Genome expansion via lineage splitting and genome reduction in the cicada endosymbiont *Hodgkinia*

Matthew A. Campbell[a,1], James T. Van Leuven[a,1], Russell C. Meister[b], Kaitlin M. Carey[a], Chris Simon[b], and John P. McCutcheon[a,c,2]

[a]Division of Biological Sciences, University of Montana, Missoula, MT 59812; [b]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269; and [c]Program in Integrated Microbial Biodiversity, Canadian Institute for Advanced Research, Toronto, ON, Canada M5G 1Z8

Edited by W. Ford Doolittle, Dalhousie University, Halifax, NS, Canada, and approved April 16, 2015 (received for review December 22, 2014)

Comparative genomics from mitochondria, plastids, and mutualistic endosymbiotic bacteria has shown that the stable establishment of a bacterium in a host cell results in genome reduction. Although many highly reduced genomes from endosymbiotic bacteria are stable in gene content and genome structure, organelle genomes are sometimes characterized by dramatic structural diversity. Previous results from *Candidatus* Hodgkinia cicadicola, an endosymbiont of cicadas, revealed that some lineages of this bacterium had split into two new cytologically distinct yet genetically interdependent species. It was hypothesized that the long life cycle of cicadas in part enabled this unusual lineage-splitting event. Here we test this hypothesis by investigating the structure of the *Ca.* Hodgkinia genome in one of the longest-lived cicadas, *Magicicada tredecim*. We show that the *Ca.* Hodgkinia genome from *M. tredecim* has fragmented into multiple new chromosomes or genomes, with at least some remaining partitioned into discrete cells. We also show that this lineage-splitting process has resulted in a complex of *Ca.* Hodgkinia genomes that are 1.1-Mb pairs in length when considered together, an almost 10-fold increase in size from the hypothetical single-genome ancestor. These results parallel some examples of genome fragmentation and expansion in organelles, although the mechanisms that give rise to these extreme genome instabilities are likely different.

symbiosis | genome evolution | nonadaptive evolution | organelles | bacteria

The first published genome from a nutritional bacterial endosymbiont of an insect was *Buchnera aphidicola* from the pea aphid *Acyrthosiphon pisum* (AP) (1). This landmark paper provided many key insights that would be repeatedly reinforced in different bacterial symbioses during the subsequent 15 y, including extreme gene loss and genome reduction, precise metabolic complementarity and interdependence with the host insect, highly biased nucleotide and amino acid compositions, and limited gene sets involved in DNA repair, gene regulation, and cell envelope biosynthesis (2). The second complete *Buchnera* genome, from the aphid *Schizaphis graminum* (SG), provided the next archetype for endosymbiont genomes: the *Buchnera* AP and SG genomes showed no rearrangements or gene acquisitions, despite large amounts of sequence evolution and 50+ million y of divergence (3). Unusual genomic structural stability has been repeatedly found in many other insect endosymbiont genera, including *Blochmannia* (4, 5), an ant endosymbiont; *Sulcia* (6–8), which forms a widespread and ancient association with sap-feeding insects, such as sharpshooters, spittlebugs, and cicadas (9), and *Carsonella*, an endosymbiont of psyllids (10, 11). A pattern thus emerged whereby the process of genome reduction in endosymbionts resulted in small and stable genomes.

However, several other examples, some recently published, have placed small cracks in the façade of genomic stability in endosymbionts. Sequencing of *Buchnera* from a third more diverged aphid genus showed two inversion rearrangements and two small translocations relative to the first two genomes (12). Genomes from various endosymbiont genera found in cockroaches (13), tsetse flies (14), mealybugs (15), leafhoppers (8), and especially whiteflies (16) also showed some structural rearrangements in otherwise completely colinear genomes (reviewed in ref. 16). Although these results do not much change the general picture of genomic stability in endosymbionts, they do suggest that highly reduced endosymbiotic genomes are not strictly fated to unalterable colinearity and stability given the right circumstances.

Some mitochondrial genomes are similar to the highly reduced genomes of endosymbionts in that they are incredibly stable in size and number of genes across a wide diversity of hosts. For example, the 4,000+ sequenced bilaterian animal mitochondrial genomes typically, but not always (17), map as a single circular 14- to 20-kb molecule and encode the same 37 genes (18) (Fig. 1). However, well before the first complete animal mitochondrial genome was sequenced (19), it was clear that some mitochondrial genomes, especially in vascular plants (20, 21), were large and variable in size and thus very different from those found in animals (22). More recent mitochondrial genome sequencing has confirmed these early studies, and efforts aimed at diverse eukaryotes have shown that their mitochondrial genomes range widely in structure (circular, linear, single chromosomes, multiple chromosomes) and size (ranging over more than three orders-of-magnitude, from about 6 kb to 11 Mb) (23–26) (Fig. 1).

Interestingly, this wild diversity in organelle genome structure and size is not reflected in coding capacity (25). All known mitochondrial genomes encode genes that are subsets of the ∼70 protein-coding genes encoded by the ∼70-kb jakobid protist mitochondrial genomes (27, 28). Even the enormous vascular plant mitochondrial genomes—which can be >100 times larger than those found in early land plants (29)—contain only ∼25 nonredundant protein-coding genes, all of which are present on jakobid mitochondrial genomes (24). It is now clear that large mitochondrial genomes are derived from smaller mitochondrial genomes through processes such as repetitive DNA expansion and foreign DNA acquisition (23, 24, 30–32).

Over the past decade, the distinction between "endosymbiont" and "organelle" has become increasingly more difficult to make (33): organelle and endosymbiont genomes overlap in size and coding capacity (Fig. 1), genome reduction in both can be

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Symbioses Becoming Permanent: The Origins and Evolutionary Trajectories of Organelles," held October 15–17, 2014, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Symbioses.

Author contributions: M.A.C., J.T.V.L., R.C.M., C.S., and J.P.M. designed research; M.A.C., J.T.V.L., R.C.M., K.M.C., C.S., and J.P.M. performed research; M.A.C., J.T.V.L., R.C.M., K.M.C., C.S., and J.P.M. analyzed data; and C.S. and J.P.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database [accession nos. CP010828 (Sulcia genome), JYFR00000000 (Hodgkinia genome), and KR607294–KR607430 (Hodgkinia 16S)]. Illumina reads were deposited in the Sequence Read Archive database, www.ncbi.nlm.nih.gov/sra (accession no. SRS824162).

[1]M.A.C. and J.T.V.L. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: john.mccutcheon@umontana.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1421386112/-/DCSupplemental.

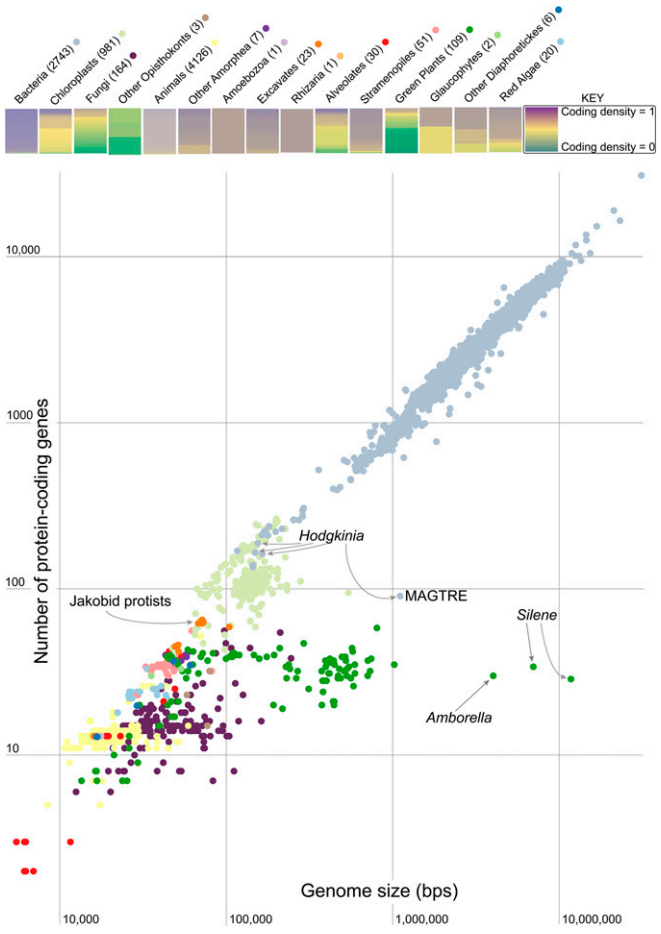

Fig. 1. Genome sizes, gene numbers, and coding density for all sequenced bacterial and organelle genomes. Number of protein coding genes are plotted as a function of genome size. Genomes from organisms called out in the main text are noted. The color-coded heat maps at the top show the coding density of every genome in each major group as defined by Eme et al. (59); the key is shown at the upper right. The number of mitochondrial genomes in the major groups of eukaryotes is show in parentheses. Genome annotations were downloaded on October 18, 2014 from ftp.ncbi.nlm.nih.gov/genomes.

enabled by horizontal gene transfer to the host (34–37), and both have been shown to import proteins produced by the host (38, 39). But what of the apparent difference in genomic stability observed between endosymbionts and organelles?

We recently reported that some lineages of the cicada endosymbiont *Candidatus* Hodgkinia cicadicola (hereafter *Hodgkinia*) showed unusual genomic structural instability (40). Like many other insect endosymbionts, the first sequenced *Hodgkinia* genome, from the cicada *Diceroprocta semicincta* (DICSEM), had a single very small circular mapping genome of 144 kb (41). However, in the cicada *Tettigades undata* (TETUND), the *Hodgkinia* lineage has split into two new lineages, each isolated to distinct *Hodgkinia* cells (40). These two new genomes are smaller and have both lost genes compared with the single-genome version of *Hodgkinia* present in DICSEM and in a closely related congener, *Tettigades ulnaria* (TETULN). Strikingly, gene loss has occurred in a complementary pattern so that all ancestral genes encoded in *Hodgkinia* TETULN are retained on at least one of the new *Hodgkinia* TETUND genomes (40). As a result, the host cicada now relies on two species of *Hodgkinia* with a combined genome size nearly twice the size of their single-genome ancestor. We proposed a model to explain the genomic instability in *Hodgkinia* (40) (see, for example, Fig. 3) based partly on the unusually long and variable cicada life cycle. Here we test this model by analyzing the genomic and cytological structure of *Hodgkinia* from one of the longest-lived cicada species, the 13-y periodical cicada *Magicicada tredecim* (MAGTRE).

Results and Discussion

Extravagant Complexity in *Hodgkinia* from *M. tredecim*. We first attempted to sequence both the *Sulcia* and *Hodgkinia* genomes from MAGTRE using short-insert Illumina sequencing methods. The *Sulcia* MAGTRE assembly reflected the structural stability of many endosymbiont genomes: it cleanly assembled into one circular-mapping 268-kb molecule that was completely colinear with all other *Sulcia* genomes. In contrast, the reads associated with the *Hodgkinia* genome assembled into an extremely complex mix of small contigs. We added sequencing reads from a 2.5-kb large-insert Illumina library with the aim of joining these small contigs into larger scaffolds. We found 233 scaffolds assembled from these combined data that contained recognizable *Hodgkinia* sequences and totaled 1.1 Mb in length. The assembled *Hodgkinia* scaffolds were present at different depths of coverage (Table 1), consistent with what would be expected if the scaffolds did not assemble from DNA fragments derived from the same physically linked DNA molecule.

We also found many cases where different versions of the same gene were present on several different scaffolds. The variation in depth of sequencing coverage combined with the existence of related stretches of sequence at various levels of similarity made it difficult for us to finish the entire 1.1-Mb *Hodgkinia* assembly into distinct molecules. However, we identified 27 scaffolds totaling 739 kb of sequence where mate-pair information suggested the two scaffold ends were joined to each other (Fig. 2 and Table 1). Of these 27, we were able to verify that 17 scaffolds were circular-mapping molecules by PCR and Sanger sequencing, or by using ∼421 Mb of PacBio long-read data. These 17 verified circles totaled 512 kb of sequence (Table 1). The remaining 10 circular scaffolds were not closed by PacBio reads and did not provide clean PCR results because of stretches of sequence that were shared by many scaffolds. However, because all of them had at least five independent gap-spanning paired-end reads, we consider these putative circular-mapping molecules. The remaining 206 scaffolds contained 424 kb of sequence, ranged in size from 200 bp to 27 kb in length (166 of these were less than 2 kb in length), were frequently broken at stretches of sequence that were shared among several different scaffolds, and were left as a draft assembly.

We next searched the entire 1.1-Mb *Hodgkinia* assembly for full-length open reading frames (ORFs). We found only 165 ORFs that were apparently functional. Ninety-seven were unique (that is, 68 were paralogs of other genes, showing on average 76% sequence identity at the amino acid level), representing about 60% of the 155 ORFs we expected based on previous *Hodgkinia* genomes (40, 41). Sixty-eight of the 165 ORFs were on the 17 closed circular molecules; 47 of these were unique (Fig. 2). Because we found no additional ORFs outside of these 165, we conclude that, like in TETUND, homologs of the ∼150 genes present on the single-genome versions of *Hodgkinia* are the only genes present in the entire MAGTRE assembly. The *Hodgkinia* assembly also contained many pseudogenes, but we restricted the analysis in this paper to the 17 verified circles because of the difficulty in identifying nonfunctional genes in draft assemblies of rapidly evolving sequence (the average percent identity at the amino acid level was ∼35% between MAGTRE and DICSEM orthologs, and 40% for MAGTRE-TETULN comparisons). The intergenic regions of these closed circular molecules contained mostly sequences that had no significant similarity to anything in sequence databases (Table 1). The coding density of these 17 molecules was thus extremely low, the most gene-dense circle being 45% coding DNA (Fig. 2 and Table 1). It is worth noting that the assembled region of the 13-kb scaffold PUTATIVE006 contains four pseudogenes but no obviously functional genes (Table 1). It is possible that a functional gene exists in the unfinished gap; if not, we would expect this circle to be under little selection to be maintained and likely to be lost in other cicada lineages.

Genes for the biosynthesis of methionine, histidine, and a vitamin B_{12}-like molecule have been found on all previous *Hodgkinia* genomes. This is thought to reflect the nutritional contribution of *Hodgkinia* to the symbiosis (6). We looked for evidence that these genes were conserved in the *Hodgkinia* MAGTRE assembly, and found that they were distributed on several scaffolds. For example, apparently functional copies of all genes in the histidine biosynthesis pathway except histidinol-phosphate aminotransferase (*hisC*) are present on at least one of the 27 circles shown in Fig. 2 (*hisC* is present as a pseudogene on a small scaffold outside of the 27 circles). It is presently unclear if functional copies of the genes missing in the histidine or B_{12} pathways are present but poorly assembled, or if like in mealybugs and psyllids, the insect host has taken over these functions (35, 36).

We noticed that the average guanine + cytosine (GC) content of the *Hodgkinia* MAGTRE assembly was quite different from other *Hodgkinia* genomes: MAGTRE was 28% GC, TETUND was 47% GC, and DICSEM was 58% GC on average. It has been shown that *Hodgkinia* DICSEM has an AT mutational bias (42), and thus the low GC content in *Hodgkinia* MAGTRE seems to reflect this mutational bias more strongly compared with other *Hodgkinia* lineages. It was proposed that selection is acting to keep the genomic GC content high in *Hodgkinia* DICSEM (42); if this is correct, it seems this selective force has been lost, reduced, or was never present in *Hodgkinia* MAGTRE.

Some of These Circles Reside in Discrete *Hodgkinia* Cells. We looked for evidence that the 17 closed MAGTRE circles arose through the lineage-splitting and reductive process that we hypothesized for the duplicated TETUND genomes (Fig. 3*B*). First, we assumed that none of the MAGTRE circles should be larger than the inferred ancestral single genome size of 144–150 kb. This assumption holds, as the largest closed circle we find is 61 kb (although it is possible that some of the unassembled scaffolds in our assembly join to form larger molecules). Second, we expect that many of the MAGTRE circles should display regions of colinearity with each other and with other *Hodgkinia* genomes. Analysis of the five most gene-rich circles (encoding between 8 and 20 genes each) reveals small blocks of 5–10 genes that are colinear with other *Hodgkinia* genomes in a background of what seems to be an extensive history of gene rearrangement. This finding is consistent with what was observed in TETUND on a much smaller scale, where one of the two new genomes had an inversion relative to the other (40). Third—and most critically—because the two *Hodgkinia* TETUND genomes were isolated into discrete cells (40), we expect that this would also be true for MAGTRE if the process driving the lineage splitting is the same. Although we could not exhaustively check all combinations of the 233 *Hodgkinia* MAGTRE scaffolds using genome-targeted fluorescence microscopy, we did find evidence that 4 of the 17 finished circles are partitioned into separate *Hodgkinia* cells that are intermixed in the same host tissue (Fig. 4). None of the four tested circles showed overlapping signal, suggesting that at least these four molecules (and perhaps many others) remain separated into discrete cells. We also find that the scaffold assembly coverage, which corresponds to the frequency with which the molecule is present in the sample, correlates with the number of cells producing signal such that lower coverage scaffolds were present in fewer cells that higher coverage scaffolds (Fig. 4).

Table 1. Statistics for the 27 assembled *Hodgkinia* MAGTRE circles

Name	Length (bp)	No. genes	Coding density (%)	Unrecognizable DNA (%)	Average coverage
MAGTRE001	61,247	6	8.6	76.4	181
MAGTRE002	59,026	7	15.3	60.0	273
MAGTRE003	58,092	20	40.9	42.8	673
MAGTRE004	43,811	4	7.7	74.8	110
MAGTRE005	41,390	5	15.3	40.0	259
MAGTRE006	40,594	2	4.9	71.4	57
MAGTRE007	32,598	4	19.1	44.7	126
MAGTRE008	25,773	3	9.7	51.8	145
MAGTRE009	25,213	4	14.3	64.3	209
MAGTRE010	24,039	8	41.5	50.9	227
MAGTRE011	19,536	2	4.8	53.9	132
MAGTRE012	19,044	3	45.1	47.4	70
MAGTRE013	13,768	2	12.9	47.7	91
MAGTRE014	12,894	1	7.2	25.5	83
MAGTRE015	12,600	3	16.1	61.9	481
MAGTRE016	11,445	2	12.4	77.2	88
MAGTRE017	11,402	1	7.6	84.2	120
PUTATIVE001	51,436	9	16.2	39.0	407
PUTATIVE002	43,303	20	45.4	37.7	863
PUTATIVE003	31,053	3	11.3	42.2	205
PUTATIVE004	23,623	4	13.6	56.5	193
PUTATIVE005	21,975	2	16.3	67.3	111
PUTATIVE006	13,053	0	0.0	57.4	92
PUTATIVE007	11,360	2	38.4	49.3	140
PUTATIVE008	10,904	2	15.2	54.5	140
PUTATIVE009	10,150	2	22.4	60.9	144
PUTATIVE010	9,970	2	15.3	24.5	110

The average coverage values reflect only the reads generated in the large-insert Illumina sequencing. The percentage of unrecognizable DNA is an estimate of the total amount of sequence for each circle that has no significant similarity (blast e-value less than 0.1) to anything in the GenBank nonredundant protein database.

These data are consistent with a process where new *Hodgkinia* lineages are created when ancestral circles split into new cytologically distinct molecules (Fig. 3*C*). It is presently unclear if all 17 finished circles we have found are present in separate cells. Work from other endosymbionts shows that small plasmid-like subgenomic molecules can stably fracture from the main chromosome (35), so it is possible that part of what we are seeing in this complex mix of molecules is a combination of genomes that have split into new lineages combined with subgenomic circles that have split off from larger chromosomes. Given this uncertainty, we use the term "circle" instead of "chromosome" or "genome" throughout this report. Because we see little evidence for tandem duplication playing a role in the evolution of *Hodgkinia* genes, we can use the number of duplicate genes as an estimate for a minimum number of splitting events. There are least 10 distinct copies of the small ribosomal subunit RNA (16S rRNA) in the assembly, each present on a different circle, indicating that the *Hodgkinia* lineage may have split at least nine times (although the presence of at least 27 circular-mapping molecules suggests the number might be substantially higher). Finally, consistent with existence of numerous *Hodgkinia* lineages, we noticed that the amount of bacteriome tissue taken up by *Hodgkinia* relative to *Sulcia* was much larger in MAGTRE than in TETUND or DICSEM (Fig. 4 *G–I*). This may be an adaptation by the insect to accommodate the numerous different cellular and genomic lineages of *Hodgkinia* MAGTRE.

All 13- and 17-y Periodical Cicada Species Encode Multiple *Hodgkinia* Lineages. To rapidly survey the genomes of other *Magicicada* species, we sequenced multiple 16S rRNA clones from single individuals of all seven species of 13- and 17-y periodical cicadas. Because we found several copies of the 16S gene that were distinguishable in our genomic data, we reasoned that if other periodical cicadas shared the same *Hodgkinia* complexity they should contain several 16S versions in the same insect. Every individual possessed sequences that cluster into the same four or five groups (Fig. 5 and Table S1), suggesting that the fragmentation process happened early in the *Magicicada* lineage or before it diverged from its last common ancestor, and that multiple *Hodgkinia* lineages have been passed down from the *Magicicada* common ancestor. It is noteworthy that neither *M. tredecim*, whose symbiont genomes

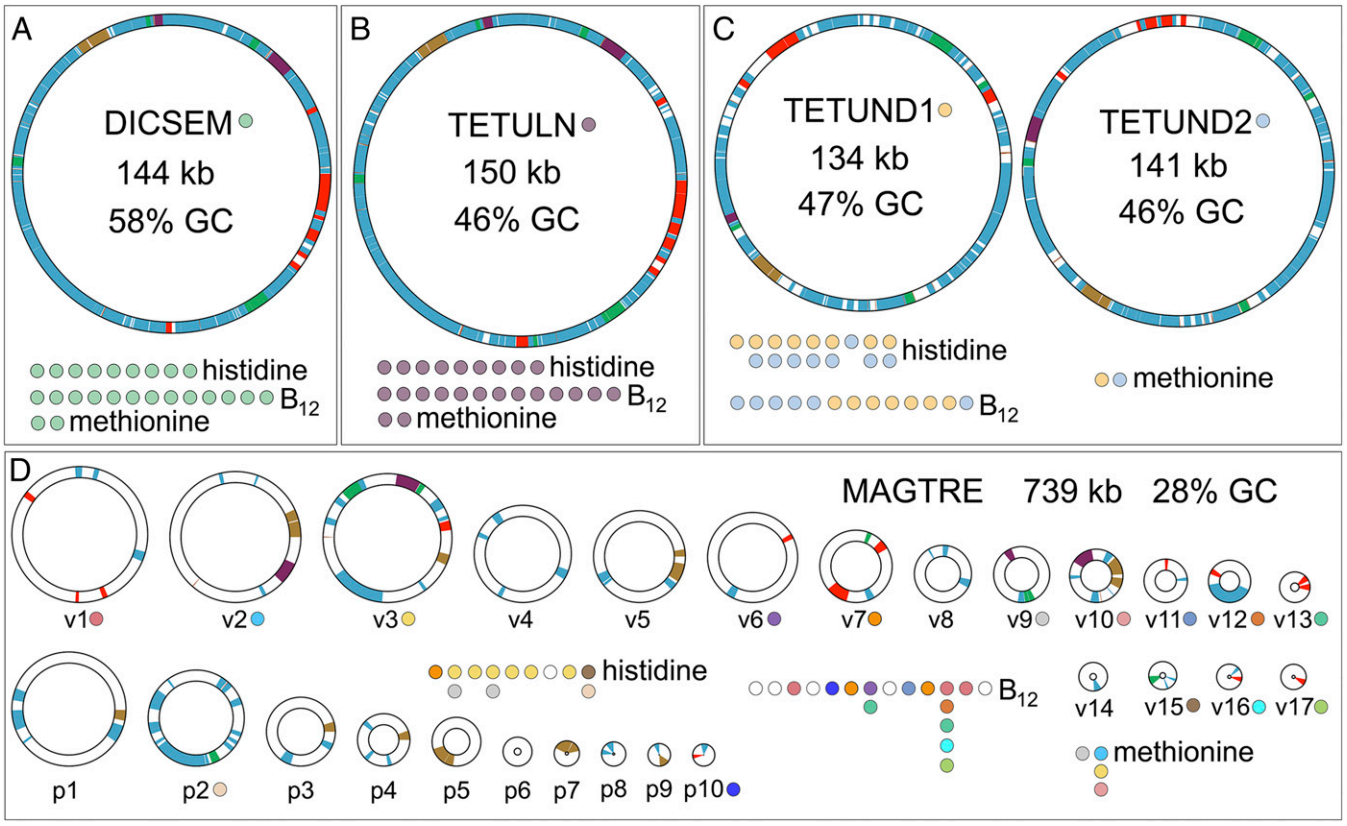

Fig. 2. Schematic representations of sequenced *Hodgkinia* genomes from (*A*) DICSEM, (*B*) TETULN, (*C*) TETUND, and (*D*) MAGTRE drawn to scale. On the genome diagrams, genes involved in methionine biosynthesis are shown in purple, vitamin B_{12} biosynthesis in red, histidine biosynthesis in green, the 16S and 23S rRNAs and tRNAs are shown in brown, and all other genes are shown in light blue. Regions of genomes encoding pseudogenes or other apparently nonfunctional DNA are shown in white. In each box, the gene homologs present on each genome from the methionine, B_{12}, and histidine pathways are shown as colored circles. The *Hodgkinia* genomes from DICSEM (green dots) and TETULN (purple dots) encode all of these genes on one genome, TETUND on two (blue and orange dots), and MAGTRE encodes these genes distributed over several circles (15 dots of different colors). In *D*, v1–v17 are the verified circles and p1–p10 are the putative circles from Table 1.

were sequenced, nor *M. tredecim* from a different locality (Table S1), whose *Hodgkinia* 16S rRNA genes were cloned and sequenced, seemed to contain 16S versions in group 3, indicating that some *Hodgkinia* lineages have been differentially lost as *Magicicada* diversified. Previous molecular clock calculations place the common ancestor of *Magicicada* at ∼3.8 Mya (43).

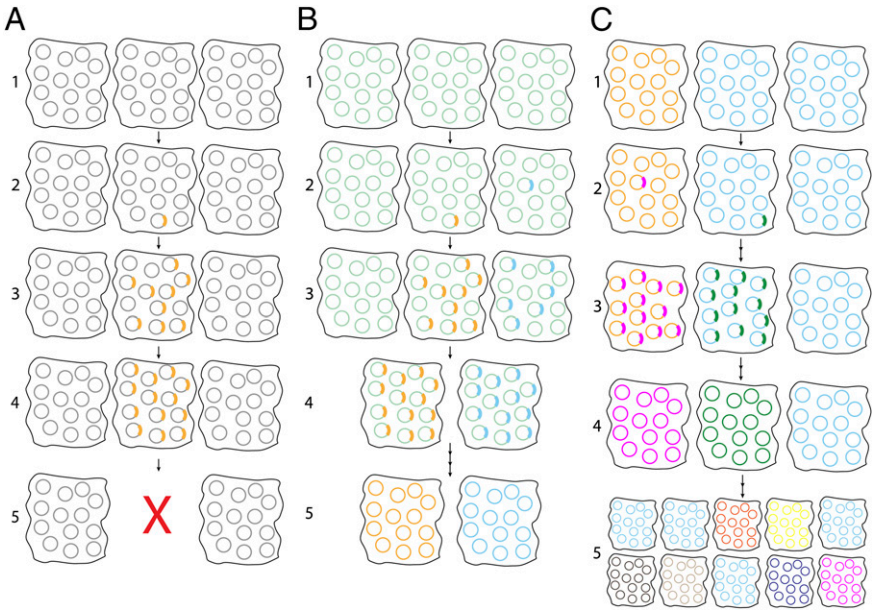

Fig. 3. Models for lineage splitting in different endosymbionts. In these models, all symbionts are polyploid with several genome copies per cell, and all experience a population bottleneck when they are distributed to eggs (40). (*A*) *Sulcia* (gray circles in *A1*) has a low mutation rate, and inactivating mutations (orange mark) arise infrequently (*A2*). These inactivating mutations may drift to high frequency (*A3* and *A4*), but the cell lineages carrying these mutations are eventually purged by selection (red X in *A5*), keeping the *Sulcia* lineage coherent. (*B*) The model for lineage splitting (adapted from ref. 40) in *Hodgkinia* TETUND. *Hodgkinia* (green circles) has a high mutation rate, and complementary inactivating mutations (orange and blue marks) can arise in the same insect (*B2*) and rise to high frequency through drift (*B3*). If the ancestral genotype is eliminated (*B4*), the new *Hodgkinia* genotypes can evolve further interdependencies through reciprocal gene loss (*B5*). Reprinted from ref. 40, with permission from Elsevier; www.sciencedirect.com/science/journal/00928674. In MAGTRE (*C*), two new lineages (*C1* orange and blue circles) could continue splitting although the same process hypothesized for TETUND (*C2*–*C4*), leading to several related interdependent genotypes, each with progressively smaller genomes (*C5*).

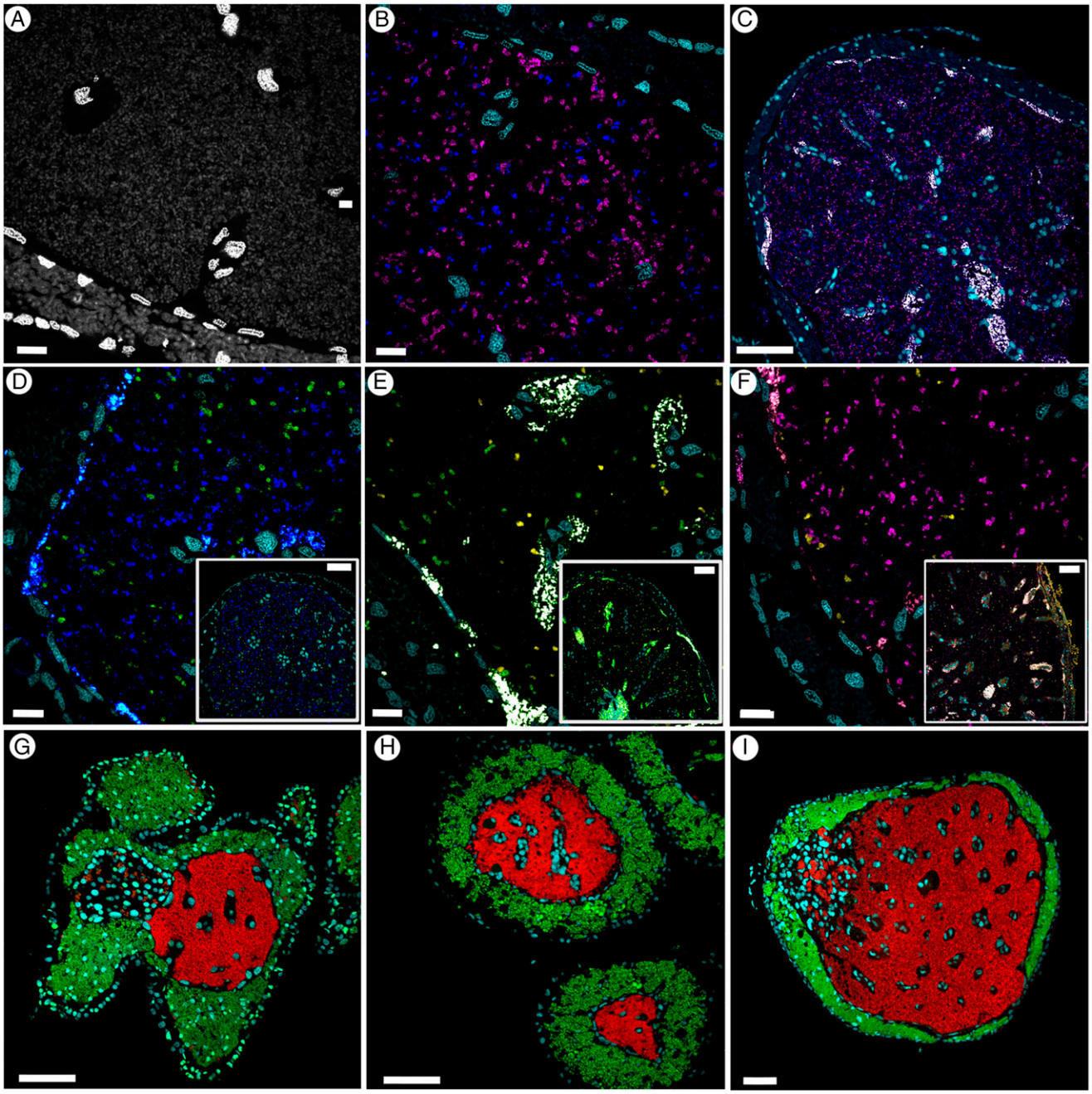

Fig. 4. FISH microscopy shows some *Hodgkinia* circles remain cytologically distinct in *M. tredecim*. (Scale bars in the main panels of *A*, *B*, and *D–F* are 20 μm; all others, including the *Insets*, are 100 μm.) (*A*) A section of bacteriome tissue from MAGTRE stained only with the general DNA Hoechst dye showing that all *Hodgkinia* cells (upper right four-fifths of image) and *Sulcia* cells (band across the lower left one-fifth of image) contain DNA. Insect nuclei are large bright punctate spots. In *B–I*, insect nuclei are teal. *B–F* show sections of MAGTRE bacteriome tissue stained with DNA probes targeting specific circles; in no case do the signals overlap. (*B*) Cells containing two high-coverage circles MAGTRE001 (blue) and MAGTRE005 (purple) are both present at high numbers. (*C*) A lower-resolution image of the tissue shown in *B*. In *D–F*, lower-resolution images of the tissue are shown in the insets. (*D*) Cells containing the high-coverage circle MAGTRE001 (blue) are more abundant than cells containing the low-coverage circle MAGTRE006 (green). (*E*) Cells containing the two low-coverage circles MAGTRE006 (green) and MAGTRE012 (orange) are both present at low numbers. (*F*) Cells containing the high-coverage circle MAGTRE005 (purple) are more abundant than cells containing the low-coverage circle MAGTRE012 (orange). *G–I* show whole bacteriome tissue sections stained for *Hodgkinia* (red) and *Sulcia* (green) small subunit rRNAs. As the number of *Hodgkinia* lineages increases from one in DICSEM (*G*), to two in TETUND (*H*), to several in MAGTRE (*I*), the relative amount of tissue volume devoted to *Hodgkinia* seems to also increase.

Why Does *Hodgkinia* Fracture into Many Lineages Whereas *Sulcia* Remains Cohesive? In all reported cicada species, the bacterial symbionts *Sulcia* and *Hodgkinia* are contained within different insect cells but have been restricted to cicada tissues (Fig. 4) for tens of millions of years. Therefore, *Sulcia* and *Hodgkinia* should be subject to the same forces imposed by their extensive gene loss and living conditions: that is, the effects of strict asexuality, intracellularity, host dependence, and transovarial transmission should be the same for both endosymbionts.

We suggest that the structural differences between the *Sulcia* and *Hodgkinia* genomes may result from differences in their mutation rates. (Although the mutation rate itself has not been measured in

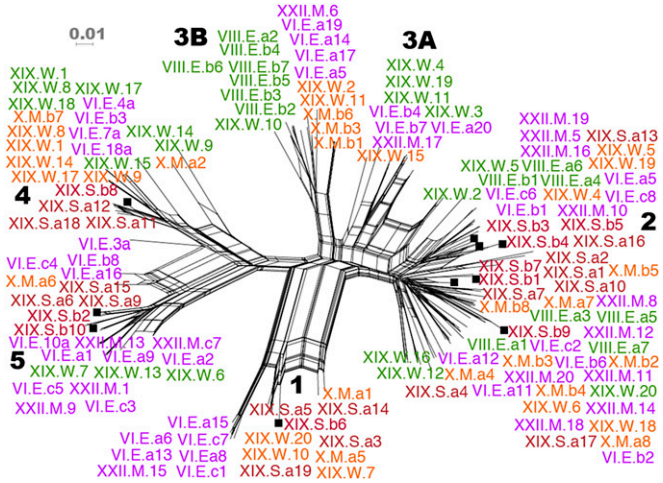

Fig. 5. All species in *Magicicada* contain many *Hodgkinia* lineages. Neighbor-net Splits Tree of *Magicicada Hodgkinia* 16S sequences suggest at least five lineages in all species. Orange font = *Hodgkinia* from Decim species group, except *M. tredecim* which are colored red. Black squares identify *M. tredecim Hodgkinia* (XIX.S.b) lineages from whole-genome sequencing; green = *Hodgkinia* from Cassini species group; purple = *Hodgkinia* from Decula species group. Each individual cicada is uniquely labeled: Roman numerals identify *Magicicada* year-class (brood); brood numbers in the range I–XVII = 17-y cicadas, XVIII–XXX = 13-y cicadas; E is eastern, M is Midwestern, W is western, S is southern phylogeographic clades as defined by Sota et al. (43); multiple individuals from the same population or brood are labeled with lowercase "a," "b," or "c"; specimen clones are sequential Arabic numbers. See Table S1 for collecting localities and clone lineage membership summary.

Sulcia or *Hodgkinia*, here we use the relative DNA substitution rates as a proxy.) *Sulcia* has been noted to have a very low DNA substitution rate in various insects, usually with its partner coprimary symbiont showing a more rapid rate of sequence evolution (2, 44, 45). For example, in sharpshooters, *Sulcia* has a five-times slower rate of DNA substitution than its partner symbiont *Baumannia cicadellinicola* (45). Thus, symbiont pairs that are present in the same host can have different rates of sequence evolution, perhaps because of mechanical differences in their DNA replication machinery (44).

The difference in DNA substitution rate between partner endosymbionts appears to be even more dramatic in the case of *Sulcia* and *Hodgkinia*. By comparing the average rates of synonymous site substitutions (d_S) in *Sulcia* and *Hodgkinia* homologs in different cicada species, we estimate that the DNA substitution rate is one to two orders-of-magnitude higher in *Hodgkinia* than in *Sulcia* (Table S2). The model we propose for the lineage-splitting events in TETUND and MAGTRE require at least two complementary and inactivating mutations to arise in different *Hodgkinia* cells (Fig. 3). If the mutation rate is much higher in *Hodgkinia* compared with *Sulcia*, then the odds of acquiring two mutations in the population for a given number of genome replication cycles is higher in *Hodgkinia*. *Sulcia* will still encounter inactivating mutations, and these may even rise to high frequency, but cell lineages that accumulate high levels of these deleterious genotypes will eventually be purged by selection (Fig. 3*A*). It is also possible that there are cell biological reasons why *Sulcia* and *Hodgkinia* are different. In particular, the patterns of genome evolution in *Hodgkinia* suggest that its cellular boundary is porous to most molecules except genomes (40), but this may not be true in *Sulcia.* In this case, it would not be possible for inactivating mutations in two different *Sulcia* cells to interact, and thus cell lineages carrying inactivating mutations would not be masked from selection by other lineages with active gene copies and would eventually be purged by host-level purifying selection.

Why Do Symbionts Fracture into Many Lineages in Cicadas, but Not in Other Insects? Aside from *Hodgkinia*, many other endosymbionts with tiny genomes have very high substitution rates (2). Why have the lineage-splitting events we observe in *Hodgkinia* not occurred in other symbionts in other insects, and why are they found in only some lineages of cicadas? We suggest that it is related to the very long and variable life cycles of cicadas. Although some exceptional insects have multiyear diapause stages that can last more than 25 y (e.g., ref. 46), the vast majority of sap-feeding insects have life cycles of 1 y or less (47–50). With known life cycles ranging from 2 to 19 y, cicadas are therefore among the longest-lived nondiapausing insects (51, 52). Most cicada species for which we have data have life cycles of 2 to 5 y, with the synchronized 13- and 17-y life cycles of periodical cicadas in the genus *Magicicada* at the long end of the spectrum (Table S3).

We hypothesize that the number of splitting events experienced by a *Hodgkinia* lineage is proportional to the life-cycle length of the cicada in which it resides. This could be the result of two factors. The first is the inferred high mutation rate in *Hodgkinia*; it could simply be that the longer an insect lives, the more genome replication cycles *Hodgkinia* undergoes and thus the likelihood of accumulating inactivating mutations is higher. The second factor relates the amount of time a cicada species exists in states of lowered metabolism, such as winter diapause (53, 54), or the waiting period (51) between when it has reached the critical fifth-instar weight and when it emerges above ground. Because *Sulcia* and *Hodgkinia* provide essential amino acids to their host cicada (6), we assume that host-level selection will test the quality of symbiont genotypes most intensely when protein synthesis is at its maximum; that is, when the insect is putting on mass during growth. Therefore, if there are periods during the cicada lifecycle

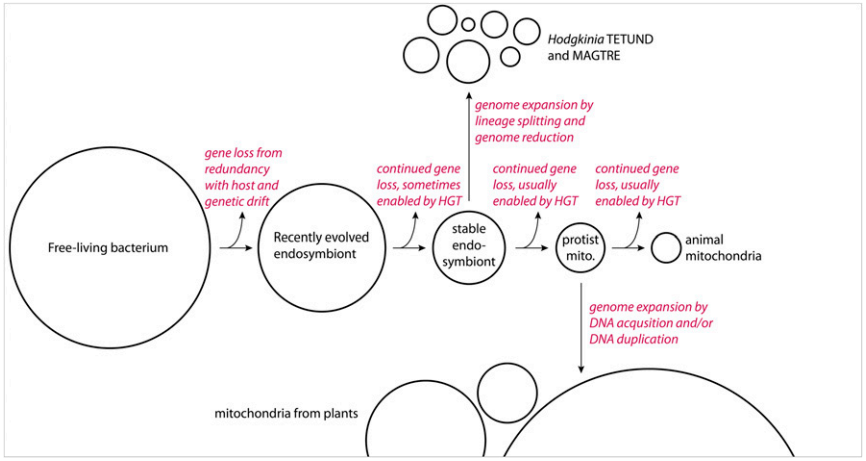

Fig. 6. Genome reduction and expansion in organelles and endosymbionts. Genomes are schematized as black circles. The reductive process experienced by bacteria as they transition from free-living lifestyles to strict endosymbiotic lifestyles is shown across the middle of the figure from left to right. Both the expanded *Hodgkinia* and plant mitochondria genomes originated from highly reduced ancestral genomes.

where the symbionts are undergoing genome replication (to be maintained and passed to the next generation) but when the cicada is not putting on mass, then it may be possible for less-fit symbiont genotypes to accumulate because their symbiotic quality would not be vigorously tested by host-level selection (40).

Two Definitions of the *Hodgkinia* "Genome." Is the *Hodgkinia* genome the sum of what is found in a single cicada, or does each cellular lineage possess its own genome? From the perspective of the *Hodgkinia* lineage, it is difficult to argue that dramatic and permanent splits have not occurred; a single *Hodgkinia* species has irreversibly split into two or more new species, each with a smaller genome encoding fewer genes. From the perspective of the insect host, the number of *Hodgkinia* lineages has multiplied, but so has the collective genome size because the host now needs most if not all of the circles encoded by these new lineages to perform the tasks originally performed by one. Genome expansion describes the outcome experienced by the host; lineage-splitting and genome reduction describes the processes happening to *Hodgkinia*. However, at some point one has to choose between the two to represent *Hodgkinia* in figures such as Fig. 1. We have chosen the genome size resulting from sum of all lineages because the genes contained in the entire *Hodgkinia* genome complex are likely the important unit of selection for the maintenence of the symbiosis.

Differences and Similarities in Endosymbiont and Organelle Genome Evolution. One important difference between mitochondria and *Hodgkinia* is the physical location of the genomes. The *Hodgkinia* genomes from TETUND (40) and at least some of the circles from MAGTRE (Fig. 4) appear to remain cytologically distinct, but this is likely not true in mitochondria because of the frequent fission and fusion events they undergo (55). Indeed, the frequency of mitochondrial fusion is the explanation proposed for the massive levels of foreign DNA acquisition seen in mitochondrial genomes from the plant genus *Amborella* (23) (Fig. 1). Thus, even when mitochondrial genomes fragment into several chromosomes, those chromosomes stay distributed throughout a cell's mitochondria because of frequent organelle fusion. In contrast, when a *Hodgkinia* lineage fragments, each new genome seems to stay sequestered into discrete cells and mixing does not occur.

Despite these cell biological differences, decades of work on organelle and endosymbiont genomes has shown that genome reduction is a strong unifying theme of intracellular symbioses (Fig. 6). Although many organelle genomes remain small and gene dense, others have undergone secondary genome expansions through DNA proliferation or acquisition that make the genome larger but add little or no coding capacity (23, 24, 30). Similarly, most insect endosymbiont genomes are small and gene dense, but here we have shown that the *Hodgkinia* genome complex has grown in size by almost an order-of-magnitude and has drastically reduced its coding density, but through a different process involving lineage splitting and reciprocal gene inactivation. These examples of secondary genome expansion have three important similarities. The first is that they have all lead to the accumulation of large amounts of "junk" DNA, inspiring arguments that these genome expansions are the result of nonadaptive evolution (23, 24, 40, 56, 57). The second is that mutation rate seems to be an important correlate in the structure and stability of organelle (24) and endosymbiont genomes. The third is that they both have evolved in the context of absolute codependency with their hosts. A eukaryotic cell is nothing without its mitochondria, just as an insect that only eats plant sap is nothing without its endosymbiotic bacteria. It is likely that strong selection on the host to maintain the symbiosis provides a fertile ground for nonadaptive processes observed in organelles and endosymbionts. If conditions arise whereby an organelle acquires several genome's worth of foreign DNA, such as in *Amborella* (23), or if an insect host is not able to stop an endosymbiont splitting its genome into tens or hundreds of discrete cells, the host—and therefore the entire symbiosis—has no choice but to cope with the changes or die.

Materials and Methods

Additional details for the genome sequencing, annotation, and 16S PCR and sequencing can be found in *SI Materials and Methods*.

Genome Sequencing. Total DNA was purified from dissected bacteriome tissue from 12 ethanol-preserved cicadas, wild-caught in 2011 from King William County, Virginia, using the Qiagen DNeasy Blood and Tissue kit. DNA libraries from individual cicadas were separately barcoded for Illumina short-insert sequencing using NEXTflex adapters and protocols (Bio Scientific). Pooled DNA from the same individuals was used to generate the Illumina Nextera large-insert and PacBio RS II DNA libraries using standard protocols from the manufacturer.

Genome Assembly. Adapter sequences were trimmed with trimmomatic (parameters: SLIDINGWINDOW:10:15 LEADING:3 TRAILING:3 MINLEN:60) (58) and quality-filtered using FASTX Toolkit v0.0.13. High-quality, paired reads were assembled using SPAdes v3.1.1 (60), with kmer sizes of 91 and 95. Uncorrected PacBio reads were used to scaffold with SSPACE-LONGREADS v1.1 (61). Putatively circular scaffolds were confirmed with manual inspection of mate-pair read mapping and Sanger sequencing of PCR products. Internal gaps in the scaffolds were closed using PacBio reads and custom Python scripts.

Microscopy. Genome- and ribosome-targeted fluorescence in situ hybridization microscopy was performed as described previously (40) on bacteriome tissue from a single female cicada from Brood XXII (C.S. Laboratory specimen no. 14.LA.EB.WWP.01) using the DNA sequences listed in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Betsey Pitts and Phil Stewart (Montana State University Center for Biofilm Engineering and the University of Montana Molecular Histology Fluorescence Imaging Core, supported by National Institutes of Health National Institute of General Medical Sciences Grant P20RR017670) for imaging assistance, and Illumina for generating the Nextera data used in this study. Genomics at the University of Montana was supported by a grant from the M. J. Murdock Charitable Trust. Cicada systematic work and 16S rRNA sequencing was supported by National Science Foundation Grant DEB-0955849 and DEB-0529679 (to C.S.). J.P.M. was supported by National Science Foundation Grant IOS-1256680, NSF-EPSCoR Award NSF-IIA-1443108 to the Montana Institute on Ecosystems, and National Aeronautics and Space Administration Astrobiology Institute Award NNA15BB04A.

1. 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(6800): 81–86.
2. McCutcheon JP, Moran NA (2012) Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* 10(1):13–26.
3. Tamas I, et al. (2002) 50 million years of genomic stasis in endosymbiotic bacteria. *Science* 296(5577):2376–2379.
4. Gil R, et al. (2003) The genome sequence of *Blochmannia floridanus*: Comparative analysis of reduced genomes. *Proc Natl Acad Sci USA* 100(16):9388–9393.
5. Degnan PH, Lazarus AB, Wernegreen JJ (2005) Genome sequence of *Blochmannia pennsylvanicus* indicates parallel evolutionary trends among bacterial mutualists of insects. *Genome Res* 15(8):1023–1033.
6. McCutcheon JP, McDonald BR, Moran NA (2009) Convergent evolution of metabolic roles in bacterial co-symbionts of insects. *Proc Natl Acad Sci USA* 106(36): 15394–15399.
7. McCutcheon JP, Moran NA (2010) Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. *Genome Biol Evol* 2:708–718.
8. 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.
9. Moran NA, Tran P, Gerardo NM (2005) Symbiosis and insect diversification: An ancient symbiont of sap-feeding insects from the bacterial phylum Bacteroidetes. *Appl Environ Microbiol* 71(12):8802–8810.
10. Nakabachi A, et al. (2006) The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* 314(5797):267.
11. 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.
12. van Ham RC, et al. (2003) Reductive genome evolution in *Buchnera aphidicola*. *Proc Natl Acad Sci USA* 100(2):581–586.
13. Sabree ZL, Degnan PH, Moran NA (2010) Chromosome stability and gene loss in cockroach endosymbionts. *Appl Environ Microbiol* 76(12):4076–4079.
14. Rio RVM, et al. (2012) Insight into the transmission biology and species-specific functional capabilities of tsetse (Diptera: Glossinidae) obligate symbiont *Wigglesworthia*. *MBio* 3(1):e00240-11.

15. McCutcheon JP, von Dohlen CD (2011) An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Curr Biol* 21(16):1366–1372.
16. Sloan DB, Moran NA (2013) The evolution of genomic instability in the obligate endosymbionts of whiteflies. *Genome Biol Evol* 5(5):783–793.
17. Shao R, Kirkness EF, Barker SC (2009) The single mitochondrial chromosome typical of animals has evolved into 18 minichromosomes in the human body louse, *Pediculus humanus*. *Genome Res* 19(5):904–912.
18. Boore JL (1999) Animal mitochondrial genomes. *Nucleic Acids Res* 27(8):1767–1780.
19. Anderson S, et al. (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290(5806):457–465.
20. Ward BL, Anderson RS, Bendich AJ (1981) The mitochondrial genome is large and variable in a family of plants (Cucurbitaceae). *Cell* 25(3):793–803.
21. Quetier F, Vedel F (1977) Heterogeneous population of mitochondrial DNA molecules in higher plants. *Nature* 268(5618):365–368.
22. Gray MW (1982) Mitochondrial genome diversity and the evolution of mitochondrial DNA. *Can J Biochem* 60(3):157–171.
23. Rice DW, et al. (2013) Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. *Science* 342(6165):1468–1473.
24. Sloan DB, et al. (2012) Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biol* 10(1): e1001241.
25. Burger G, Gray MW, Lang BF (2003) Mitochondrial genomes: Anything goes. *Trends Genet* 19(12):709–716.
26. Vaidya AB, Arasu P (1987) Tandemly arranged gene clusters of malarial parasites that are highly conserved and transcribed. *Mol Biochem Parasitol* 22(2-3):249–257.
27. Burger G, Gray MW, Forget L, Lang BF (2013) Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biol Evol* 5(2):418–438.
28. Lang BF, et al. (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* 387(6632):493–497.
29. Liu Y, Medina R, Goffinet B (2014) 350 My of mitochondrial genome stasis in mosses, an early land plant lineage. *Mol Biol Evol* 31(10):2586–2591.
30. Smith DR, et al. (2013) Organelle genome complexity scales positively with organism size in volvocine green algae. *Mol Biol Evol* 30(4):793–797.
31. Chaw S-M, et al. (2008) The mitochondrial genome of the gymnosperm *Cycas taitungensis* contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. *Mol Biol Evol* 25(3):603–615.
32. Smith DR, Lee RW (2009) The mitochondrial and plastid genomes of *Volvox carteri*: Bloated molecules rich in repetitive DNA. *BMC Genomics* 10:132.
33. McCutcheon JP, Keeling PJ (2014) Endosymbiosis: Protein targeting further erodes the organelle/symbiont distinction. *Curr Biol* 24(14):R654–R655.
34. Adams KL, Palmer JD (2003) Evolution of mitochondrial gene content: Gene loss and transfer to the nucleus. *Mol Phylogenet Evol* 29(3):380–395.
35. 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.
36. 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.
37. Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* 5(2):123–135.
38. 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.
39. Nakabachi A, Ishida K, Hongoh Y, Ohkuma M, Miyagishima S-Y (2014) Aphid gene of bacterial origin encodes a protein transported to an obligate endosymbiont. *Curr Biol* 24(14):R640–R641.
40. 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.
41. McCutcheon JP, McDonald BR, Moran NA (2009) Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *PLoS Genet* 5(7): e1000565.
42. Van Leuven JT, McCutcheon JP (2012) An AT mutational bias in the tiny GC-rich endosymbiont genome of *Hodgkinia*. *Genome Biol Evol* 4(1):24–27.
43. Sota T, et al. (2013) Independent divergence of 13- and 17-y life cycles among three periodical cicada lineages. *Proc Natl Acad Sci USA* 110(17):6919–6924.
44. Bennett GM, McCutcheon JP, MacDonald BR, Romanovicz D, Moran NA (2014) Differential genome evolution between companion symbionts in an insect-bacterial symbiosis. *MBio* 5(5):e01697-14.
45. Takiya DM, Tran PL, Dietrich CH, Moran NA (2006) Co-cladogenesis spanning three phyla: Leafhoppers (Insecta: Hemiptera: Cicadellidae) and their dual bacterial symbionts. *Mol Ecol* 15(13):4175–4191.
46. Powell JA (2001) Longest insect dormancy: Yucca moth larvae (Lepidoptera: Prodoxidae) metamorphose after 20, 25, and 30 years in diapause. *Ann Entomol Soc Am* 94(5):677–680.
47. Denno RF, Roderick GK (1990) Population biology of planthoppers. *Annu Rev Entomol* 35:489–520.
48. Heliövaara K, Väisänen R, Simon C (1994) Evolutionary ecology of periodical insects. *Trends Ecol Evol* 9(12):475–480.
49. Nickel H, Remane R (2002) Check list of the planthoppers and leafhoppers of Germany, with notes on food plants, diet width, life cycles, geographic range and conservation status. *Beiträge zur Zikadenkunde* 5:27–64.
50. Hodkinson ID (2009) Life cycle variation and adaptation in jumping plant lice (Insecta: Hemiptera: Psylloidea): A global synthesis. *J Nat Hist* 43(1-2):65–179.
51. Williams KS, Simon C (1995) The ecology, behavior, and evolution of periodical cicadas. *Annu Rev Entomol* 40:269–295.
52. Karban R (1986) in *The Evolution of Insect Life Cycles*, eds Taylor F, Karban R (Springer, New York), pp 222–235.
53. Flannagan RD, et al. (1998) Diapause-specific gene expression in pupae of the flesh fly *Sarcophaga crassipalpis*. *Proc Natl Acad Sci USA* 95(10):5616–5620.
54. Hahn DA, Denlinger DL (2007) Meeting the energetic demands of insect diapause: Nutrient storage and utilization. *J Insect Physiol* 53(8):760–773.
55. Westermann B (2010) Mitochondrial fusion and fission in cell life and death. *Nat Rev Mol Cell Biol* 11(12):872–884.
56. Lynch M, Koskella B, Schaack S (2006) Mutation pressure and the evolution of organelle genomic architecture. *Science* 311(5768):1727–1730.
57. Boussau B, Brown JM, Fujita MK (2011) Nonadaptive evolution of mitochondrial genome size. *Evolution* 65(9):2706–2711.
58. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.
59. Eme L, Sharpe SC, Brown MW, Roger AJ (2014) On the age of eukaryotes: Evaluating evidence from fossils and molecular clocks. *Cold Spring Harb Perspect Biol* 6(8): a016139.
60. 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.
61. Boetzer M, Pirovano W (2014) SSPACE-LongRead: Scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinf* 15:211.