School of Doctoral Studies in Biological Sciences
University of South Bohemia in České Budějovice
Faculty of Science

Genomic and Cellular Integration in the Tripartite Nested Mealybug Symbiosis

Ph.D. Thesis

RNDr. Filip Husník

Supervisor: Prof. John McCutcheon, Ph.D.

Division of Biological Sciences, University of Montana, Missoula, USA &
Program in Integrated Microbial Biodiversity, Canadian Institute for Advanced Research, Toronto, Canada

University guarantor: Prof. Ing. Miroslav Oborník, Ph.D.

Institute of Parasitology, Biology Centre of the Czech Academy of Sciences &
Department of Molecular Biology and Genetics, Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic

České Budějovice 2017

This thesis should be cited as:
Husník, F, 2016: Genomic and Cellular integration in the Tripartite Nested Mealybug Symbiosis. Ph.D. Thesis. University of South Bohemia, Faculty of Science, School of Doctoral Studies in Biological Sciences, České Budějovice, Czech Republic, 82 pp.

Annotation
The PhD thesis is composed of three publications on genomic, metabolic, and cellular integration between the host and its symbionts in the tripartite nested mealybug system. The articles revealed a path to an intimate endosymbiosis that can be compared to what we think happened before (and to some extent after) bacterial ancestors of key eukaryotic organelles, mitochondria and plastids, became highly integrated into their host cells. I argue that these much younger symbioses may tell us something about how the mitochondria and plastids came to be, at the very least by revealing what types of evolutionary events are possible as stable intracellular relationships proceed along the path of integration.

Declaration [in Czech]
Prohlašuji, že svoji disertační práci jsem vypracoval samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své disertační práce, a to v nezkrácené podobě elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

České Budějovice, 06/02/2017

....
Filip Husník

This thesis originated from a partnership of Faculty of Science, University of South Bohemia, and Institute of Parasitology, Biology Centre of the ASCR, supporting doctoral studies in the Molecular and Cell Biology and Genetics study programme.

BIOLOGY CENTRE CAS

Přírodovědecká fakulta
Faculty of Science

Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Financial support

Filip Husnik was funded by the Fulbright Commission and Grant Agency of the University of South Bohemia Grant 04-001/2014/P.

All the projects of this thesis were further supported by funding to John P. McCutcheon: National Science Foundation (NSF) Grants IOS-1256680 and IOS-1553529, National Aeronautics and Space Administration Astrobiology Institute Award NNA15BB04A, and NSF-Experimental Program to Stimulate Competitive Research Award NSF-IIA-1443108 (to the Montana Institute on Ecosystems).

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor John McCutcheon for his continuous support of my research, for his patience, motivation, enthusiasm, and for numerous fascinating discussions about not only science. The joy and enthusiasm he has for discovering of the unknown was contagious for me and I could not have imagined having a better advisor, mentor, collaborator, and friend during my PhD study.

Second, Miroslav Oborník has made this thesis possible by being its university guarantor and I thank him for his generous help and support in unconventional times.

Third, none of the work presented in this thesis would originate without numerous colleagues and collaborators from all over the world. Diversity of people and ideas was incredibly stimulating in all the projects I had pleasure to be part of during my PhD research. My PhD study was also made enjoyable in large part due to many friends and colleagues creating friendly atmosphere in the labs I worked in Budweis, Missoula, Liverpool, Edinburgh, and Krakow. Thank you all!

Last but not least, I thank my family, close friends, and my partner Petra for their continuous support.

List of papers and author's contribution

The thesis is based on the following papers (listed chronologically):

I.

<u>Husnik, F.</u>, Nikoh, N., Koga, R., Ross, L., Duncan, R.P., Fujie, M., Tanaka, M., Satoh, N., Bachtrog, D., Wilson, A.C.C., von Dohlen, C.D., Fukatsu, T., McCutcheon, J.P., 2013. Horizontal Gene Transfer from Diverse Bacteria to an Insect Genome Enables a Tripartite Nested Mealybug Symbiosis. **Cell** 153(7), 1567-1578.

F.H. participated in the study design, experimental data generation and analysis, and drafting and editing of the manuscript.

Commentaries in scientific journals:
Gerardo N. 2013. The Give and Take of Host-Microbe Symbioses. <u>Cell Host & Microbe</u> 14(1): 1-3.
Molloy S. 2013. A symbiotic mosaic. <u>Nature Reviews Microbiology</u> 11: 510-511.

The article was also featured <u>on the cover of Cell</u>, and was highlighted for example at <u>National Geographic</u>, <u>the New York Times</u>, <u>the LA Times</u>, <u>Scientific American</u>, <u>and the ASM blog Small Things Considered</u>.

II.

Duncan R.P., **<u>Husnik, F.</u>**, Van Leuven, J.T., Gilbert, D.G., Dávalos, L.M., McCutcheon, J.P., Wilson, A.C.C., 2014. Dynamic Recruitment of Amino Acid Transporters to the Insect/Symbiont Interface. **Molecular Ecology** 23(6), 1608-1623.

F.H. assembled the mealybug bacteriocyte transcriptome, conducted the mealybug differential expression analysis, and edited the manuscript.

III.

<u>Husnik, F.</u> and McCutcheon, J., 2016. Repeated Replacement of an Intrabacterial Symbiont in the Tripartite Nested Mealybug Symbiosis. **Proceedings of the National Academy of Sciences of the United States of America** 113(37), E5416–E5424.

F.H. and J.P.M. designed research, performed research, analyzed data, and wrote the paper.

The article was highlighted at <u>The Atlantic</u> and recommended by the <u>Peer Community in Evolutionary Biology</u>. The manuscript was also preprinted at the <u>bioRxiv server</u>.

Co-author agreement

John P. McCutcheon, the supervisor of this Ph.D. thesis and co-author of all presented papers, fully acknowledges the contribution of Filip Husnik.

....
John P. McCutcheon, Ph.D.

Contents

Introduction

Endosymbiont-organelle transition: better three hours too soon than a minute too late.

Filip Husnik

Faculty of Science, University of South Bohemia & Institute of Parasitology, Biology Centre ASCR, Ceske Budejovice 370 05, Czech Republic.

Abstract:

Mitochondria and plastids are now the cellular organelles of eukaryotes, but they were derived early in eukaryotic history from bacterial symbionts. Numerous recent studies show that similar bacterial symbionts are found across eukaryotic lineages and some of these symbionts rival organelles in genome reduction. Do these endosymbionts also rival organelles in cellular integration? Are these symbionts on the path to becoming organelles, or are there any other evolutionary processes in play? In this introduction, I focus on the transition period between an endosymbiont and an organelle. I review recent developments in both the endosymbiont and organelle fields, paying particular attention to how the endosymbiont-organelle transition is affected by time. I conclude that most of the evolutionary processes that have shaped bacterial endosymbionts are similar to the processes that shaped the plastid and mitochondrial ancestors. The differences between endosymbionts and organelles most likely reflect their different age, the stochastic nature of endosymbiosis, and the simple fact that mitochondria were first and thus paved the way for subsequent endosymbioses between eukaryotic cells and bacteria.

Keywords: eukaryogenesis, protein import, endosymbiosis, horizontal gene transfer

I. Is there any difference between an endosymbiont and an organelle? Should we care about this transition and its precise timing?

If we replayed the tape of life and observed the origin of the essential eukaryotic organelles, mitochondria and plastids, would there be a period of time when we would call these organisms bacterial symbionts? Surely, yes. That these organelles originated from bacterial endosymbionts is no longer questioned (Gray & Doolittle 1982). But when would we start calling them organelles, and how much time did this endosymbiont-organelle transition take? Our perception of these transitions is very limited because they likely took millions of years and happened billions of years ago. However, we can try to infer the key innovation that would lead us to change our label from 'endosymbiont' to 'organelle'. This innovation is often suggested to be protein import from the host cell **(Box 1)** and is perhaps the most widely accepted definition of organelles. With functional protein import, host proteins from the host cytoplasm make endosymbiont homologs obsolete, and eventually lead to losses of genes coding even the most essential components such as DNA and RNA polymerases from symbiont genomes. The endosymbiont then becomes part of its host cell – an organelle.

The extreme age of the transition from endosymbiont to a highly integrated cellular component has resulted in relatively sparse and weak data, and as such has generated extensive debates (Archibald 2006; Theissen & Martin 2006; Keeling & Archibald 2008; Keeling *et al.* 2015; Booth & Doolittle 2015b; Lane & Martin 2015; McCutcheon 2016). Luckily, there are much younger symbiotic systems that allow us to see the timing of genetic, cellular, and metabolic integration in both unicellular and multicellular symbiotic systems more clearly **(Figure 1, Figure 2)**. Mitochondrion and plastid acquisition each happened only once, so these fields will always lack the power of comparative analysis for primary symbioses. But these younger symbioses have originated many times independently in various host systems, and can therefore provide us with hints about the possible time frames and outcomes of these processes **(Figure 2, Figure 3)**. For example, how long does it take for an endosymbiont to lose majority of its ancestral genome? How long does it take to establish metabolic integration? Have some of the younger, but still quite old (e.g. ~300 Mya in some insects) symbioses had time to establish protein import? If not, why not? If yes, how and when? Can highly integrated endosymbionts be replaced, and how long does it take for the new partner to itself become integrated? Is this process faster thanks to the pre-existing symbiosis? When and for how long does endosymbiotic gene transfer (EGT) influence the transition? And how does the contribution of horizontal gene transfer (HGT) from other organisms affect the timing and process of integration?

For at least two decades, the level of integration in symbiotic systems of arthropods, protists, marine animals, and other eukaryotes was viewed as less than that of the classic cellular organelles, and the questions I outlined above were rarely considered (Dubilier *et al.* 2008; Moran *et al.* 2008; Nowack & Melkonian 2010; McCutcheon & Moran 2011; Hentschel *et al.* 2012; Moran & Bennett 2014; Douglas 2016). For example, there was little evidence for host genes (either eukaryotic or HGTs) interacting with endosymbionts in any obvious or meaningful way, and endosymbiont lability and replacement, although sometimes observed, was mostly interpreted as rare and ancient. However, recent developments in the field strongly suggest that most, if not all, features previously used to define organelles occur in these much younger systems **(Table 1, Table 2, Figure 2)**. Diverse bacterial symbionts of eukaryotes were shown to be extremely tightly integrated at the genetic, cellular, and metabolic level, some of them even crossing the River Styx to the 'organelle world' by protein import from the host cell (McCutcheon & Keeling 2014). Overall, the mechanistic and genetic parallels between these symbionts and organelles make clear distinctions hard to see.

From the organelle and eukaryotic perspective, several findings related to the transition stage have emerged as well. Perhaps the most important finding is that the ancestral cell that acquired mitochondria about 2.5 billion years ago was very likely archaeal and related to the recently named Asgard superphylum (Williams *et al.* 2013; Williams & Embley 2014; Spang *et al.* 2015; Koonin 2015; Martin *et al.* 2015; Zaremba-Niedzwiedzka *et al.* 2017). It is hotly debated whether the mitochondrial 'symbiont' came in rather late in the evolution of a cell that already looked eukaryote-like (Spang *et al.* 2015; Pittis & Gabaldón 2016; Zaremba-Niedzwiedzka *et al.* 2017) or whether the mitochondrion acquisition was early and was the main stimulus for the origin of eukaryotes **(Figure 3)** (Lane & Martin 2015; Martin *et al.* 2016). Gene transfer from other bacteria was clearly involved before and after the acquisition of mitochondria, but the taxonomic diversity of these transfers makes it impossible to infer phylogenies with high confidence using single gene sequences for a variety of both biological and methodical reasons (Kurland & Andersson 2000; Qiu *et al.* 2013; Gray 2015). Dating of the deeply-branching eukaryotic lineages (supergroups) is unfortunately also very unclear (Dacks *et al.* 2016), but it seems that the major lineages have diverged rather quickly after mitochondrion acquisition (and its genome reduction). Several deeply-branching lineages such as jakobid protists harbor gene-rich mitochondrial genomes (Burger *et al.* 2013), providing interesting data about the genes that were likely present in this ancestor of mitochondria. The very same situation, although with different levels of genome reduction, is also observed in plastids of mostly glaucophytes and red algae (Smith & Keeling 2015; Lee *et al.* 2016). But interpreting the order of events in these systems is further blurred by the shuffling of plastid genes due to secondary, tertiary, and higher-level endosymbioses (Keeling 2013).

II. Our view of genetic, cellular, and metabolic integration of eukaryotic endosymbionts has quite dramatically changed over the last five years.

Genome reduction of insect endosymbionts is much more extensive than originally imagined. In recent years, numerous endosymbiont genomes were sequenced from diverse eukaryotes revealing a range of genome sizes. However, the smallest genomes are almost always found in hemipteran insects **(Figure 1, Figure 2, Table 2)**. The smallest reported genome from a non-organelle bacterium is from the leafhopper endosymbiont *Nasuia deltocephalinicola* (Bennett & Moran 2013). Its genome size of 112 kb and total number of protein-coding genes (137) is even smaller than in some red algal plastid genomes such as *Porphyridium purpureum* (218 kbp; 224 protein-coding genes) (Bhattacharya *et al.* 2013; Lee *et al.* 2016). How old is the leafhopper symbiosis? It is not so easy to tell, but it co-resides in the insect with one more symbiont, *Sulcia muelleri*, and this co-residence was estimated up to the origin of Auchenorrhyncha, i.e. around 260-280 Mya (Moran *et al.* 2005; Bennett & Moran 2013). The Auchenorrhyncha lineage includes also other sap-feeding insects such as spittlebugs, cicadas, planthoppers, treehoppers, or lanternflies. The majority of these insects house *Sulcia* with one or more additional co-symbionts. This long-term co-symbiosis of *Sulcia* has been followed by both ancient and recent replacements of the second partner (*Hodgkinia*, *Zinderia/Nasuia/Vidania, Sodalis*, and likely others), and thus presents a fascinating system to study the speed of genome reduction in symbionts of various ages (Bennett & Moran 2015). Other tiny endosymbiont genomes are found in hemipteran insects such as psyllids, whiteflies, moss bugs, and scale insects (Sloan & Moran 2012a; b; Sabree *et al.* 2012; Rosas-Pérez *et al.* 2014; Santos-Garcia *et al.* 2014; Husnik & McCutcheon 2016). Importantly, all of these symbioses were estimated to originate more than 100 million years ago **(Figure 3)**, but they often involve also other (much younger) obligate co-symbionts or show replacements of the partners. The idea that time is needed to establish an intimate organelle-like symbiosis is rarely questioned, but numerous examples of endosymbiont losses and replacements show that the time required to adapt to symbiosis may be initially required by the host, but once established the endosymbiont population can sometimes change relatively rapidly. One fascinating model system supporting this hypothesis is the mealybug-*Tremblaya-gammaproteobacteria* symbiosis examined in great detail by the manuscripts of my thesis (Husnik *et al.* 2013; Duncan *et al.* 2014; Husnik & McCutcheon 2016).

Endosymbionts from unicellular eukaryotes show less genome reduction than those from insect systems, but both symbioses show high levels of integration with their hosts. Simple logic would suggest that we should most often find organelle-like endosymbionts in unicellular eukaryotes. These eukaryotes are commonly bacterivorous and domestication of an endosymbiont through EGT and protein import should be more straightforward inside their single cells than in multicellular eukaryotes with highly protected germline cells. Moreover, we know that such a transition happened at least once when the archaeplastidal ancestor (already harboring a mitochondrion) acquired cyanobacterial symbionts that later became plastids. Unicellular protists should have had plenty of time to develop these symbioses, because the major eukaryotic supergroups have diverged early after the origin of LECA (Knoll 2014).

So why do we find no such novel organelles in protists? Perhaps we have not sampled hard enough, especially in comparison to the insect lineages discussed above, but several endosymbionts with severe genome reduction do exist in single-celled eukaryotes. These symbioses are in most cases nutritional in nature, such as the cyanobacterial symbionts (called spheroid bodies) in Rhopalodiaceae diatoms (Prechtl *et al.* 2004; Kneip *et al.* 2008; Nakayama *et al.* 2014), cyanobacterial symbionts (called UCYN-A lineage, *Atelocyanobacterium thalassa*) in haptophytes (Zehr *et al.*

2008; Tripp *et al.* 2010; Thompson *et al.* 2012; Bombar *et al.* 2014; Cornejo-Castillo *et al.* 2016), two independent betaproteobacterial symbioses in trypanosomatids (*Kinetoplastibacterium* and *Pandoraea* species) (Alves *et al.* 2013a; b; Klein *et al.* 2013; Kostygov *et al.* 2016), numerous bacterial symbionts in ciliates (e.g. *Polynucleobacter necessarius* in *Euplotes* spp. or a lineage called TC1 in *Trimyema compressum*) (Boscaro *et al.* 2013; Shinzato *et al.* 2016), and numerous endosymbioses of protists inhabiting termite guts (Brune & Dietrich 2015) such as *Endomicrobium trichonymphae* (Hongoh *et al.* 2008; Izawa *et al.* 2016) or *Desulfovibrio trichonymphae* (Kuwahara *et al.* 2016).

Strikingly, there is one example where the symbiotic cyanobacterium (called chromatophore or cyanelle) is kept for exactly the same reason as the ancient archaeplastidal symbiosis – for photosynthesis. The host organism, *Paulinella chromatophora*, is an amoeboid protist from the Rhizaria lineage. Similarly to the other protist symbioses described above, it has acquired the symbiont (relatively) recently, about 60-200 million years ago. Despite its young age and modest amount of genome reduction (1,021,616 bp), it seems to be on the path to becoming highly integrated into its host. For example, it is already dependent on proteins imported from the host cytoplasm (Marin *et al.* 2005; Nowack *et al.* 2008, 2011, 2016; Nowack & Grossman 2012; Nowack 2014).

Endosymbionts of eukaryotes are often dependent on various compounds, including proteins, imported from the host cell. There is a growing body of both genomic and experimental evidence that various endosymbionts rely on their hosts for provisioning of essential compounds. When any compound is freely available from the host cytoplasm, metabolic pathways encoded on the symbiont genome are no longer under strong selection and 'use it or loose it' strategy of bacterial genome evolution is inevitable. For example, aphids provide to their *Buchnera* endosymbiont almost all non-essential amino acids (Shigenobu & Wilson 2011; Hansen & Moran 2011; Poliakov *et al.* 2011; Macdonald *et al.* 2012) and likely several vitamins and co-factors needed by the endosymbiont enzymes (Charles *et al.* 2011), so these pathways were eventually lost from the *Buchnera* genome (Shigenobu *et al.* 2000). In a similar manner, the most extremely reduced endosymbiont genomes such as *Nasuia* or *Tremblaya* no longer code genes for ATP synthase, NADH dehydrogenase, cytochrome oxidase, TCA cycle, lipid metabolism, sugar metabolism, nucleotide metabolism, etc. because compounds from these pathways are provided from either their host/mitochondria or their obligate co-symbionts (von Dohlen *et al.* 2001; Gottlieb *et al.* 2008; Bennett & Moran 2013; Moran & Bennett 2014).

Transporters were so far studied predominantly for the aphid-*Buchnera* model system, where they seem to play a central role in the aphid/*Buchner*a symbiosis. Published studies conclude that *Buchnera* retains only a few general transporters, some of which very likely lost their substrate specificity (Charles *et al.* 2011). Among the aphid transporters, 82 genes were reported to be up-regulated in bacteriocytes (Hansen & Moran 2011), amino acid transporters were found to be extensively duplicated and specialized for bacteriocyte transfer (Price *et al.* 2011; Duncan *et al.* 2014), and some of them implicated to be essential for endocytosis of *Buchnera* during transmission (Lu *et al.* 2016).

Indirect evidence from different animal and protist symbioses implies that there is an immense flow of both small and large compounds from and to symbiont cells, but the precise mechanical functioning of this transport is poorly understood. A major transport role is likely played by the outermost host-derived ‘symbiosomal’ membrane covering every symbiont cell (although not present in all systems). The membrane likely incorporates transporters and controls which compounds are provided to the symbiont and how often (Price *et al.* 2013). That the most highly integrated endosymbionts are engulfed by completely host-derived cell envelopes further supports the hypothesis that any machinery for transport is host-derived and inside the cell envelope. Apart from exchange of various metabolites, protein exchange is likely needed for some endosymbioses, but experimental data testing protein import/export to and from symbionts are extremely scarce due to methodological difficulties facing experimental work with non-model species **(Box 1)**.

Genes of bacterial origin on the host genome compensate for genome reduction of endosymbionts. Five years ago, there was little evidence for HGT interacting with endosymbionts in any obvious or meaningful way, although numerous genes believed to be essential were found to be missing from the symbiont genomes. This situation started to change after the discovery of several likely functional bacterial genes in the aphid genome (Nikoh *et al.* 2010). Since then, bacterial genes have been found in many eukaryotes harboring intracellular symbionts such as mealybugs (Husnik *et al.* 2013; Husnik & McCutcheon 2016), psyllids (Sloan *et al.* 2014), whiteflies (Luan *et al.* 2015; Chen *et al.* 2016), *Angomonas* and *Strigomonas* trypanosomatids (Alves *et al.* 2013a), and *Paulinella chromatophora* (Nowack *et al.* 2016). In most cases (except in aphids), the bacterial genes seem to fill in gaps in pathways predicted to be carried out in cooperation between the host and its symbiont. The host thus takes over enzymatic steps originally coded by the symbiont genome. Importantly, very few of these bacterial genes found in eukaryotic genomes come from the current endosymbiont, but rather from bacteria common in the environment, i.e. for multicellular eukaryotes mostly from bacteria infecting oocytes. It now seems that the role of gene transfer from diverse bacteria to eukaryotes with symbionts is to compensate for gene loss in extant symbionts to maintain function in the symbiosis (Husnik *et al.* 2013; Nowack *et al.* 2016).

III. A few hints about timing and evolution of mitochondria and plastids have appeared in the last five years

The cell that became an eukaryotic ancestor was an archaeon. Eukaryotes are cellular and genetic chimeras of two or more organisms. The last eukaryotic common ancestor from which all contemporary eukaryotes descend originated roughly 2 billion years ago from a symbiotic event between an archaeal host cell and an alphaproteobacterial endosymbiont (Gray & Doolittle 1982; Embley & Martin 2006; Koonin 2010, 2015; Booth & Doolittle 2015a; Zaremba-Niedzwiedzka *et al.* 2017). The phylogenetic origin of the archaeal host is now consistently being resolved to be close to or within the Asgard superphylum (Williams *et al.* 2013; Williams & Embley 2014; Spang *et al.* 2015; Zaremba-Niedzwiedzka *et al.* 2017), but cellular complexity of the host cell and mitochondria-early or mitochodria-late timing of the symbiosis keeps to be hotly debated (Ettema 2016; Pittis & Gabaldon 2016; Pittis & Gabaldón 2016; Martin *et al.* 2016). There are therefore only two domains of life, Bacteria and Archaea, not three as suggested by ribosomal RNA trees (Woese *et al.* 1990), and eukaryotes are deeply nested inside Archaea. Interestingly, several lines of evidence suggest that the proto-eukaryote host cell already contained many genes and functions previously considered to be eukaryote-specific innovations such as cytoskeletal components, membrane-trafficking machinery components, and coat proteins involved in vesicle biogenesis (Zaremba-Niedzwiedzka *et al.* 2017).

Unlike for the host cell ancestor, the exact present-day closest relative of the alphaproteobacterial lineage from which mitochondria descent remains elusive

(Wang & Wu 2014, 2015), but a recent study has shed some light on the origin of primary plastids. Interestingly, a freshwater cyanobacterium *Gloeomargarita lithophora* was inferred as the most closely related lineage to plastids suggesting that the first photosynthetic eukaryote most likely evolved in terrestrial-freshwater settings, not in oceans (Ponce-Toledo *et al.* 2017).

Mitochondrial and plastid evolution: from stability to craziness. Genomes of mitochondria and plastids can be both immensely stable and remarkably dynamic. Different organelle lineages show large ranges of genome size, GC content, coding density, structure, and content. Some genomes expand, such as plant mitochondrial genomes (Sloan *et al.* 2012), while other genomes shrink, such as plastid genomes of non-photosynthetic plants (Logacheva *et al.* 2016) or mitochondrial genomes of dinoflagellates, apicomplexans, and their relatives (Waller & Jackson 2009; Oborník & Lukeš 2015). Mitochondrial genomes can be lost and the remaining organelles then rely solely on imported proteins (Stairs *et al.* 2015) and one recent study suggests that even the entire organelle can be lost (Karnkowska *et al.* 2016). Very similar evolutionary history of genome loss also likely affected plastid evolution (Smith & Lee 2014).

This diversity (and sometimes eccentricity) of mitochondria and plastids can be explained by combination of their age, DNA repair processes, mutation rates, and population genetic structure (Smith & Keeling 2015). Importantly, the diversity also provides us with almost unbelievable examples of what is possible in organelle evolution and shows that 'anything goes' for both mitochondria and plastids (Burger *et al.* 2003; Archibald & Richards 2010). When stripped to the bone, the omnipresent function of mitochondria (and various mitochondria like organelles) seems to be iron sulfur assembly (Lill *et al.* 2012). This process is present in the majority of endosymbionts as well (McCutcheon & Moran 2011), but is not likely as crucial because iron-sulfur clusters are already available from the host mitochondrion.

Perhaps the most relevant genomes for this review are the gene-rich mitochondrial genomes of jakobid protists (Burger *et al.* 2013) and the gene-rich plastid genomes of red algae (Janouškovec *et al.* 2013; Lee *et al.* 2016) **(Table 2).** In terms of gene content, these genomes are akin to the tiniest endosymbiont genomes such as *Tremblaya* or *Nasuia* as they still retain four genes encoding bacterial RNA polymerase (*rpoABC*) together with its sigma factor (*rpoD*), large portion of ribosomal proteins, and even some translational factors **(Figure 4)**. But there are several significant differences related to the bacterial genetic machinery. First is that unlike endosymbionts, no organellar genomes retain genes for even a minimal DNA polymerase nor any aminoacyl tRNA synthetases. They are completely dependent on their hosts for replication and translation (and transcription in mitochondria other than from jakobid protists). The only endosymbiont lineage that has lost all of its aminoacyl tRNA synthetases is *Tremblaya princeps* with its own bacterial symbionts likely supplementing this lost function (McCutcheon & von Dohlen 2011). If there is any bacterial essence remaining in these tiny symbiont genomes that differentiates them from organelles, it is their ability to independently replicate their genomes (McCutcheon 2010).

Complex plastid acquisitions across the tree of life

Acquiring a photosynthetic ability was crucial for the diversification of many eukaryotic lineages. Since the origin of primary plastids, several lineages of algae have been acquired as multi-genome symbiotic sets to form secondary and tertiary endosymbioses. Secondary plastids are known from euglenids (Excavata) and chlorarachniophytes (Rhizaria) which both secondarily acquired green algae. Haptophyta, Cryptomonada, and several lineages in Alveolata and Stramenopila acquired red-algal plastids in symbiotic events that remain unresolved (Keeling 2013; Ševčíková *et al.* 2015). Interestingly, two lineages with complex plastids, chlorarachniophytes and cryptomonads, still keep highly reduced nuclei between two sets of plastid membranes (Curtis *et al.* 2012). Several additional layers of symbiotic complexity are known from dinoflagellates (Alveolata). Although they harbor an ancestral plastid of red-algal origin, some dinoflagellate lineages have acquired new plastids by subsequent serial endosymbioses of green algae (serial secondary endosymbiosis) or haptophytes and diatoms (tertiary endosymbiosis) which in some cases still retain their own nuclei and even mitochondria (Dorrell & Howe 2015).

Proteomes of organelles are incredibly mosaic. Endosymbiotic gene transfer from mitochondria and plastids to the nucleus and re-targeting of proteins back to the organelles has long been viewed as one of the major steps in eukaryogenesis (Timmis *et al.* 2004). A recent taxon-rich (55 eukaryotes) analysis of gene clustering and phylogenetic analyses of eukaryotic gene families with prokaryotic gene homologs detected 2,585 gene clusters containing sequences from at least two eukaryotic and five prokaryotic lineages. While cyanobacterial EGT signal was clearly detected by the analyses, alphaproteobacterial signal was basically absent. However, all these 2,585 clusters were determined to be putative EGTs by the authors, a conclusion which in my view is quite unconservative (1,525 clusters from the mitochondrial ancestor and 1,060 from the plastid ancestor) (Ku *et al.* 2015b). When contrasting these results to several previous analyses (Kurland & Andersson 2000; Gabaldón & Huynen 2004, 2005, 2007; Esser *et al.* 2004; Cotton & McInerney 2010; Thiergart *et al.* 2012; Reyes-Prieto & Moustafa 2012; Huynen *et al.* 2013; Qiu *et al.* 2013; Gray 2015), it becomes abundantly clear that such analyses are extremely method and sampling dependent and that the bacterial part of nucleus-encoded mitochondrial and plastid proteomes shows striking taxonomic diversity when evaluated by single-gene trees (Kurland & Andersson 2000; Qiu *et al.* 2013; Gray 2015).

This discrepancy between different studies has been suggested to result from poor phylogenetic signal in single-gene matrices, inherited chimerism of bacterial ancestors of organelles, lineage-specific gene losses combined with poor taxon sampling, and previous and ongoing horizontal gene transfers from diverse sources such as unsuccessful symbionts (Larkum *et al.* 2007; Ku *et al.* 2015a; b; Gray 2015). Of course, simple models will likely never fully reconstruct the evolutionary history of eukaryotes, and so most of the processes mentioned above (and possibly a few more) have probably occurred in distinct eukaryotic clades with different frequencies. The presence of numerous bacterial-like genes in the Asgard archaea genomes might in the near future clarify the importance of horizontal and endosymbiotic gene transfer for mitochondrial evolution (Zaremba-Niedzwiedzka *et al.* 2017).

IV. On the importance of protein import from the host preceding massive genome reduction (<100 kbp)

How far can endosymbiont genome reduction go? Six years ago, it was calculated that a theoretical minimal genome size for an intracellular symbiont of insects was approximately in the range of 70–80 kb (McCutcheon & Moran 2011). In terms of gene context, such a genome would be almost indistinguishable from the most gene-rich mitochondrial genomes from jakobid protists (Burger *et al.* 2013). However, after more than six years and very comprehensive sampling of insect lineages with intracellular endosymbionts, no data suggest such highly reduced genomes actually occur.

Although it is still possible that such an extremely degenerate endosymbiont will be discovered in the near future, it is perhaps appropriate to start asking questions. If we are not finding these tiny genomes, why not? It has been shown repeatedly that the initial stages of genome reduction can be extremely fast. For example, it has been estimated that 55% percent of an ancestral endosymbiont genome was lost in only ~28,000 years (Oakeson *et al.* 2014). However, once the symbiont genome is reduced to approximately 250 kbp, the host might be more likely to face extinction because of its reliance on such a degenerate symbiont, so gene loss is very likely much slower at this stage and relies upon first evolving complementarity with the host. Complementarity can be achieved in several different ways, but this period of slow gradual increase of interdependence (observable in some endosymbiont systems) likely coincides with the beginning of symbiont-organelle transition.

Why do we find no novel organelles in unicellular eukaryotes? Several scenarios can be put forward to explain why unicellular eukaryotes have not formed any other highly integrated symbioses since mitochondrion and plastid origins. Putting aside that it is still possible that we did not find them yet, another likely scenario is that they were not stable over evolutionary history and either were replaced or the lineage went extinct before fixed (Keeling *et al.* 2015). In principle, the transfer of both too few and too many of essential genes can lead to symbiont extinction. With too many transfers, the symbiont (or at least its genome) may no longer be needed by the host. On the other hand, genes kept on the symbiont genome drive the symbiosis into the symbiotic rabbit hole **(Box 2)**. Eukaryotic genomes contain genes from bacteria (Keeling & Palmer 2008; Alsmark *et al.* 2013; Wybouw *et al.* 2016), and these genes often code enzymes involved in nutrition. These HGTs can thus be thought of as ‘ghosts’ of symbiosis past. When a specialized compartment is not needed for the symbiont function (as has been shown for mitochondria and plastids) and some proteins do not have to be translated in the organelle (e.g. hydrophobic membrane proteins would likely be targeted to the endoplasmic reticulum if they were nuclear-encoded (Björkholm *et al.* 2015), the symbiont can ‘dissolve’ in its host (Karnkowska *et al.* 2016), for exampe after donating genes originally essential for the symbiosis such as genes for biosynthetic pathways shown to be crucial in almost all protist symbioses. It is therefore interesting that in most cases, plastids have not evolved independently and *de-novo* (as in *Paulinella*), but rather acquired in the form of a plastid-containing lineage.

Another scenario explaining the lack of 'novel' organelles in eukaryotes might be that eukaryotes already contain hundreds to thousands of genes (EGT and HGT) from bacteria transferred to their chromosomes. Perhaps there is no need for novel organelles as horizontal gene transfer or alternative ways of adaptive evolution such as acquisition of a co-symbiont or symbiont replacement allow much faster innovations. Mitochondrion-generated ATP allowed eukaryotes to grow large and complex cells (Martin & Müller 1998). But how does the presence of mitochondria decrease a chance to establish novel symbioses? For example, leakage of mitochondria-targeted proteins into plastids and rapid establishment of dual targeting can be hypothesized as mechanisms causing parallel evolution of plastid genomes (Smith & Keeling 2015), but how mitochondria-targeted proteins influence symbiont evolution has never been tested.

Timing is essential for an endosymbiont to become an organelle. Endosymbiont genome reduction has been shown to be extremely fast. In some cases it can take only thousands-to-millions of years to lose several thousand endosymbiont genes (Clayton *et al.* 2012; Oakeson *et al.* 2014). After this initial massive genome reduction, the reductive evolution seems to often slow down for tens of millions of years with approximately 500-1000 functional genes left in the symbiont genome (**Figure 2**). It is possible that a similar pace of gene loss also affected the ancestors of mitochondria and plastids. If so, it is unlikely that concurrent functional EGT coupled with fine-tuning of protein import could manage to compensate for such extremely fast gene loss. Numerous genes complementing the organelles were needed for the symbiont to survive and become the organelle, but the rate of gene loss would mean that many of them were likely not there yet.

It also seems unlikely that both organelles were successful on the first try. Endosymbiont dynamism has long been observed, but most of it seemed rare and ancient (Moran *et al.* 2008). However, recent findings from many endosymbiont-housing eukaryotes (Douglas 2016) point towards extreme instability and dynamism of symbioses, especially when reaching near-organelle genome sizes. Symbiosis loss, complementation, and replacement were shown to occur even when the current symbiont is extremely highly integrated into its host cells (Husnik & McCutcheon 2016). This dynamism sometimes also leads to irremediable complexity ('craziness') at the genomic and cellular levels, paralleling what is observed in organelles (Gray *et al.* 2010; Wu *et al.* 2015). For example, a single circular genome of the cicada endosymbiont *Hodgkinia cicadicola* MAGTRE has been split into numerous genomes present in separate cells over evolutionary history (Van Leuven *et al.* 2014; Campbell *et al.* 2015). Somehow, these new lineages seem to share even the most essential proteins such as for DNA and RNA polymerases.

The evolution of the cellular organelles was probably not a neat and tidy process. The orderly transfer of massive numbers of EGTs combined with rapid co-evolution of protein-targeting seems incredibly unlikely. Rather, I argue that the process was an inefficient and chaotic one, involving failed endosymbioses and HGT from numerous sources. In my view, this transition required previous and late HGTs to allow the final 'evolutionarily lucky' symbiont to survive the symbiont-organelle transition. Further adjustments to the cell biology of the host took hundreds of millions of years, and explains why other examples of endosymbionts in diverse eukaryotes differ mainly by the level of integration in the host cell, not by genome reduction.

Display items

Box 1: ***The most important piece of the puzzle is missing: protein import into endosymbionts.*** Many endosymbiont and organelle researchers would agree that the point when an endosymbiont becomes organelle-like is when there is a well-established protein import from the host. This reasoning is based on the current situation of organelles – a majority of their proteins come from the host cytoplasm and protein complexes importing them (such as TIM/TOM in mitochondria and TIC/TOC in plastids) (Soll & Schleiff 2004; Doležal *et al.* 2006; Balsera *et al.* 2009).

Are there many cases of proteins being shown to be imported into an endosymbiont from the host cytoplasm? No, there are not. Whether it is a result of methodologically challenging experiments, or a true biological state, there is only a handful of examples, including chromatophores in *Paulinella* protists (Nowack & Grossman 2012), bacterial symbionts of trypanosomatids (Morales *et al.* 2016), and *Buchnera* symbionts in aphids (Nakabachi *et al.* 2014). However, no protein silencing experiments are presently available for these organisms, so it is still not clear how important protein import is for these symbioses or if it is more akin to a host mechanism used to manipulate the symbionts such as in plant-*Rhizobium* (Van de Velde *et al.* 2010) or weevil-*Sodalis* systems (Login *et al.* 2011).

Importantly, one significant difference between organelles and recent endosymbionts might be the status of the eukaryotic endomembrane system at the establishment of symbiosis. If it was not present in the eukaryotic ancestor, evolution of protein import complexes was crucial for eukaryogenesis. On the other hand, if late-coming symbioses could use an already established endomembrane system, this might obviate the need for a specific import system, especially given that entirely host-derived outer membranes of some of these symbionts are likely highly similar to membranes of other cellular compartments (such as mitochondria) (Husnik & McCutcheon 2016). In addition, outer membrane vesicles (OMV) were shown to be critical elements in many extracellular host-microbe interactions such as the squid-*Vibrio* (Aschtgen *et al.* 2016) or human-gut microbiota (Elhenawy *et al.* 2014), but their role in intracellular symbioses remains enigmatic. Comprehensive analysis of metabolite and protein exchange at the host-symbiont interfaces in diverse systems, although methodologically challenging, is thus needed to answer in our view the most important question of the field. How are proteins imported into organelle-like endosymbionts?

Box 2: ***The symbiotic rabbit hole: when your population genetic structure brings you to the verge of extinction but selection keeps you there for over a billion of years.*** The total population of heritable symbiotic bacteria in a single individual is subsampled every generation (for example into eggs in multicellular animals) and maternally transmitted to offspring. This bottlenecking leads to extremely small effective population sizes of endosymbiotic bacteria and random genetic drift accumulating deleterious mutations in their genomes (Moran 1996; Lambert & Moran 1998; Woolfit & Bromham 2003). Since the lineages are asexual and often missing DNA repair and recombination genes, these changes are irreversible due to Muller's ratchet (Moran 1996). Features of endosymbiotic bacteria such as rapid sequence evolution, gene loss, lower thermal stability of proteins and RNAs, and extreme biases in nucleotide composition root from this population structure (McCutcheon & Moran 2011).

Over evolutionary time, this process eventually ends in a state where the host is incredibly dependent on a symbiont that is degenerating and, in some cases, seems clumsily balanced on the verge of extinction. This irreversible host-symbiont co-dependence resulting from population genetics structure of symbionts was described as the ‘symbiotic rabbit hole’ (Bennett & Moran 2015). Any of these detrimental changes potentially leading to extinction of both partners can be slowed down by selection acting either on the symbiont or host level (Wernegreen 2002), but selection can be dangerously inefficient when acting on populations of polyploid symbiont cells (Van Leuven *et al.* 2014; Campbell *et al.* 2015). Is there any other way out for the host from this degenerative ratchet? It seems that there is. Endosymbiont replacement can rescue the host by providing an endosymbiont with a ‘fresh’ genome (Husnik & McCutcheon 2016), but this rescue is, of course, only temporary. Transferring endosymbiont genes out of the reach of deleterious mutations, i.e. to the host genome from either the current symbiont (EGT) or from other organisms (HGT), or adjusting native genes to carry out symbiont functions is the solution that allowed eukaryotes to keep their quintessential symbionts, mitochondria, for almost two billions years (Timmis *et al.* 2004).

Table 1: Various genomic and cellular features usually characterizing organelles and their presence in diverse endosymbiont lineages. Features never reported from endosymbionts include for example loss of genes for DNA and RNA polymerases, group II catalytic introns, and RNA editing. RNA and DNA polymerase genes were lost in individual *Hodgkinia* lineages co-residing in bacteriomes of some cicadas, but this example is not included here for simplicity (Campbell *et al.* 2015). #I am not aware of any manuscripts examining cell division in animal symbioses.

Feature	**'Endosymbiont' lineages**	**References**
Massive genome reduction (<250 kbp) and associated changes (highly gene-dense genomes with overlapping genes, increased ortholog length variation, and loss of large accessory proteins)	*Tremblaya, Hodgkinia, Nasuia/Zinderia Carsonella*	(McCutcheon & Moran 2011; Kenyon & Sabree 2014; Moran & Bennett 2014)
Origin of an alternative genetic code	*Hodgkinia, Nasuia/Zinderia*	(McCutcheon *et al.* 2009a; McCutcheon & Moran 2010; Bennett & Moran 2013)
Loss of genes for translation, i.e. translation factors, tRNAs, rRNAs, RNA modification genes and ribosomal proteins	*Tremblaya, Hodgkinia, Nasuia/Zinderia, Carsonella*	(McCutcheon *et al.* 2009b; McCutcheon & Von Dohlen 2011; Bennett & Moran 2013; Husnik & McCutcheon 2016)
Import of some compounds and intermediate products (amino acids, vitamins, ATP, sugars, nucleotides, etc.) from the host cytoplasm	All obligate symbionts of insects 'spheroid body' in diatoms 'chromatophore' in *Paulinella Kinetoplastibacterium*	(McCutcheon & Moran 2011; Hansen & Moran 2011; Poliakov *et al.* 2011; Duncan *et al.* 2014; Moran & Bennett 2014; Douglas 2016)
Reliance on proteins from the host genome that are of bacterial origin (HGT)	*Tremblaya*, *Buchnera*, *Carsonella*, *Portiera*, 'chromatophore', *Kinetoplastibacterium*	(Nikoh *et al.* 2010; Husnik *et al.* 2013; Sloan *et al.* 2014; Nakabachi *et al.* 2014; Luan *et al.* 2015; Chen *et al.* 2016; Nowack *et al.* 2016; Husnik & McCutcheon 2016; Morales *et al.* 2016)
Endosymbiotic gene transfer from the current symbiont to the host genome (EGT)	*Paulinella*-chromatophore (~58 genes), psyllids-*Carsonella* (1 gene)	(Sloan *et al.* 2014; Nowack *et al.* 2016)
Import of proteins from the host cytoplasm to the symbiont cell	*Buchnera,* chromatophore, *Kinetoplastibacterium*	(Alves *et al.* 2013a; Klein *et al.* 2013; Nakabachi *et al.* 2014; Nowack *et al.* 2016)
Loss of peptidoglycan and phospholipid pathways and thus reliance on host-derived cell envelopes (often with an outermost 'symbiosomal' membrane)	*Tremblaya, Hodgkinia, Nasuia/Zinderia, Carsonella*	(McCutcheon *et al.* 2009b; McCutcheon & Von Dohlen 2011; Bennett & Moran 2013; Husnik & McCutcheon 2016)
Reliance on the host cell for division#	chromatophore *Kinetoplastibacterium*	(Nowack *et al.* 2008; Motta *et al.* 2010; Brum *et al.* 2014)

Table 2: Genome features of the most highly reduced genomes of animal endosymbionts (*Carsonella, Hodgkinia, Tremblaya, Nasuia*), the most gene-rich organelle genomes (mitochondrial genomes of Jakobida and plastid genomes of glaucophyta and red algae), and several selected endosymbionts of unicellular eukaryotes.

Lineage	**Genome size (bp)**	**CDS (pseudo)**	**G + C (%)**	**tRNAs | rRNAs**
Endosymbionts of unicellular eukaryotes				
‘chromatophore’ (*P. chromatophora)*	1,021,616 bp	867 (NA)	39.0	42 | 6
Kinetoplastibacterium oncopeltii	810,172 bp	694 (NA)	31.2	43 | 9
Atelocyanobacterium thalassa	1,443,806 bp	1133 (NA)	31.1	37 | 6
‘spheroid body’ (*Epithemia turgida*)	2,794,318 bp	1720 (225)	33.4	39 | 6
Endomicrobium trichonymphae Rs-D17	1,125,857 bp	761 (121)	35.2	45 | 3
Highly reduced genomes of animal (insects in all cases) endosymbionts				
Carsonella ruddii HT	157,543 bp	180 (NA)	14.6	28 | 3
Tremblaya phenacola PAVE	170,756 bp	178 (3)	42.2	31 | 4
Tremblaya princeps PCIT	138,927 bp	125 (16)	58.8	10 | 6
Nasuia deltocephalinicola ALF	112,091 bp	137 (NA)	17.1	29 | 3
Hodgkinia cicadicola DSEM	143,795 bp	169 (NA)	58.4	15 | 3
Gene-rich chloroplast genomes (from Glaucophyta and Rhodophyta)				
Cyanophora paradoxa	135,599 bp	149 (NA)	30.5	36 | 6
Cyanidioschyzon merolae	149,987 bp	207 (NA)	37.6	31 | 3
Porphyridium purpureum	217,694 bp	224 (NA)	30.0	30 | 6
Porphyra purpurea	191,028 bp	209 (NA)	33.0	37 | 6
Hildenbrandia rivularis	189,725 bp	184 (NA)	32.4	31 | 3
Gene-rich mitochondrial genomes (from Jakobida)				
Reclinomonas americana	69,034 bp	67 (NA)	26.1	26 | 4
Andalucia godoyi	67,656 bp	72 (NA)	36.3	29 | 3
Histiona aroides	70,177 bp	72 (NA)	35.4	26 | 3
Jakoba libera	100,252 bp	84 (NA)	32.0	26 | 3
Jakoba bahamiensis	65,327 bp	68 (NA)	32.2	26 | 3

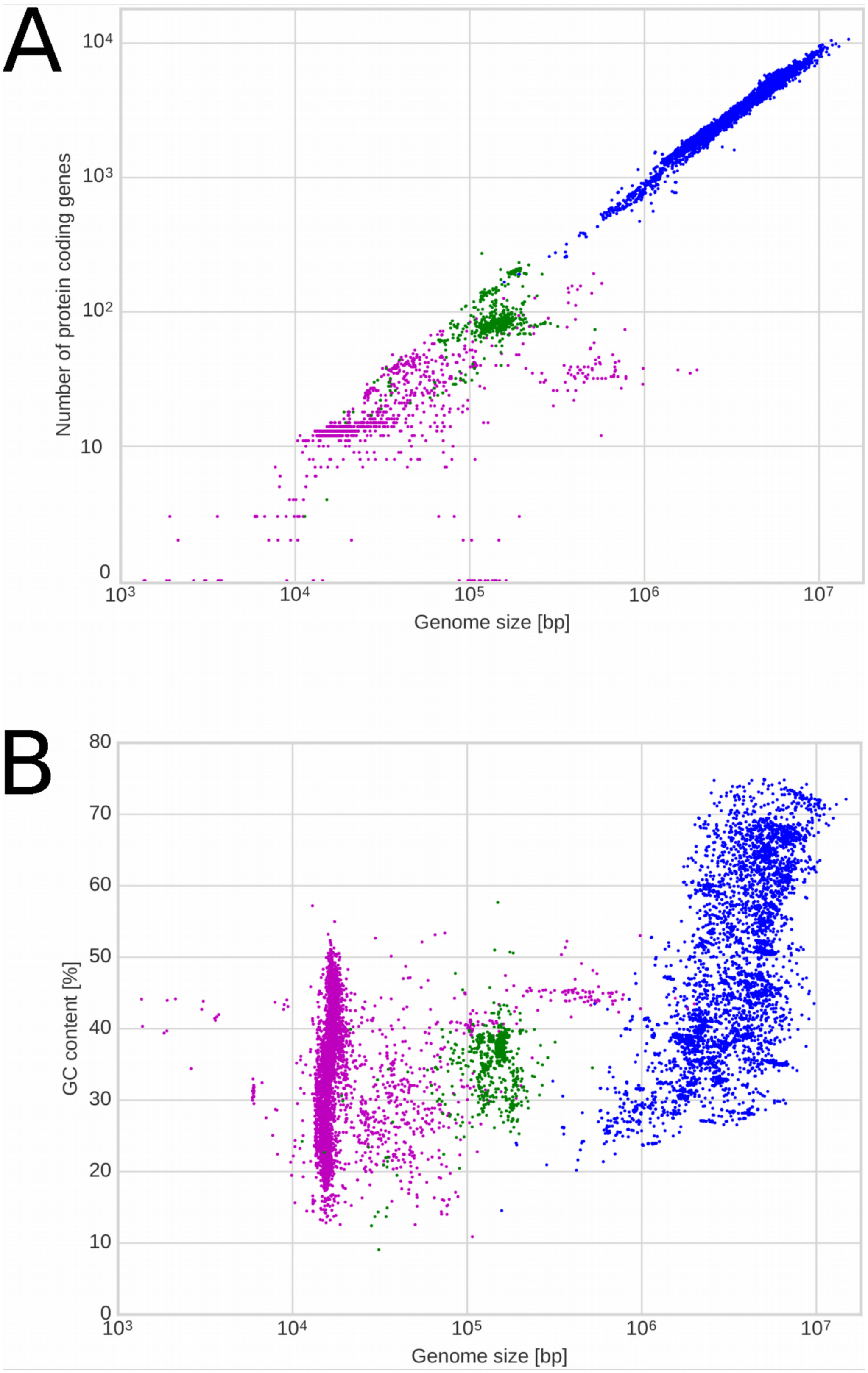

Figure 1: (A): Bacterial genome size correlates to total number of protein-coding genes. The X axis represents genome size and the Y axis represents total number of protein coding genes. **(B):** Bacterial genome sizes compared to GC content. The X axis represents genome size and the Y axis represents GC content. Bacteria are in blue, plastids in green, and mitochondria in magenta.

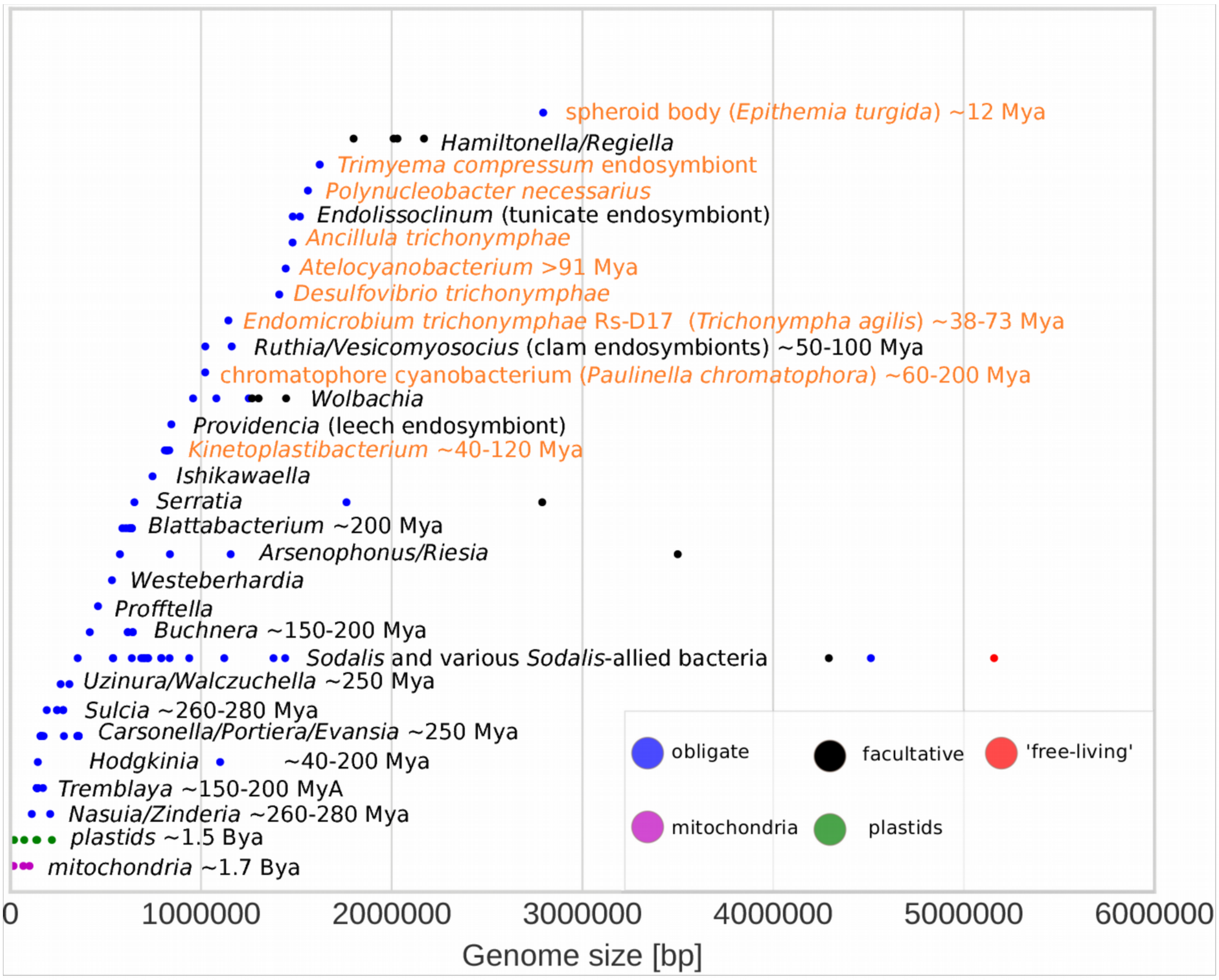

Figure 2: Selected lineages of symbiotic bacteria and organelles sorted according to genome sizes and annotated with estimates of their age. Note that early obligate endosymbionts such as several *Sodalis* lineages or 'spheroid bodies' of diatoms have large genome sizes. Several lineages with a different genus name, but originating from the same ancestor (e.g. *Wolbachia*, *Sodalis* and *Arsenophonus*) are collapsed into a single row to highlight genome reduction associated with facultative or obligate lifestyle. Endosymbionts of animals are in black and endosymbionts of unicellular eukaryotes are in orange. Secondarily expanded gene-poor genomes of mitochondria and plastids are not shown for simplicity.

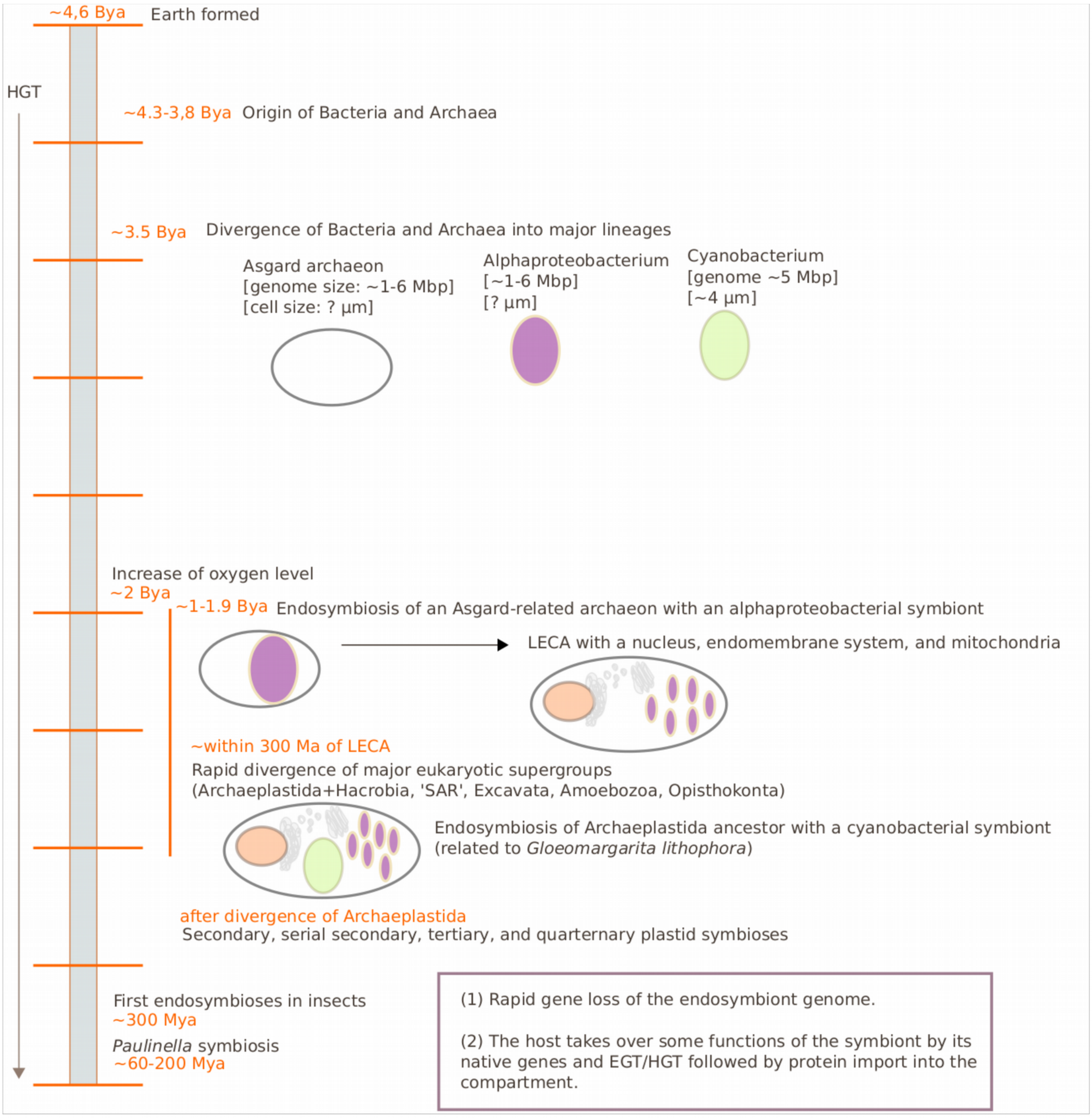

Figure 3: A schematic timeline of almost two billion years of mitochondrial and plastid evolution contrasted to much shorter evolution of the oldest known and most cellular integrated symbioses in multicellular (insects) and unicellular (*Paulinella chromatophora*) eukaryotes. Numerous acquisitions of complex plastids are not shown for simplicity.

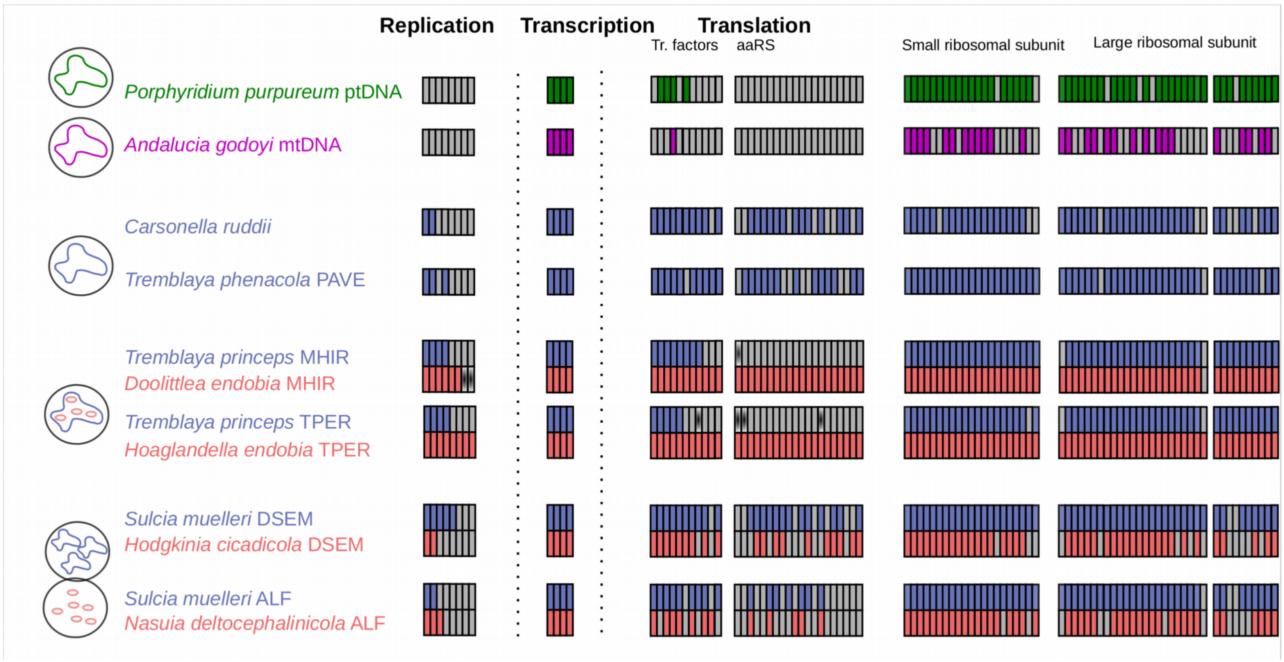

Figure 4: Genetic machinery (replication, transciption, and translation) genes shared by the most highly reduced endosymbiont genomes (*Carsonella, Hodgkinia, Tremblaya, Nasuia*) in comparison to two gene-rich organelle genomes (the mitochondrial genome of *Andalucia godoyi* and the plastid genome of *Porphyridium purpureum*). Three different cellular organizations found in insect endosymbionts are shown: single species symbiosis, obligate 'intrabacterial' co-symbiosis (one endosymbiont inside another), and obligate co-symbiosis with both symbionts present in their own bacteriocytes. Note that that all of the endosymbiont genomes have retained at least a minimal set of DNA polymerase proteins and that the only endosymbiont lineage missing all aminoacyl-tRNA synthetases is *Tremblaya princeps* with intrabacterial symbionts likely supplementing this function.

References:

Alsmark C, Foster PG, Sicheritz-Ponten T *et al.* (2013) Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. *Genome Biology*, **14**, R19.

Alves JMP, Klein CC, da Silva FM *et al.* (2013a) Endosymbiosis in trypanosomatids: the genomic cooperation between bacterium and host in the synthesis of essential amino acids is heavily influenced by multiple horizontal gene transfers. *BMC Evolutionary Biology*, **13**, 190.

Alves JMP, Serrano MG, da Silva FM *et al.* (2013b) Genome evolution and phylogenomic analysis of *Candidatus* Kinetoplastibacterium, the betaproteobacterial endosymbionts of *Strigomonas* and *Angomonas*. *Genome Biology and Evolution*, **5**, 338–350.

Archibald JM (2006) Endosymbiosis: double-take on plastid origins. *Current Biology*, **16**, R690–R692.

Archibald JM, Richards TA (2010) Gene transfer: anything goes in plant mitochondria. *BMC Biology*, **8**, 147.

Aschtgen M-S, Wetzel K, Goldman W, McFall-Ngai M, Ruby E (2016) *Vibrio fischeri*-derived outer membrane vesicles trigger host development. *Cellular Microbiology*, **18**, 488–499.

Balsera M, Soll J, Bölter B (2009) Protein import machineries in endosymbiotic organelles. *Cellular and Molecular Life Sciences*, **66**, 1903–1923.

Bennett GM, Moran NA (2013) Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a phloem-feeding insect. *Genome Biology and Evolution*, **5**, 1675–1688.

Bennett GM, Moran NA (2015) Heritable symbiosis: the advantages and perils of an evolutionary rabbit hole. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 10169–10176.

Bhattacharya D, Price DC, Chan CX *et al.* (2013) Genome of the red alga *Porphyridium purpureum*. *Nature Communications*, **4**, 1941.

Björkholm P, Harish A, Hagström E, Ernst AM, Andersson SGE (2015) Mitochondrial genomes are retained by selective constraints on protein targeting. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 10154–10161.

Bombar D, Heller P, Sanchez-Baracaldo P, Carter BJ, Zehr JP (2014) Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria. *The ISME Journal*, **8**, 2530–2542.

Booth A, Doolittle WF (2015a) Eukaryogenesis, how special really? *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 10278–10285.

Booth A, Doolittle WF (2015b) Reply to Lane and Martin: Being and becoming eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, E4824.

Boscaro V, Felletti M, Vannini C *et al.* (2013) *Polynucleobacter necessarius*, a model for genome reduction in both free-living and symbiotic bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 18590–18595.

Brum FL, Catta-Preta CMC, de Souza W *et al.* (2014) Structural characterization of the cell division cycle in *Strigomonas culicis*, an endosymbiont-bearing trypanosomatid. *Microscopy and Microanalysis*, **20**, 228–237.

Brune A, Dietrich C (2015) The gut microbiota of termites: digesting the diversity in the light of ecology and evolution. *Annual Review of Microbiology*, **69**, 145–166.

Burger G, Gray MW, Forget L, Lang BF (2013) Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biology and Evolution*, **5**, 418–438.

Burger G, Gray MW, Lang BF (2003) Mitochondrial genomes: anything goes. *Trends in Genetics*, **19**, 709–716.

Campbell MA, Leuven JT Van, Meister RC *et al.* (2015) Genome expansion via lineage splitting and genome reduction in the cicada endosymbiont *Hodgkinia*. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 10192–10199.

Charles H, Balmand S, Lamelas A *et al.* (2011) A genomic reappraisal of symbiotic function in the aphid/*Buchnera* symbiosis: reduced transporter sets and variable membrane organisations. *Plos One*, **6**, e29096.

Chen W, Hasegawa DK, Kaur N *et al.* (2016) The draft genome of whitefly *Bemisia tabaci* MEAM1, a global crop pest, provides novel insights into virus transmission, host adaptation, and insecticide resistance. *BMC Biology*, **14**, 110.

Clayton AL, Oakeson KF, Gutin M *et al.* (2012) A novel human-infection-derived bacterium provides insights into the evolutionary origins of mutualistic insect-bacterial symbioses. *Plos Genetics*, **8**, e1002990.

Cornejo-Castillo FM, Cabello AM, Salazar G *et al.* (2016) Cyanobacterial symbionts diverged in the late Cretaceous towards lineage-specific nitrogen fixation factories in single-celled phytoplankton. *Nature Communications*, **7**, 11071.

Cotton JA, McInerney JO (2010) Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 17252–17255.

Curtis BA, Tanifuji G, Burki F *et al.* (2012) Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature*, **492**, 59–65.

Dacks JB, Field MC, Buick R *et al.* (2016) The changing view of eukaryogenesis – fossils, cells, lineages and how they all come together. *Journal of Cell Science*, 10.1242/jcs.178566.

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

Doležal P, Likic V, Tachezy J, Lithgow T (2006) Evolution of the molecular machines for protein import into mitochondria. *Science*, **313**, 314–318.

Dorrell RG, Howe CJ (2015) Integration of plastids with their hosts: Lessons learned from dinoflagellates. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 10247–10254.

Douglas AE (2016) How multi-partner endosymbioses function. *Nature Reviews Microbiology*, **14**, 731–743.

Dubilier N, Bergin C, Lott C (2008) Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. *Nature Reviews Microbiology*, **6**, 725–740.

Duncan RP, Husnik F, Van Leuven JT *et al.* (2014) Dynamic recruitment of amino acid transporters to the insect/symbiont interface. *Molecular Ecology*, **23**, 1608–1623.

Elhenawy W, Debelyy MO, Feldman MF (2014) Preferential packing of acidic glycosidases and proteases into *Bacteroides* outer membrane vesicles. *mBio*, **5**, e00909-14.

Embley TM, Martin W (2006) Eukaryotic evolution, changes and challenges. *Nature*, **440**, 623–630.

Esser C, Ahmadinejad N, Wiegand C *et al.* (2004) A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Molecular Biology and Evolution*, **21**, 1643–1660.

Ettema TJG (2016) Evolution: Mitochondria in the second act. *Nature*, **531**, 39–40.

Gabaldón T, Huynen MA (2004) Shaping the mitochondrial proteome. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, **1659**, 212–220.

Gabaldón T, Huynen MA (2005) Lineage-specific gene loss following mitochondrial endosymbiosis and its potential for function prediction in eukaryotes. *Bioinformatics*, **21**, 144-150.

Gabaldón T, Huynen MA (2007) From endosymbiont to host-controlled organelle: the hijacking of mitochondrial protein synthesis and metabolism. *PLoS Computational Biology*, **3**, e219.

Gottlieb Y, Ghanim M, Gueguen G *et al.* (2008) Inherited intracellular ecosystem: symbiotic bacteria share bacteriocytes in whiteflies. *FASEB Journal*, **22**, 2591–2599.

Gray MW (2015) Mosaic nature of the mitochondrial proteome: implications for the origin and evolution of mitochondria. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 10133–10138.

Gray MW, Doolittle WF (1982) Has the endosymbiont hypothesis been proven? *Microbiological Reviews*, **46**, 1–42.

Gray MW, Lukes J, Archibald JM, Keeling PJ, Doolittle WF (2010) Cell biology. Irremediable complexity? *Science*, **330**, 920–921.

Hansen AK, Moran NA (2011) Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 2849–2854.

Hentschel U, Piel J, Degnan SM, Taylor MW (2012) Genomic insights into the marine sponge microbiome. *Nature Reviews Microbiology*, **10**, 641–654.

Hongoh Y, Sharma VK, Prakash T *et al.* (2008) Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 5555–5560.

Husnik F, McCutcheon JP (2016) Repeated replacement of an intrabacterial symbiont in the tripartite nested mealybug symbiosis. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, E5416–E5424.

Husnik F, Nikoh N, Koga R *et al.* (2013) Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell*, **153**, 1567–1578.

Huynen MA, Duarte I, Szklarczyk R (2013) Loss, replacement and gain of proteins at the origin of the mitochondria. *Biochimica et Biophysica Acta*, **1827**, 224–231.

Izawa K, Kuwahara H, Kihara K *et al.* (2016) Comparison of intracellular *Ca.* Endomicrobium trichonymphae genomovars illuminates the requirement and decay of defense systems against foreign DNA. *Genome Biology and Evolution*, **8**, 3099–3107.

Janouškovec J, Liu S-L, Martone PT *et al.* (2013) Evolution of red algal plastid genomes: ancient architectures, introns, horizontal gene transfer, and taxonomic utility of plastid markers. *PloS One*, **8**, e59001.

Karnkowska A, Vacek V, Zubáčová Z *et al.* (2016) A eukaryote without a mitochondrial organelle. *Current Biology*, **26**, 1274–1284.

Keeling PJ (2013) The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annual Review of Plant Biology*, **64**, 583–607.

Keeling PJ, Archibald JM (2008) Organelle evolution: what's in a name? *Current Biology*, **18**, R345-7.

Keeling PJ, McCutcheon JP, Doolittle WF (2015) Symbiosis becoming permanent: Survival of the luckiest. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 10101–10103.

Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, **9**, 605–618.

Kenyon LJ, Sabree ZL (2014) Obligate insect endosymbionts exhibit increased ortholog length variation and loss of large accessory proteins concurrent with genome shrinkage. *Genome Biology and Evolution*, **6**, 763–775.

Klein CC, Alves JMP, Serrano MG *et al.* (2013) Biosynthesis of vitamins and cofactors in bacterium-harbouring trypanosomatids depends on the symbiotic association as revealed by genomic analyses. *PloS One*, **8**, e79786.

Kneip C, Voss C, Lockhart PJ, Maier UG (2008) The cyanobacterial endosymbiont of the unicellular algae *Rhopalodia gibba* shows reductive genome evolution. *BMC Evolutionary Biology*, **8**, 30.

Knoll AH (2014) Paleobiological perspectives on early eukaryotic evolution. *Cold Spring Harbor Perspectives in Biology*, **6**, a016121.

Koonin EV (2010) The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biology*, **11**, 209.

Koonin EV (2015) Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier? *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, **370**, 20140333.

Kostygov AY, Dobáková E, Grybchuk-Ieremenko A *et al.* (2016) Novel trypanosomatid-bacterium association: evolution of endosymbiosis in action. *mBio*, **7**, e01985.

Ku C, Nelson-Sathi S, Roettger M *et al.* (2015a) Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 10139–10146.

Ku C, Nelson-Sathi S, Roettger M *et al.* (2015b) Endosymbiotic origin and differential loss of eukaryotic genes. *Nature*, **524**, 427–432.

Kurland CG, Andersson SG (2000) Origin and evolution of the mitochondrial proteome. *Microbiology and Molecular Biology Reviews*, **64**, 786–820.

Kuwahara H, Yuki M, Izawa K, Ohkuma M, Hongoh Y (2016) Genome of '*Ca.* Desulfovibrio trichonymphae', an H2-oxidizing bacterium in a tripartite symbiotic system within a protist cell in the termite gut. *The ISME Journal,* 10.1038/ismej.2016.143.

Lambert JD, Moran NA (1998) Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 4458–4462.

Lane N, Martin WF (2015) Eukaryotes really are special, and mitochondria are why. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, E4823.

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

Lee J, Cho CH, Park SI *et al.* (2016) Parallel evolution of highly conserved plastid genome architecture in red seaweeds and seed plants. *BMC Biology*, **14**, 75.

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.

Lill R, Hoffmann B, Molik S *et al.* (2012) The role of mitochondria in cellular iron–sulfur protein biogenesis and iron metabolism. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, **1823**, 1491–1508.

Logacheva MD, Schelkunov MI, Shtratnikova VY *et al.* (2016) Comparative analysis of plastid genomes of non-photosynthetic Ericaceae and their photosynthetic relatives. *Scientific Reports*, **6**, 30042.

Login FH, Balmand S, Vallier A *et al.* (2011) Antimicrobial peptides keep insect endosymbionts under control. *Science*, **334**, 362–365.

Lu H, Chang C, Wilson ACC *et al.* (2016) Amino acid transporters implicated in endocytosis of *Buchnera* during symbiont transmission in the pea aphid. *EvoDevo*, **7**, 24.

Luan J-B, Chen W, Hasegawa DK *et al.* (2015) Metabolic coevolution in the bacterial symbiosis of whiteflies and related plant sap-feeding insects. *Genome Biology and Evolution*, **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. *Proceedings of the Royal Society B-Biological Sciences*, **279**, 2965–2973.

Marin B, Nowack ECM, Melkonian M (2005) A plastid in the making: Evidence for a second primary endosymbiosis. *Protist*, **156**, 425–432.

Martin W, Müller M (1998) The hydrogen hypothesis for the first eukaryote. *Nature*, **392**, 37–41.

Martin WF, Roettger M, Ku C *et al.* (2016) Late mitochondrial origin is pure artefact. *bioRxiv*, 10.1101/055368055368.

McCutcheon JP (2010) The bacterial essence of tiny symbiont genomes. *Current Opinion in Microbiology*, **13**, 73–78.

McCutcheon JP (2016) From microbiology to cell biology: when an intracellular bacterium becomes part of its host cell. *Current Opinion in Cell Biology*, **41**, 132–136.

McCutcheon JP, Von Dohlen CD (2011) An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Current Biology*, **21**, 1366–1372.

McCutcheon JP, Keeling PJ (2014) Endosymbiosis: protein targeting further erodes the organelle/symbiont distinction. *Current Biology*, **24**, R654-655.

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

McCutcheon J, McDonald B, Moran N (2009b) Convergent evolution of metabolic roles in bacterial co-symbionts of insects. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 15394–15399.

McCutcheon JP, Moran NA (2010) Functional convergence in reduced genomes of bacterial symbionts spanning 200 million years of evolution. *Genome Biology and Evolution*, **2**, 708–718.

McCutcheon JP, Moran NA (2011) Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, **10**, 13–26.

Morales J, Kokkori S, Weidauer D *et al.* (2016) Development of a toolbox to dissect host-endosymbiont interactions and protein trafficking in the trypanosomatid *Angomonas deanei*. *BMC Evolutionary Biology*, **16**, 247.

Moran NA (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 2873–2878.

Moran NA, Bennett GM (2014) The tiniest tiny genomes. *Annual Review of Microbiology*, **68**, 195–215.

Moran NA, McCutcheon JP, Nakabachi A (2008) Genomics and evolution of heritable bacterial symbionts. *Annual Review of Genetics*, **42**, 165–190.

Moran NA, Tran P, Gerardo NM (2005) Symbiosis and insect diversification: an ancient symbiont of sap-feeding insects from the bacterial phylum Bacteroidetes. *Applied and Environmental Microbiology*, **71**, 8802–8810.

Motta MCM, Catta-Preta CMC, Schenkman S *et al.* (2010) The bacterium endosymbiont of *Crithidia deanei* undergoes coordinated division with the host cell nucleus. *PloS One*, **5**, e12415.

Nakabachi A, Ishida K, Hongoh Y, Ohkuma M, Miyagishima S (2014) Aphid gene of bacterial origin encodes a protein transported to an obligate endosymbiont. *Current Biology*, **24**, R640-641.

Nakayama T, Kamikawa R, Tanifuji G *et al.* (2014) Complete genome of a nonphotosynthetic cyanobacterium in a diatom reveals recent adaptations to an intracellular lifestyle. *Proceedings of the National Academy of Sciences of the United States of America*, **111**.

Nikoh N, McCutcheon JP, Kudo T *et al.* (2010) Bacterial genes in the aphid genome: absence of functional gene transfer from Buchnera to its host. *Plos Genetics*, **6**, e1000827.

Nowack ECM (2014) *Paulinella chromatophora* – rethinking the transition from endosymbiont to organelle. *Acta Societatis Botanicorum Poloniae*, **83**, 387–397.

Nowack ECM, Grossman AR (2012) Trafficking of protein into the recently established photosynthetic organelles of *Paulinella chromatophora*. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 5340–5345.

Nowack ECM, Melkonian M (2010) Endosymbiotic associations within protists. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, **365**, 699–712.

Nowack ECM, Melkonian M, Glöckner G (2008) Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. *Current Biology*, **18**, 410–418.

Nowack ECM, Price DC, Bhattacharya D *et al.* (2016) Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of *Paulinella chromatophora*. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 12214–12219.

Nowack ECM, Vogel H, Groth M *et al.* (2011) Endosymbiotic gene transfer and transcriptional regulation of transferred genes in *Paulinella chromatophora*. *Molecular Biology and Evolution*, **28**, 407–422.

Oakeson KF, Gil R, Clayton AL *et al.* (2014) Genome degeneration and adaptation in a nascent stage of symbiosis. *Genome Biology and Evolution*, **6**, 76–93.

Oborník M, Lukeš J (2015) The organellar genomes of *Chromera* and *Vitrella*, the phototrophic relatives of apicomplexan parasites. *Annual Review of Microbiology*, **69**, 129–144.

Pittis AA, Gabaldon T (2016) On phylogenetic branch lengths distribution and the late acquistion of mitochondria. *bioRxiv*, 10.1101/064873.

Pittis AA, Gabaldón T (2016) Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature*, **531**, 101–104.

Poliakov A, Russell CW, Ponnala L *et al.* (2011) Large-scale label-free quantitative proteomics of the pea aphid-Buchnera symbiosis. *Molecular & Cellular Proteomics*, **10**, M110.007039.

Ponce-Toledo RI, Deschamps P, López-García P *et al.* (2017) An early-branching freshwater cyanobacterium at the origin of plastids. *Current Biology*, **27**, 386–391.

Prechtl J, Kneip C, Lockhart P, Wenderoth K, Maier UG (2004) Intracellular spheroid bodies of *Rhopalodia gibba* have nitrogen-fixing apparatus of cyanobacterial origin. *Molecular Biology and Evolution*, **21**, 1477–1481.

Price DRG, Duncan RP, Shigenobu S, Wilson ACC (2011) Genome expansion and differential expression of amino acid transporters at the aphid/*Buchnera* symbiotic interface. *Molecular Biology and Evolution*, **28**, 3113–3126.

Price DRG, Feng H, Baker JD *et al.* (2013) Aphid amino acid transporter regulates glutamine supply to intracellular bacterial symbionts. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 320–325.

Qiu H, Price DC, Weber APM *et al.* (2013) Assessing the bacterial contribution to the plastid proteome. *Trends in Plant Science*, **18**, 680–687.

Reyes-Prieto A, Moustafa A (2012) Plastid-localized amino acid biosynthetic pathways of Plantae are predominantly composed of non-cyanobacterial enzymes. *Scientific Reports*, **2**, 955.

Rosas-Pérez T, Rosenblueth M, Rincón-Rosales R, Mora J, Martínez-Romero E (2014) Genome sequence of *Candidatus* Walczuchella monophlebidarum the flavobacterial endosymbiont of *Llaveia axin axin* (Hemiptera: Coccoidea: Monophlebidae). *Genome Biology and Evolution*, **6**, 714–726.

Sabree ZL, Huang CY, Okusu A, Moran NA, Normark BB (2012) The nutrient supplying capabilities of *Uzinura*, an endosymbiont of armoured scale insects. *Environmental Microbiology*, **15**, 1988–1999.

Santos-Garcia D, Latorre A, Moya A *et al.* (2014) Small but powerful, the primary endosymbiont of moss bugs, *Candidatus* Evansia muelleri, holds a reduced genome with large biosynthetic capabilities. *Genome Biology and Evolution*, **6**, 1875–1893.

Ševčíková T, Horák A, Klimeš V *et al.* (2015) Updating algal evolutionary relationships through plastid genome sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte? *Scientific Reports*, **5**, 10134.

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.

Shigenobu S, Wilson ACC (2011) Genomic revelations of a mutualism: the pea aphid and its obligate bacterial symbiont. *Cellular and Molecular Life Sciences*, **68**, 1297–1309.

Shinzato N, Aoyama H, Saitoh S *et al.* (2016) Complete genome sequence of the intracellular bacterial symbiont TC1 in the anaerobic ciliate *Trimyema compressum*. *Genome Announcements*, **4**, e01032-16.

Sloan DB, Alverson AJ, Chuckalovcak JP *et al.* (2012) Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biology*, **10**, e1001241.

Sloan DB, Moran NA (2012a) Endosymbiotic bacteria as a source of carotenoids in whiteflies. *Biology Letters*, **8**, 986–989.

Sloan DB, Moran NA (2012b) Genome reduction and co-evolution between the primary and secondary bacterial symbionts of psyllids. *Molecular Biology and Evolution*, **29**, 3781–3792.

Sloan DB, Nakabachi A, Richards S *et al.* (2014) Parallel histories of horizontal gene transfer facilitated extreme reduction of endosymbiont genomes in sap-feeding insects. *Molecular Biology and Evolution*, **31**, 857–871.

Smith DR, Keeling PJ (2015) Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 10177–10184.

Smith DR, Lee RW (2014) A plastid without a genome: evidence from the nonphotosynthetic green algal genus *Polytomella*. *Plant Physiology*, **164**, 1812–1819.

Soll J, Schleiff E (2004) Protein import into chloroplasts. *Nature Reviews Molecular Cell Biology*, **5**, 198–208.

Spang A, Saw JH, Jørgensen SL *et al.* (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, **521**, 173–179.

Theissen U, Martin W (2006) The difference between organelles and endosymbionts. *Current Biology*, **16**, 1016–1017.

Thiergart T, Landan G, Schenk M, Dagan T, Martin WF (2012) An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. *Genome Biology and Evolution*, **4**, 466–85.

Thompson AW, Foster RA, Krupke A *et al.* (2012) Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science*, **337**, 1546-1550.

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

Tripp HJ, Bench SR, Turk KA *et al.* (2010) Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature*, **464**, 90–94.

Van de Velde W, Zehirov G, Szatmari A *et al.* (2010) Plant peptides govern terminal differentiation of bacteria in symbiosis. *Science*, **327**, 1122–1126.

Waller RF, Jackson CJ (2009) Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. *BioEssays*, **31**, 237–245.

Wang Z, Wu M (2014) Phylogenomic reconstruction indicates mitochondrial ancestor was an energy parasite. *PLoS One*, **9**, e110685.

Wang Z, Wu M (2015) An integrated phylogenomic approach toward pinpointing the origin of mitochondria. *Scientific Reports*, **5**, 7949.

Wernegreen JJ (2002) Genome evolution in bacterial endosymbionts of insects. *Nature Reviews Genetics*, **3**, 850–861.

Williams T a, Embley TM (2014) Archaeal dark matter and the origin of eukaryotes. *Genome Biology and Evolution*, **6**, 474–481.

Williams TA, Foster PG, Cox CJ, Embley TM (2013) An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, **504**, 231–236.

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 4576–4579.

Woolfit M, Bromham L (2003) Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Molecular Biology and Evolution*, **20**, 1545–1555.

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. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 10185–10191.

Wybouw N, Pauchet Y, Heckel DG, Van Leeuwen T (2016) Horizontal gene transfer contributes to the evolution of arthropod herbivory. *Genome Biology and Evolution*, **8**, 1785–1801.

Zaremba-Niedzwiedzka K, Caceres E, Saw J *et al.* (2017) Metagenomic exploration of Asgard arcahea illuminates the origin of eukaryotic cellular complexity. *Nature,* **541**, 353–358.

Zehr JP, Bench SR, Carter BJ *et al.* (2008) Globally distributed uncultivated oceanic N2-fixing cyanobacteria lack oxygenic photosystem II. *Science*, **322**, 1110-1112.

Chapter I

Cell

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

Filip Husnik,[1] Naruo Nikoh,[2] Ryuichi Koga,[3] Laura Ross,[4] Rebecca P. Duncan,[5] Manabu Fujie,[6] Makiko Tanaka,[7] Nori Satoh,[7] Doris Bachtrog,[8] Alex C.C. Wilson,[5] Carol D. von Dohlen,[9] Takema Fukatsu,[3] and John P. McCutcheon[10,*]
[1]Faculty of Science, University of South Bohemia and Institute of Parasitology, Biology Centre ASCR, České Budějovice 370 05, Czech Republic
[2]Department of Liberal Arts, The Open University of Japan, Chiba 261-8586, Japan
[3]National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba 305-8566, Japan
[4]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK
[5]Department of Biology, University of Miami, Coral Gables, FL 33146, USA
[6]DNA Sequencing Section
[7]Marine Genomics Unit
Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan
[8]Department of Integrative Biology, University of California Berkeley, Berkeley, CA 94720, USA
[9]Department of Biology, Utah State University, Logan, UT 84322, USA
[10]Division of Biological Sciences, University of Montana, Missoula, MT 59812, USA
*Correspondence: john.mccutcheon@umontana.edu
http://dx.doi.org/10.1016/j.cell.2013.05.040

SUMMARY

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

INTRODUCTION

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

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

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

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

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

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

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

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

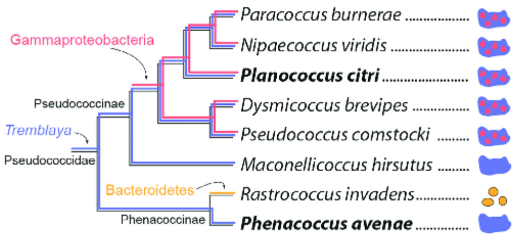

Figure 1. Cladogram of Selected Mealybugs and Their Obligate Symbionts

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

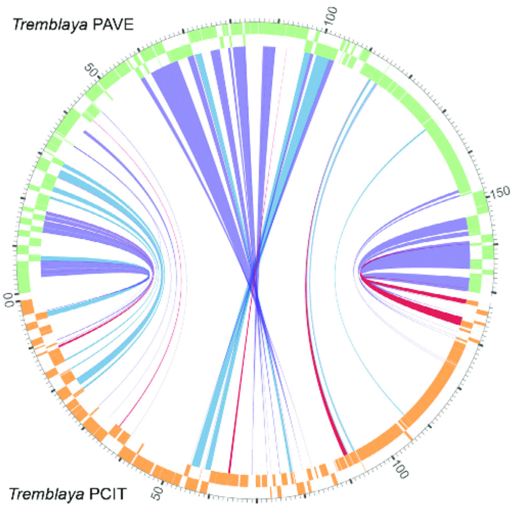

Figure 2. The *Tremblaya* PCIT Genome Is Largely a Subset of the *Tremblaya* PAVE Genome

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

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

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

RESULTS

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

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

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

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

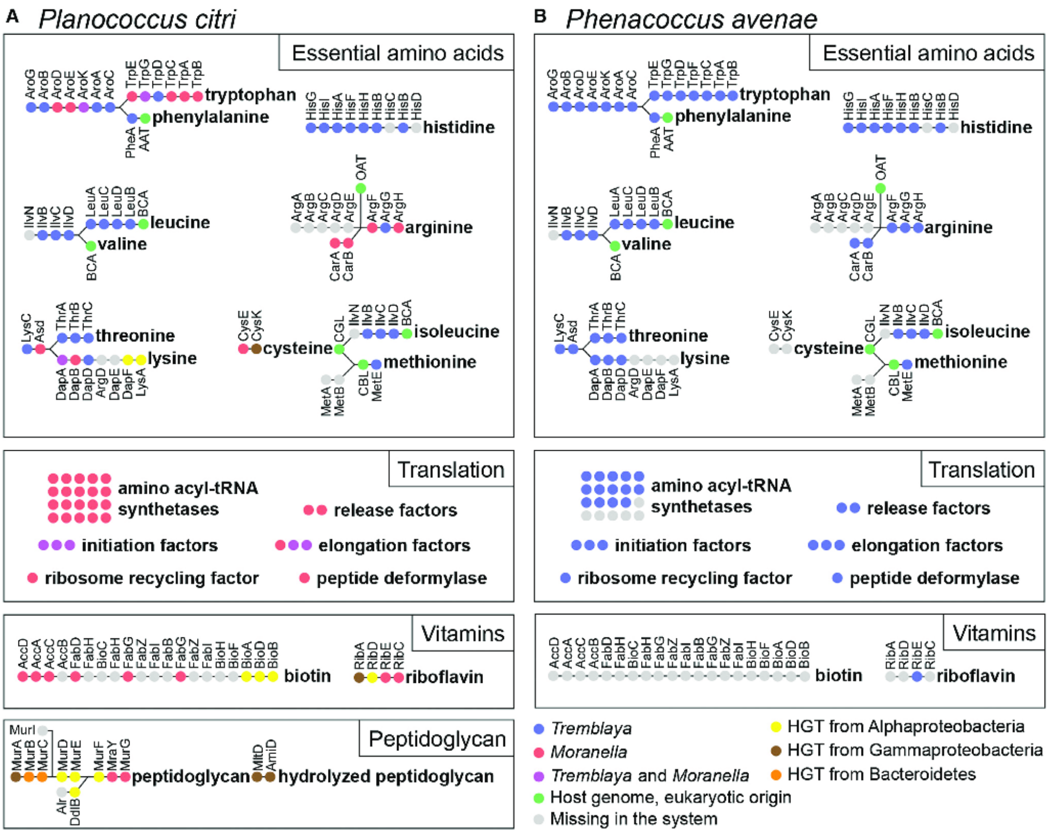

Figure 3. Symbiont Gene Retention and HTG Expression Patterns for the *Pl. citri* and *Ph. avenae* Symbioses

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

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

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

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

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

Cell

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

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

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

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

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

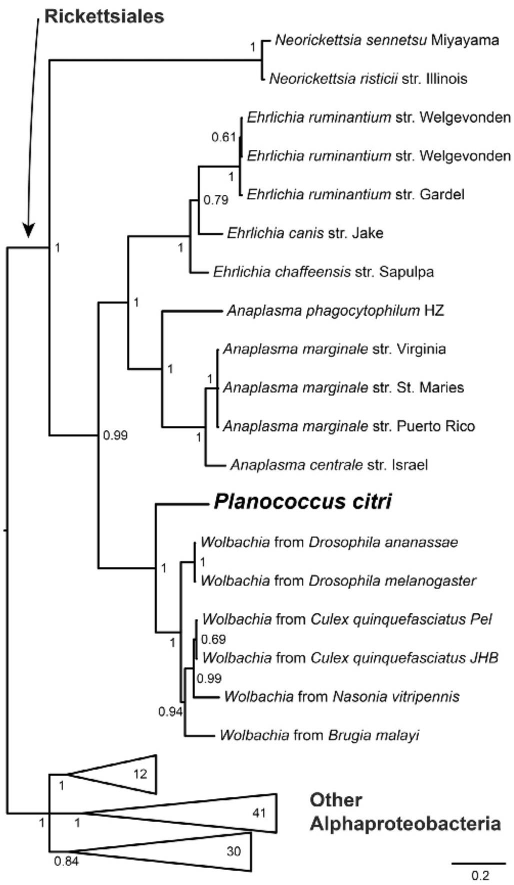

Figure 4. A Representative Phylogenetic Tree Confirming that RibD Is the Result of HGT

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

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

Phylogenetic Analyses Suggest the Source of Most ExHTGs Are Facultative Symbionts

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

Verification that ExHTGs Are Encoded on the Insect Genome

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

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

Probable but Unconfirmed ExHTGs

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

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

DISCUSSION

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

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

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

Table 2. Expression Values for Selected Insect Transcripts

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

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

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

Several Pathways Are Composed of Genes from Multiple Phylogenetic Sources

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

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

Host Genes of Eukaryotic Origin Overexpressed in Bacteriome Tissue

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

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

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

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

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

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

The smallest reported bacterial genomes, which are all from nutritional symbionts of sap-feeding insects, are indistinguishable from organelles when considered only in terms of genome size and gene number (McCutcheon and Moran, 2012). Unlike organelles, however, they tend to retain a certain set of the most critical genes involved in DNA replication, transcription, and translation (McCutcheon, 2010). *Tremblaya* PCIT is strikingly different, as it has lost many genes involved in translation that are retained in other highly reduced genomes (López-Madrigal et al., 2011; McCutcheon and von Dohlen, 2011). This degeneracy, along with its extensive interdependency on *Moranella* and the insect host, makes it difficult to apply an appropriate label to *Tremblaya* PCIT—is it still a bacterium or has it transitioned to something more akin to an organelle? This labeling problem is complicated by the lack of a generally accepted definition of "organelle" (Keeling, 2011; Keeling and Archibald, 2008; Theissen and Martin, 2006). In any case, more important than applying an appropriate label to *Tremblaya* is understanding how the *Pl. citri* symbiosis came to be and how it currently works, as this may provide insight on how host-organelle relationships formed in the general sense of being highly integrated mosaic organisms.

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

EXPERIMENTAL PROCEDURES

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

Insect Strains, DNA and RNA Isolation, and Sequencing

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

ACCESSION NUMBERS

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

SUPPLEMENTAL INFORMATION

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

ACKNOWLEDGMENTS

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

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Supplemental Information

Cell

EXTENDED EXPERIMENTAL PROCEDURES

Phenacoccus avenae Genome Sequencing and Annotation

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

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

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

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

Planococcus citri RNA Preparation and Sequencing

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

RNA-Seq and Differential Expression Analyses

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

RT-qPCR Verification of ExHTG Enrichment in Bacteriome Tissue

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

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

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

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

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

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

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

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

Phylogenetic Analyses

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

Planococcus citri DNA Preparation, Sequencing, and Genome Assembly

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

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

SUPPLEMENTAL REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

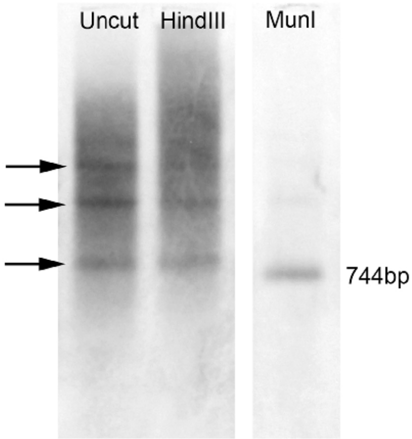

Figure S1. Southern Blot of *Tremblaya* PAVE Plasmid-Like Molecule, Related to Experimental Procedures

Chapter II

MOLECULAR ECOLOGY

Molecular Ecology (2014) 23, 1608–1623 doi: 10.1111/mec.12627

SPECIAL ISSUE: NATURE'S MICROBIOME

Dynamic recruitment of amino acid transporters to the insect/symbiont interface

REBECCA P. DUNCAN,* FILIP HUSNIK,† JAMES T. VAN LEUVEN,‡ DONALD G. GILBERT,§ LILIANA M. DÁVALOS,¶ JOHN P. MCCUTCHEON‡ and ALEX C. C. WILSON*

*Department of Biology, University of Miami, Coral Gables, FL 33146, USA, †Faculty of Science, University of South Bohemia & Institute of Parasitology, Czech Academy of Sciences, Ceske Budejovice 37005, Czech Republic, ‡Division of Biological Sciences, University of Montana, Missoula, MT 59812, USA, §Department of Biology, Indiana University, Bloomington, IN 47405, USA, ¶Department of Ecology and Evolution, and Consortium for Inter-Disciplinary Environmental Research, State University of New York at Stony Brook, Stony Brook, NY 11794, USA

Abstract

Symbiosis is well known to influence bacterial symbiont genome evolution and has recently been shown to shape eukaryotic host genomes. Intriguing patterns of host genome evolution, including remarkable numbers of gene duplications, have been observed in the pea aphid, a sap-feeding insect that relies on a bacterial endosymbiont for amino acid provisioning. Previously, we proposed that gene duplication has been important for the evolution of symbiosis based on aphid-specific gene duplication in amino acid transporters (AATs), with some paralogs highly expressed in the cells housing symbionts (bacteriocytes). Here, we use a comparative approach to test the role of gene duplication in enabling recruitment of AATs to bacteriocytes. Using genomic and transcriptomic data, we annotate AATs from sap-feeding and non sap-feeding insects and find that, like aphids, AAT gene families have undergone independent large-scale gene duplications in three of four additional sap-feeding insects. RNA-seq differential expression data indicate that, like aphids, the sap-feeding citrus mealybug possesses several lineage-specific bacteriocyte-enriched paralogs. Further, differential expression data combined with quantitative PCR support independent evolution of bacteriocyte enrichment in sap-feeding insect AATs. Although these data indicate that gene duplication is not necessary to initiate host/symbiont amino acid exchange, they support a role for gene duplication in enabling AATs to mediate novel host/symbiont interactions broadly in the sap-feeding suborder Sternorrhyncha. In combination with recent studies on other symbiotic systems, gene duplication is emerging as a general pattern in host genome evolution.

Keywords: aphid, bacteriocyte, functional evolution, gene duplication, mealybug, sap-feeding insect

Received 10 July 2013; revision received 3 December 2013; accepted 8 December 2013

Introduction

Interspecific interactions fundamentally impact the evolutionary trajectory of species and have long been known to influence characteristics such as morphology (Schemske & Bradshaw 1999), colour patterns (Sandoval 1994), community structure (Kennedy 2010) and even behaviour (Eberhard 2000). Furthermore, interactions between species shape an organism's genome in ways that are only just beginning to be appreciated. Not only do species interactions influence the genes and pathways directly involved in those interactions, but overall genome content, organization, expression, size and even base composition are influenced by interspecific interactions. The most intriguing examples of how genome evolution is shaped by interspecific interactions are found in obligate, endosymbiotic mutualists. For

Correspondence: Alex C. C. Wilson, Fax: 305 284 3039;
E-mail: acwilson@bio.maimi.edu

example, bacterial nutritional endosymbionts have undergone drastic genome reduction and gene loss in response to evolving an obligate endosymbiotic lifestyle (Shigenobu *et al.* 2000; Nakabachi *et al.* 2006; McCutcheon & Moran 2007, 2012; McCutcheon *et al.* 2009; Sabree *et al.* 2009, 2012a; McCutcheon & von Dohlen 2011; Nikoh *et al.* 2011; Sloan & Moran 2012; Bennett & Moran 2013). Historically, symbiont genomes have received more attention than the genomes of their hosts, but as deep sequencing becomes cheaper and assembly technology advances, host genomes are providing insight into how symbiosis shapes genomes in eukaryotic hosts (International Aphid Genomics Consortium 2010; Kirkness *et al.* 2010; Nygaard *et al.* 2011; Young *et al.* 2011; Husnik *et al.* 2013). Four interesting and novel (given current sampling) features of host genomes include (i) metabolic complementarity with symbionts in essential nutrient biosynthesis (Shigenobu *et al.* 2000; Wilson *et al.* 2010; Hansen & Moran 2011; McCutcheon & von Dohlen 2011; Nygaard *et al.* 2011; Husnik *et al.* 2013), (ii) loss or modulation of immune pathways (Gerardo *et al.* 2010; Kim *et al.* 2011b; Ratzka *et al.* 2013), (iii) maintenance and expression of functional genes acquired horizontally from bacteria other than the obligate symbiont (Nikoh & Nakabachi 2009; Nikoh *et al.* 2010; Husnik *et al.* 2013) and (iv) duplication of genes with functions that may facilitate symbiosis (Ganot *et al.* 2011; Price *et al.* 2011; Young *et al.* 2011; Shigenobu & Stern 2013). Although these features suggest a role for symbiosis in shaping host genomes, some genomic attributes of eukaryotic hosts come from isolated examples and a role for symbiosis in their evolutionary origin remains untested. One way to test the role of symbiosis in shaping host genome evolution is by evaluating specific genomic traits within an evolutionary framework.

An evolutionary framework is especially powerful in evaluating the extensive gene duplication and differential expression of amino acid transporters (AATs) observed in the genome of the pea aphid, *Acyrthosiphon pisum*. This evolutionary pattern may be influenced by the relationship between *A. pisum* and its obligate bacterial endosymbiont, *Buchnera aphidicola*. *A. pisum*, a member of the insect order Hemiptera, feeds on plant phloem sap, a diet deficient in key nutrients such as essential amino acids (Douglas 1993, 2006; Sandstrom & Pettersson 1994; Wilkinson & Douglas 2003). Essential amino acids – that is, amino acids that animals are unable to synthesize *de novo* – are provided to aphids by *Buchnera* in exchange for nonessential amino acids (Shigenobu *et al.* 2000). Supply of nonessential amino acids to *Buchnera* and distribution of essential amino acids from *Buchnera* to host tissues is mediated by amino acid transport across three key membrane barriers that we collectively refer to as the symbiotic interface: (i) the plasma membrane of the specialized aphid cells that house *Buchnera* (bacteriocytes), (ii) the host-derived symbiosomal membrane surrounding individual *Buchnera* cells and (iii) the bacterial inner and outer membranes of individual *Buchnera* (Shigenobu & Wilson 2011). Analyses of transcripts (Hansen & Moran 2011; Price *et al.* 2011; Macdonald *et al.* 2012) and proteins (Poliakov *et al.* 2011) enriched in aphid bacteriocytes suggest that amino acid flux at the aphid/*Buchnera* symbiotic interface is mediated by several aphid AATs from two gene families: the amino acid polyamine organocation (APC) family (transporter classification (TC) #2.A.3) and the amino acid/auxin permease (AAAP) family (TC #2.A.18) (Castagna *et al.* 1997; Saier 2000; Saier *et al.* 2006, 2009). These two AAT families play important nutritional roles in insects (Martin *et al.* 2000; Dubrovsky *et al.* 2002; Colombani *et al.* 2003; Jin *et al.* 2003; Goberdhan *et al.* 2005; Attardo *et al.* 2006; Evans *et al.* 2009). Some aphid AATs enriched in bacteriocytes are paralogs derived from within an aphid-specific gene expansion. The membership of bacteriocyte-enriched AATs to an aphid-specific expansion intrigues us because gene duplication can be a critical source of raw genetic material for evolutionary innovation. While gene duplication is random, duplicates can be maintained in a genome for many reasons, including the evolution of novel functions and/or the spatiotemporal partitioning of ancestral function across paralogs (reviewed in Kondrashov 2012). Finding AAT gene duplicates with enriched expression in bacteriocytes is consistent with the hypothesis that gene duplication plays an important, possibly adaptive, role in recruiting AATs to the symbiotic interface of aphids and other sap-feeders (Price *et al.* 2011). This hypothesis predicts that other sap-feeders with obligate bacterial endosymbionts also maintain duplicated AATs with similar patterns of bacteriocyte enrichment.

Most sap-feeding insects are hemipterans, and thus, testing the role of gene duplication in recruiting AATs to the symbiotic interface can be facilitated with genomic data from sap-feeding and non sap-feeding hemipteran taxa. Despite difficulties resolving higher-level hemipteran relationships (Campbell *et al.* 1995; von Dohlen & Moran 1995; Grimaldi & Engel 2005; Cryan & Urban 2011; Song *et al.* 2012), current understanding of hemipteran suborders can facilitate the selection of appropriate taxa to evaluate whether symbiosis influences AAT evolution. Ideal taxon sampling will span the three major hemipteran suborders of Sternorrhyncha, Auchenorrhyncha and Heteroptera (see Fig. 1). Sternorrhyncha, the suborder that includes aphids, will enable the determination of whether the AAT duplications we discovered in the pea aphid (Price *et al.* 2011)

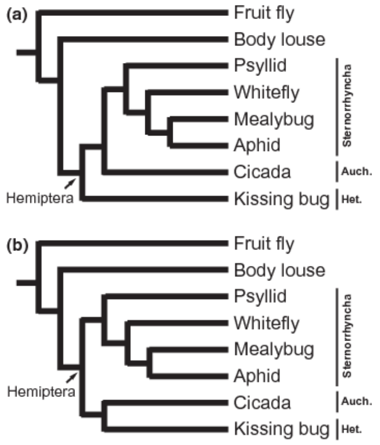

Fig. 1 Alternative hypotheses for phylogenetic relationships among sampled hemipterans. (a) Sternorrhyncha + cicada sister to kissing bug (consistent with Hennig 1981; Song *et al.* 2012). (b) Sternorrhyncha sister to cicada + kissing bug (consistent with Zrzavy 1992; Campbell *et al.* 1995; von Dohlen & Moran 1995; Grimaldi & Engel 2005). Suborders are indicated to the right of taxon names: Sternorrhyncha (aphids, mealybugs, whiteflies and psyllids), Auchenorrhyncha (cicadas) and Heteroptera (kissing bugs).

pre- or postdate diversification of the Sternorrhyncha. Draft genomes and transcriptomes are available for four sternorrhynchan lineages including the pea aphid *A. pisum* (International Aphid Genomics Consortium 2010), the whitefly *Bemisia tabaci* (Wang *et al.* 2010), the potato psyllid *Bactericera cockerelli* (Nachappa *et al.* 2012) and the citrus mealybug *Planococcus citri* (Husnik *et al.* 2013). Auchenorrhyncha, a suborder that independently evolved sap-feeding (Zrzavy 1990, 1992), will provide tests of independence (Weber & Agrawal 2012) in AAT evolutionary patterns. Here, we generate a transcriptome for an auchenorrhynchan, the cicada *Diceroprocta semicincta*. Lastly, Heteroptera comprises mostly non-sap-feeders, and inclusion of this suborder will provide a test of whether gene duplication in AATs is influenced by a general aspect of hemipteran biology unrelated to diet. A transcriptome is available for a blood-feeding heteropteran, the kissing bug *Rhodnius prolixus* (Ribeiro *et al.* 2014).

In this study, we use comparative transcriptomics and gene expression analyses to test the role of gene duplication in recruiting AATs to the sap-feeder symbiotic interface by pinpointing the relative timing of gene duplication in hemipteran AATs and quantifying the expression of AATs in bacteriocytes. Importantly, the sap-feeding taxa we sampled have comparable symbiotic interfaces to the aphid/*Buchnera* system: one or more obligate, bacterial symbionts residing within host-derived membrane-bound compartments inside bacteriocytes (Table 1). Remarkably, we find that numerous gene duplications took place independently in sap-feeders of the suborder Sternorrhyncha. Consistent with our observations of aphid AATs (Price *et al.* 2011), we find that citrus mealybug paralogous AATs are also differentially expressed at the symbiotic interface, with some paralogs enriched in bacteriocytes. Together, these data indicate that gene duplication has broadly played a role in recruiting amino acid transporters to operate at the symbiotic interface of sternorrhynchans.

Materials and methods

Insect collection and cultivation

Adult female cicadas (*Diceroprocta semicincta*) were collected in Tucson, AZ, and preserved in RNAlater (Ambion). Citrus mealybugs (*Planococcus citri*) were collected from coleus plants in the Utah State University greenhouse in Logan, Utah (von Dohlen *et al.* 2001; McCutcheon & von Dohlen 2011), and raised on coleus plants in the laboratory at 25 °C. Pea aphids (*Acyrthosiphon pisum*) from the genome line LSR1 (Caillaud *et al.* 2002) were raised on fava plants at 20 °C. Both insect colonies were maintained under a photoperiodicity of 16:8 (L:D).

Transcriptome sequencing and assembly

For cicada transcriptomes, total RNA was purified from either (i) bacteriocytes or (ii) a combination of head, legs and wing muscles (hereafter referred to as 'insect') dissected from RNAlater-preserved adult female cicadas according to the manufacturer's protocols (MoBio PowerBiofilm RNA Isolation Kit). RNA was sent to Hudson Alpha Institute for Biotechnology for barcoded library preparation and Illumina HiSeq sequencing. Paired-end 100-nt reads were filtered to a minimum quality of 20 over 95% of the read, and 5 nt were trimmed from the 5′ end. Insect (42 688 895 read pairs) and bacteriocyte (53 510 432 read pairs) reads were assembled into separate insect and bacteriocyte transcriptomes in TRINITY (25 January 2012 release) (Haas *et al.* 2013) using kmer_length = 25 and min_contig_length = 48.

Mealybug whole body, paired-end, 100-nt reads (Husnik *et al.* 2013) from a mixed population of adult and penultimate instar females were filtered to a minimum quality of 30 over 95% of the read. The resulting 58 812 530 read pairs were assembled with two different assembly packages. First, reads were assembled in

Table 1 Hemipteran taxa and associated symbionts

Taxon	Diet	Obligate symbiont(s)	Symbiont classification	Symbiont localization	Symbiosomal membrane
Sternorrhyncha					
Pea aphid	Phloem sap	*Buchnera aphidicola*[a]	γ-Proteobacteria	Bacteriocytes	Yes
Citrus mealybug	Phloem sap	*Tremblaya princeps*[b]	β-Proteobacteria	Bacteriocytes	Yes
		Moranella endobia[c]	γ-Proteobacteria	Nested within *Tremblaya*[d]	Not applicable
Whitefly	Phloem sap	*Portiera aleyrodidarum*[e]	γ-Proteobacteria	Bacteriocytes	Yes
Potato psyllid	Phloem sap	*Carsonella ruddii*[f]	γ-Proteobacteria	Bacteriocytes	Yes
Auchenorrhyncha					
Cicada	Xylem sap	*Sulcia muelleri*[g]	Bacteroidetes	Bacteriocytes	Yes
		Hodgkinia cicadicola[h]	α-Proteobacteria	Bacteriocytes	
Heteroptera					
Kissing bug	Vertebrate blood	*Rhodococcus rhodnii*[i]	Actinobacteria	Gut lumen	No

References: (Munson *et al.* 1991)[a]; (Thao *et al.* 2002)[b]; (McCutcheon & von Dohlen 2011)[c]; (von Dohlen *et al.* 2001)[d]; (Thao & Baumann 2004)[e]; (Thao *et al.* 2000)[f], (Moran *et al.* 2005)[g]; (McCutcheon *et al.* 2009)[h]; (Goodfellow & Alderson 1977)[i].

VELVET (v.1.2) (Zerbino & Birney 2008) and OASES (v.0.2) (Schulz *et al.* 2012) using variable k-mer lengths (between 33 and 63 nt), and resulting assemblies were merged into one master assembly. Second, reads were assembled in TRINITY using default parameters (kmer_length = 25).

The whitefly (*Bemisia tabaci*) transcriptome was re-assembled using 170 884 234 RNA-seq read pairs from a mixed population of adult males and females (NCBI BioProject PRJNA89143). Reads were assembled with RNA assemblers VELVET/OASES v.1.2.03/v.0.2.06 (2012.02), SOAPDENOVO-TRANS v.2011.12.22 and TRINITY (17 March 2012 release), using multiple options. EVIDENTIALGENE TR2AACDS pipeline software was used to process the many resulting assemblies by coding sequences, translate to proteins, score gene evidence and classify/reduce to a biologically informative transcriptome of primary and alternate transcripts. The gene set is publicly available at http://arthropods.eugenes.org/EvidentialGene/arthropods/whitefly/whitefly1eg6/.

De novo identification of hemipteran amino acid transporters

Amino acid transporters (AATs) were identified using HMMER (v.3.0) (Eddy 2009) from transcriptomes of cicada, mealybug, whitefly, the potato psyllid *Bactericera cockerelli* (Nachappa *et al.* 2012; mixed population of adult males and females) and the kissing bug *Rhodnius prolixus* (NCBI BioProject PRJNA191820; mixed developmental stages and sexes). Briefly, using a stand-alone PERL script underlying the open reading frame (ORF) prediction webserver hosted by the Proteomics/Genomics Research Group at Youngstown State University (http://proteomics.ysu.edu/tools/OrfPredictor.html), transcripts were translated into all six reading frames. As described previously (Price *et al.* 2011), translated transcripts were searched for conserved functional domains associated with the APC (TC # 2.A.3) and AAAP (TC # 2.A.18) families of amino acid transporters (Castagna *et al.* 1997; Saier 2000; Saier *et al.* 2006, 2009) in HMMER v.3.0 (Eddy 2009; Finn *et al.* 2011). Transcripts significantly matching APC or AAAP domains ($e \leq 0.001$) were verified by BLASTX searches against the NCBI refseq database and retained for further analyses if they showed a significant ($e \leq 0.001$) similarity to the APC or AAAP sequences from the fruit fly *Drosophila melanogaster* and/or *A. pisum*.

Alleles and splice variants were collapsed into a conservative set of representative transcripts for each insect by one of the following two methods depending on availability of genome sequence data: (i) draft genome assemblies are available for mealybugs (Husnik *et al.* 2013) and kissing bugs (unpublished; hosted at vectorbase.org and NCBI), so we validated loci by mapping transcripts to genomic scaffolds by BLASTN searches. Of the transcripts mapping to the same region of a particular scaffold(s), the transcript encoding the longest protein was kept to represent the gene locus. In a few cases, 2-3 partial transcripts were merged into a single locus for phylogenetic analyses (Tables S1–S4 in Appendix S1, Supporting Information). In all cases, partial transcripts mapped side by side to genomic scaffolds on the same strand. Additionally in all cases, the partial transcript mapping upstream in the genome aligned to the 5′ end of other, full-length AAT loci and the downstream partial transcript aligned to the 3′ end. (ii) In contrast, whiteflies, psyllids and cicadas lack draft genome assemblies. In these insects, transcripts that have been diverging for a short period of time were collapsed into representative loci. Time of transcript divergence was determined by estimating the pairwise rate of synonymous substitutions (dS) by the Goldman and Yang method (Goldman & Yang 1994), a common proxy for relative age of homologous gene pairs (e.g. paralogs within a species or orthologs between species) (Lynch & Conery 2000). We collapsed all transcripts with a dS of less than 0.25, keeping the longest sequence to represent the locus (Appendix S2, Supporting Information). This cut-off dS (0.25) is the average dS between orthologs of two aphid species (*A. pisum* and *Myzus persicae*) that diverged between 32 and 53 million years ago (International Aphid Genomics Consortium 2010; Kim *et al.* 2011a). When closely related transcripts for a particular taxon were partial and nonoverlapping or had a very short region of overlap (50 bp or less), we removed the shortest of the pair to ensure conservative estimates of locus number. To confirm the accuracy of using pairwise dS to collapse related transcripts into loci, we performed the same analysis on related aphid paralogs in the APC gene family (Appendix S2, Supporting Information), all of which map to unique regions of the aphid genome (Price *et al.* 2011). We found three aphid-specific paralogs with pairwise dS measurements below 0.25 (Appendix S2, Supporting Information), indicating that our approach to estimate locus number may collapse true paralogs that duplicated relatively recently. Thus, importantly, our estimation of locus number is conservative.

Phylogenetic analyses

Gene phylogenies for the APC and AAAP amino acid transporter families were estimated using sequences from citrus mealybug, potato psyllid, whitefly, cicada and kissing bug as well as previously annotated AATs (Price *et al.* 2011) from the pea aphid, the human body louse (*Pediculus humanus*), the fruit fly (*D. melanogaster*), a tick (*Ixodes scapularis*) and humans (*Homo sapiens*). Outgroup sequences were aphid and/or fruit fly genes closely related to APC and AAAP gene families and members of the same transporter superfamily (Price *et al.* 2011). Full-length protein sequences were aligned in MAFFT (Katoh *et al.* 2002) using default parameters, and resulting alignments were trimmed in TRIMAL v.1.2 (Capella-Gutiérrez *et al.* 2009) using a gap threshold of 25%.

Phylogenies were estimated using maximum-likelihood (ML) and Bayesian methods. ML phylogenies were estimated in RAXML v.7.2.8 (Stamatakis 2006; Ott *et al.* 2007) using the protein evolution model LG+G [the best-fit model as determined by PROTTEST v.2.4 using the Akaike Information Criterion (Abascal *et al.* 2005)] and the fast bootstrap option. The number of bootstrap replicates for each analysis was chosen by the bootstrap convergence criterion 'autofc' implemented in RAXML. Bayesian phylogenies were reconstructed in MRBAYES v.3.1.2 (Huelsenbeck & Ronquist 2001; Ronquist & Huelsenbeck 2003) using two runs with 4 chains per run. The LG protein substitution matrix is not available in MRBAYES, so phylogenies were inferred using WAG+G. Analyses were allowed to run until the standard deviation of split frequencies between runs dropped below 0.05. Convergence of estimated parameters was confirmed in TRACER v.1.5 (Rambaut & Drummond 2007) and of topology in AWTY (Nylander *et al.* 2008), assuming a burn-in of 10% of generations. The criteria supported convergence, so the first 10% of generations were discarded and phylogenies sampled in the remaining generations were used to estimate a 50% majority-rule consensus tree.

The AAAP family contained a large amount of sequence divergence, preventing convergence of the Markov chain Monte Carlo (MCMC) in Bayesian phylogenetic analyses. Therefore, we estimated the phylogeny of a reduced set of AAAP genes corresponding to a monophyletic clade supported in a preliminary maximum-likelihood analysis (Fig. S1 in Appendix S1, Supporting Information).

Gene conversion analyses

Lineage-specific AAT expansions were assessed for the possibility of gene conversion using the program GENECONV (Sawyer 1989). Codon alignments were produced by the CLUSTALW plugin of SEAVIEW (Gouy *et al.* 2010) and run in GENECONV using three different mismatch penalties, g0, g1 and g2. Applying different mismatch penalties to the analysis facilitates the identification of recent gene conversion and ancient gene conversion that may be partially masked by the accumulation of different substitutions between paralogs.

Expression analysis by quantitative reverse transcriptase PCR

Expression profiles of select AATs were measured by quantitative reverse transcriptase PCR (qRT–PCR) in whole bodies and bacteriocytes of adult female LSR1 pea aphids, a mixture of adult and penultimate female citrus mealybugs and adult female potato psyllids [from the same colonies used for the potato psyllid transcriptome (Nachappa *et al.* 2012)]. Bacteriocytes were dissected from 100 female aphids, mealybugs or psyllids in 0.9% RNase-free NaCl and immediately stabilized by placing in TRI Reagent (Ambion). Total RNA was extracted from dissected bacteriocytes and whole female bodies of each insect using the TRI Reagent procedure (Ambion), treated with DNase I in solution and cleaned up using the RNeasy Mini Kit (Qiagen). First-strand cDNA was synthesized from 500 ng of RNA from each tissue, using qScript cDNA Supermix (Quanta Biosciences) and following the manufacturer's protocol.

qRT–PCR assays were performed as previously described (Price *et al.* 2011) using one biological replicate and three technical replicates for each gene/tissue. Primers were subject to BLASTN searches against genomic and/or transcriptomic data sets using a word length of 7, an expect threshold of 1000 and without the low-complexity filter. In all cases, only the target sequence was returned as a hit for each pair of forward and reverse primers. To confirm that primers amplified only one locus, we analysed melt curves from our qRT–PCR results. With the exception of one gene, which was discarded from analysis, all melt curves showed one clear peak, indicating a single product. No template controls and no reverse transcriptase controls (controlling for RNA contaminated with gDNA) were run in parallel with unknown samples. Identifiers, sequences, amplification efficiency and optimization details for primers used in qRT–PCR assays are listed in Table S6 (Appendix S1, Supporting Information). Expression for target genes within a particular insect was compared between whole insect and bacteriocytes using $2^{-\Delta\Delta CT}$ methodology (Livak & Schmittgen 2001) with expression normalized to either glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) in aphids or the 60S ribosomal protein L7 (*RPL7*) in mealybugs and psyllids. Expression data within each insect were collectively normalized by converting ΔC_T to z-scores as follows:

$$z = -10 \times \left(\frac{\Delta C_T - \overline{\Delta C_T}}{\sigma_{\Delta C_T}}\right)$$

Normalized expression values were compiled into a heat map where $z > 0$ (high expression) is represented as yellow and $z < 0$ (low expression) is represented as blue.

Differential expression quantification

Global differential expression between mealybug insect and bacteriocyte tissues was quantified for mealybug AATs using the whole body transcriptome data from this study and previously published bacteriocyte transcriptome data (Husnik *et al.* 2013). Differential expression analyses were conducted with the PERL script pipeline implemented in TRINITY. Briefly, raw RNA-seq reads were mapped to transcripts using BOWTIE v.0.12.7 (Langmead *et al.* 2009), and mapped reads were counted by RSEM v.1.1.18 (Li & Dewey 2011). Data were normalized by TMM (trimmed mean of M values), and transcripts significantly differentially expressed between whole body and bacteriocytes were identified using the Bioconductor package EDGER v.2.10 (Robinson *et al.* 2010). Digital expression values of differentially expressed transcripts are presented in Appendix S3 (Supporting Information) as 'fragments per kilobase of exon per million fragments mapped' (FPKM). Differential expression was not quantified for cicada bacteriome vs. insect tissues because cicadas lacked gene duplications.

Results and Discussion

Nutrient amino acid transporter families are expanded in the Sternorrhyncha

Consistent with our pea aphid work (Price *et al.* 2011), all sternorrhynchan hemipterans we sampled (Table 1, Fig. 1) possessed expanded amino acid transporter (AAT) families relative to non sap-feeding insects (kissing bug, human body louse and the fruit fly) (Table 2). In particular, citrus mealybugs, potato psyllids and whiteflies possessed 36-38 AAT loci across both gene families; relatively large AAT numbers compared with the 20 AAT loci in the non sap-feeding hemipteran annotated here (kissing bug; Table 2) and 22-28 AAT loci in other insects annotated by Price *et al.* (2011) (the fruit fly *D. melanogaster*, the body louse *P. humanus*, the honey bee *Apis mellifera*, the flour beetle *Tribolium castaneum*, the silkworm moth *Bombyx mori*, the wasp *Nasonia vitripennis* and the mosquito *Anopheles gambiae*). In contrast, we identified only 26 AAT loci in cicada, a sap-feeder belonging to the hemipteran suborder Auchenorrhyncha (Table 1, Fig. 1).

Table 2 Amino acid transporters in sampled insects

	APC Loci	AAAP Loci	Total
Pea aphid	18[a]	22[a]	40
Citrus mealybug	10[b]	28[b]	38
Whitefly	12[c]	24[c]	36
Potato psyllid	13[c]	25[c]	38
Cicada	10[c]	16[c]	26
Kissing bug	7[b]	13[b]	20
Human body louse	8[a]	13[a]	21
Fruit fly	10[a]	17[a]	27

[a]Distinct loci confirmed by mapping transcripts to genomic scaffolds (Price *et al.* 2011).
[b]Distinct loci confirmed by mapping transcripts to genomic scaffolds (this study).
[c]Estimated number of loci based on the rate of synonymous substitutions (dS) between paralogs being greater than 0.25.

Amino acid transporter expansions in sap-feeding insects result from both ancient and recent gene duplication events

To clarify the evolutionary mechanism and timing of events that led to AATs expanding in sap-feeding insects, we estimated phylogenies for the APC and AAAP amino acid transporter families (Figs 2a and 3a). The phylogenies revealed that gene duplications occurred on two timescales. First, two ancient gene duplication events (one in each gene family) pre-date hemipteran diversification (marked by pale green boxes in Figs 2a and 3a). Second, consistent with our previous observation in aphids (Price *et al.* 2011), multiple, more recent, gene duplications occurred independently in sternorrhynchan taxa following their divergence from a common ancestor (marked by grey boxes in Figs 2a and 3a). In contrast, our analyses failed to support any Auchenorrhyncha-specific gene duplications in either AAT family. That said, in four instances (marked in Figs 2a and 3a with asterisks), we found phylogenetic support for close relationships between 2-3 cicada (auchenorrhynchan) loci and one kissing bug (heteropteran) locus. Two scenarios could explain these close relationships. First, AAT duplication could have taken place independently in the linage leading to cicadas and no gene duplication took place in kissing bugs, but sequence similarity among orthologs prevents resolution of cicada-specific clades. Second, assuming species tree B (Fig. 1b), gene duplications took place in the common ancestor of cicadas and kissing bugs, but paralogs were only retained in cicada. Of the two scenarios, the second is the least parsimonious, requiring that cicadas retain their paralogs and that kissing bugs lose all but one paralog in three independent instances.

In our pea aphid work (Price *et al.* 2011), we found that aphid AAT paralogs were tandemly arrayed in the genome. Although new AATs in this study were annotated from transcriptome data, a draft genome assembly for the citrus mealybug (Husnik *et al.* 2013) enabled us to preliminarily assess paralog arrangement in that genome. In the mealybug AAAP expansion (Fig. 3a), three pairs of paralogs map to different regions of the same scaffold within ~4 kbp or less of each other (Fig. 4, Table S2 in Appendix S1, Supporting Information), indicating that these paralogs are tandemly arrayed in the mealybug genome. These tandemly arrayed paralogs thus resulted from localized gene duplication (as opposed to whole genome duplication). No other mealybug AAT loci shared a scaffold (Tables S1-S2 in Appendix S1, Supporting Information), which could at least in part be due to the poor quality of the assembly (Husnik *et al.* 2013). In the kissing bug genome, several transcripts mapped to the same scaffold (Tables S3-S4 in Appendix S1, Supporting Information), but were usually separated by large genomic regions between 19 kbp and 1.2 Mbp. The only exception was that two loci were separated by 5.7 kbp (Tables S3-S4 in Appendix S1, Supporting Information). Despite the short distance between those two loci, our phylogeny (Fig. 3a) indicates that they did not result from a recent gene duplication event in the lineage leading to kissing bugs.

Amino acid transporter evolution within the Sternorrhyncha

One unexpected result of this work is finding that AATs have undergone gene family expansions independently in each of the sternorrhynchans we sampled (aphids, mealybugs, psyllids and whiteflies). Consistent molecular and morphological phylogenetic support for the monophyly of Sternorrhyncha (Hennig 1981; Campbell *et al.* 1995; von Dohlen & Moran 1995; Grimaldi & Engel 2005; Cryan & Urban 2011; Song *et al.* 2012) indicates that aphids, mealybugs, whiteflies and psyllids inherited sap-feeding from their common ancestor. We thus assume that the common ancestor also had an amino acid-provisioning symbiont that was later replaced in three, or perhaps all four, lineages we sampled (explaining why each lineage has a different symbiont, Table 1). Importantly, this common ancestor required that AATs mediate host/symbiont amino acid exchange. Retention of independently duplicated paralogs in sternorrhynchans could be explained in four ways. First, our understanding that symbiosis pre-dates sternorrhynchan diversification could be wrong, and each lineage independently evolved symbiosis and comparable symbiotic interfaces. Second, the importance of different transporters could depend on the symbiont lineage. Third, some AAT gene duplications in these taxa could appear to be more recent than they truly are if tandem arrays of paralogs have undergone concerted evolution through gene conversion or nonhomologous crossing-over after the major sternorrhynchan lineages (aphids, mealybugs and other scale insects, whiteflies and psyllids) began diversifying (e.g. Colbourne *et al.* 2011). We found evidence of gene conversion only among a few paralogs in aphids and whiteflies (Table S5 in Appendix S1, Supporting Information), indicating that AAT paralogs are largely evolving independently of one another. However, we cannot rule out the possibility that gene conversion played a more important role in paralog diversification at some time in the past. Fourth, consistent with our previous discovery of male-biased AAT paralogs in aphids (Duncan *et al.* 2011), many paralogs are probably retained in sternorrhynchan genomes for lineage-specific roles not related to symbiosis. Most aphid and mealybug AAT paralogs are

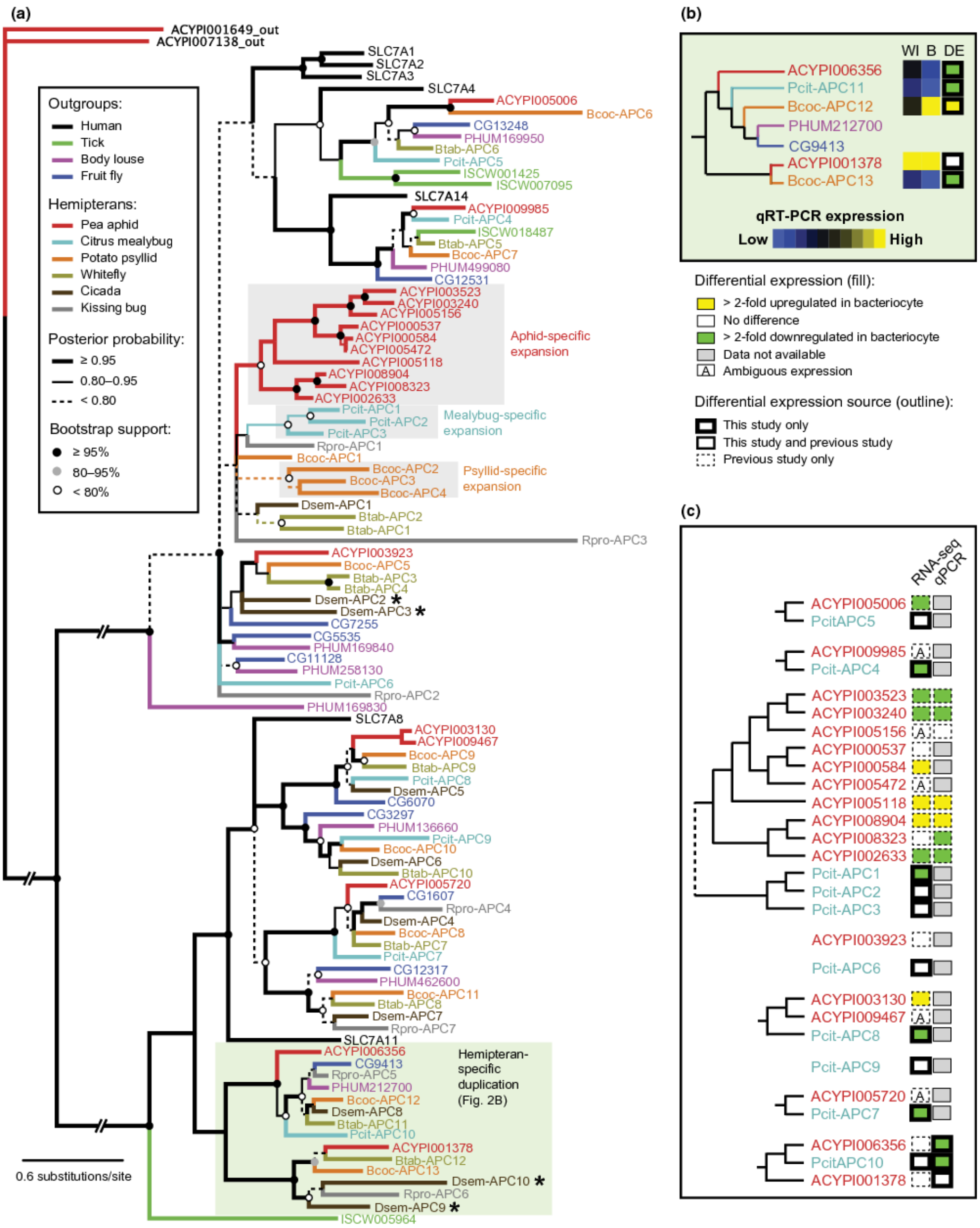
(a)
ACYPI001649_out
ACYPI007138_out
Outgroups:
Human
Tick
Body louse
Fruit fly
Hemipterans:
Pea aphid
Citrus mealybug
Potato psyllid
Whitefly
Cicada
Kissing bug
Posterior probability:
≥ 0.95
0.80–0.95
< 0.80
Bootstrap support:
≥ 95%
80–95%
< 80%
0.6 substitutions/site
Aphid-specific expansion
Mealybug-specific expansion
Psyllid-specific expansion
Hemipteran-specific duplication (Fig. 2B)
(b)
WI B DE
ACYPI006356
Pcit-APC11
Bcoc-APC12
PHUM212700
CG9413
ACYPI001378
Bcoc-APC13
qRT-PCR expression
Low
High
Differential expression (fill):
> 2-fold upregulated in bacteriocyte
No difference
> 2-fold downregulated in bacteriocyte
Data not available
Ambiguous expression
Differential expression source (outline):
This study only
This study and previous study
Previous study only
(c)
RNA-seq
qPCR
ACYPI005006
PcitAPC5
ACYPI009985
Pcit-APC4
ACYPI003523
ACYPI003240
ACYPI005156
ACYPI000537
ACYPI000584
ACYPI005472
ACYPI005118
ACYPI008904
ACYPI008323
ACYPI002633
Pcit-APC1
Pcit-APC2
Pcit-APC3
ACYPI003923
Pcit-APC6
ACYPI003130
ACYPI009467
Pcit-APC8
Pcit-APC9
ACYPI005720
Pcit-APC7
ACYPI006356
PcitAPC10
ACYPI001378

Fig. 2 APC (TC # 2.A.3) phylogeny and bacteriocyte expression. (a) Bayesian gene phylogeny for amino acid transporters (AATs) in the APC family. Hemipteran-specific gene duplications and taxon-specific expansions are highlighted with green or grey boxes, respectively. Asterisks denote possible cicada-specific paralogs. Branches are colour-coded based on taxon and clade support ≥ 50% (posterior probability and ML bootstrap support) is indicated on branches/nodes as described in the key. (b) qRT–PCR expression data generated in this study for Hemiptera-specific gene duplication are presented both as a heat map for whole insect ('WI') and bacteriocyte ('B') and as differential expression ('DE') between bacteriocyte and whole insect. Heat map expression data are normalized across all tissues and genes within each insect, but not across insects. (c) Differential expression between whole insect and bacteriocyte is indicated for aphid and mealybug genes in boxes to the right of gene IDs, as indicated in the key. RNA-seq differential expression data for aphids are from Hansen and Moran (2011) and Macdonald *et al.* (2012). qRT–PCR data not generated here are from Price *et al.* (2011). Expression is marked as ambiguous ('A') if different transcripts or data sets show inconsistent relative bacteriocyte expression.

not enriched in bacteriocytes, supporting a role for nonsymbiotic factors in driving the maintenance of AAT paralogs in these insects. Given the important role that AATs play broadly in animals, it is not surprising that AAT paralog maintenance in sternorrhynchan genomes is not only driven by symbiosis. For example, nutrient AATs also mediate amino acid uptake from the gut into hemolymph (insect blood) (Colombani *et al.* 2003; Morris *et al.* 2009; Price *et al.* 2011). Further, some nutrient AATs play a role in nutrient sensing (Colombani *et al.* 2003; Attardo *et al.* 2006). Lastly, some AATs transport neurotransmitters, likely explaining their expression in aphid heads (Price *et al.* 2011). Accepting their many roles, it is not surprising that some insects without intracellular, amino acid-provisioning symbionts maintain lineage-specific AAT duplications (e.g. Fig 3 and Price *et al.* 2011). However, that sternorrhynchan sap-feeding insects maintain more AAT duplications in their genomes than other, non sap-feeding insects is compelling and suggests that gene duplication has facilitated AAT recruitment to bacteriocytes – at least in the Sternorrhyncha. Nevertheless, the absence of duplicates in cicada indicates that gene duplication is not a prerequisite for initiation of host/symbiont amino acid exchange in sap-feeding insects, an interpretation consistent with the fact that some single-copy AATs also operate at the symbiotic interface in aphids (Price *et al.* 2011) and mealybugs (Fig. 3c).

Amino acid transporter recruitment to the symbiotic interface is dynamic

We measured the expression of paralogs resulting from both ancient and recent gene duplication events because both could play a role in recruiting AATs to the symbiotic interface. Although most examples of gene duplication giving rise to novelty involve gene duplication evolving concurrently with or after the origin of new traits, there are some examples of gene duplication predating the evolution of novelty (Ben Trevaskis *et al.* 1997; Arnegard *et al.* 2010). Expression patterns in both the anciently and recently duplicated AATs (Figs 2 and 3) indicate that AATs were recruited independently to the bacteriocytes of different sap-feeding insect lineages. In the ancient duplications pre-dating hemipteran diversification, qRT–PCR results for aphids, mealybugs and psyllids indicate that bacteriocyte enrichment in one psyllid AAT (Bcoc-APC12; Fig. 2b and Fig. S2, Supporting Information) and one aphid AAT (ACYPI000536; Fig. 3b and Fig. S2 in Appendix S1, Supporting Information) is derived. This finding is consistent with sap being a derived diet within Hemiptera (Cobben 1979; Zrzavy 1990, 1992), requiring that AATs be independently recruited to the symbiotic interface after hemipteran suborders (and these orthologs) diverged from their common ancestor. Biological replication within each sternorrhynchan lineage (psyllids, mealybugs and aphids) would provide finer resolution of the extent to which expression is or is not conserved in these taxa. However, lack of within-species biological replication does not compromise our finding that expression of orthologous AATs is not conserved.

Similarly with respect to the recent taxon-specific gene duplications, qRT–PCR from this study (Fig. 2 and 3; Appendix S1, Supporting Information) and Price *et al.* (2011; adult female aphids) together with RNA-seq differential expression data from this study (Fig. 2, 3; Appendix S3, Supporting Information), Hansen & Moran (2011; fourth-instar female aphids) and Macdonald *et al.* (2012; 7-day-old female aphids) support independent AAT recruitment to the symbiotic interface in pea aphids and citrus mealybugs (Figs 2c and 3c). Notably, bacteriocyte expression in aphid AATs is remarkably consistent across qRT–PCR and RNA-seq studies that together include data from four different pea aphid lineages at different developmental stages (this present study; Price *et al.* 2011; Hansen & Moran 2011; Macdonald *et al.* 2012). Similar to what was previously reported for pea aphids (Hansen & Moran 2011; Price *et al.* 2011), six mealybug-specific paralogs have enriched bacteriocyte expression (Fig. 3c; Appendix S3, Supporting Information). As we reported previously, expression profiles among aphid APC paralogs (Fig. 2c) are most parsimoniously explained by bacteriocyte enrichment evolving after (and potentially being enabled by) gene duplication, an argument based on

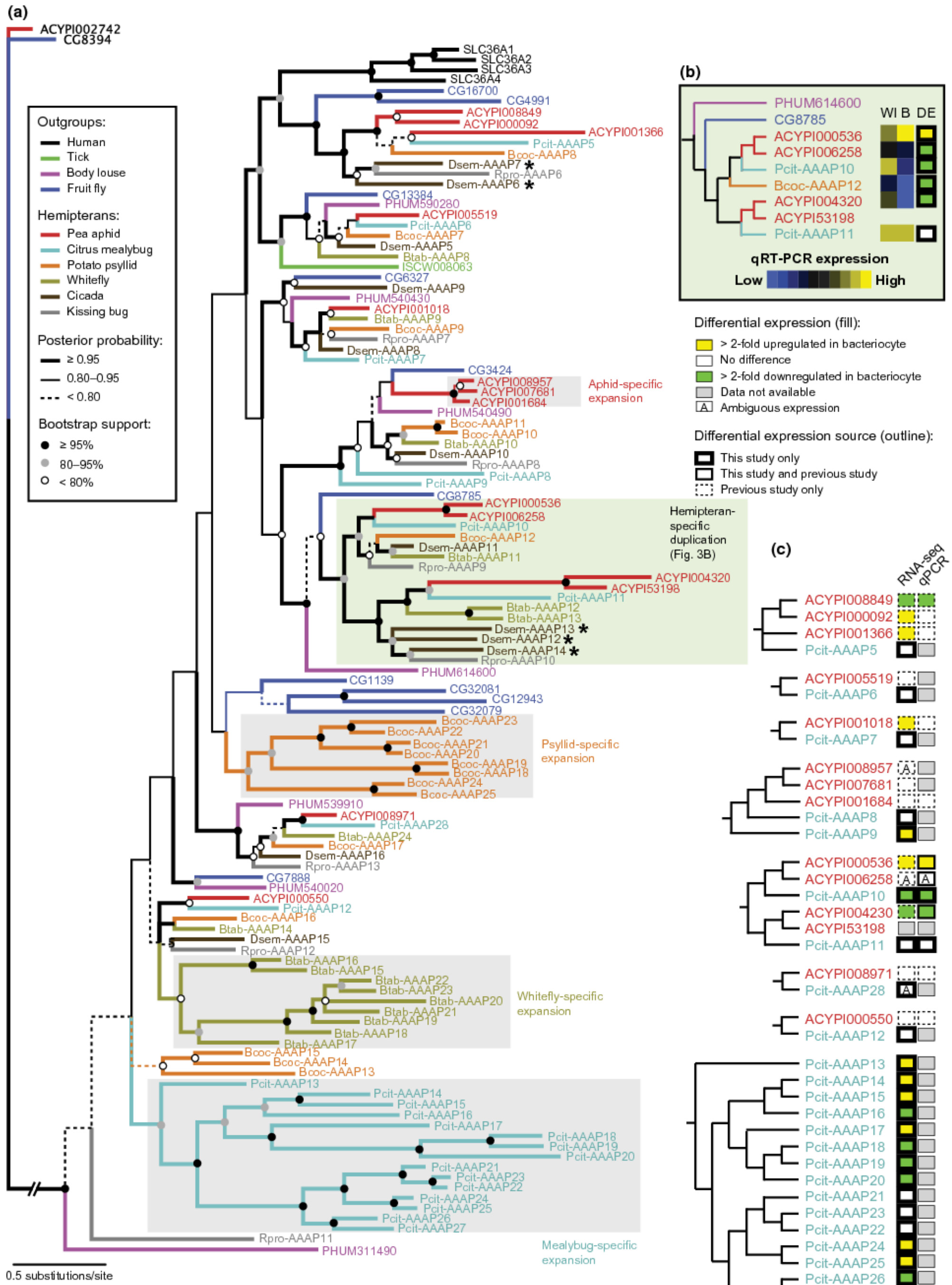
(a)
Outgroups:
Human
Tick
Body louse
Fruit fly
Hemipterans:
Pea aphid
Citrus mealybug
Potato psyllid
Whitefly
Cicada
Kissing bug
Posterior probability:
≥ 0.95
0.80–0.95
< 0.80
Bootstrap support:
≥ 95%
80–95%
< 80%
Aphid-specific expansion
Hemipteran-specific duplication (Fig. 3B)
Psyllid-specific expansion
Whitefly-specific expansion
Mealybug-specific expansion
0.5 substitutions/site
(b)
WI B DE
qRT-PCR expression
Low
High
Differential expression (fill):
> 2-fold upregulated in bacteriocyte
No difference
> 2-fold downregulated in bacteriocyte
Data not available
Ambiguous expression
Differential expression source (outline):
This study only
This study and previous study
Previous study only
(c)
RNA-seq
qPCR

Fig. 3 Partial AAAP (TC # 2.A.18) phylogeny and bacteriocyte expression. (a) Bayesian gene phylogeny for amino acid transporters (AATs) in the AAAP family. Hemipteran-specific gene duplications and taxon-specific expansions are highlighted with green or grey boxes, respectively. Asterisks denote possible cicada-specific paralogs. Branches are colour-coded based on taxon and clade support ≥ 50% (posterior probability and ML bootstrap support) is indicated on branches/nodes as described in the key. (b) qRT–PCR expression data generated in this study for Hemiptera-specific gene duplication are presented both as a heat map for whole insect ('WI') and bacteriocyte ('B') and as differential expression ('DE') between bacteriocyte and whole insect. Heat map expression data are normalized across all tissues and genes within each insect, but not across insects. (c) Differential expression between whole insect and bacteriocyte is indicated for aphid and mealybug genes in boxes to the right of gene IDs, as indicated in the key. RNA-seq differential expression data for aphids are from Hansen and Moran (2011) and Macdonald *et al.* (2012). qRT–PCR data not generated here are from Price *et al.* (2011). Expression is marked as ambiguous ('A') if different transcripts or data sets show inconsistent relative bacteriocyte expression.

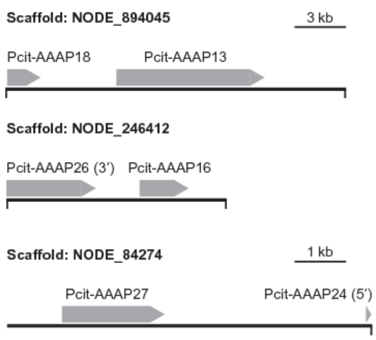

Fig. 4 Paralogs in mealybug-specific AAAP expansion are tandemly arrayed in the genome. Schematic illustrating the arrangement of mealybug AAAP paralogs along genomic scaffolds. Grey arrows depict the position and 5′–3′ direction of representative transcripts (including introns) along three mealybug genomic scaffolds. Each row represents a different scaffold. The top two scaffolds are depicted at the same scale (upper scale bar), and the bottom scaffold is depicted at a different scale (bottom scale bar).

the fact that bacteriocytes are a novel, derived tissue and most aphid APC paralogs, like their insect orthologs, are highly expressed in gut (Price *et al.* 2011). In contrast, the distribution of bacteriocyte enrichment among mealybug AAAP paralogs (Fig. 3c) lacks a clear most parsimonious explanation. Bacteriocyte enrichment/expression could be derived or ancestral, consistent with either neofunctionalization or subfunctionalization of duplicated paralogs. Furthermore, some paralogs may be functionally redundant and are maintained for dosage reasons or are differentially expressed across time and space. Indeed, some aphid AAT paralogs are enriched in head and gut tissues (Price *et al.* 2011), and others have male-biased expression (Duncan *et al.* 2011). Distinguishing between these explanations will be facilitated with functional data for paralogs of this expansion and their orthologs in other insects. However, that multiple paralogs show bacteriocyte enrichment together with substantial sequence divergence among paralogs (indicated by long branches) strongly suggests that at least some mealybug paralogs have evolved novel functional roles. Our results are thus consistent with the hypothesis that gene duplication played a role in recruiting mealybug AATs to the symbiotic interface, enabling them to carry out novel, symbiotic functions.

Interestingly, expression patterns indicate that aphids and mealybugs use different AATs at their symbiotic interface. For example, bacteriocyte-enriched AATs are not orthologous between aphids and mealybugs (Figs 2 and 3). Recruitment of different AATs in aphids and mealybugs could reflect differences in nutritional demand between these insects or could simply result from chance. Alternatively, AATs could be functionally dynamic, with similar environmental pressures experienced by aphids and mealybugs resulting in distinct AAT loci converging upon common functional roles.

Differential AAT expansion among sap-feeding hemipterans is consistent with co-evolutionary patterns of host/symbiont metabolic collaboration

Despite evidence that gene duplication has facilitated the recruitment of AATs to the symbiotic interface in the Sternorrhyncha, cicadas demonstrate that gene duplication is not necessary to initiate novel sap-feeder/symbiont amino acid exchange. Cicadas did not experience expansions in their AATs, a pattern that may relate to a dietary difference between cicadas and sternorrhynchan sap-feeders. While sternorrhynchans feed on plant phloem sap (Gullan *et al.* 2003), the source of sap for cicadas is the plant xylem (White & Strehl 1978), a more dilute source of nitrogen than phloem (Redak *et al.* 2004). It is unclear how amino acid concentration *per se* could influence host insect AAT evolution and recruitment to the symbiotic interface. However, differences in individual amino acid content could potentially influence the nutritional demands of different sap-feeding insects and thus the evolutionary trajectory of amino acid transporters operating at the symbiotic interface. However, recent sequencing of the symbiont genomes of a phloem-feeding auchenorrhynchan suggests that differences in AAT copy number between sternorrhynchans and auchenorrhynchans are not driven by diet.

Bennett and Moran (2013) recently sequenced *Sulcia muelleri* and *Nasuia deltocephalinicola*, the obligate symbionts of the phloem-feeding auchenorrhynchan *Macrosteles quadrilineatus*. Their work highlights an important genomic difference between the obligate symbioses of sternorrhynchans and auchenorrhynchans. Obligate symbionts of both sternorrhynchans and auchenorrhynchans play a major role in providing their hosts with essential amino acids. However, while symbionts of both phloem-feeding and xylem-feeding auchenorrhynchans retain relatively autonomous metabolic pathways (Wu *et al*. 2006; McCutcheon & Moran 2007; McCutcheon *et al*. 2009; Bennett & Moran 2013), sternorrhynchan symbionts lack some genes for crucial metabolic steps – metabolic steps that the host has been demonstrated to complement (Russell *et al*. 2013). For example, sternorrhynchan symbionts typically lack genes necessary to complete the terminal steps in branch-chain amino acid and phenylalanine biosynthesis as well as the step required to synthesize homocysteine for methionine biosynthesis (Shigenobu *et al*. 2000; Nakabachi *et al*. 2006; McCutcheon & von Dohlen 2011; Sabree *et al*. 2012b; Sloan & Moran 2012; Husnik *et al*. 2013). These missing steps are carried out by host insect enzymes (Wilson *et al*. 2010; Hansen & Moran 2011; McCutcheon & von Dohlen 2011; Poliakov *et al*. 2011; Shigenobu & Wilson 2011; Macdonald *et al*. 2012; Husnik *et al*. 2013; Russell *et al*. 2013). This within-metabolic pathway host/symbiont collaboration likely necessitates host/symbiont exchange of intermediate metabolites, a step that is not required in auchenorrhynchans that possess metabolically autonomous symbionts (Wu *et al*. 2006; McCutcheon & Moran 2007; McCutcheon *et al*. 2009; Bennett & Moran 2013). Therefore, gene duplication could have enabled, through neofunctionalization of paralogs, the evolution of novel transporters capable of transporting intermediate metabolites in amino acid biosynthesis pathways, facilitating pathway partitioning between sternorrhynchan hosts and their symbionts. Once functional data are available for these transporters, this hypothesis can be tested. Thus, current evidence suggests that differences in AAT copy number between sternorrhynchans and auchenorrhynchans are driven by differences in the extent of host/symbiont metabolic independence.

Gene duplication and the evolution of novel, symbiotic interactions

The generation of genomic resources for nonmodel organisms, including the partners of symbiotic systems, makes it possible to understand how intimate symbiotic relationships have influenced genome evolution in both symbionts (Shigenobu *et al*. 2000; Nakabachi *et al*. 2006; McCutcheon & Moran 2007; McCutcheon *et al*. 2009; Sabree *et al*. 2009; McCutcheon & von Dohlen 2011; Nikoh *et al*. 2011; McCutcheon & Moran 2012; Sabree *et al*. 2012a; a) and hosts (International Aphid Genomics Consortium 2010; Kirkness *et al*. 2010; Nygaard *et al*. 2011; Young *et al*. 2011; Husnik *et al*. 2013). The pea aphid/*Buchnera* symbiosis was the first symbiotic system to have both host and symbiont genomes sequenced, providing the first insights into how host genomes are shaped by symbiosis. Here, we provide evidence that one of those insights applies more broadly to sternorrhynchan sap-feeding insects: gene duplication plays a role in recruiting amino acid transporters to operate at the host/symbiont interface. Further, recent studies in other, very divergent, symbiotic systems also invoke gene duplication in the evolution of genes with symbiotic functions. For example, in legumes, an ancient whole-genome duplication event in the ancestor of the major papilionoid subfamily was followed by some paralogs evolving enriched expression in symbiotic root nodules. This pattern correlates with the evolution of many important Nod factor signalling components that are critical for legume/*Rhizobium* recognition and the initiation of nodulation in this subfamily (Young *et al*. 2011). Additionally, gene duplication may have facilitated the origin of leghaemoglobin, a special haemoglobin protein that legumes use to remove oxygen from symbiotic root nodules, facilitating symbiotic nitrogen fixation (Anderson *et al*. 1996; Ben Trevaskis *et al*. 1997). Similarly, in an anemone/dinoflagellate symbiosis, cnidarian-specific paralogs gave rise to three genes proposed to function in symbiosis. All three of these cnidarian-specific paralogs are both enriched in individuals hosting symbionts (as opposed to individuals lacking symbionts) and preferentially expressed in the gastroderm, where symbionts are housed (Ganot *et al*. 2011). Together with the results presented here, these plant and cnidarian studies suggest that gene duplication facilitates the recruitment of nonsymbiotic genes to play a role in symbiosis broadly across symbiotic systems. The independent evolution in diverse symbiotic systems of gene duplication followed by expression in tissues that host symbionts, however intriguing, does not in itself provide insight into the potential adaptive significance of gene duplication in the evolution of symbiosis-related genes. The crucial next step to deciphering the role of gene duplication in the evolution of symbiotic interactions will be functional characterization within a phylogenetic framework, which will reveal whether paralogs preferentially expressed at the host/symbiont interface have also evolved novel symbiotic functions.

Acknowledgements

We thank Carol von Dohlen for helpful discussion and for supplying mealybugs. Cecilia Tambourindeguy and Joseph Hancock helped with potato psyllid dissections. The assembled kissing bug transcriptome was provided by Pedro Lagerblad de Oliveira and Gloria Braz, and the assembled potato psyllid transcriptome was provided by Cecilia Tambourindeguy. Jack Min kindly provided the stand-alone ORF prediction PERL script. We are also grateful to Dan Price, Angela Douglas, Jacob Russell and three anonymous reviewers for helpful discussion and feedback on the manuscript. This research was supported by a University of Miami Department of Biology Evoy award (RPD), a Sigma Xi Grant-in-Aid of Research (RPD), NSF Graduate Research Fellowship DG1E-0951782 (RPD), the Czech Science Foundation 13-01878S (FH), NSF DEB-0949759 (LMD), NSF DBI-0640462 (DGG), NSF IOS-1256680 (JPM), NSF IOS-1121847 (ACCW) and start-up funds from the University of Miami (ACCW).

References

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, **21**, 2104–2105.

Anderson CR, Jensen EO, LLewellyn DJ, Dennis ES, Peacock WJ (1996) A new hemoglobin gene from soybean: a role for hemoglobin in all plants. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 5682–5687.

Arnegard ME, Zwickl DJ, Lu Y, Zakon HH (2010) Old gene duplication facilitates origin and diversification of an innovative communication system—twice. *Proceedings of the National Academy of Sciences of the United States of America*, **51**, 22172–22177.

Attardo GM, Hansen IA, Shiao S-H, Raikhel AS (2006) Identification of two cationic amino acid transporters required for nutritional signaling during mosquito reproduction. *Journal of Experimental Biology*, **209**, 3071–3078.

Bennett GM, Moran NA (2013) Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a phloem-feeding insect. *Genome Biology and Evolution*, **5**, 1675–1688.

Caillaud M, Boutin M, Braendle C, Simon J-C (2002) A sex-linked locus controls wing polymorphism in males of the pea aphid, *Acyrthosiphon pisum* (Harris). *Heredity*, **89**, 346–352.

Campbell BC, Steffen-Campbell JD, Sorensen JT, Gill RJ (1995) Paraphyly of Homoptera and Auchenorrhyncha inferred from 18S rDNA nucleotide sequences. *Systematic Entomology*, **20**, 175–194.

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

Castagna M, Shayakul C, Trotti D *et al.* (1997) Molecular characteristics of mammalian and insect amino acid transporters: implications for amino acid homeostasis. *Journal of Experimental Biology*, **200**, 269–286.

Cobben RH (1979) On the Original Feeding Habits of the Hemiptera (Insecta): a Reply to Merrill Sweet. *Annals of the Entomological Society of America*, **72**, 711–715.

Colbourne JK, Pfrender ME, Gilbert D *et al.* (2011) The Ecoresponsive Genome of Daphnia pulex. *Science*, **331**, 555–561.

Colombani J, Raisin S, Pantalacci S *et al.* (2003) A nutrient sensor mechanism controls Drosophila growth. *Cell*, **114**, 739–749.

Cryan JR, Urban JM (2011) Higher-level phylogeny of the insect order Hemiptera: is Auchenorrhyncha really paraphyletic? *Systematic Entomology*, **37**, 7–21.

von Dohlen CD, Moran NA (1995) Molecular phylogeny of the Homoptera: a paraphyletic taxon. *Journal of Molecular Evolution*, **41**, 211–223.

von Dohlen CD, Kohler S, Alsop ST, McManus WR (2001) Mealybug beta-proteobacterial endosymbionts contain gamma-proteobacterial symbionts. *Nature*, **412**, 433–436.

Douglas AE (1993) The Nutritional Quality of Phloem Sap Utilized by Natural Aphid Populations. *Ecological Entomology*, **18**, 31–38.

Douglas AE (2006) Phloem-sap feeding by animals: problems and solutions. *Journal of Experimental Botany*, **57**, 747–754.

Dubrovsky EB, Dubrovskaya VA, Berger EM (2002) Juvenile hormone signaling during oogenesis in Drosophila melanogaster. *Insect Biochemistry and Molecular Biology*, **32**, 1555–1565.

Duncan RP, Nathanson L, Wilson ACC (2011) Novel male-biased expression in paralogs of the aphid *slimfast* nutrient amino acid transporter expansion. *BMC Evolutionary Biology*, **11**, 253.

Eberhard WG (2000) Spider manipulation by a wasp larva. *Nature*, **406**, 255–256.

Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. *Genome Informatics*, **23**, 205–211.

Evans AM, Aimanova KG, Gill SS (2009) Characterization of a blood-meal-responsive proton-dependent amino acid transporter in the disease vector, Aedes aegypti. *The Journal of Experimental Biology*, **212**, 3263–3271.

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, **39**, W29–W37.

Ganot P, Moya A, Magnone V *et al.* (2011) Adaptations to endosymbiosis in a cnidarian-dinoflagellate association: differential gene expression and specific gene duplications. *PLoS Genetics*, **7**, e1002187.

Gerardo NM, Altincicek B, Anselme C *et al.* (2010) Immunity and other defenses in pea aphids, *Acyrthosiphon pisum*. *Genome Biology*, **11**, R21.

Goberdhan DCI, Meredith D, Boyd CAR, Wilson C (2005) PAT-related amino acid transporters regulate growth via a novel mechanism that does not require bulk transport of amino acids. *Development*, **132**, 2365–2375.

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, **11**, 725–736.

Goodfellow M, Alderson G (1977) The actinomycete-genus Rhodococcus: a home for the "rhodochrous" complex. *Journal of General Microbiology*, **100**, 99–122.

Gouy M, Guindon S, Gascuel O (2010) SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Molecular Biology and Evolution*, **27**, 221–224.

Grimaldi D, Engel MS (2005) *Evolution of the Insects*. Cambridge University Press, New York.

Gullan PJ, Downie DA, Steffan SA (2003) A New Pest Species of the Mealybug Genus Ferrisia Fullaway (Hemiptera:

Pseudococcidae) from the United States. *Annals of the Entomological Society of America*, **96**, 723–737.

Haas BJ, Papanicolaou A, Yassour M *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–1512.

Hansen AK, Moran NA (2011) Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 2849–2854.

Hennig W (1981) *Insect Phylogeny*. John Wiley & Sons, New York.

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

Husnik F, Nikoh N, Koga R *et al.* (2013) Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell*, **153**, 1567–1578.

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

Jin X, Aimanova K, Ross LS, Gill SS (2003) Identification, functional characterization and expression of a LAT type amino acid transporter from the mosquito Aedes aegypti. *Insect Biochemistry and Molecular Biology*, **33**, 815–827.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.

Kennedy P (2010) Ectomycorrhizal fungi and interspecific competition: species interactions, community structure, coexistence mechanisms, and future research directions. *The New Phytologist*, **187**, 895–910.

Kim H, Lee S, Jang Y (2011a) Macroevolutionary patterns in the Aphidini aphids (Hemiptera: Aphididae): diversification, host association, and biogeographic origins. *PLoS ONE*, **6**, e24749.

Kim JH, Min JS, Kang JS *et al.* (2011b) Comparison of the humoral and cellular immune responses between body and head lice following bacterial challenge. *Insect Biochemistry and Molecular Biology*, **41**, 332–339.

Kirkness EF, Haas BJ, Sun W *et al.* (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 12168–12173.

Kondrashov FA (2012) Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 5048–5057.

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

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

Livak K, Schmittgen T (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta CT}$ method. *Methods*, **25**, 402–408.

Lynch M, Conery J (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Macdonald SJ, Lin GG, Russell CW, Thomas GH, Douglas AE (2012) The central role of the host cell in symbiotic nitrogen metabolism. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 2965–2973.

Martin JF, Hersperger E, Simcox A, Shearn A (2000) *minidiscs* encodes a putative amino acid transporter subunit required non-autonomously for imaginal cell proliferation. *Mechanisms of Development*, **92**, 155–167.

McCutcheon JP, Moran NA (2007) Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 19392–19397.

McCutcheon JP, Moran NA (2012) Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, **10**, 13–26.

McCutcheon JP, von Dohlen CD (2011) An Interdependent Metabolic Patchwork in the Nested Symbiosis of Mealybugs. *Current Biology*, **21**, 1366–1372.

McCutcheon JP, McDonald BR, Moran NA (2009) Convergent evolution of metabolic roles in bacterial co-symbionts of insects. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 15394–15399.

Moran NA, Tran P, Gerardo NM (2005) Symbiosis and insect diversification: an ancient symbiont of sap-feeding insects from the bacterial phylum Bacteroidetes. *Applied and Environmental Microbiology*, **71**, 8802–8810.

Morris K, Lorenzen MD, Hiromasa Y *et al.* (2009) Tribolium castaneum larval gut transcriptome and proteome: a resource for the study of the coleopteran gut. *Journal of Proteome Research*, **8**, 3889–3898.

Munson MA, Baumann P, Kinsey MG (1991) Buchnera gen. nov. and Buchnera aphidicola sp. nov., a taxon consisting of the mycetocyte-associated primary endosymbionts of aphids. *International Journal of Systematic Bacteriology*, **41**, 566–568.

Nachappa P, Levy J, Tamborindeguy C (2012) Transcriptome analyses of Bactericera cockerelli adults in response to "Candidatus Liberibacter solanacearum" infection. *Molecular Genetics and Genomics*, **287**, 803–817.

Nakabachi A, Yamashita A, Toh H *et al.* (2006) The 160-kilobase genome of the bacterial endosymbiont Carsonella. *Science*, **314**, 267.

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

Nikoh N, McCutcheon JP, Kudo T *et al.* (2010) Bacterial genes in the aphid genome: absence of functional gene transfer from Buchnera to its host. *PLoS Genetics*, **6**, e1000827.

Nikoh N, Hosokawa T, Oshima K, Hattori M, Fukatsu T (2011) Reductive evolution of bacterial genome in insect gut environment. *Genome Biology and Evolution*, **3**, 702–714.

Nygaard S, Zhang G, Schiøtt M *et al.* (2011) The genome of the leaf-cutting ant Acromyrmex echinatior suggests key adaptations to advanced social life and fungus farming. *Genome Research*, **21**, 1339–1348.

Nylander JAA, Wilgenbusch JC, Warren DL, Swofford DL (2008) AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics*, **24**, 581–583.

Ott M, Zola J, Stamatakis A, Aluru S (2007) Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. In: *High Performance Computing in Science and Engineering, Garching*, Springer Verlag, Berlin Heidelberg.

Poliakov A, Russell CW, Ponnala L *et al.* (2011) Large-scale label-free quantitative proteomics of the pea aphid-Buchnera symbiosis. *Molecular & Cellular 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. *Molecular Biology and Evolution*, **28**, 3113–3126.

Rambaut A, Drummond AJ (2007) Tracer v1.5, Available from http://beast.bio.ed.ac.uk/Tracer.

Ratzka C, Gross R, Feldhaar H (2013) Gene expression analysis of the endosymbiont-bearing midgut tissue during ontogeny of the carpenter ant Camponotus floridanus. *Journal of Insect Physiology*, **59**, 611–623.

Redak RA, Purcell AH, Lopes JRS *et al.* (2004) The biology of xylem fluid-feeding insect vectors of Xylella fastidiosa and their relation to disease epidemiology. *Annual Review of Entomology*, **49**, 243–270.

Ribeiro JMC, Genta FA, Sorgine MHF *et al.* (2014) An Insight into the Transcriptome of the Digestive Tract of the Bloodsucking Bug, *Rhodnius prolixus*. *PLoS Neglected Tropical Diseases*, **8**, e2594.

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Russell CW, Bouvaine S, Newell PD, Douglas AE (2013) Shared metabolic pathways in a coevolved insect-bacterial symbiosis. *Applied and Environmental Microbiology*, **79**, 6117–6123.

Sabree ZL, Kambhampati S, Moran NA (2009) Nitrogen recycling and nutritional provisioning by Blattabacterium, the cockroach endosymbiont. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 19521–19526.

Sabree ZL, Huang CY, Arakawa G *et al.* (2012a) Genome shrinkage and loss of nutrient-providing potential in the obligate symbiont of the primitive termite Mastotermes darwiniensis. *Applied and Environmental Microbiology*, **78**, 204–210.

Sabree ZL, Huang CY, Okusu A, Moran NA, Normark BB (2012b) The nutrient supplying capabilities of Uzinura, an endosymbiont of armoured scale insects. *Environmental Microbiology*, **15**, 1988–1999.

Saier MH (2000) A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiology and Molecular Biology Reviews*, **64**, 354–411.

Saier MHJ, Tran CV, Barabote RD (2006) TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Research*, **34**, D181–D186.

Saier MH, Yen MR, Noto K, Tamang DG, Elkan C (2009) The Transporter Classification Database: recent advances. *Nucleic Acids Research*, **37**, D274–D278.

Sandoval CP (1994) Differential visual predation on morphs of Timema cristinae (Phasmatodeae:Timemidae) and its consequences for host range. *Biological Journal of the Linnean Society*, **52**, 341–356.

Sandstrom 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.

Sawyer S (1989) Statistical tests for detecting gene conversion. *Molecular Biology and Evolution*, **6**, 526–538.

Schemske DW, Bradshaw HD (1999) Pollinator preference and the evolution of floral traits in monkeyflowers (Mimulus). *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 11910–11915.

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**, 1086–1092.

Shigenobu S, Stern DL (2013) Aphids evolved novel secreted proteins for symbiosis with bacterial endosymbiont. *Proceedings of the Royal Society B: Biological Sciences*, **280**, 20121952.

Shigenobu S, Wilson ACC (2011) Genomic revelations of a mutualism: the pea aphid and its obligate bacterial symbiont. *Cellular and Molecular Life Sciences*, **68**, 1297–1309.

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 (2012) Endosymbiotic bacteria as a source of carotenoids in whiteflies. *Biology Letters*, **8**, 986–989.

Song N, Liang A-P, Bu C-P (2012) A molecular phylogeny of Hemiptera inferred from mitochondrial genome sequences. *PLoS ONE*, **7**, e48778.

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Thao ML, Baumann P (2004) Evolutionary relationships of primary prokaryotic endosymbionts of whiteflies and their hosts. *Applied and Environmental Microbiology*, **70**, 3401–3406.

Thao ML, Moran NA, Abbot P *et al.* (2000) Cospeciation of psyllids and their primary prokaryotic endosymbionts. *Applied and Environmental Microbiology*, **66**, 2898–2905.

Thao ML, Gullan PJ, Baumann P (2002) Secondary (gamma-Proteobacteria) endosymbionts infect the primary (beta-Proteobacteria) endosymbionts of mealybugs multiple times and coevolve with their hosts. *Applied and Environmental Microbiology*, **68**, 3190–3197.

Trevaskis B, Watts RA, Andersson CR *et al.* (1997) Two hemoglobin genes in Arabidopsis thaliana: the evolutionary origins of leghemoglobins. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 12230–12234.

Wang X-W, Luan J-B, Li J-M *et al.* (2010) De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics*, **11**, 400.

Weber MG, Agrawal AA (2012) Phylogeny, ecology, and the coupling of comparative and experimental approaches. *Trends in Ecology & Evolution*, **27**, 394–403.

White J, Strehl CE (1978) Xylem feeding by periodical cicada nymphs on tree roots. *Ecological Entomology*, **3**, 323–327.

Wilkinson TL, Douglas AE (2003) Phloem amino acids and the host plant range of the polyphagous aphid, Aphis fabae. *Entomologia Experimentalis Et Applicata*, **106**, 103–113.

Wilson ACC, Ashton PD, Calevro F *et al.* (2010) Genomic insight into the amino acid relations of the pea aphid, *Acyrthosiphon pisum*, with its symbiotic bacterium *Buchnera aphidicola*. *Insect Molecular Biology*, **19**(Suppl 2), 249–258.

Wu D, Daugherty SC, Van Aken SE *et al.* (2006) Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. *PLoS Biology*, **4**, e188–e1092.

Young ND, Debellé F, Oldroyd GED *et al.* (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, **480**, 520–524.

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

Zrzavy J (1990) Evolution of Hemiptera: An attempt at synthetic approach. *Proceedings (Part II) of the sixth international symposium of scale insect studies*, Krakow, August 6-12, 1990. Agricultural University Press, Krakow, Poland, 19–22.

Zrzavy J (1992) Evolution of antennae and historical ecology of the hemipteran insects(Paraneoptera). *Acta Entomol Bohemoslov*, **89**, 77–86.

R.P.D. and A.C.C.W. conceived of and designed the project. R.P.D. assembled the citrus mealybug whole insect transcriptome, mapped transcripts to genome scaffolds and performed dS analyses, gene conversion analyses and qRT–PCR experiments. F.H. assembled the mealybug bacteriocyte transcriptome and conducted the mealybug differential expression analysis. J.P.M. collected cicadas and generated the cicada RNA-seq data. J.T.V.L. assembled cicada whole insect and bacteriocyte transcriptomes. R.P.D., A.C.C.W., F.H. and L.M.D. designed the phylogenetic analyses, and R.P.D. conducted the phylogenetic analyses. D.G.G. conducted the whitefly transcriptome re-assembly. R.P.D. and A.C.C.W. drafted the manuscript, and all authors edited the manuscript. All authors approved the final version of the manuscript.

Data accessibility

Raw sequence reads and assemblies: Mealybug – raw sequence reads for transcriptomes and genome are available under NCBI BioProject PRJNA196641. Psyllid – the full psyllid transcriptome assembly is publicly available at http://psyllid.org/download. Whitefly – raw sequence reads are available under NCBI BioProject PRJNA89143. The full transcriptome assembly is publicly available at http://arthropods.eugenes.org/EvidentialGene/arthropods/whitefly/whitefly1eg6/. Cicada – raw sequence reads are available in the NCBI Sequence Read Archive (SRR952383). Insect and bacteriocyte transcripts are pooled and can be separated by the index sequences CAGATC (insect) and ACTTGA (bacteriocyte). Kissing bug – raw sequence reads are available under NCBI BioProject PRJNA191820. Assembled contigs: transcripts for amino acid transporters are available in Appendix S6 (Supporting Information). Mealybug genome scaffolds associated with amino acid transporters are available in Appendix S7 (Supporting Information). Kissing bug genome scaffolds are available on vectorbase.org under the scaffold IDs reported in Appendix S1 (Supporting Information). Data for phylogenetic analyses: protein sequences used for phylogenetic analyses: Appendix S4 (APC) and S5 (AAAP) (Supporting Information).

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Maximum-likelihood phylogeny of full AAAP family.

Fig. S2 qRT–PCR expression results for hemipteran-specific gene duplications.

Table S1 Mealybug loci for ACP family and representative transcripts.

Table S2 Mealybug loci for AAAP family and representative transcripts.

Table S3 Kissing bug loci for ACP family and representative transcripts.

Table S4 Kissing bug loci for AAAP family and representative transcripts.

Table S5 Gene conversion results.

Table S6 qRT–PCR primers.

Appendix S1 Tables S1–S6, Figures S1–S2.

Appendix S2 dS analysis results.

Appendix S3 Mealybug bacteriocyte-whole insect differential expression results.

Appendix S4 APC protein sequences used for phylogenetic analyses.

Appendix S5 AAAP protein sequences used for phylogenetic analyses.

Appendix S6 Assembled contigs for transcripts referenced in this study.

Appendix S7 Assembled mealybug genome scaffolds referenced in this study.

Chapter III

Repeated replacement of an intrabacterial symbiont in the tripartite nested mealybug symbiosis

Filip Husnik[a,b,c,1] and John P. McCutcheon[a,d,1]

[a]Division of Biological Sciences, University of Montana, Missoula, MT 59812; [b]Institute of Parasitology, Biology Centre of the Czech Academy of Sciences, Ceske Budejovice 37005, Czech Republic; [c]Faculty of Science, University of South Bohemia, Ceske Budejovice 37005, Czech Republic; and [d]Canadian Institute for Advanced Research, Program in Integrated Microbial Biodiversity, Toronto, ON, Canada M5G 1Z8

Edited by Jeffrey D. Palmer, Indiana University, Bloomington, IN, and approved July 19, 2016 (received for review March 8, 2016)

Stable endosymbiosis of a bacterium into a host cell promotes cellular and genomic complexity. The mealybug *Planococcus citri* has two bacterial endosymbionts with an unusual nested arrangement: the γ-proteobacterium *Moranella endobia* lives in the cytoplasm of the β-proteobacterium *Tremblaya princeps*. These two bacteria, along with genes horizontally transferred from other bacteria to the *P. citri* genome, encode gene sets that form an interdependent metabolic patchwork. Here, we test the stability of this three-way symbiosis by sequencing host and symbiont genomes for five diverse mealybug species and find marked fluidity over evolutionary time. Although *Tremblaya* is the result of a single infection in the ancestor of mealybugs, the γ-proteobacterial symbionts result from multiple replacements of inferred different ages from related but distinct bacterial lineages. Our data show that symbiont replacement can happen even in the most intricate symbiotic arrangements and that preexisting horizontally transferred genes can remain stable on genomes in the face of extensive symbiont turnover.

Sodalis | organelle | horizontal gene transfer | scale insect

Many organisms require intracellular bacteria for survival. The oldest and most famous example is the eukaryotic cell, which depends on mitochondria (and in photosynthetic eukaryotes, the chloroplasts or plastids) for the generation of biochemical energy (1–4). However, several more evolutionarily recent examples exist, where intracellular bacteria are involved in nutrient production from unbalanced host diets. For example, deep sea tube worms, some protists, and many sap-feeding insects are completely dependent on intracellular bacteria for essential nutrient provisioning (5–7). Some of these symbioses can form highly integrated organismal and genetic mosaics that, in many ways, resemble organelles (8–11). Like organelles, these endosymbionts have genomes encoding few genes (12, 13), rely on gene products of bacterial origin that are encoded on the host genome (9–11, 14, 15), and in some cases, import protein products encoded by these horizontally transferred genes back into the symbiont (16, 17). The names given to these bacteria—endosymbiont, protoorganelle, or bona fide organelle—are a matter of debate (18–21). What is not in doubt is that long-term interactions between hosts and essential bacteria generate highly integrated and complex symbioses.

Establishment of a nutritional endosymbiosis is beneficial for a host by allowing access to previously inaccessible food sources. However, strict dependence on intracellular bacteria can come with a cost: endosymbionts that stably associate with and provide essential functions to hosts often experience degenerative evolution (22–25). This degenerative process is thought to be driven by long-term reductions in effective population size (N_e) caused by the combined effects of asexuality [loss of most recombination and lack of new DNA through horizontal gene transfer (HGT)] and host restriction (e.g., frequent population bottlenecks at transmission in vertically transmitted bacteria) (26). The outcomes of these processes are clearly reflected in the genomes of long-term endosymbionts. These genomes are the smallest of any bacterium that is not an organelle, have among the fastest rates of evolution measured for any bacterium (12, 13), and are predicted to encode proteins and RNAs with decreased structural stability (26, 27). In symbioses where the endosymbiont is required for normal host function, such as in the bacterial endosymbionts of sap-feeding insects, this degenerative process can trap the host in a symbiotic "rabbit hole," where it depends completely on a symbiont which is slowly degenerating (28).

Unimpeded, the natural outcome of this degenerative process would seem to be extinction of the entire symbiosis. However, extinction, if it does happen, is difficult to observe, and surely is not the only solution to dependency on a degenerating symbiont. For example, organelles are bacterial endosymbionts that have managed to survive for billions of years (2). Despite the reduced N_e of organelle genomes relative to nuclear genomes, eukaryotes are able to purge deleterious mutations that arise on organelle genomes, perhaps through a combination of host-level selection and the strong negative selective effects of substitutions on gene-dense organelle genomes (29, 30). Extant organelle genomes also encode few genes relative to most bacteria, and it is also likely that a long history of moving genes to the nuclear genome has helped slow or stop organelle degeneration (21, 31). Some of the most degenerate insect endosymbionts also seem to have adopted a gene transfer strategy, although the number of transferred genes is far smaller compared with organelles. In aphids, mealybugs, psyllids, and whiteflies, some genes related to endosymbiont function are encoded on the nuclear genome, although in most cases, these genes have been transferred from other bacteria and not the

Significance

Mealybugs are plant sap-sucking insects with a nested symbiotic arrangement, where one bacterium lives inside another bacterium, which together live inside insect cells. These two bacteria, along with genes transferred from other bacteria to the insect genome, allow the insect to survive on its nutrient-poor diet. Here, we show that the innermost bacterium in this nested symbiosis was replaced several times over evolutionary history. These results show that highly integrated and interdependent symbiotic systems can experience symbiont replacement and suggest that similar dynamics could have occurred in building the mosaic metabolic pathways seen in mitochondria and plastids.

Author contributions: F.H. and J.P.M. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The nine complete endosymbiont genomes, five draft assemblies of insect genomes, and raw data have been deposited into the European Nucleotide Archive (ENA; accession nos.: *Maconellicoccus hirsutus*: PRJEB12066; *Ferrisia virgata*: PRJEB12067; *Pseudococcus longispinus*: PRJEB12068; *Paracoccus marginatus*: PRJEB12069; and *Trionymus perrisii*: PRJEB12071). Unannotated draft genomes of two Enterobacteriaceae symbionts from *P. longispinus* mealybugs and a B-supergroup *Wolbachia* strain sequenced from *M. hirsutus* mealybugs were deposited in Figshare (accession nos. 10.6084/m9.figshare.2010393 and 10.6084/m9.figshare.2010390).

[1]To whom correspondence may be addressed. Email: filip.husnik@gmail.com or john.mccutcheon@umontana.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1603910113/-/DCSupplemental.

symbionts themselves (9–11, 14). Another solution to avoid host extinction is to replace the degenerating symbiont with a fresh one or supplement it with a new partner. Examples of symbiont replacement and supplementation are replete in insects, occurring in at least the sap-feeding Auchenorrhyncha (23, 32–34), psyllids (22, 35), aphids (25, 36, 37), lice (38), and weevils (39, 40). When viewed over evolutionary time, it becomes clear that endosymbioses can be dynamic—both genes and organisms come and go. It follows that any view of a symbiotic system established from just one or a few host lineages might provide only a snapshot of the complexity that built the observed relationship.

Mealybugs (Hemiptera: Cocoidea: Pseudococcidae) are a group of phloem sap-sucking insects that contain most of the symbiotic complexity described above. All of these insects depend on bacterial endosymbionts to provide them with essential amino acids missing from their diets, but nutrient provisioning is accomplished in dramatically different ways in different mealybug lineages. One subfamily, the Phenacoccinae, has a single β-proteobacterial endosymbiont called *Tremblaya phenacola*, which provides essential amino acids and vitamins to the host insect (9, 41). In the other subfamily of mealybugs, the Pseudococcinae, *Tremblaya* has been supplemented with a second bacterial endosymbiont, a γ-proteobacterium named *Moranella endobia* in the mealybug *Planococcus citri* (PCIT). Although symbiont supplementation is not uncommon, what makes this symbiosis unique is its structure: *Moranella* stably resides in the cytoplasm of its partner bacterial symbiont, *Tremblaya princeps* (42–45).

The organisms in the nested three-way *P. citri* symbiosis are intimately tied together at the metabolic level. *T. princeps* PCIT has one of the smallest bacterial genomes ever reported, totaling 139 kb in length, encoding only 120 protein-coding genes, and lacking many translation-related genes commonly found in the most extremely reduced endosymbiont genomes (42). Many metabolic genes missing in *Tremblaya* are present on the *M. endobia* PCIT genome. Together with their host insect, these two symbionts are thought to work as a "metabolic patchwork" to produce nutrients needed by all members of the consortium (42). The symbiosis in *P. citri* is further supported by numerous HGTs from several different bacterial donors to the insect genome, but not from *Tremblaya* or *Moranella*. These genes are upregulated in the insect's symbiotic tissue (the bacteriome) and fill in many of the remaining metabolic gaps inferred from the bacterial endosymbiont genomes (9).

Other data suggest additional complexity in the mealybug symbiosis. Phylogenetic analyses of the intra-*Tremblaya* endosymbionts show that, although different lineages of mealybugs in the Pseudococcinae all possess γ-proteobacterial endosymbionts related to *Sodalis*, these bacteria do not show the coevolutionary patterns typical of many long-term endosymbionts (43, 44, 46). Developmental studies suggest that *Tremblaya* and its resident γ-proteobacteria can be differentially regulated by the host (44, 47). These data raise the possibility that the innermost bacterium of this symbiosis is labile and may have resulted from separate acquisitions, or that the original intra-*Tremblaya* symbiont has been replaced in different mealybug lineages. What is not clear is when these acquisitions may have occurred and what effect they have had on the symbiosis. Here, we use host and symbiont genome sequencing from seven mealybug species (five generated for this study) to better understand how complex interdependent symbioses may develop over time in the context of gene and organism acquisition and loss.

Results

Overview of Our Sequencing Efforts. We generated genome data for five diverse Pseudococcinae mealybug species, in total closing nine symbiont genomes into single circular-mapping molecules (five genomes from *Tremblaya* and four from the *Sodalis*-allied γ-proteobacterial symbionts) (Table 1). Unexpectedly, we detected γ-proteobacterial symbionts in *Maconellicoccus hirsutus* (MHIR), which was not previously reported to harbor intrabacterial symbionts inside *Tremblaya* cells (Figs. 1–3 and Fig. S1). We also found that *Pseudococcus longispinus* (PLON) harbored two γ-proteobacterial symbionts, each with a complex genome larger than 4 Mbp; these genomes were left as a combined draft assembly of 231 contigs with a total size of 8,191,698 bp and an *N*50 of 82.6 kbp (Table 1).

We also assembled five mealybug draft genomes (Table 1). Because our assemblies were generated only from short-insert paired end data, the insect draft genomes consisted primarily of numerous short scaffolds (Fig. S2 and Table S1).

Verifying the Intra-*Tremblaya* Location for the γ-Proteobacterial Endosymbionts. The intra-*Tremblaya* location of the γ-proteobacterial symbionts has been established for mealybugs in the genera *Planococcus* (44, 45), *Pseudococcus* (44, 48), *Crisicoccus* (49), *Antonina*, *Antoniella*, *Rhodania*, *Trionymus*, and *Ferrisia* (50). However, to our knowledge, the organization of *Tremblaya* and its partner γ-proteobacteria has never been investigated in *Maconellicoccus* or *Paracoccus*. We therefore verified that both *M. hirsutus* and *Paracoccus marginatus* (PMAR) had the expected γ-proteobacteria inside *Tremblaya* structure using FISH microscopy (Fig. S3).

***Tremblaya* Genomes Are Stable in Size and Structure; the γ-Proteobacterial Genomes Are Not.** Genomes from all five *T. princeps* species (those that have a γ-proteobacterial symbiont) are completely syntenic and similar in size, ranging from 138 to 143 kb (Fig. 1). The gene contents are also similar, with 107 protein-coding genes shared in all five *Tremblaya* genomes. All differences in gene content come from gene loss or nonfunctionalization in different lineages (Fig. 1). Four pseudogenes (*argS*, *mnmG*, *lpd*, and *rsmH*) are shared in all five *T. princeps* genomes, indicating that some pseudogenes can be retained in *Tremblaya* for long periods of time. Pseudogene numbers were notably higher and coding densities were lower in *T. princeps* genomes from *P. marginatus* and *Trionymus perrisii* (TPER) (Fig. 1 and Table 1).

In contrast to the genomic stability observed in *Tremblaya*, the genomes of the γ-proteobacterial symbionts vary dramatically in size, coding density, and gene order (Figs. 1 and 3 and Table 1). These genomes range in size from 353 to ∼4,000 kb (*P. longispinus* contains two ∼4,000-kb genomes from different γ-proteobacteria) and are all notably different from the 539-kb *Moranella* genome of *P. citri* (42).

Phylogenetic Analyses Confirm the Intra-*Tremblaya* γ-Proteobacterial Symbionts Result from Multiple Infections. The lack of conservation in γ-proteobacterial genome size and structure, combined with data showing that their phylogeny does not mirror that of their mealybug or *Tremblaya* hosts (43, 44) (Fig. S1), supports early hypotheses that the γ-proteobacterial symbionts of diverse mealybug lineages result from multiple unrelated infections (43, 44). Although the *Sodalis*-allied clade is extremely hard to resolve because of low taxon sampling of facultative and free-living relatives, nucleotide bias, and rapid evolution in obligate symbionts, none of our analyses indicate a monophyletic group of mealybug symbionts congruent with the host and *Tremblaya* trees (Fig. 2 and Fig. S1).

Draft Insect Genomes Reveal the Timing of Mealybug HGTs. Gene annotation of low-quality draft genome assemblies is known to be problematic (51). We therefore verified that our mealybug assemblies were sufficient for our purpose of establishing gene presence or absence by comparing our gene sets with databases containing core eukaryotic [Core Eukaryotic Genes Mapping Approach (CEGMA)] and Arthropod [Benchmarking Universal Single-Copy Orthologs (BUSCO)] gene sets. CEGMA scores surpass 98% in all of our assemblies, and BUSCO Arthropoda scores range from 66 to 76% (Table S1). We note that the low scores against the BUSCO database likely reflect the hemipteran origin of mealybugs rather than our fragmented assembly; the high-quality

Table 1. Genome statistics for mealybug endosymbionts and draft mealybug genomes

Mealybug species	*P. avenae*	*M. hirsutus*	*F. virgata*	*P. citri*	*P. longispinus*	*T. perrisii*	*P. marginatus*
Mealybug abbreviation	PAVE	MHIR	FVIR	PCIT	PLON	TPER	PMAR
Total assembly size (bp)	NA	163,044,544	304,570,832	377,829,872	284,990,201	237,582,518	191,208,351
Total o. of scaffolds	NA	12,889	32,723	167,514	66,857	80,386	60,102
N50 | *N75*	NA	47,025 | 22,300	25,562 | 12,551	7,078 | 3,639	10,126 | 4,908	4,681 | 2,689	6,799 | 3,788
BUSCOs Arthropoda (*n*=2,675)	NA	76%	76%	71%	70%	66%	72%
BUSCOs Eukaryota (*n*=429)	NA	85%	84%	80%	78%	77%	82%
CEGMA (*n*=248; including partial)	NA	99.19%	97.98%	98.79%	98.39%	99.6%	98.79%
Tremblaya symbiont	*T. phenacola*	*T. princeps*	*T. princeps*	*T. princeps*	*T. princeps*	*T. princeps*	*T. princeps*
Genome size (plasmid size if present)	170,756 bp (744 bp)	138,415 bp	141,620 bp	138,927 bp	144,042 bp	143,340 bp	140,306 bp
Average fragment coverage	NA (454 data)	795	663	374	1,326	2,364	787
G + C (%)	42.2	61.8	58.3	58.8	58.9	57.8	58.3
CDS (pseudogenes)	178 (3)	136 (7)	132 (13)	125 (16)	134 (15)	116 (31)	124 (17)
CDS coding density (%)	86.3	77.2	69.3	66.0	70.7	59.2	67.0
rRNAs | tRNAs | ncRNAs	4 | 31 | 3	6 | 14 | 3	6 | 14 | 3	6 | 10 | 3	6 | 16 | 3	6 | 12 | 3	6 | 17 | 3
γ-Proteobacterial symbiont	Not present	*D. endobia*	*G. endobia*	*Mo. endobia*	PLON1 and PLON2	*H. endobia*	*Mi. endobia*
Genome size (plasmid size)	NA	834,723 bp (11,828 bp)	938,041 bp	538,294 bp	8,190,816*	628,221 bp (8,492 bp)	352,837 bp
Average fragment coverage	NA	121 (38)	372	827	30	559 (312; 1,750)	620
G + C (%)	NA	44.2	28.9	43.5	53.9	42.8	30.6
CDS (pseudogenes)	NA	564 (99)	461 (30)	419 (24)	NA (NA)	510 (16)	273 (8)
CDS coding density (%)	NA	59.8	48.1	77.4	NA	80.4	75.5
rRNAs | tRNAs | ncRNAs	NA	3 | 40 | 14	3 | 39 | 8	5 | 41 | 9	NA	3 | 41 | 10	3 | 41 | 5
Reference	9	This study	This study	42	This study	This study	This study

H. endobia codes two plasmids of 3,244 and 5,248 bp. Extended assembly metrics for draft mealybug genomes are available as Table S2.
*Combined assembly size for both γ-proteobacterial symbionts in PLON. CDS, protein-coding DNA sequence; NA, not applicable; ncRNA, noncoding RNA; PAVE, *Phenacoccus avenae*.

pea aphid genome (52) scores 72% using identical settings. We conclude that our mealybug draft assemblies are sufficient for determining the presence or absence of bacterial HGTs.

We first sought to confirm that the HGTs found previously in the *P. citri* genome (9) were present in other mealybug species (Tables S2 and S3) and establish the timing of these transfers. [Consistent with our previous findings (9), there were no well-supported HGTs of *Tremblaya* origin detected in any of our mealybug assemblies.] Our data show that the acquisition of some HGTs [*bioABD*, *ribAD*, *dapF*, *lysA*, tryptophan 2-monooxygenase oxidoreductase (*tms*), and ATPases associated with diverse cellular activities (AAA-ATPases)] predated the Phenacoccinae/Pseudococcinae divergence and thus the acquisition of any γ-proteobacterial endosymbiont (Fig. 3). These old HGTs mostly involve amino acid and B vitamin metabolism, are usually found on longer insect scaffolds that contain several essential insect genes, and are syntenic across mealybug species (Fig. 4). In each of these cases, no other bacterial genes or pseudogenes were found within the scaffolds (Tables S2 and S3), suggesting that these HGTs resulted from the transfer of small DNA fragments or that flanking bacterial DNA from larger fragments was lost after the transfer was established. The origin of some of these transfers [7,8-diaminopelargonic acid synthase and biotin synthase (*bioAB*)] likely predates the entire mealybug lineage, because they are found in the genome of the whitefly *Bemisia tabaci* (11).

We find that several HGTs were likely acquired after the divergence of the *Maconellicoccus* clade [cysteine synthase A (*cysK*), beta-lactamase (*b-lact*), type III effector (*T3ef*), and D-alanine-D-alanine ligase B (*ddlB*)]. One of these genes, *cysK*, clusters with sequences from other *Sodalis*-allied bacteria, consistent with a possible origin from an early γ-proteobacterial intrabacterial symbiont (Dataset S1*F*). We note that *cysK* has undergone tandem duplication in *P. longispinus*, *Ferrisia virgata* (FVIR), and *P. citri* (Fig. S2*A* and Tables S2 and S3), which was also observed for several other HGTs (*tms*, *b-lact*, *T3ef*, *chiA*, ankyrin repeat proteins, and AAA-ATPases). Most of the HGTs found in only one or two mealybug species are related to peptidoglycan metabolism and were assembled on shorter scaffolds with few insect genes on them. Possible HGT losses of *tms* in FVIR and *ddlB* in *P. marginatus* were detected based on our assemblies. Except in three cases (*amiD*, *murC*, and DUR1), HGT candidates detected from several mealybug species shared a significant amount of sequence similarity and clustered as a single clade in our phylogenies (Dataset S1), suggesting that these transfers resulted from single events.

Evolution of the Metabolic Patchwork. We previously found complementary patterns of gene loss and retention between *Tremblaya*, *Moranella*, and the mealybug host in the *P. citri* symbiosis (9, 42). Our comparative genomic data allow us to see how genes are retained or lost in different genomes in multiple lineages that have γ-proteobacterial symbionts of different inferred ages (Fig. 3). These data also allow us to observe how new symbionts evolve in response to the presence of both preexisting symbionts and horizontally transferred genes.

Overall, our data point to an extremely complex pattern of gene loss and retention in the mealybug symbiosis (Fig. 3). Some pathways, such as those for the production of lysine, phenylalanine, and methionine, show a relatively similar patchwork pattern in all mealybugs, with gene retention interspersed between *Tremblaya*, its γ-proteobacterial endosymbiont, and/or the host. Gene retention patterns from many other pathways, however, show much less

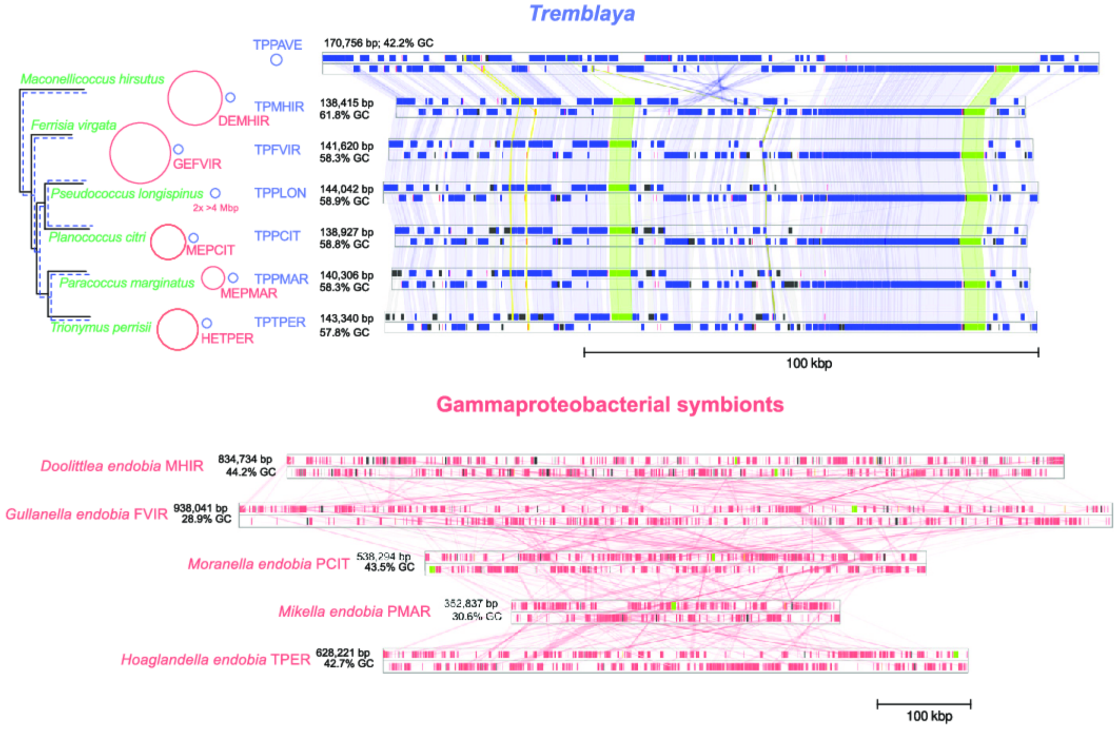

Fig. 1. Genome size and structure of the mealybug endosymbionts. Linear genome alignments of (*Upper*) seven *Tremblaya* genomes (blue) are contrasted with linear genome alignments of (*Lower*) five genomes of their respective γ-proteobacterial symbionts (red). The *T. princeps* genomes are perfectly collinear and similar in size, whereas the γ-proteobacterial genomes are highly rearranged and different in size. Alignments are ordered based on a schematic mealybug/*Tremblaya* phylogeny (original phylogenies are in Fig. S1) and accompanied by basic genome statistics (detailed genome statistics are in Table 1). Gene boxes are colored according to their category: proteins in blue, pseudogenes in gray, rRNAs in green, noncoding RNAs in yellow, and tRNAs in red.

predictable patterns. The isoleucine, valine, leucine, threonine, and histidine pathways show a tendency toward *Tremblaya*-dominated biosynthesis in *M. hirsutus*, *F. virgata*, and *P. citri* (that is, gene retention in *Tremblaya* and gene loss in the γ-proteobacterial symbiont) but with a clear shift toward γ-proteobacterial–dominated biosynthesis in *P. marginatus* and *T. perrisii*. Other pathways, such as tryptophan, show γ-proteobacterial dominance in all mealybug symbioses but with reliance on at least one *Tremblaya* gene in *P. citri*, *P. marginatus*, and *T. perrisii*. In the arginine pathway, gene retention is dominated by *Tremblaya* in *M. hirsutus* but by the γ-proteobacterial endosymbiont in all other lineages, with sporadic loss of *Tremblaya* genes in different lineages. Overall, *M. hirsutus* encodes the most *Tremblaya* genes and the fewest γ-proteobacterial genes, whereas TPER shows the opposite pattern.

Gene Retention Patterns for Translation-Related Genes in *Tremblaya*. In contrast to metabolic genes involved in nutrient production, the retention patterns for genes involved in translation vary little between mealybug species (Fig. 3). As first shown in *Tremblaya* PCIT (42), none of the additional *Tremblaya* genomes that we report here encode any functional aminoacyl tRNA synthetase, with an exception of one likely functional gene (*cysS*) in *T. princeps* PLON, which is present as a pseudogene in several other lineages of *Tremblaya*. Furthermore, all *Tremblaya* genomes have lost key translational control proteins that are typically retained even in the smallest endosymbiont genomes, such as ribosome recycling factor, L-methionyl-tRNA^{fMet} *N*-formyltransferase, and peptide deformylase. The translational release factors RF-1 and RF-2 (*prfAB*) and elongation factor (EF) EF-Ts (*tsf*) are present only in the gene-rich *T. princeps* MHIR genome and absent or pseudogenized in all other *T. princeps* genomes. Initiation factors (IFs) IF-1, IF-2, and IF-3 (*infABC*) and EFs EF-Tu and EF-G (*tufA* and *fusA*) are retained in all *Tremblaya* genomes, as are most ribosomal proteins (Dataset S24).

Taxonomy of Mealybug Endosymbionts. The naming convention in the field of insect endosymbiosis has been to keep the species names constant for lineages of endosymbiotic bacteria resulting from single infections, even if they exist in different species of host insects. The host is denoted by appending a specific abbreviation to the end of the endosymbiont name (e.g., *T. princeps* PCIT for *T. princeps* from *P. citri*). However, our data show that the intra-*Tremblaya* γ-proteobacterial symbionts are not from the same infection; they result from independent endosymbiotic events from clearly discrete lineages within the *Sodalis* clade (Fig. 2). Following convention, we have chosen to give these γ-proteobacteria different genus names but unite them by retaining the "*endobia*" species denomination for each one (such as in *Moranella endobia*).

We propose the following Candidatus status names for four lineages of intra-*Tremblaya* γ-proteobacterial symbionts of mealybugs for which we have completed a genome. First, *Candidatus* Doolittlea endobia MHIR is for the endosymbiont from *M. hirsutus*. This name honors the American evolutionary biologist W. Ford Doolittle (1941–) for his contributions to our understanding of HGT and endosymbiosis. Second, *Candidatus* Gullanella endobia FVIR is for the endosymbiont from *F. virgata*. This name honors the Australian entomologist Penny J. Gullan

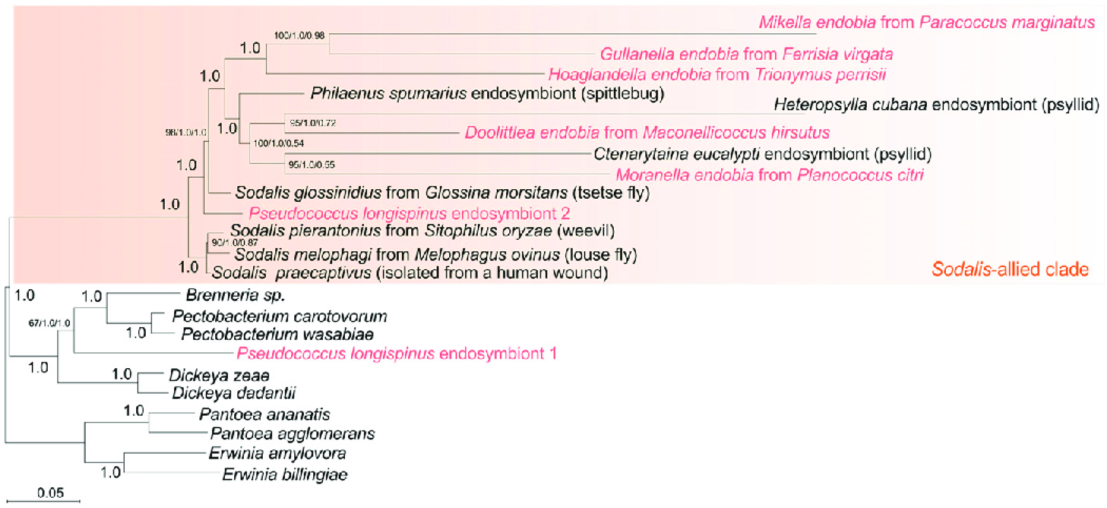

Fig. 2. The intra-*Tremblaya* mealybug symbionts are members of the *Sodalis* clade of γ-proteobacteria. A multigene phylogeny of *Sodalis*-allied insect endosymbionts and closely related Enterobacteriaceae (γ-proteobacteria) was inferred from 80 concatenated proteins under the LG + G evolutionary model in RaxML v8.2.4. Mealybug endosymbionts are highlighted in red. Values at nodes represent bootstrap pseudoreplicates from the maximum likelihood (ML) analysis, posterior probabilities from Bayesian inference (BI) topology inferred under the LG + I + G model, and posterior probabilities from BI topology inferred from the Dayhoff6 recoded dataset under the CAT + GTR + G model in PhyloBayes, respectively.

(1952–) for her contributions to numerous aspects of mealybug biology and taxonomy. Third, *Candidatus* Mikella endobia PMAR is for the endosymbiont from *P. marginatus*. This name honors the Canadian biochemist Michael W. Gray (1943–) for his contributions to our understanding of organelle evolution. Fourth, *Candidatus* Hoaglandella endobia TPER is for the endosymbiont from *T. perrisii*. This name honors the American biochemist Mahlon B. Hoagland (1921–2009) for his contributions to our understanding of the genetic code, including the codiscovery of tRNA. All of the names that we propose could be extendible to related mealybugs species (e.g., *G. endobia* for other members of the *Ferrisia* clade) if future phylogenetic analyses show that these symbionts result from the same infection. For simplicity, we use all endosymbiont names without the *Candidatus* denomination.

Discussion

Diversity of Intra-*Tremblaya* Symbiont Genomes Suggests Multiple Replacements. Phylogenetic analyses based on rRNA and protein-coding genes from the γ-proteobacterial endosymbionts of mealybugs first indicated their origins from multiple unrelated bacteria (43, 44). What was unclear from these data was the order and timing of the γ-proteobacterial infections and how these infections affected the other members of the symbiosis. We imagine three possible scenarios that could explain these phylogenetic and genomic data (Fig. 5). The first is that there was a single γ-proteobacterial acquisition in the ancestor of the Pseudococcinae that has evolved idiosyncratically as mealybugs diversified over time, leading to seemingly unrelated genome structures and coding capacities (the "idiosyncratic" scenario) (Fig. 5*A*). The second is that the γ-proteobacterial infections occurred independently, each establishing symbioses inside *Tremblaya* in completely unrelated and separate events (the "independent" scenario) (Fig. 5*B*). The third is that there was a single γ-proteobacterial acquisition in the Pseudococcinae ancestor that has been replaced in some mealybug lineages over time (the "replacement" scenario) (Fig. 5*C*). The idosyncratic scenario is easy to disregard, because although acquisition of a symbiont followed by rapid diversification of the host might result in different patterns of genome evolution in different lineages, it should result in monophyletic clustering in phylogenetic trees. Previous phylogenetic work as well as our phylogenomic data (Fig. 2) show that the γ-proteobacteria that have infected different mealybugs have originated from clearly distinct (and well-supported) bacterial lineages.

The independent and replacement scenarios are more difficult to tell apart with our data, and the true history of the symbiosis may have involved both. However, we favor symbiont replacement as the main mechanism that generated the complexity that we see in mealybugs, primarily because of the large differences in size observed in the γ-proteobacterial genomes (Fig. 1 and Table 1). Genome size is strongly correlated to endosymbiotic age in bacteria, especially at the onset of symbiosis, when genome reduction can be rapid (53–57). Most relevant to our argument here is the speed with which genome reduction has been shown to take place in *Sodalis*-allied bacteria closely related to the γ-proteobacterial symbionts of mealybugs (34, 58, 59). It has been estimated that as much as 55% of an ancestral *Sodalis* genome was lost on the transition to endosymbiosis in a mere ∼28,000 y, barely enough time for 1% sequence divergence to accumulate between the new symbiont and a free-living relative (58). Our general assumption is, therefore, that recently established endosymbionts should have larger genomes than older symbionts. However, we note that genome reduction is not a deterministic process related to time, especially as the symbiosis ages. It is clear that, in some insects housing pairs of ancient symbionts with highly reduced genomes, the older endosymbiont can have a larger genome than the newer symbiont (60).

The evidence for recent replacement is most obvious in *P. longispinus* (Fig. 3 and Table 1). This symbiosis harbors two related γ-proteobacterial symbionts (61), each with a rod-like cell shape, although it is currently unclear if both bacteria reside within *Tremblaya* (48). Both of these genomes are about 4 Mb in size (Table 1), approximately the same size as the recently acquired *Sodalis* symbionts from tsetse fly (4.3 Mb) (62) and rice weevil (4.5 Mb) (59). These morphological and genomic features as well as their relatively short branches in Fig. 2 all suggest that

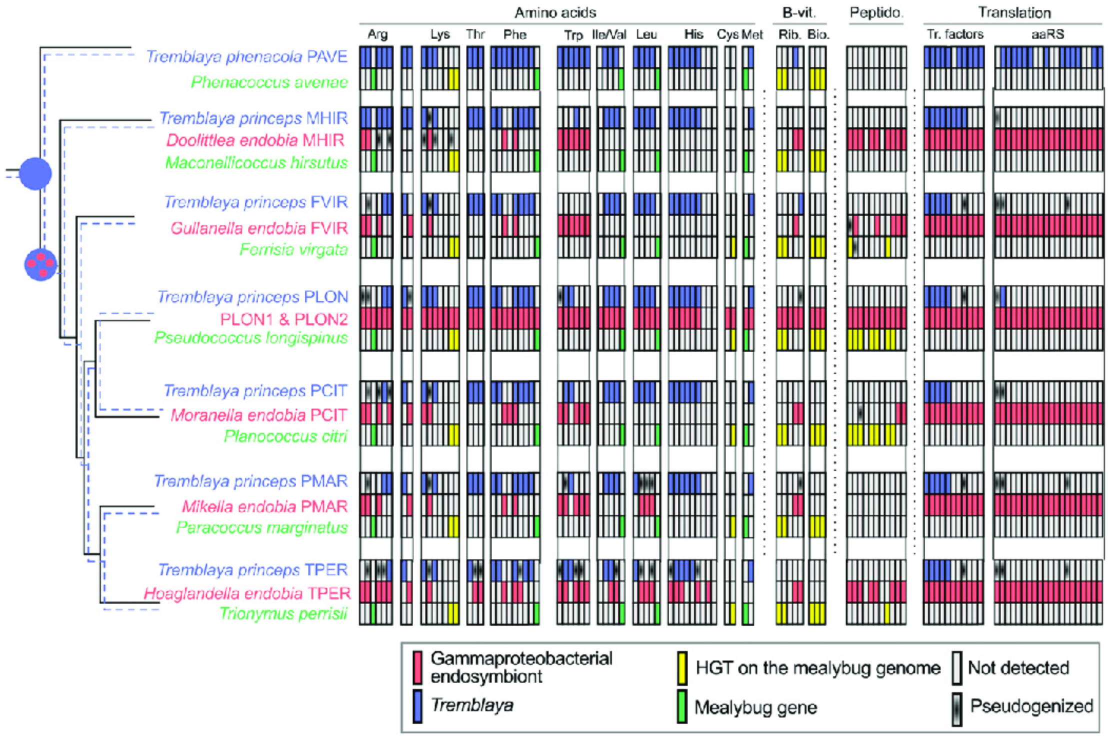

Fig. 3. A complex history of gene retention, loss, and acquisition in the mealybug symbiosis. Retention of selected biosynthetic pathways, such as amino acids, B vitamins (B-vit.), peptidoglycan (Peptido.), translation-related genes [various initiation, elongation, and termination factors (Tr. factors)], and HGTs. For each of seven mealybug species, boxes in row 1 represent *Tremblaya* genes (blue), row 2 represents its γ-proteobacterial symbionts (red), and row 3 represents the host genome (insect genes in green and HGTs in yellow). Missing genes are shown in gray, and recognizable pseudogenes are shown with black radial gradient. Raw data used here (including gene names) are available in Dataset S2B. Bio., biotin; Rib., riboflavin.

the γ-proteobacterial symbionts are recent acquisitions in the *P. longispinus* symbiosis. The *P. longispinus* replacement seems so recent that the stereotypical complementary patterns of gene loss and retention have not had time to accumulate between the γ-proteobactia and *Tremblaya* (Fig. 3). However, *Tremblaya* PLON is missing the same translation-related genes (aside from *cysS*) as all other *Tremblaya*, indicating that it has long ago adapted to the presence of a (now eliminated) bacterium living in its cytoplasm. Comprehensive analyses of the two γ-proteobacterial genomes from *P. longispinus* are ongoing and will be published elsewhere.

We hypothesize that the larger, gene-rich γ-proteobacterial genomes that we describe here are the result of symbiont replacements of an ancestral γ-proteobacterial endosymbiont rather than completely independent infections in different mealybug lineages. We suspect that the massive loss in key translation-related genes

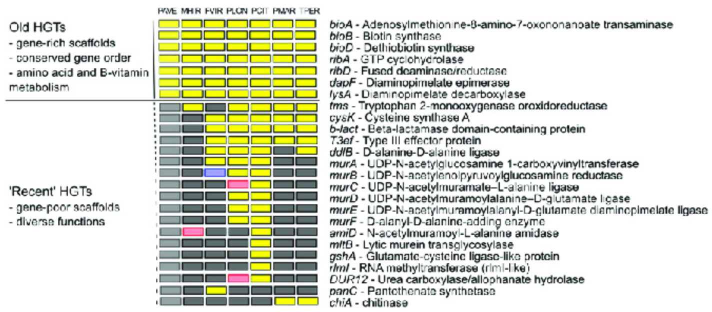

Fig. 4. HGTs detected in individual mealybug species. Retention of HGT candidates detected across all mealybug species (blue, possible pseudogene; gray, gene not detected; red, different phylogenetic origin; yellow, gene present).

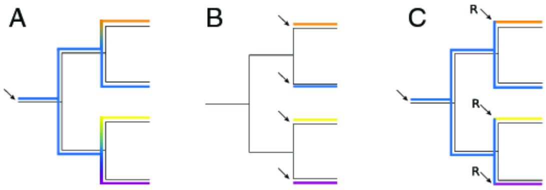

Fig. 5. Three possible scenarios that built the mealybug symbiosis. Independent γ-proteobacterial acquisitions are shown as arrows, and replacements are noted with Rs above the arrow. Colors represent the different γ-proteobacterial genomes shown in Fig. 1. (*A*) The idiosyncratic scenario, where a single γ-proteobacterial acquisition evolved differently as mealybugs diverged, leading to different genome sizes and structures in extant mealybugs. (*B*) The independent scenario, where the different sizes and structures of the γ-proteobacterial genomes shown in Fig. 1 result from completely independent acquisitions. (*C*) The replacement scenario, where the different sizes and structures of the γ-proteobacterial genomes shown in Fig. 1 result from several replacements of an ancestral γ-proteobacterial symbiont.

(Fig. 3) in *Tremblaya* occurred in response to the first γ-proteobacterial infection, which then required all subsequent replacement events to also reside within the *Trembalay* cytoplasm. It is tempting to speculate that the 353-kb *Mikella* PMAR genome is the ancestral intra-*Tremblaya* symbiont lineage that has not been replaced or at least has not been recently replaced. However, because the relevant clades split right after the Phenacoccinae/Pseudococcinae divergence—that is, right at the acquisition of the first γ-proteobacterial symbiont—much richer taxon sampling would be needed to test the hypothesis that this was, in fact, the original symbiont lineage (Fig. 2). We also note that, in at least one other case, bacteria from the *Sodalis* group have established multiple repeated infections in a replacement-like pattern (38).

How Did the Bacteria Within a Bacterium Structure Start, and Why Does It Persist? In extreme cases of endosymbiotic genome reduction, genes required for the generation of a cell envelope, along with other fundamental processes, are lost (12, 13). This phenomenon is seen in *Tremblaya*, where even the largest genome (from *Phenacoccus avenae*, which lacks a γ-proteobacterial symbiont) encodes no genes for the production of fatty acids or peptidoglycan (9). We assume that the envelope that defines the *Tremblaya* cytoplasm is made by the host, because it cannot be made by *Tremblaya*. These data suggest that when the first γ-proteobacterial endosymbiont established residence in *Tremblaya*, it invaded a membrane system that was perhaps more eukaryotic than bacterial in nature (even if it ultimately ended up in a "bacterial" cytoplasm). Bacteria in the *Sodalis* group are very good at establishing intracellular infections in insect cells (38, 63, 64), and we suggest that their propensity to infect *Tremblaya* might simply reflect this ability. The cytoplasm vs. envelope distinction is important, because the mealybug symbiosis has been held up by many—including us—as a rare example of a stable bacteria within a bacterium symbiosis. Although this description might be apt if one considers the *Tremblaya* cytoplasm bacterial in nature, it may not be if one considers the types of membranes that the innermost bacteria had to cross to get there.

But why did the first γ-proteobacterial endosymbiont end up inside *Tremblaya*? We can think of two related possibilities. The first is that it was easier to use the established transport system between the insect cell and *Tremblaya* (65) than to evolve a new one. The second is that the insect immune system likely does not target *Tremblaya* cells, and so the *Tremblaya* cytoplasm is an ideal hiding place for a newly arrived symbiont. After the loss of critical translation-related genes in *Tremblaya*, the symbiosis would persist with a bacteria within a bacterium structure because no other structure is possible. We note that *Sodalis*- and *Arsenophonus*-allied symbionts were recently suggested to sometimes reside within *Sulcia* cells in the leafhoppers *Cicadella viridis* and *Macrosteles laevis* (66, 67). Although these studies were based only on EM imaging and not confirmed by specific probes (e.g., with FISH), it is possible that symbioses formed by bacteria taking up residence inside of degenerate symbionts with host-derived cell envelopes are not uncommon.

Evolution of Organelles and the Timing of HGT. It is widely accepted that the mitochondria found across eukaryotes are related back to a single common α-proteobacterial ancestor (68) and that the plastids resulted from a single cyanobacterial infection (69). What is less clear is what happened before these endosymbiont lineages were fixed into organelles. The textbook concept is that a bacterium was taken up by a host cell, transferred most of its genes, and became the mitochondrion or plastid (70). This idea becomes more complicated when the taxonomic affiliation of bacterial genes on eukaryotic genomes is examined (71–74). For example, only about 20% of mitochondria-related horizontally transferred genes have strong α-proteobacterial phylogenetic affinities (72). The signals for the remaining 80% are either too weak to confidently place the gene or show clear affiliation with other bacterial groups (71, 72). Hypotheses that explain these data fall roughly into two camps. Some imagine a gradual process where multiple taxonomically diverse endosymbioses may have occurred—and transferred genes—before the final α-proteobacterial symbiont was fixed. That is, the mitochondria arrived rather late in the evolution of a eukaryotic-like cell that already contained many bacterial genes resulting from HGT of previous symbionts (75, 76). Others favor a more abrupt "mitochondria early" scenario, where an endosymbiont with a taxonomically diverse mosaic genome made the transition to becoming the mitochondrion in a single endosymbiotic event, transferring its genes during the process. In this scenario, the mosaic nature of the extant eukaryotic genomes resulted from the "inherited chimerism" of the lone mitochondria bacterial ancestor because of the propensity of bacteria to participate in HGT with distantly related groups (73, 77, 78).

We suggest that the data reported here indirectly support the gradualist or mitochondria late view of organelle evolution. We find that the majority of nutrient-related HGTs occurred before the divergence of the Phenacoccinae and Pseudococcinae (Figs. 3 and 4) and therefore before the establishment of any γ-proteobacterial symbiont. In particular, HGTs in the riboflavin and lysine pathways were retained on the insect genomes as the first γ-proteobactieral symbiont was established and new γ-proteobacterial symbionts replaced old ones (Figs. 2 and 3). Our results make it clear that HGTs can remain stable on host genomes for millions of years, even after the addition or replacement of symbionts that share pathways with these genes, and directly show how mosaic metabolic pathways can be built gene by gene as symbionts come and go over time. We note that the "shopping bag" hypothesis (79), which argues that establishment of an endosymbiosis should be regarded as a continuous process involving a number of partners rather than a single event involving two partners, fits our data remarkably well. Of course, our data do not rule out inherited chimerism as a contributor to the taxonomic diversity of genes that support organelle function, because many bacterial genomes are taxonomically mosaic because of HGT (73). As with most solutions to endosymbiotic problems, the true answer is likely a complicated mixture of both processes.

Using Symbiont Supplementation and Replacement to Claw Out of the Rabbit Hole. At the onset of a nutritional symbiosis, a new organism comes on board and allows access to a previously inaccessible food source. Rapid adaptation and diversification can occur—the new symbiont adapts to the host, the host adapts to the symbiont, and the entire symbiosis expands in the newly available ecological niche. However, cases where a bacterial symbiont takes up stable residence in a host cell also seem to lead to irreversible degeneration and codependence between host and symbiont (26, 28, 80, 81). What HGT, symbiont supplementation, and symbiont replacement may offer is a way out—at least temporarily, but perhaps permanently—of this degenerative ratchet.

However, new symbionts may also provide ecological opportunity in addition to evolutionary reinvigoration. We note that the mealybug with one of the broadest host ranges is also the species with the most recent γ-proteobacterial replacement, *P. longispinus*. *P. longispinus* is an important agricultural pest and known to feed on plants from 82 families (scalenet.info/catalogue/pseudococcus%20longispinus/). It seems possible that fresh symbionts with large genomes could provide novel functions unavailable in more degenerate symbionts, again propelling the symbioses into new niches.

Materials and Methods

Samples of the mealybug species *M. hirsutus* (pink hibiscus mealybug; MHIR; collection locality: Helwan, Egypt), *F. virgata* (striped mealybug; FVIR; collection locality: Helwan, Egypt), and *P. marginatus* (papaya mealybug; PMAR; collection locality: Mayotte, Comoro Islands) were identified and provided by Thibaut Malausa, Institut National de la Recherche Agronomique, Sophia, France. *T. perrisii* (TPER; collection locality: Poland) samples were provided by Małgorzata Kalandyk-Kołodziejczyk, University of Silesia, Katowice, Poland. *P. longispinus* samples (long-tailed mealybug; PLON) were collected by F.H. in a winter garden of the Faculty of Science, University of South Bohemia. DNA vouchers and insect vouchers of adult females for slide mounting are available from F.H. DNA was isolated from three to eight whole insects of all species by the Qiagen QIAamp DNA Micro Kit, and each library was multiplexed on two-thirds of an Illumina HiSeq 2000 Lane and sequenced as 100-bp paired end reads. The *M. hirsutus* sample was sequenced on an entire MiSeq lane with v3 chemistry and 300-bp paired end mode. Both approaches generated sufficient coverage for both symbiont genomes and draft insect genomes. Adapter clipping and quality filtering were carried out in the Trimmomatic package (82) using default settings. Read error correction (BayesHammer), de novo assembly (k-mers K21, K33, K55, and K77 for 100-bp data and K99 and K127 for 300-bp data), and mismatch/short-indel correction were performed by the SPAdes assembler, v3.5.0 (83). Additional endosymbiont-targeted long k-mer (91 and 241 bp) assemblies generated by the Ray v2.3.1 (84) and PRICE v1.2 (85) assemblers were used to improve assemblies of complex endosymbiont regions.

Additional information on the computational and microscopy methods can be found in *SI Materials and Methods*. General *Tremblaya* primers are shown in Table S4.

ACKNOWLEDGMENTS. We thank the Genomics Core Facility at the University of Montana, the DNA Sequencing Facility at the University of Utah, and the European Molecular Biology Laboratory Genomics Core Facility in Heidelberg for sequencing services. F.H. was funded by the Fulbright Commission and Grant Agency of the University of South Bohemia Grant 04-001/2014/P. J.P.M. was funded by National Science Foundation (NSF) Grants IOS-1256680 and IOS-1553529, National Aeronautics and Space Administration Astrobiology Institute Award NNA15BB04A, and NSF-Experimental Program to Stimulate Competitive Research Award NSF-IIA-1443108 (to the Montana Institute on Ecosystems).

1. Gray MW, Doolittle WF (1982) Has the endosymbiont hypothesis been proven? *Microbiol Rev* 46(1):1–42.
2. Palmer JD (1997) Organelle genomes: Going, going, gone! *Science* 275(5301):790–791.
3. Martin W, Müller M (1998) The hydrogen hypothesis for the first eukaryote. *Nature* 392(6671):37–41.
4. Embley TM, Martin W (2006) Eukaryotic evolution, changes and challenges. *Nature* 440(7084):623–630.
5. Douglas AE (1989) Mycetocyte symbiosis in insects. *Biol Rev Camb Philos Soc* 64(4): 409–434.
6. Nowack ECM, Melkonian M (2010) Endosymbiotic associations within protists. *Philos Trans R Soc Lond B Biol Sci* 365(1541):699–712.
7. Stewart FJ, Newton ILG, Cavanaugh CM (2005) Chemosynthetic endosymbioses: Adaptations to oxic-anoxic interfaces. *Trends Microbiol* 13(9):439–448.
8. Nakayama T, Ishida K (2009) Another acquisition of a primary photosynthetic organelle is underway in *Paulinella chromatophora*. *Curr Biol* 19(7):R284–R285.
9. Husnik F, et al. (2013) Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell* 153(7):1567–1578.
10. Sloan DB, et al. (2014) Parallel histories of horizontal gene transfer facilitated extreme reduction of endosymbiont genomes in sap-feeding insects. *Mol Biol Evol* 31(4):857–871.
11. Luan J-B, et al. (2015) Metabolic coevolution in the bacterial symbiosis of whiteflies and related plant sap-feeding insects. *Genome Biol Evol* 7(9):2635–2647.
12. McCutcheon JP, Moran NA (2011) Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* 10(1):13–26.
13. Moran NA, Bennett GM (2014) The tiniest tiny genomes. *Annu Rev Microbiol* 68: 195–215.
14. Nikoh N, et al. (2010) Bacterial genes in the aphid genome: Absence of functional gene transfer from *Buchnera* to its host. *PLoS Genet* 6(2):e1000827.
15. Nowack ECM, et al. (2011) Endosymbiotic gene transfer and transcriptional regulation of transferred genes in *Paulinella chromatophora*. *Mol Biol Evol* 28(1):407–422.
16. Nowack ECM, Grossman AR (2012) Trafficking of protein into the recently established photosynthetic organelles of *Paulinella chromatophora*. *Proc Natl Acad Sci USA* 109(14):5340–5345.
17. Nakabachi A, Ishida K, Hongoh Y, Ohkuma M, Miyagishima SY (2014) Aphid gene of bacterial origin encodes a protein transported to an obligate endosymbiont. *Curr Biol* 24(14):R640–R641.
18. Theissen U, Martin W (2006) The difference between organelles and endosymbionts. *Curr Biol* 16(24):R1016–R1017.
19. Keeling PJ, Archibald JM (2008) Organelle evolution: What's in a name? *Curr Biol* 18(8):R345–R347.
20. McCutcheon JP, Keeling PJ (2014) Endosymbiosis: Protein targeting further erodes the organelle/symbiont distinction. *Curr Biol* 24(14):R654–R655.
21. Keeling PJ, McCutcheon JP, Doolittle WF (2015) Symbiosis becoming permanent: Survival of the luckiest. *Proc Natl Acad Sci USA* 112(33):10101–10103.
22. Sloan DB, Moran NA (2012) Genome reduction and co-evolution between the primary and secondary bacterial symbionts of psyllids. *Mol Biol Evol* 29(12):3781–3792.
23. 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(9):1675–1688.
24. Nakabachi A, et al. (2013) Defensive bacteriome symbiont with a drastically reduced genome. *Curr Biol* 23(15):1478–1484.
25. 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(7):1683–1698.
26. Moran NA (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci USA* 93(7):2873–2878.
27. Fares MA, Barrio E, Sabater-Muñoz B, Moya A (2002) The evolution of the heat-shock protein GroEL from *Buchnera*, the primary endosymbiont of aphids, is governed by positive selection. *Mol Biol Evol* 19(7):1162–1170.
28. Bennett GM, Moran NA (2015) Heritable symbiosis: The advantages and perils of an evolutionary rabbit hole. *Proc Natl Acad Sci USA* 112(33):10169–10176.
29. Popadin KY, Nikolaev SI, Junier T, Baranova M, Antonarakis SE (2013) Purifying selection in mammalian mitochondrial protein-coding genes is highly effective and congruent with evolution of nuclear genes. *Mol Biol Evol* 30(2):347–355.
30. 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 (Bethesda)* 5(10):2165–2176.
31. Smith DR, Keeling PJ (2015) Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *Proc Natl Acad Sci USA* 112(33):10177–10184.
32. McCutcheon JP, Moran NA (2007) Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proc Natl Acad Sci USA* 104(49):19392–19397.
33. Koga R, Bennett GM, Cryan JR, Moran NA (2013) Evolutionary replacement of obligate symbionts in an ancient and diverse insect lineage. *Environ Microbiol* 15(7): 2073–2081.
34. Koga R, Moran NA (2014) Swapping symbionts in spittlebugs: Evolutionary replacement of a reduced genome symbiont. *ISME J* 8(6):1237–1246.
35. Thao ML, et al. (2000) Secondary endosymbionts of psyllids have been acquired multiple times. *Curr Microbiol* 41(4):300–304.
36. Lamelas A, et al. (2011) *Serratia symbiotica* from the aphid *Cinara cedri*: A missing link from facultative to obligate insect endosymbiont. *PLoS Genet* 7(11):e1002357.
37. Vogel KJ, Moran NA (2013) Functional and evolutionary analysis of the genome of an obligate fungal symbiont. *Genome Biol Evol* 5(5):891–904.
38. Smith WA, et al. (2013) Phylogenetic analysis of symbionts in feather-feeding lice of the genus *Columbicola*: Evidence for repeated symbiont replacements. *BMC Evol Biol* 13(1):109.
39. Lefèvre C, et al. (2004) Endosymbiont phylogenesis in the dryophthoridae weevils: Evidence for bacterial replacement. *Mol Biol Evol* 21(6):965–973.
40. Toju H, Tanabe AS, Notsu Y, Sota T, Fukatsu T (2013) Diversification of endosymbiosis: Replacements, co-speciation and promiscuity of bacteriocyte symbionts in weevils. *ISME J* 7(7):1378–1390.
41. 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(22):7521–7525.
42. McCutcheon JP, von Dohlen CD (2011) An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Curr Biol* 21(16):1366–1372.
43. Thao ML, Gullan PJ, Baumann P (2002) Secondary (gamma-Proteobacteria) endosymbionts infect the primary (beta-Proteobacteria) endosymbionts of mealybugs multiple times and coevolve with their hosts. *Appl Environ Microbiol* 68(7):3190–3197.
44. Kono M, Koga R, Shimada M, Fukatsu T (2008) Infection dynamics of coexisting beta- and gammaproteobacteria in the nested endosymbiotic system of mealybugs. *Appl Environ Microbiol* 74(13):4175–4184.

45. von Dohlen CD, Kohler S, Alsop ST, McManus WR (2001) Mealybug β-proteobacterial endosymbionts contain γ-proteobacterial symbionts. *Nature* 412(6845):433–436.
46. López-Madrigal S, et al. (2014) Molecular evidence for ongoing complementarity and horizontal gene transfer in endosymbiotic systems of mealybugs. *Front Microbiol* 5:449.
47. Parkinson JF, Gobin B, Hughes WOH (2016) Heritability of symbiont density reveals distinct regulatory mechanisms in a tripartite symbiosis. *Ecol Evol* 6(7):2053–2060.
48. Gatehouse LN, Sutherland P, Forgie SA, Kaji R, Christeller JT (2012) Molecular and histological characterization of primary (betaproteobacteria) and secondary (gammaproteobacteria) endosymbionts of three mealybug species. *Appl Environ Microbiol* 78(4):1187–1197.
49. Koga R, Nikoh N, Matsuura Y, Meng XY, Fukatsu T (2013) Mealybugs with distinct endosymbiotic systems living on the same host plant. *FEMS Microbiol Ecol* 83(1): 93–100.
50. Buchner P (1965) *Endosymbiosis of Animals with Plant Microorganisms* (Interscience Publishers, New York), p 909.
51. Denton JF, et al. (2014) Extensive error in the number of genes inferred from draft genome assemblies. *PLOS Comput Biol* 10(12):e1003998.
52. International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol* 8(2):e1000313.
53. Moran NA, Mira A (2001) The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biol* 2(12):H0054.
54. Frank AC, Amiri H, Andersson SG (2002) Genome deterioration: Loss of repeated sequences and accumulation of junk DNA. *Genetica* 115(1):1–12.
55. Moran NA (2002) Microbial minimalism: Genome reduction in bacterial pathogens. *Cell* 108(5):583–586.
56. Moran NA, McCutcheon JP, Nakabachi A (2008) Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet* 42:165–190.
57. Moya A, Peretó J, Gil R, Latorre A (2008) Learning how to live together: Genomic insights into prokaryote-animal symbioses. *Nat Rev Genet* 9(3):218–229.
58. Clayton AL, et al. (2012) A novel human-infection-derived bacterium provides insights into the evolutionary origins of mutualistic insect-bacterial symbioses. *PLoS Genet* 8(11):e1002990.
59. Oakeson KF, et al. (2014) Genome degeneration and adaptation in a nascent stage of symbiosis. *Genome Biol Evol* 6(1):76–93.
60. McCutcheon JP, Moran NA (2010) Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. *Genome Biol Evol* 2:708–718.
61. Rosenblueth M, Sayavedra L, Sámano-Sánchez H, Roth A, Martínez-Romero E (2012) Evolutionary relationships of flavobacterial and enterobacterial endosymbionts with their scale insect hosts (Hemiptera: Coccoidea). *J Evol Biol* 25(11):2357–2368.
62. Toh H, et al. (2006) Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Res* 16(2):149–156.
63. Hosokawa T, Kaiwa N, Matsuura Y, Kikuchi Y, Fukatsu T (2015) Infection prevalence of *Sodalis symbionts* among stinkbugs. *Zoological Lett* 1(1):5.
64. Dale C, Young SA, Haydon DT, Welburn SC (2001) The insect endosymbiont *Sodalis glossinidius* utilizes a type III secretion system for cell invasion. *Proc Natl Acad Sci USA* 98(4):1883–1888.
65. Duncan RP, et al. (2014) Dynamic recruitment of amino acid transporters to the insect/symbiont interface. *Mol Ecol* 23(6):1608–1623.
66. Michalik A, Jankowska W, Kot M, Gołas A, Szklarzewicz T (2014) Symbiosis in the green leafhopper, *Cicadella viridis* (Hemiptera, Cicadellidae). Association *in statu nascendi? Arthropod Struct Dev* 43(6):579–587.
67. Kobiałka M, Michalik A, Walczak M, Junkiert Ł, Szklarzewicz T (2016) *Sulcia* symbiont of the leafhopper *Macrosteles laevis* (Ribaut, 1927) (Insecta, Hemiptera, Cicadellidae: Deltocephalinae) harbors *Arsenophonus* bacteria. *Protoplasma* 253(3):903–912.
68. Wang Z, Wu M (2014) Phylogenomic reconstruction indicates mitochondrial ancestor was an energy parasite. *PLoS One* 9(10):e110685.
69. Ochoa de Alda JAG, Esteban R, Diago ML, Houmard J (2014) The plastid ancestor originated among one of the major cyanobacterial lineages. *Nat Commun* 5:4937.
70. Booth A, Doolittle WF (2015) Eukaryogenesis, how special really? *Proc Natl Acad Sci USA* 112(33):10278–10285.
71. Kurland CG, Andersson SG (2000) Origin and evolution of the mitochondrial proteome. *Microbiol Mol Biol Rev* 64(4):786–820.
72. Gray MW (2015) Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. *Proc Natl Acad Sci USA* 112(33):10133–10138.
73. Ku C, et al. (2015) Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes. *Proc Natl Acad Sci USA* 112(33):10139–10146.
74. Zimorski V, Ku C, Martin WF, Gould SB (2014) Endosymbiotic theory for organelle origins. *Curr Opin Microbiol* 22:38–48.
75. Ettema TJG (2016) Evolution: Mitochondria in the second act. *Nature* 531(7592): 39–40.
76. Pittis AA, Gabaldón T (2016) Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature* 531(7592):101–104.
77. Ku C, et al. (2015) Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* 524(7566):427–432.
78. Koonin EV (2015) Archaeal ancestors of eukaryotes: Not so elusive any more. *BMC Biol* 13(1):84.
79. Larkum AWD, Lockhart PJ, Howe CJ (2007) Shopping for plastids. *Trends Plant Sci* 12(5):189–195.
80. Fares MA, Ruiz-González MX, Moya A, Elena SF, Barrio E (2002) Endosymbiotic bacteria: groEL buffers against deleterious mutations. *Nature* 417(6887):398.
81. Andersson JO, Andersson SG (1999) Insights into the evolutionary process of genome degradation. *Curr Opin Genet Dev* 9(6):664–671.
82. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.
83. Bankevich A, et al. (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19(5):455–477.
84. Boisvert S, Laviolette F, Corbeil J (2010) Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J Comput Biol* 17(11):1519–1533.
85. Ruby JG, Bellare P, Derisi JL (2013) PRICE: Software for the targeted assembly of components of (Meta) genomic sequence data. *G3 (Bethesda)* 3(5):865–880.
86. Walker BJ, et al. (2014) Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9(11):e112963.
87. Hunt M, et al. (2013) REAPR: A universal tool for genome assembly evaluation. *Genome Biol* 14(5):R47.
88. Seemann T (2014) Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30(14):2068–2069.
89. Konwar KM, Hanson NW, Pagé AP, Hallam SJ (2013) MetaPathways: A modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics* 14(1):202.
90. Karp PD, et al. (2010) Pathway Tools version 13.0: Integrated software for pathway/genome informatics and systems biology. *Brief Bioinform* 11(1):40–79.
91. Jones P, et al. (2014) InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30(9):1236–1240.
92. Rutherford K, et al. (2000) Artemis: Sequence visualization and annotation. *Bioinformatics* 16(10):944–945.
93. Segata N, Börnigen D, Morgan XC, Huttenhower C (2013) PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* 4:2304.
94. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 9(4):286–298.
95. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973.
96. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
97. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572–1574.
98. Lartillot N, Rodrigue N, Stubbs D, Richer J (2013) PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol* 62(4):611–615.
99. Darling AE, Mau B, Perna NT (2010) progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5(6):e11147.
100. Li L, Stoeckert CJ, Jr, Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* 13(9):2178–2189.
101. Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M (2013) Blobology: Exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front Genet* 4:237.
102. Koutsovoulos G, et al. (2016) No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proc Natl Acad Sci USA* 113(18): 5053–5058.
103. Delmont TO, Eren AM (2016) Identifying contamination with advanced visualization and analysis practices: Metagenomic approaches for eukaryotic genome assemblies. *PeerJ* 4:e1839.
104. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* 29(8):1072–1075.
105. Parra G, Bradnam K, Korf I (2007) CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23(9):1061–1067.
106. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210–3212.
107. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* 33(20):6494–6506.
108. Wheeler TJ, Eddy SR (2013) nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29(19):2487–2489.
109. Huerta-Cepas J, Dopazo J, Gabaldón T (2010) ETE: A python environment for tree exploration. *BMC Bioinformatics* 11(1):24.
110. 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(6):1270–1280.
111. Schindelin J, et al. (2012) Fiji: An open-source platform for biological-image analysis. *Nat Methods* 9(7):676–682.

Summary

This PhD thesis unfolds a path to an intimate endosymbiosis that can be compared to what we think happened before (and to some extent after) bacterial ancestors of eukaryotic organelles, mitochondria and plastids, became highly integrated into their host cells. First, the extreme genome reduction of mealybug symbionts has not been enabled by endosymbiotic gene transfer to the host nucleus, but rather by very intimate host-symbiont-symbiont cooperation and horizontal gene transfer from diverse bacteria infecting the host oocytes. Second, the marked fluidity over evolutionary time in the mealybug system implies that serial symbiont replacement can happen even in the most intricate symbiotic arrangements, and that pre-existing horizontally transferred genes can remain stable on genomes in the face of extensive symbiont turnover. Do these results allow us to say that insect endosymbionts are comparable to mitochondria and plastids? They do not if you define organelles as organisms that originated early in the eukaryotic clade and dramatically shaped its evolution. But if we put aside age and perceived specialness, many of the mechanistic and evolutionary outcomes of intimate endosymbiosis discussed in this thesis seem similar between organelles and insect endosymbionts. I argue that these other, much younger symbioses may tell us something about how the mitochondria and plastids came to be, at the very least by revealing what types of evolutionary events are possible as stable intracellular relationships proceed along the path of integration.

CURRICULUM VITAE

Filip Husník

CONTACT INFORMATION

Biology Centre of the Czech Academy of Sciences, Institute of Parasitology &
University of South Bohemia, Faculty of Science
Branišovská 31, České Budějovice 370 05, Czech Republic
E-mail: filip.husnik@gmail.com or filip@paru.cas.cz
Personal website: www.filiphusnik.com
Google Scholar profile: https://scholar.google.com/citations?user=cMf9LXwAAAAJ

EDUCATION

since 2012
Ph.D. student of Molecular and Cell Biology and Genetics
Department of Molecular Biology and Genetics, Faculty of Science, **University of South Bohemia**, Czech Republic &
Institute of Parasitology, **Czech Academy of Sciences**, Czech Republic
Thesis: *Genomic and Cellular Integration in the Tripartite Nested Mealybug Symbiosis*
Supervisor: John McCutcheon (University of Montana)
University guarantor: Miroslav Oborník

2012
RNDr., Parasitology
Faculty of Science, **University of South Bohemia**, Czech Republic

2010-2012
M.S., Parasitology
Department of Parasitology, Faculty of Science, **University of South Bohemia**, Czech Republic. Thesis: *Evolutionary origins of intracellular symbionts in arthropods.* Supervisors: Tomáš Chrudimský, Václav Hypša.

2007-2010
B.S., Biology
Department of Parasitology, Faculty of Science, **University of South Bohemia**, Czech Republic. Thesis: *Molecular phylogeny of intracellular symbiotic Gammaproteobacteria in insects.* Supervisors: Tomáš Chrudimský, Václav Hypša.

PROFESSIONAL EXPERIENCE

Employment:
2012-2017
Graduate student, Institute of Parasitology, **Biology Centre of the Czech Academy of Sciences**

2010-2015
Research worker, Faculty of Science, University of South Bohemia

Research stays:
2016
Visiting Student (03/07-14/07)
Anna Michalik, **Jagiellonian University**, Krakow, Poland

2015
Visiting Student (06/10-31/10)

Laura Ross lab, **University of Edinburgh**, Edinburgh, UK

2014-2015

Fulbright Visiting Student Researcher (18/08-25/05)
John McCutcheon lab, **University of Montana**, Missoula, USA

2012-2013

Erasmus Visiting Student Researcher (31/10-04/02)
Alistair Darby lab, **University of Liverpool**, Liverpool, UK

2011

Visiting Student (03/06-17/08)
John McCutcheon lab, **University of Montana**, Missoula, USA

PEER-REVIEWED PUBLICATIONS

Husnik F, McCutcheon JP: Repeated replacement of an intrabacterial symbiont in the tripartite nested mealybug symbiosis. ***Proceedings of the National Academy of Sciences of the United States of America*** 2016, 113(3): E5416-5424.

Nováková E, Hypša V, Nguyen P, **Husník F**, Darby AC: Genome sequence of *Candidatus* Arsenophonus lipopteni, the exclusive symbiont of a blood sucking fly *Lipoptena cervi* (Diptera: Hippoboscidae). ***Standards in Genomic Sciences*** 2016, 11: 72.

Kyselková M, Chrudimský T, **Husník F**, Chroňáková A, Heuer H, Smalla K, Elhottová D: Characterization of tet (Y)-carrying LowGC plasmids exogenously captured from cow manure at a conventional dairy farm. ***FEMS Microbiology Ecology*** 2016, 92(6): fiw075.

Nováková E, **Husnik F**, Šochová E, Hypša V: *Arsenophonus* and *Sodalis* symbionts in louse flies: an analogy to the *Wigglesworthia* and *Sodalis* system in tsetse flies. ***Applied and Environmental Microbiology*** 2015, 81 (18): 6189-6199.

Duncan RP, **Husnik F**, Van Leuven JT, Gilbert DG, Dávalos LM, McCutcheon JP, Wilson ACC: Dynamic recruitment of amino acid transporters to the insect/symbiont interface. ***Molecular Ecology*** 2014, 23(6): 1608-1623.

Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, Tanaka M, Satoh N, Bachtrog D, Wilson ACC, von Dohlen CD, Fukatsu T, McCutcheon JP: Horizontal Gene Transfer from Diverse Bacteria to an Insect Genome Enables a Tripartite Nested Mealybug Symbiosis. ***Cell*** 2013, 153(7): 1567-1578.

Chrudimský T, **Husník F**, Nováková E, Hypša V: *Candidatus* Sodalis melophagi sp. nov.: phylogenetically independent comparative model to the tsetse fly symbiont *Sodalis glossinidius*. ***PLoS ONE*** 2012, 7(7): e40354.

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

PRESENTATIONS AT CONFERENCES

2016

XIV International Symposium on Scale Insect Studies, Catania, Italy (13-16/06). Talk presentation.

2015

8th International Symbiosis Society Congress. Lisbon, Portugal (12-18/07). Talk presentation.

2014

Symbioses becoming permanent: The origins and evolutionary trajectories of organelles. Irvine, CA, USA (15-17/10). Poster presentation.

8th International Wolbachia Conference. Innsbruck, Igls, Austria (06-11/06). Talk presentation.

2013

12th International Colloquium on Endocytobiology and Symbiosis. Dalhousie University, Halifax, Nova Scotia, Canada (18-22/08). Talk presentation.

2012

7th International Symbiosis Society Congress. Krakow, Poland (22-28/07). Poster presentation.

7th International Wolbachia Conference and Final Meeting of the EU COST Action FA0701 "Arthropod Symbiosis: from fundamental studies to pest and disease management". La Vieille Perrotine, Ile d'Oléron, France (07-14/06). Poster presentation.

OTHER LECTURES

2015

Charles University, Faculty of Science, Prague, Czech Republic (10/11)
University of Edinburgh, Institute of Evolutionary Biology, Edinburgh, UK (15/10)

2014

University of Ostrava, Faculty of Science, Ostrava, Czech Republic (18/02)

2013

Charles University, Faculty of Science, Prague, Czech Republic (28/11)

HONORS, AWARDS, AND FUNDING

2017-2019

EMBO long-term postdoctoral fellowship (laboratory of Patrick Keeling, University of British Columbia, Vancouver, Canada)

2015

30 Under 30, Forbes Czech Republic

2014-2015

Fulbright visiting student fellowship (laboratory of John McCutcheon, University of Montana, Missoula, USA)

2014-2015

Grant Agency of the University of South Bohemia (001/2014/P, PI: Filip Husník)
Evolution of intrabacterial symbiosis in mealybugs: from mosaic pathways to mosaic organisms.

2012-2013

Erasmus visiting student fellowship (laboratory of Alistair Darby, University of Liverpool, Liverpool, UK)

2012

Dean's award for excellent research results presented in master thesis.

2009

Student Grant Agency of the University of South Bohemia (SGA2009002, PI: Filip Husník)
Molecular phylogeny of symbiotic Gammaproteobacteria in insects

---

ATTENDED WORKSHOPS

Workshop on Genomics, Český Krumlov, Czech Republic (08-21/01/2012).

Host-microbe symbioses – old friends and foes. Instituto Gulbenkian de Ciencia, Oeiras, Lisbon, Portugal (19-31/07/2015).

SKILLS

Laboratory methods: more than ten years of experience with basic molecular biology methods (incl. genomic and transcriptomic library preparations), microscopy methods (light, fluorescence/confocal, and TEM), and insect cell culture and symbiotic bacteria cultivation.

Bioinformatics: good knowledge of Unix and phylogenomics, genomics, and transcriptomics programs

Programming: basic experience in Bash, Perl, Processing/Java, Python, R, and C++.

Languages: English (full professional proficiency), French (limited working proficiency), Czech (native).

TEACHING AND STUDENTS

2016

Teaching assistant at the Workshop on Population and Speciation Genomics [http://evomics.org/] , Cesky Krumlov, Czech Republic

Courses taught at the University of South Bohemia:
2014
Introduction to Genomics (selected lectures and exercises on Unix, databases and mapping, genomics, and transcriptomics)
2013
Biology of Parasites (selected lectures on phylogeny and evolution of parasites)
2012, 2013
Molecular Phylogenetics (selected lectures and exercises on probabilistic methods)

Mentoring students at the University of South Bohemia:

Kamila Machová, Bioinformatics cross-border B.S. student (University of South Bohemia/Johannes Kepler University of Linz). RNA biology of symbiotic bacteria in insects [defended in 2016].

Eva Šochová, Biology B.S. and Parasitology M.S. student. Genomics of symbiotic bacteria in bloodsucking insects [defended in 2014 and 2016].

PEER-REVIEW

Molecular Biology and Evolution, ISME J, Genome Biology and Evolution, Applied and Environmental Microbiology, PeerJ, PloS One, Physiological Entomology, Scientific Reports

---