEHCTESVNNSCAIPKNFPIQVSSKITEKLKKFAVGTSKAHAST
DGSCSKTDIESRDTSLNKSDTNVFHIGTSLLNDSTVSTSENEMHKIPEKFGITANKNCEVPVTAGGFSRTVQGAALQKTIKNDPQNESIQSL
TINHNESQEQHNSVDRGTSQKTEIKTSSQLFSE
>m.2 g.2 type:5prime_partial len:161
LGQPLKGYKITMVSVSEPHTLKSSFIVASAIKQLGGDVQCVTDQNWEKVDFIEDLGRFHSLFSDAIVVQGRFHSSLCLFAKGATVPVFSA
DCLRFRPFHGLGALMTIQEYFGGLKNLTLTWIGPVSAMLNTYTFLLPKVGMNIKYNTAPTPVNMHFILHK*
>m.3 g.3 type:5prime_partial len:100
PRAHIDGKVWHLDVGSSHPGAGEGPKGSAVRRVKWYVSWVKYVVRQYGSYLLMGLESDKNLPIVREDLGGLISGIPVIVSFLIGIVLYG
WKATLGQDNC*
>m.4 g.4 type:complete len:638
MGDFITTAGDGLEDLTEGLVWIESHKTVPPLRREDKIVYGFKFDWLNATGMQVVLFLHRCVEELIKNKSLFAQVDMLRSCHQSQSQTEK
NFSNLIQQFHNIIGNIVDNDFTVADTQTRTESHVCTMWMSPLHVLTCLVVPAFSLKVNMFVETSAESSYVFQLFGEICGKIGPYFSVTEQS
AVTLPVTFKMHHGSAHMFVYEDADVHSAVSVIVEYLWNMADQELCELTEVYVQESIHSKFSFLLKSKLAIKAENHKWVKHCSEDMESF
TKYKDYIQCAVSLATSKGMDVWKHWEHSETFVPTVIFGKVERKQTDIPLPVICIDSFRTIEEGILLSEKSQKIRFASIWTESGPTAQYIAGQL
KADLVWVNIYGLFSTKVPFHLTVSQRDCIGCIRGCCVSGQKWFSPIWQSHHVPLGFWKSMRETNDIKNVYKLAEESHMKWGIRSSSESR
CEVLLKIVNSISCNKEQYSELLETHDCIEECVKLLYLFAYKCKEDSESRNVDNMLSITTFRPAGVVTIVCTSQTKTVDYLKLIFGMIAYGNS
VVLFHGKNDKLAECAKAFCQHIDLPKGSVNFLECDHVISAKDCFHGKSYFLQYPYGSHILDHNVIETVFESDVTKFNTTMFRWFTEPKS
VFIPVK*
>m.5 g.5 type:complete len:198
MILQKMSIENREMNNKLENKMEKLNGLEVKIEKLNELETRMMENNNKSKMQLKQSMIEINERLESNKMEIKMEINKVDEKISTLDKKL
DCEIEKLKQDFEDLEKRQQTQQDIVEVINTEVERIKEHQRVHEDTIRGVGVEIGDLKEKLMKNEIRMGTAEGKIEVMETEMKTHTKKME
ILENLNVQRTEESSWTCSRE*
>m.6 g.6 type:complete len:198
MILQKMSIENREMNNKLENKMEKLNGLEVKIEKLNELETRMMENNNKSKMQLKQSMIEINERLESNKMEIKMEINKVDEKISTLDKKL
DCEIEKLKQDFEDLEKRQQTQQDIVEVINTEVERIKEHQRVHEDTIRGVGVEIGDLKEKLMKNEIRMGTAEGKIEVMETEMKTHTKKME
ILENLNVQRTEESSWTCSRE*
>m.7 g.7 type:complete len:198
MILQKMSIENREMNNKLENKMEKLNGLEVKIEKLNELETRMMENNNKSKMQLKQSMIEINERLESNKMEIKMEINKVDEKISTLDKKL
DCEIEKLKQDFEDLEKRQQTQQDIVEVINTEVERIKEHQRVHEDTIRGVGVEIGDLKEKLMKNEIRMGTAEGKIEVMETEMKTHTKKME
ILENLNVQRTEESSWTCSRE*
>m.8 g.8 type:3prime_partial len:148
MLFILSFIFKLMSVNCSLTCSLILSILLFSTLTFKSTSFIMSFFTSILNSLNLLFIWSIFSFNFKLFSRFSALILSMFCSIFSLSPFFIFTIFSRFSLFIF
SIFSRISALIFSLFSFTVLISALIFSMFSRFSTLILFMFSRFSALIL
>m.9 g.9 type:complete len:252
MAGHSKFKNVQHRKGRQDSKRSKLFNKLIREITTAVKTGSTDVRCNPRLRHALIVARSNNLPKERIDRIIKSARESTNSEDYDEVRYEGY
APQGIGIIVEALTDNRHRTASSVRAAFTKYGGSLGETGTVSYMFKRRGIVQYPLKIASKDEILERVLECGALDASSDDVSHIIYTSVENFTK
TVDHFNEKYGPPEESYIGWVPNTTVIIHDKVRAQKLLDLVDLLEDNDDVQRVFGNYELSDAVYEALKNS*
>m.10 g.10 type:5prime_partial len:494
EADAAVTEDENGPSEPDDGETDADTGTDPLTVSSEADAAVTEDEIGSSEPDDGEMDVSIDTDPGRDPLTGCSEADAAVAEDDTGSSDPDD
GDTEASAEADIGTDPLTGSSEADDAVAVGDTDSSDPDDGETDSDTGTDPLTGSSVGDSAVTEDDIGSCDPDTDTDTSVNDVTGTDPLAGS
SEDDTAVTEDDTGSSDPDDGVTDASTDADTGIEPLTYSSEVEMAVIDDNIGSSVPDDADASEETAGDPLAGSSEAEAAVTEADTDSSELED
GDTDCETGRDPLTGSSEAVTAVTESDSGSSEADVGDTEASTDDDTGTDPVTGSSEAEAAVTEGDAGSSVPEDGVREASTDSDTGTDPLTV
SSEAEDERGSSDSGDCDSETTSDVEAILTTDDSDDDETLPCSVAELAVAAPDGCSDVAPSVPDVGPDSEIGSWVVVSSPVGSSDSLEVACD
DSSVDGCGDSVVNVVVFLFFVVRFSSYSELIPNLSFSSEWI*
>m.11 g.11 type:complete len:252
MAGHSKFQNIKFRKERQDKRRSNVFEKLVREISAAAKDGGTDPKSNSRLRHALQKARSQNLPKDKIEKALKKGQDKKDTTYSEERFEA
FIGAGACIIIETLTDNKNRTVGEIRKVFNKNGANLTNAGCVTHKFHRRGIIQFPLSVASAEQMLETAVEAGALDTVSENDVHCIYTEVQDF
WKVLDFMSKTYGDPLESHIGWTPKEYVIIDDKTIAKTALKFVEDLEDLDDVQHVFVNYEITDKVYDALKSNL*
>m.12 g.12 type:internal len:132
FQHRAGVSPYTSPCGFAQTCVFAKQSLGPFHCGPLGLFTLPRHPFSRSYGVILPSSLTRVAPRALECSSCLPVSVSGTGTYDLARGFSWQC
EIMTFATVIFTPHHSPALRLADLPTNQPHCLDEHPSARVTI
>m.13 g.13 type:internal len:155
IIDNLLYKNLCVERAFLTGTLPLIVESEYARHGSYIHIYSFMDSHYLSKYYGLSSRNFEKILSLIIHDDAEKNVARGAIDEFYSGYVTGSHSI
HLCNTWSVLHYLHRGKARCYWSGCERLQRLEPFFKNSQIRDSIEKLLLGESQLVDRFHELSS

Supplementary table 2. Similarity search between all HGT candidates using blastP. Only m11 and m9 were combined for phylogenetic analyses. Candidates in gray cells have no significant blast hits in the protein database.

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13
m1	4e-126	-	-	-	-	-	-	-	-	3.7	-	-	-

m2	-	1e-95	-	6.8	-	-	-	-	3.5	-	0.26	-	-
m3	-	-	1e-45	-	-	-	-	-	-	0.49	-	-	-
m4	-	-	-	0	-	-	-	-	-	-	-	-	-
m5	-	-	-	-	1e-94	1e-94	1e-94	-	-	-	-	-	-
m6	-	-	-	-	1e-94	1e-94	1e-94	-	-	-	-	-	-
m7	-	-	-	-	1e-94	1e-94	1e-94	-	-	-	-	-	-
m8	-	-	-	2.9	-	-	-	2e-31	-	-	-	-	-
m9	-	-	-	4.1	-	-	-	-	7e-142	-	7e-68	3.4	-
m10	-	-	1.8	-	-	-	-	-	-	5e-158	-	-	-
m11	-	0.43	-	6.3	-	-	-	-	7e-68	6.9	6e-152	-	3.4
m12	-	-	-	-	-	-	-	-	1.6	-	-	3e-76	-
m13	-	-	-	-	-	-	-	-	-	-	2.0	-	5e-91

5.8 References

Braendle C, Miura T, Bickel R, Shingleton AW, Kambhampati S, Stern DL. 2003. Developmental Origin and Evolution of Bacteriocytes in the Aphid–Buchnera Symbiosis. PLoS Biol 1:e21.

Campbell MA, Van Leuven JT, Meister RC, Carey KM, Simon C, McCutcheon JP. 2015. Genome expansion via lineage splitting and genome reduction in the cicada endosymbiont Hodgkinia. Proc. Natl. Acad. Sci. 112:10192–10199.

Duncan RP, Husnik F, Van Leuven JT, Gilbert DG, Dávalos LM, McCutcheon JP, Wilson ACC. 2014. Dynamic recruitment of amino acid transporters to the insect/symbiont interface. Mol. Ecol. 23:1608–1623.

Gagnon P, Nian R, Lee J, Tan L, Latiff SMA, Lim CL, Chuah C, Bi X, Yang Y, Zhang W, et al. 2014. Nonspecific interactions of chromatin with immunoglobulin G and protein A, and their impact on purification performance. J. Chromatogr. A 1340:68–78.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, 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.

Hansen AK, Moran NA. 2011. Aphid genome expression reveals host–symbiont cooperation in the production of amino acids. Proc. Natl. Acad. Sci. 108:2849–2854.

Hudson CM, Williams KP. 2015. The tmRNA website. Nucleic Acids Res. 43:D138–D140.

Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, Tanaka M, Satoh N, Bachtrog D, Wilson ACC, et al. 2013. Horizontal Gene Transfer from Diverse Bacteria to an Insect Genome Enables a Tripartite Nested Mealybug Symbiosis. Cell 153:1567–1578.

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. 9:605–618.

Van Leuven JT, Meister RC, Simon C, McCutcheon JP. 2014. Sympatric Speciation in a Bacterial Endosymbiont Results in Two Genomes with the Functionality of One. Cell 158:1270–1280.

Luan J-B, Chen W, Hasegawa DK, Simmons AM, Wintermantel WM, Ling K-S, Fei Z, Liu S-S, Douglas AE. 2015. Metabolic Coevolution in the Bacterial Symbiosis of Whiteflies and Related Plant Sap-Feeding Insects. Genome Biol. Evol. 7:2635–2647.

Macdonald SJ, Lin GG, Russell CW, Thomas GH, Douglas AE. 2012. The central role of the host cell in symbiotic nitrogen metabolism. Proc. R. Soc. B Biol. Sci. 279:2965–2973.

McCutcheon JP, von Dohlen CD. 2011. An Interdependent Metabolic Patchwork in the Nested Symbiosis of Mealybugs. Curr. Biol. 21:1366–1372.

McCutcheon JP, McDonald BR, Moran NA. 2009. Origin of an Alternative Genetic Code in the Extremely Small and GC–Rich Genome of a Bacterial Symbiont. PLoS Genet 5:e1000565.

McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. 10:13–26.

Moran NA, Bennett GM. 2014. The Tiniest Tiny Genomes. Annu. Rev. Microbiol. 68:195–215.

Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. Annu. Rev. Genet. 42:165–190.

Nakabachi A, Ishida K, Hongoh Y, Ohkuma M, Miyagishima S. 2014. Aphid gene of bacterial origin encodes a protein transported to an obligate endosymbiont. Curr. Biol. 24:R640–R641.

Nikoh N, McCutcheon JP, Kudo T, Miyagishima S, Moran NA, Nakabachi A. 2010. Bacterial Genes in the Aphid Genome: Absence of Functional Gene Transfer from Buchnera to Its Host. PLoS Genet 6:e1000827.

Nowack ECM, Grossman AR. 2012. Trafficking of protein into the recently established photosynthetic organelles of Paulinella chromatophora. Proc. Natl. Acad. Sci. 109:5340–5345.

Poliakov A, Russell CW, Ponnala L, Hoops HJ, Sun Q, Douglas AE, van Wijk KJ. 2011. Large-

Scale Label-Free Quantitative Proteomics of the Pea aphid-Buchnera Symbiosis. Mol. Cell. Proteomics 10:M110.007039.

Price DRG, Duncan RP, Shigenobu S, Wilson ACC. 2011. Genome Expansion and Differential Expression of Amino Acid Transporters at the Aphid/Buchnera Symbiotic Interface. Mol. Biol. Evol. 28:3113–3126.

Price DRG, Feng H, Baker JD, Bavan S, Luetje CW, Wilson ACC. 2014. Aphid amino acid transporter regulates glutamine supply to intracellular bacterial symbionts. Proc. Natl. Acad. Sci. 111:320–325.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinforma. Oxf. Engl. 26:139–140.

Sloan DB, Nakabachi A, Richards S, Qu J, Murali SC, Gibbs RA, Moran NA. 2014. Parallel Histories of Horizontal Gene Transfer Facilitated Extreme Reduction of Endosymbiont Genomes in Sap-Feeding Insects. Mol. Biol. Evol. 31:857–871.

Wigglesworth VB. 1967. Polyploidy and Nuclear Fusion in the Fat Body of Rhodnius(Hemiptera). J. Cell Sci. 2:603–616.

Chapter 6: General conclusions and future outlook

My goal in writing this chapter is to summarize my contributions to symbiosis research, to reflect upon unpublished and underdeveloped results, and to put into words what I think is one of the most important results of insect symbiosis research—to shed light on the process of host cell integration.

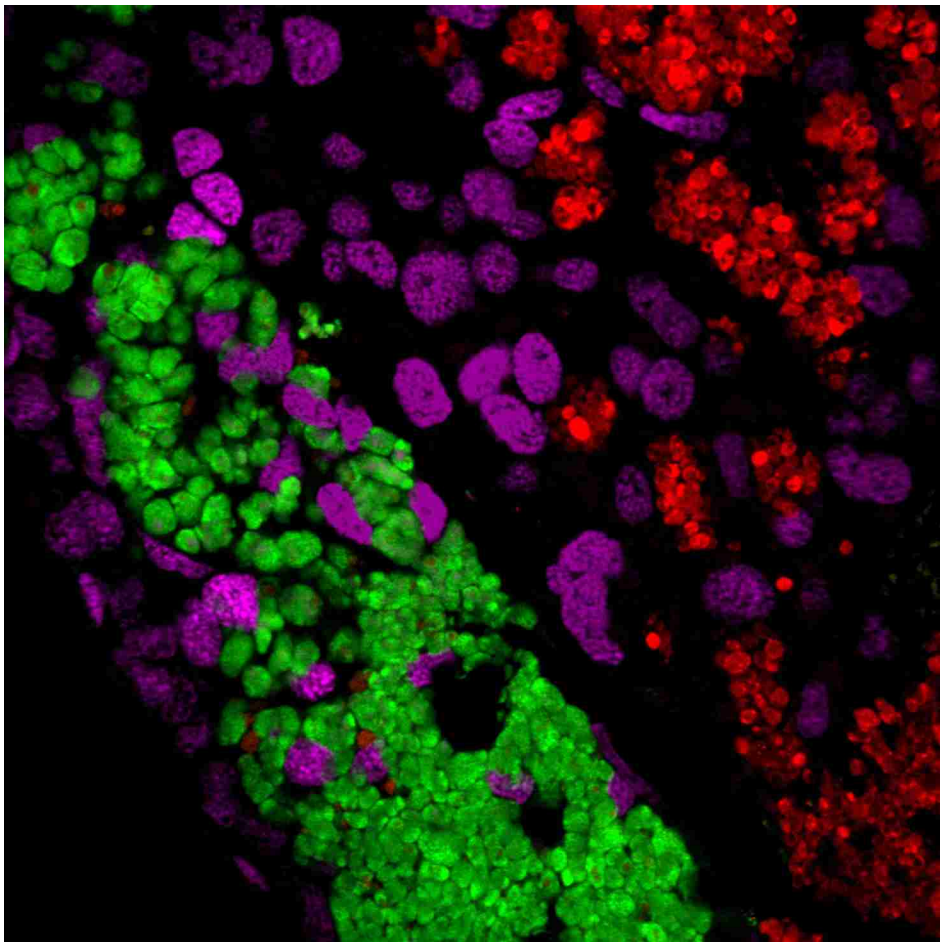


Image by James Van Leuven
appears in
Kiers ET, West SA. 2015. Science 348:392–394.

6.1 Recent advances in understanding insect nutritional endosymbionts

In 2010, when I started my PhD, a single cicada metagenome was sequenced (McCutcheon et al. 2009). We have since published three more and we have about a dozen that are complete enough to understand the basic genome structure and evolution of *Hodgkinia* and *Sulcia* in these cicada species. This set of *Hodgkinia* genomes has changed the way we think about endosymbiont genome evolution. Beforehand, endosymbiont genomes were though to be static in structure, but rapidly evolving in nucleotide sequence. Over the past 5 years, this picture has changed quite a bit. The few exceptions to this rule provided subtle hints of an alternative viewpoint; *Tremblaya* has a small plasmid containing one gene and an inversion that exists in both the inverted and non-inverted conformations, *Portiera* has a ~6.5kb fragment that is sub-genomic, missing from the main chromosome, or present in 1-3 tandem copies. As review in Sloan and Moran 2013, a handful of other endosymbiont genomes show similar, minor structural variations. The *Hodgkinia* genomes, however, have revealed incredible genome complexity. While the structural diversity that we observe seems primarily driven by only genome reduction, there is certainly some level of recombination occurring within a cicada host. Given these data, we must recognize the potential for genome structural variation to occur, despite the fact that the *Hodgkinia* genomes are missing recombinational genes. Currently, the limited evidence we have suggests that a combination of relaxed selection and severe generational bottlenecking contributes to the fixation of genomes with complementary inactivating mutations. We proposed this idea because of the direct relationship observed between the complexity of the *Hodgkinia* genome and the length of time between cicada generations (when host fitness is tested).

Regardless of what causes genome fragmentation, the events create a powerful system for studying genome evolution. Unpublished *Hodgkinia* genome data from other *Tettigades* cicadas are poised to be most informative in this pursuit. While the *Hodgkinia* TETAUR genome presented in chapter 3 was portrayed as a two variant genome complex like TETUND, the true *Hodgkinia* genome structure in TETAUR lies somewhere between *Hodgkinia* TETUND and *Hodgkinia* MAGTRE. Most *Tettigades* lineages surveyed so far contain 5-6 *Hodgkinia* circular molecules, where each chromosome falls on a spectrum of degradation. Some circular molecules retain the majority of the genes encoded on the single ancestral version, but some are highly degenerate with only a few genes remaining. We can see that lineage-splitting is not a rare occurrence; within the *Hodgkinias* in the *Tettigades* clade, we see multiple independent origins of new pairs of circular molecules from the single version ancestor. Gene complementarity between the circular molecules is evident, but the distribution of genes on the molecules seems random. It does not matter what gene copy is retained where, so long as one is encoded somewhere in the complex of genomes. These recurring gene losses reveal the variation in substitution rates across the genome, and even between pseudogenes. It is generally observed that a gene fated to be pseudogenized experiences statistically increased rates of substitution prior to a frame shift on the 3' part of the gene.

In many cases, multiple copies of the same gene are retained in *Hodgkinia* genome complexes. Under the model of reductive evolution that we proposed, this redundancy is unnecessary. And while we see evidence that some of these redundant copies are in the process of being purged from the genome, the high conservation of multiple copies is perplexing. We theorize that there is a selective advantage via dosage effect to keep some gene duplicates. The

likelihood of this depends on two things: the selective advantage of the increased dosage, and the factors limiting the efficiency of transcription/translation. Between species measures of selection show that the *Hodgkinia* genome is evolving under very weak purifying selection (Van Leuven et al. 2014). Also, the loss of important genes, the rapid rate of sequence evolution, and the high expression of chaperone proteins tell the same story (McCutcheon et al. 2009). However, our population polymorphism data shows that selection is purging mutations from *Hodgkinia* populations, and we observe a conserved frequency of *Hodgkinia* genome circular molecules between individual cicadas; all indicative of purifying selection (Van Leuven and McCutcheon 2012; Campbell et al. 2015). Selection seems only able to act on traits that have a large impact on the fitness of *Hodgkinia* and the host cicada. This suggests that there is a very high benefit to retaining redundant gene copies.

The other factor influencing the importance of gene duplicate retention is if selection can actually see the effects of gene duplication above all the background evolutionary "noise" present in *Hodgkinia*. There is no codon bias in *Hodgkinia*, there are no recognizable promoters, there is only one sigma factor, there are only 13 tRNAs, and the ribosome is missing about a dozen ribosomal proteins. To me, this suggests that transcription and translation in *Hodgkinia* are not working very well. Is cellular transcription and translation in *Hodgkinia* really precise enough so that a gene duplicate truly results in a 2-fold increase in protein abundance? Are mRNA and protein abundances consistent between *Hodgkinia* cells? If so, perhaps lineage splitting is *Hodgkinia's* attempt to control protein expression without needing to retain the mechanisms to regulate transcription and translation. Alternative mechanisms of translational control were found in *Buchnera*, where small, interfering RNAs likely alter protein levels (Hansen and Degnan 2014).

My overall view of *Hodgkinia* evolution steers me towards another explanation, where the *Hodgkinia* genome is just falling apart and the host is doing what it can to avoid extinction. The presence of gene copies indicates that these genes are important and are tenaciously resisting inactivating mutations, but given enough time it seems that these mutations will become fixed. What does the end game look like for *Hodgkinia* and other insect endosymbionts undergoing severe genome reduction? The frequency of replacements in various insect lineages suggests that it will be replaced by a bacterium with a larger genome, but at what point does this happen? What does the *Hodgkinia* genome look like when this happens? As we have found cicada species apparently lacking *Hodgkinia* species, perhaps we will find out.

6.2 Endosymbionts and organelles: convergent reduction evolution

The first organelle genome sequenced was from humans (Anderson et al. 1981). Since that first example, mitochondrial genome sequencing efforts have been bias towards animals, which all have ~14kb genome with essentially the same gene content. Similarly, the first and most commonly sequenced chloroplast genomes are from green plants (Ohyama et al. 1986). However, like with insect endosymbionts, the sequencing of many organelle genomes among diverse eukaryotes is beginning to reveal a complex picture of organelle evolution (Burger et al. 2003; Smith and Keeling 2015). In fact, more and more parallels between endosymbionts and organelles arise as additional sequences become available. Both endosymbionts and organelles have undergone genome reduction. Both have horizontally transferred genes to the host genome.

Both have host proteins localized in their cells. Both are reliant on the host for translation and transcription, but generally conserve the most crucial components in their own genomes. And lastly, we now know that both can experience secondary (after genome reduction) and massive variation in genome structure. In the case of organelles, it is clear how this variation arises: recombination between mitochondria (B. Wu et al. 2015), horizontal gene transfer via chloroplast fusion (Rice et al. 2013), and the gain and/or loss of entire mitochondrial chromosomes (Z. Wu et al. 2015). This differs somewhat from the process that leads to genome complexity in *Hodgkinia*, which seems to results from lineage-splitting followed by genome reduction (Van Leuven et al. 2014; Campbell et al. 2015). Perhaps through understanding the selective pressures that influence genome structure variation in organelles, we can better understand the evolution of *Hodgkinia* (Piganeau and Eyre-Walker 2009; Sloan et al. 2012; Cooper et al. 2015).

Certainly, there remain differences between organelles and endosymbionts. Primarily, organelles are distributed in most cells of an organism, while endosymbionts are not, and organelle genes are almost entirely encoded for in the host genome, while endosymbiont genes are not. This last point raises some interesting questions on the process of endosymbiont-host integration and the formation of organelles. Eukaryotes arose 1-2 billion years ago, likely from the fusion of an ancient archaeal cell belonging to the TACK superphylum with an alphaproteobacterium from the group Rickettsiales (Gray 2012; Williams et al. 2012; Eme et al. 2014; Martin et al. 2015). However, the details of this event remain unclear. One main point of contention arises from the phylogenetic discordance of many of the nucleus encoded, mitochondrial genes. Of the hundreds to thousands of mitochondrial genes in the nuclear genome only 10-20% can be definitively classified at alphaproteobacterial, suggesting that they are either from different bacterial donors, or were already in the proto-mitochondria endosymbiont genome at the time of endosymbiosis. An even smaller proportion of the proteins that are localized to mitochondria are alphaproteobacterial, the remainder being comprised of genes from diverse prokaryotic lineages, or entirely unique to eukaryotes. In plants, some of these proteins are dually targeted, functioning in both mitochondria and chloroplasts. The phylogenetic diversity of the entire mitochondrial proteome, combined with an inability to confidently confine all mitochondrial genes to a single bacterial progenitor has lead to the pre-mitochondrial hypothesis, where eukaryotes were formed during a series of associations between the pre-eukaryote and many transitional bacterial symbionts (Gray 2014). During these numerous associations, horizontal gene transfer and adaptation occurs that eventually facilitated the final endosymbiotic event with the proto-mitochondria. Given that these transfers would have occurred millions or billions of years ago, the phylogenetic signal needed to distinguish the pre-endosybiont theory from the one-time event would have been lost (Groussin et al. 2015; Ku et al. 2015). However, evidence from plastids lend support to the pre-mitochondria theory, where a complex mosaic of genetic material resulted from primary, secondary, and tertiary endosymbiotic events (Keeling 2010; Curtis et al. 2012). The recurring integration of the nuclear, mitochondrial, and plastid genomes suggests that it is just not that hard to form new, intimate symbiosis that are metabolically and genetically dependent on one another (Larkum et al. 2007). Insect-bacterial symbiosis, I think, also provide insight into how mitochondria may have formed over a billion years ago. Although the cellular mechanisms of nutrient transfer are not generally known for insect-bacterial symbiosis, the genetic evidence for cellular integration is certain there, especially

for *Hodgkinia*, where so much of the translational machinery is almost certainly missing (chapters 4 and 5). It is clear that horizontally transferred genes from other bacteria to the insect host support nutritional endosymbionts (Nikoh et al. 2010; Husnik et al. 2013; Sloan et al. 2014; Luan et al. 2015). However, the degraded state of the *Hodgkinia*, *Tremblaya*, and *Zinderia* raises the question of how long the symbiosis can continue without the replacement of these bacteria, as has so often occurred in the history of hemipterians (Koga et al. 2013; Bennett and Moran 2015). Given similar patterns of genome reduction in organelles and nutritional endosymbionts, the high amount of endosymbiont to host HGT that has occurred despite insects being multicelluar organisms, and the frequent turnover of insect endosymbionts in hemipterians, it is compelling to think that a process similar to what is happening now in sap-feeding insects occurred billions of years ago during the origin of eukaryotes.

References

Anderson S, Bankier AT, Barrell BG, Bruijn MHL de, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, et al. 1981. Sequence and organization of the human mitochondrial genome. Publ. Online 09 April 1981 Doi101038290457a0 290:457–465.

Bennett GM, Moran NA. 2015. Heritable symbiosis: The advantages and perils of an evolutionary rabbit hole. Proc. Natl. Acad. Sci.:201421388.

Burger G, Gray MW, Franz Lang B. 2003. Mitochondrial genomes: anything goes. Trends Genet. 19:709–716.

Campbell MA, Van Leuven JT, Meister RC, Carey KM, Simon C, McCutcheon JP. 2015. Genome expansion via lineage splitting and genome reduction in the cicada endosymbiont Hodgkinia. Proc. Natl. Acad. Sci. 112:10192–10199.

Cooper BS, Burrus CR, Ji C, Hahn MW, Montooth KL. 2015. Similar Efficacies of Selection Shape Mitochondrial and Nuclear Genes in Both Drosophila melanogaster and Homo sapiens. G3 GenesGenomesGenetics 5:2165–2176.

Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, Arias MC, Ball SG, Gile GH, Hirakawa Y, et al. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature 492:59–65.

Eme L, Sharpe SC, Brown MW, Roger AJ. 2014. On the Age of Eukaryotes: Evaluating Evidence from Fossils and Molecular Clocks. Cold Spring Harb. Perspect. Biol. 6:a016139.

Gray MW. 2012. Mitochondrial Evolution. Cold Spring Harb. Perspect. Biol. 4:a011403.

Gray MW. 2014. The Pre-Endosymbiont Hypothesis: A New Perspective on the Origin and Evolution of Mitochondria. Cold Spring Harb. Perspect. Biol. 6:a016097.

Groussin M, Boussau B, Szöllősi G, Eme L, Gouy M, Brochier-Armanet C, Daubin V. 2015.

Gene acquisitions from bacteria at the origins of major archaeal clades are vastly overestimated. Mol. Biol. Evol.:msv249.

Hansen AK, Degnan PH. 2014. Widespread expression of conserved small RNAs in small symbiont genomes. ISME J. 8:2490–2502.

Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, Tanaka M, Satoh N, Bachtrog D, Wilson ACC, et al. 2013. Horizontal Gene Transfer from Diverse Bacteria to an Insect Genome Enables a Tripartite Nested Mealybug Symbiosis. Cell 153:1567–1578.

Keeling PJ. 2010. The endosymbiotic origin, diversification and fate of plastids. Philos. Trans. R. Soc. B Biol. Sci. 365:729–748.

Koga R, Bennett GM, Cryan JR, Moran NA. 2013. Evolutionary replacement of obligate symbionts in an ancient and diverse insect lineage. Environ. Microbiol.:n/a – n/a.

Ku C, Nelson-Sathi S, Roettger M, Garg S, Hazkani-Covo E, Martin WF. 2015. Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes. Proc. Natl. Acad. Sci.:201421385.

Larkum AWD, Lockhart PJ, Howe CJ. 2007. Shopping for plastids. Trends Plant Sci. 12:189–195.

Van Leuven JT, McCutcheon JP. 2012. An AT Mutational Bias in the Tiny GC-Rich Endosymbiont Genome of Hodgkinia. Genome Biol. Evol. 4:24–27.

Van Leuven JT, Meister RC, Simon C, McCutcheon JP. 2014. Sympatric Speciation in a Bacterial Endosymbiont Results in Two Genomes with the Functionality of One. Cell 158:1270–1280.

Luan J-B, Chen W, Hasegawa DK, Simmons AM, Wintermantel WM, Ling K-S, Fei Z, Liu S-S, Douglas AE. 2015. Metabolic Coevolution in the Bacterial Symbiosis of Whiteflies and Related Plant Sap-Feeding Insects. Genome Biol. Evol. 7:2635–2647.

Martin WF, Garg S, Zimorski V. 2015. Endosymbiotic theories for eukaryote origin. Phil Trans R Soc B 370:20140330.

McCutcheon JP, McDonald BR, Moran NA. 2009. Origin of an Alternative Genetic Code in the Extremely Small and GC–Rich Genome of a Bacterial Symbiont. PLoS Genet 5:e1000565.

Nikoh N, McCutcheon JP, Kudo T, Miyagishima S, Moran NA, Nakabachi A. 2010. Bacterial Genes in the Aphid Genome: Absence of Functional Gene Transfer from Buchnera to Its Host. PLoS Genet 6:e1000827.

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M,

Chang Z, et al. 1986. Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. Nature 322:572–574.

Piganeau G, Eyre-Walker A. 2009. Evidence for Variation in the Effective Population Size of Animal Mitochondrial DNA. PLoS ONE 4:e4396.

Rice DW, Alverson AJ, Richardson AO, Young GJ, Sanchez-Puerta MV, Munzinger J, Barry K, Boore JL, Zhang Y, dePamphilis CW, et al. 2013. Horizontal Transfer of Entire Genomes via Mitochondrial Fusion in the Angiosperm Amborella. Science 342:1468–1473.

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR. 2012. Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates. PLoS Biol 10:e1001241.

Sloan DB, Moran NA. 2013. The Evolution of Genomic Instability in the Obligate Endosymbionts of Whiteflies. Genome Biol. Evol. 5:783–793.

Sloan DB, Nakabachi A, Richards S, Qu J, Murali SC, Gibbs RA, Moran NA. 2014. Parallel Histories of Horizontal Gene Transfer Facilitated Extreme Reduction of Endosymbiont Genomes in Sap-Feeding Insects. Mol. Biol. Evol. 31:857–871.

Smith DR, Keeling PJ. 2015. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. Proc. Natl. Acad. Sci. 112:10177–10184.

Williams TA, Foster PG, Nye TMW, Cox CJ, Embley TM. 2012. A congruent phylogenomic signal places eukaryotes within the Archaea. Proc. R. Soc. B Biol. Sci. 279:4870–4879.

Wu B, Buljic A, Hao W. 2015. Extensive Horizontal Transfer and Homologous Recombination Generate Highly Chimeric Mitochondrial Genomes in Yeast. Mol. Biol. Evol. 32:2559–2570.

Wu Z, Cuthbert JM, Taylor DR, Sloan DB. 2015. The massive mitochondrial genome of the angiosperm Silene noctiflora is evolving by gain or loss of entire chromosomes. Proc. Natl. Acad. Sci.:201421397.