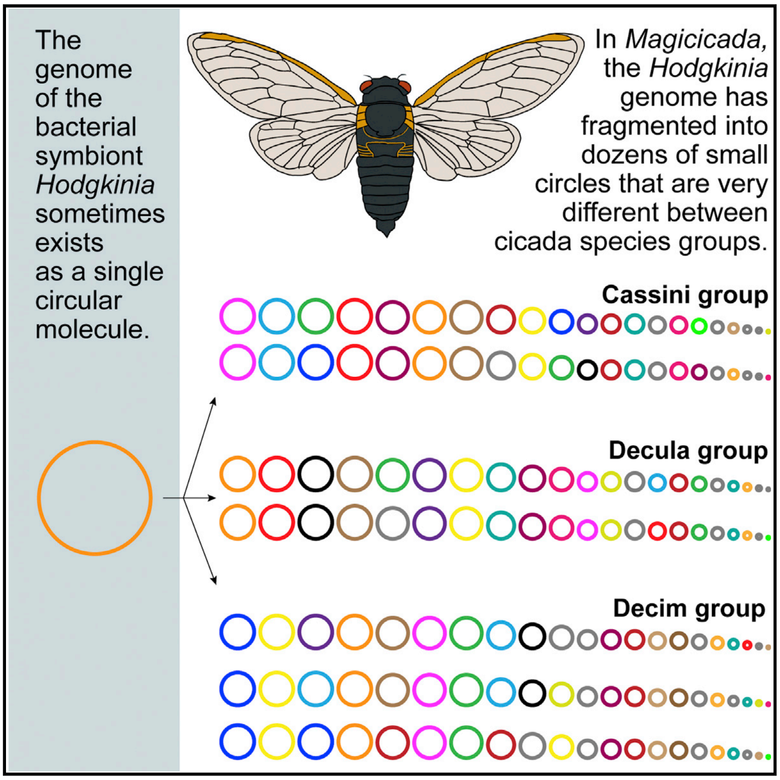
Current Biology

Report

Idiosyncratic Genome Degradation in a Bacterial Endosymbiont of Periodical Cicadas

Graphical Abstract


The genome of the bacterial symbiont *Hodgkinia* sometimes exists as a single circular molecule.
In *Magicicada,* the *Hodgkinia* genome has fragmented into dozens of small circles that are very different between cicada species groups.
Cassini group
Decula group
Decim group

Highlights

- The *Hodgkinia* genome in all *Magicicada* exist as complexes of $\geq$20 circular molecules
- Together, these circles contain most of the ancestral *Hodgkinia* gene set
- The gene dosage of *Hodgkinia* genes is wildly different in different cicada species
- The genomic complexity of *Hodgkinia* is most likely nonadaptive for the host cicada

Authors

Matthew A. Campbell, Piotr Łukasik, Chris Simon, John P. McCutcheon

Correspondence

john.mccutcheon@umontana.edu

In Brief

The stability of nutritional endosymbiont genomes reflects their importance to their hosts. Campbell et al. show that this stability has dramatically eroded in an endosymbiont of the 13- and 17-year periodical cicadas and that the outcome of this instability is wildly different in different cicadas.

Campbell et al., 2017, Current Biology *27*, 1–8
November 20, 2017 © 2017 Elsevier Ltd.
https://doi.org/10.1016/j.cub.2017.10.008

CellPress

Please cite this article in press as: Campbell et al., Idiosyncratic Genome Degradation in a Bacterial Endosymbiont of Periodical Cicadas, Current Biology (2017), https://doi.org/10.1016/j.cub.2017.10.008

Current Biology
Report

CellPress

Idiosyncratic Genome Degradation in a Bacterial Endosymbiont of Periodical Cicadas

Matthew A. Campbell,[1] Piotr Łukasik,[1] Chris Simon,[2] and John P. McCutcheon[1,3,*]
[1]Division of Biological Sciences, University of Montana, 32 Campus Drive, Missoula, MT 59812, USA
[2]Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N Eagleville Road Unit 3043, Storrs, CT 06269, USA
[3]Lead Contact
*Correspondence: john.mccutcheon@umontana.edu
https://doi.org/10.1016/j.cub.2017.10.008

SUMMARY

When a free-living bacterium transitions to a host-beneficial endosymbiotic lifestyle, it almost invariably loses a large fraction of its genome [1, 2]. The resulting small genomes often become stable in size, structure, and coding capacity [3–5], as exemplified by *Sulcia muelleri*, a nutritional endosymbiont of cicadas. *Sulcia*'s partner endosymbiont, *Hodgkinia cicadicola*, similarly remains co-linear in some cicadas diverged by millions of years [6, 7]. But in the long-lived periodical cicada *Magicicada tredecim*, the *Hodgkinia* genome has split into dozens of tiny, gene-sparse circles that sometimes reside in distinct *Hodgkinia* cells [8]. Previous data suggested that all other *Magicicada* species harbor complex *Hodgkinia* populations, but the timing, number of origins, and outcomes of the splitting process were unknown. Here, by sequencing *Hodgkinia* metagenomes from the remaining six *Magicicada* and two sister species, we show that each *Magicicada* species harbors *Hodgkinia* populations of at least 20 genomic circles. We find little synteny among the 256 *Hodgkinia* circles analyzed except between the most closely related cicada species. Gene phylogenies show multiple *Hodgkinia* lineages in the common ancestor of *Magicicada* and its closest known relatives but that most splitting has occurred within *Magicicada* and has given rise to highly variable *Hodgkinia* gene dosages among species. These data show that *Hodgkinia* genome degradation has proceeded down different paths in different *Magicicada* species and support a model of genomic degradation that is stochastic in outcome and nonadaptive for the host. These patterns mirror the genomic instability seen in some mitochondria.

RESULTS

Hodgkinia Is Complex in All *Magicicada* Species

Our new sequencing data confirm [8] that *Hodgkinia* comprises many distinct genomic circles in all species of *Magicicada* (Figure 1; Tables 1 and S1). We refer to individual circular-mapping *Hodgkinia* genomic contigs as "circles" because, though we know that some reside in distinct *Hodgkinia* cells [8], we currently do not know whether most of these molecules are chromosomes that share the same cell or genomes representing different cell types [8]. We refer to the total complement of *Hodgkinia* contigs assembled in a single species of cicada as that species' *Hodgkinia* genome complex (HGC). The smallest HGC is found in *M. tredecula* and consists of at least 152 contigs totaling 1.20 Mb of DNA, and the largest is from *M. neotredecim* and consists of 212 contigs totaling 1.68 Mb (Table 1). In each *Magicicada* species, we identified between 26 and 42 contigs with large-insert mate-pair data suggesting that they were circular DNA molecules. We were able to fully close these contigs into circular molecules in at least 20 instances in all cicada species (Figure 1; Table 1). Contigs with mate-pair data supporting their circularity were considered putative circles if they were not fully closed. The combination of confirmed and putative circular molecules comprise between 51.4% and 72.5% of the total DNA in each HGC (the remaining contigs lack end-joining data; Figure 1). Individual completed circles range in size from 0.69 kb to 70.5 kb, contain a maximum of 27 genes, a minimum of one gene (with a single exception, a circle encoding only a single pseudogene), and span as much as a 653-fold range of sequencing coverage (Table S1). Inclusion of contigs that did not assemble into circular molecules reveals an even higher range in coverages, with a minimum 2,500-fold span in each *Magicicada* species (Table 1).

In each *Magicicada* HGC, we identified between 135 and 145 of the 186 unique protein- and RNA-coding genes annotated as functional in *Hodgkinia* genomes from other cicada species [7, 9]. In all HGCs, several additional genes were identified as truncated fragments or obvious pseudogenes. Because of the very low coverage of some contigs and the extremely rapid rate of *Hodgkinia* sequence evolution [8], it is likely that at least some of the remaining genes are present but either were not fully assembled or are not recognizable due to their low sequence similarity to other annotated *Hodgkinia* genes.

We also sequenced *Hodgkinia* from two cicada species that are closely related to *Magicicada* [10, 11] (Figure 1; Table 1). The HGCs from Australian cicada species *Aleeta curvicosta* and *Tryella crassa* are similar to those in *Magicicada* but somewhat less complex. However, we generated less sequencing data for these species (~8.7 Gb for *A. curvicosta* and ~1.5 Gb for *T. crassa*, compared with an average of ~30.5 Gb per species of *Magicicada*), and so this relative simplicity may be due in part to sequencing effort.

Please cite this article in press as: Campbell et al., Idiosyncratic Genome Degradation in a Bacterial Endosymbiont of Periodical Cicadas, Current Biology (2017), https://doi.org/10.1016/j.cub.2017.10.008

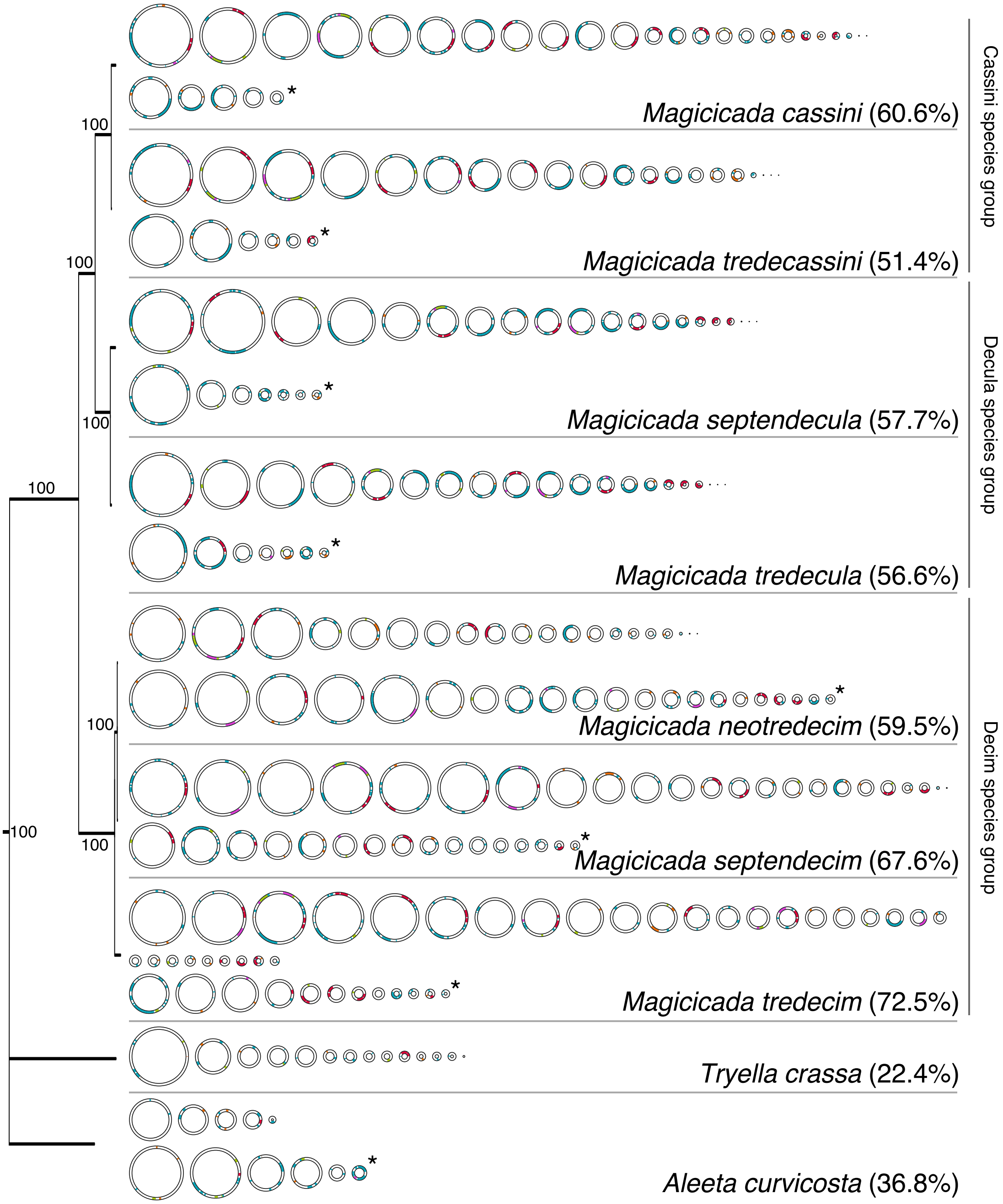

100
100
100
100
100
100
100
*
Magicicada cassini (60.6%)
Magicicada tredecassini (51.4%)
Magicicada septendecula (57.7%)
Magicicada tredecula (56.6%)
Magicicada neotredecim (59.5%)
Magicicada septendecim (67.6%)
Magicicada tredecim (72.5%)
Tryella crassa (22.4%)
Aleeta curvicosta (36.8%)
Cassini species group
Decula species group
Decim species group

Figure 1. *Hodgkinia* Genomic Complexity in All Study Species

Left: phylogeny of the cicada species used in this study, based on the 13 protein-coding genes and both large and small subunit ribosomal genes from each mitochondrial genome. The cicada *Diceroprocta semicincta* was used to root the tree but was not included in the figure. Bootstrap support values are shown on each resolved node. Right: diagrams representing the confirmed and putatively circular molecules of the *Hodgkinia* genome complex (HGC) in all study species. Rows with an asterisk at the end represent putative circular molecules. On each circle, red regions indicate rRNA genes, green regions indicate histidine synthesis genes, orange regions indicate cobalamin synthesis genes, purple regions indicate methionine synthesis genes, blue regions indicate all other genes, and white space represents noncoding DNA. Values in parentheses indicate the proportion of total *Hodgkinia* DNA from each cicada species represented by circular molecules. The three cicada species groups are shown vertically next to the species labels. See also Figures S1 and S2 and Tables S1 and S2.

The Origin of *Hodgkinia* Lineage Splitting Predates the Diversification of the Genus *Magicicada*

To determine whether *Hodgkinia* lineage splitting started in the genus *Magicicada* or predated its origin, we reconstructed phylogenetic trees for 126 *Hodgkinia* genes with at least three copies present in all *Magicicada* HGCs (Table S2). For 111 of these genes, all *Magicicada* gene copies form a single, well-supported clade, suggesting that splitting happened after

Table 1. Summary Statistics for All *Hodgkinia* Genome Complexes Described in This Work

Species	Species Abbreviation	Number of Contigs	Total HGC Size (Mb)	Total Number of Circles	Cumulative Size of Circles (Mb)	Unique Genes	Total Genes	Fold Coverage Difference
M. cassini	MAGCAS	117	1.27	29	0.77	142	306	6,376
M. tredecassini	MAGTCS	198	1.42	26	0.73	145	316	5,494
M. septendecula	MAGSDC	117	1.21	27	0.71	139	297	4,827
M. tredecula	MAGTDC	152	1.20	27	0.68	140	317	3,189
M. neotredecim	MAGNEO	212	1.68	41	1.00	138	332	3,379
M. septendecim	MAGSEP	163	1.63	39	1.11	136	313	5,723
M. tredecim	MAGTRE	118	1.57	42	1.11	135	300	2,500
T. crassa	TRYCRA	106	1.16	14	0.26	135	200	947
A. curvicosta	ALECUR	138	0.95	11	0.35	136	198	830

Total Hodgkinia genome complex (HGC) size is a sum of all *Hodgkinia* contigs, whether or not they are closed into circular molecules. The number of unique genes (protein coding, rRNA, and tRNA) found in other *Hodgkinia* genomes range from 168–183. See also Table S1.

the divergence of *Magicicada* and *Tryella*/*Aleeta* (Figure S1A). Six trees show two or three well-supported clades that group *Magicicada* genes with at least one copy from *Tryella* and/or *Aleeta* (for example, see Figures S1B–S1D), consistent with splitting that occurred in the common ancestor of all three cicada genera. Both patterns are possible because not all redundant genes from split lineages are retained in the new lineages [7]. Phylogenies for nine genes were difficult to interpret. Overall, these patterns show that at least some lineage splitting in *Hodgkinia* began before *Magicicada*, *Tryella*, and *Aleeta* diverged from one another. We estimate that the last common ancestor of these genera had a minimum of three *Hodgkinia* lineages (Figures S1B–S1D).

Hodgkinia Lineage Splitting Is Ongoing in *Magicicada* Species Groups

Having found evidence that *Hodgkinia* splitting had started prior to the divergence of *Magicicada* from its common ancestor with *Tryella* and *Aleeta*, we tried to assess whether most circular molecules were formed prior to the diversification of *Magicicada* and were conserved throughout the genus or whether lineage splitting is a process that has been ongoing in *Magicicada*. We find that phylogenies for 13 *Hodgkinia* genes show multiple (up to five) well-supported clades with representatives of all three *Magicicada* species groups (Figure S2; Table S2), consistent with splitting that occurred in the common ancestor of *Magicicada*. Using these phylogenies, we estimate that a minimum of five distinct *Hodgkinia* lineages existed in the last common ancestor of *Magicicada* (Figure S2).

Together, these phylogenetic data suggest that most of the splitting shown in Figure 1 happened after *Magicicada* started to diversify. If this is true, we expect that the similarity of HGCs should diminish as a function of cicada phylogenetic distance. In comparing extant circular molecules between cicada species groups, we find few clearly homologous circles with identical gene sets conserved in all *Magicicada* species. Because comparative genomic methods are generally based on sequence similarity and synteny comparisons and we found little obvious synteny to compare, we developed a metric based on the Jaccard index [12] to quantify the similarity in gene content of the finished circles between cicada species. We call this metric the circle similarity index (CSI; Figure 2). We calculate the CSI as follows, for hypothetical circular molecules A and B:

$$\mathbf{CSI} = \frac{Genes\ in\ A \cap Genes\ in\ B}{Genes\ in\ A \cup Genes\ in\ B} \times \frac{Length(in\ bp)of\ smaller\ of\ A\ and\ B}{Length(in\ bp)of\ larger\ of\ A\ and\ B} \qquad \text{(Equation 1)}$$

In brief, a finished circular molecule of one cicada species is compared to a circular molecule of another cicada species. We calculate the Jaccard index of the two gene sets (the intersection of gene sets divided by the union, the left half of Equation 1) and multiply that by the ratio of the length of the smaller circle divided by the length of the larger one (right half of Equation 1). We calculate this pairwise value for all circles of a species. The pair with the highest CSI score was kept for each circle, and we report the average CSI score between the pair of cicada species. We then repeat this for all pairwise comparisons of cicada species. A CSI value of one indicates that the two circles have the same functional genes and are the same length, whereas a value of zero indicates that they share no common genes. Because the circles have on average very low coding densities and have apparently undergone rearrangements in some cases (Figure 2A), this metric does not take gene co-linearity into account. The CSI is not (necessarily) a true measure of homology since it does not distinguish between conservation of an ancestral circle and convergent evolution to a similar state. Rather, it is a rough metric to score the overall similarity of HGCs between cicada species in the absence of much calculable similarity (Figure 2B).

We find a strong phylogenetic signal in CSI scores, where HGCs between species pairs (*M. cassini*-*M. tredecassini*, *M. septendecula*-*M. tredecula*, and *M. septendecim*-*M. neotredecim*) are highly similar to one another (0.80 CSI on average; Figure 2C). This is expected given that each of these species pairs are estimated to have diverged from each other less than 50 thousand years ago each [10]. The CSI scores degrade quickly with increased phylogenetic distance (Figure 2C), dropping to 0.29 in species diverged ∼4 million years ago [10]. This lack of similarity is remarkable given that

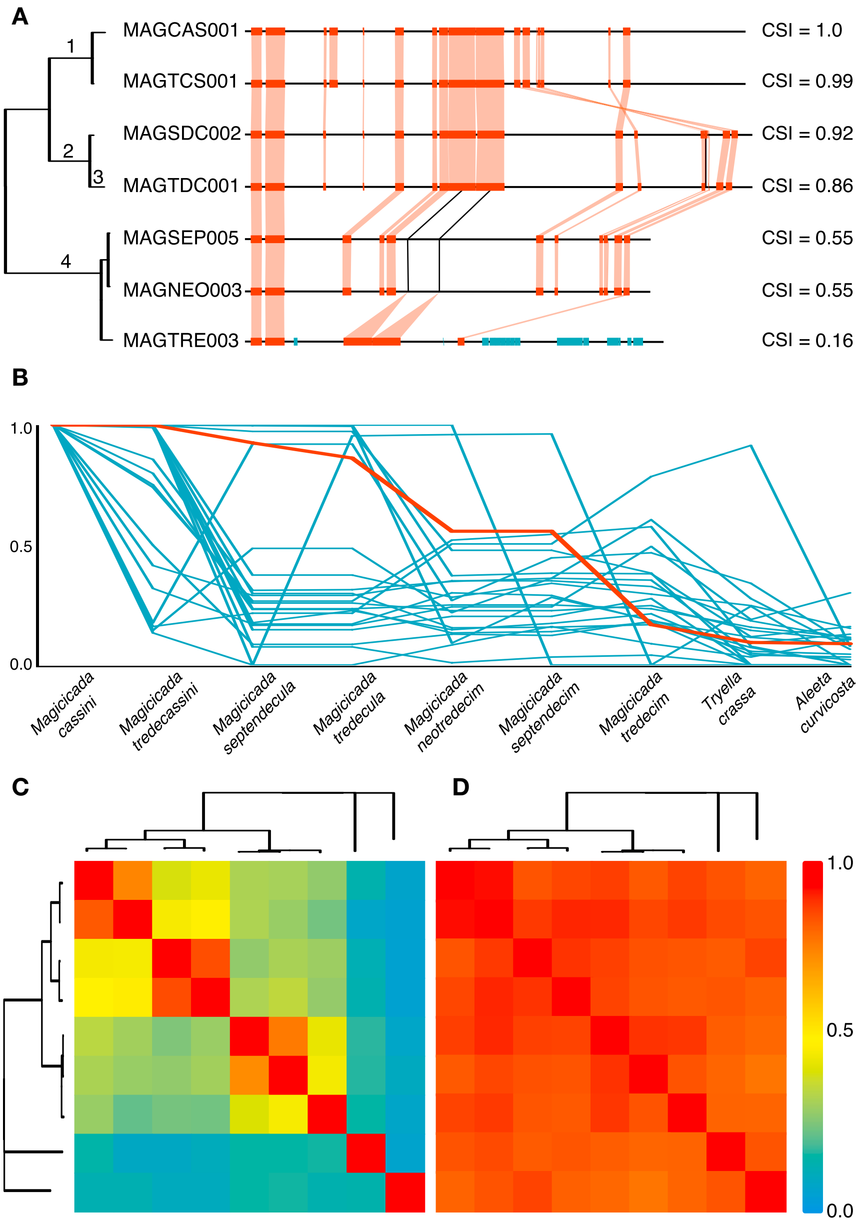


Figure 2. CSI Scores for Individual Circles and HGCs

(A) Illustration of conservation between circular molecules. Shown is the reference circle MAGCAS001 and the circle most similar to it from all other *Magicicada* species (abbreviations are taken from Table 1). Horizontal black lines represent the genome backbone, and orange boxes are genes shared between a circular molecule and MAGCAS001. Blue bars represent genes present in a given circular molecule, but not present on MAGCAS001. Shaded vertical lines show gene homologs present on different circles, and black lines connect putative homologs over gaps in some genomes. Circle similarity index (CSI) scores between MAGCAS001 and all other circular molecules are shown on the right. Numbers on the phylogenies represent inferred mutational events on the respective lineage: genome rearrangement (1), individual gene loss events (2 and 3), and loss of five genes (4). The exact branch on which (4) occurred is ambiguous. Three contigs from *M. tredecim* seem to be homologous to the reference circle when joined together, but we could not close them to a single circle so they were not included in the CSI analysis.

(B) Distribution of all CSI scores for *M. cassini*. Shown on the x axis are the species to which *M. cassini* was compared; the y axis shows the CSI score. The bold orange line represents the circles shown in (A) and shows that these circles are among the most conserved in all *M. cassini* comparisons.

(C) Heatmap showing pairwise average CSI scores between all species. Pairwise comparisons between *M. tredecim-M. neotredecim* and *M. tredecim-M. septendecim* (500 thousand years diverged [10]), *M. -cassini* species with *M. -decula* species (2.5 million years diverged [10]), and *M. -cassini* and *M. -decula* species with *M. -decim* species (4 million years diverged [10]) give average CSI scores of 0.43, 0.46, and 0.29, respectively.

(D) Heatmap showing pairwise average Jaccard index of the whole *Hodgkinia* gene set in each species.

In both (B) and (C), a score of one indicates that the two species are identical, and zero indicates that they share no genes in common. The trees in (A), (C), and (D) are taken from Figure 1.

the CSI between the single *Hodgkinia* genomes of *Diceroprocta semicincta* and *Tettigades ulnaria*, which diverged more than 60 million years ago [13–16], is 0.88.

Our combined phylogenetic and CSI analyses suggest that splitting began in the ancestor of *Magicicada*, *Tryella*, and *Aleeta* (into at least three circles) and continued somewhat in the ancestor of all *Magicicada* (into at least five circles) but that splitting accelerated (into at least 20 circles) after *Magicicada* began diversifying.

Hodgkinia's Overall Function Mostly Remains Intact

The long-term stability of endosymbiont genomes is often attributed to the importance of their function to host survival [3, 17, 18]. Because *Hodgkinia* is clearly experiencing dramatic genomic instability, we wanted to test whether the complete ancestral *Hodgkinia* gene set was retained in HGCs in different *Magicicada* species. To directly compare gene complements between *Hodgkinia* HGCs and to be consistent with the CSI, we calculated the Jaccard index of each gene set for all pairwise comparisons of all *Magicicada* species. Similar to the CSI, a score of 1 would indicate that two cicada species have identical *Hodgkinia* gene sets, and a score of 0 would indicate that no genes are shared. We find that HGC gene sets within closely related species pairs are very similar (0.90 on average; Figure 2D). Pairwise comparisons between *M. tredecim-M. neotredecim* and *M. tredecim-M. septendecim* (0.86), *M. -cassini* species with *M. -decula* species (0.87), and *M. -cassini* and *M. -decula* species with *M. -decim* species (0.86) also remain very similar, in contrast

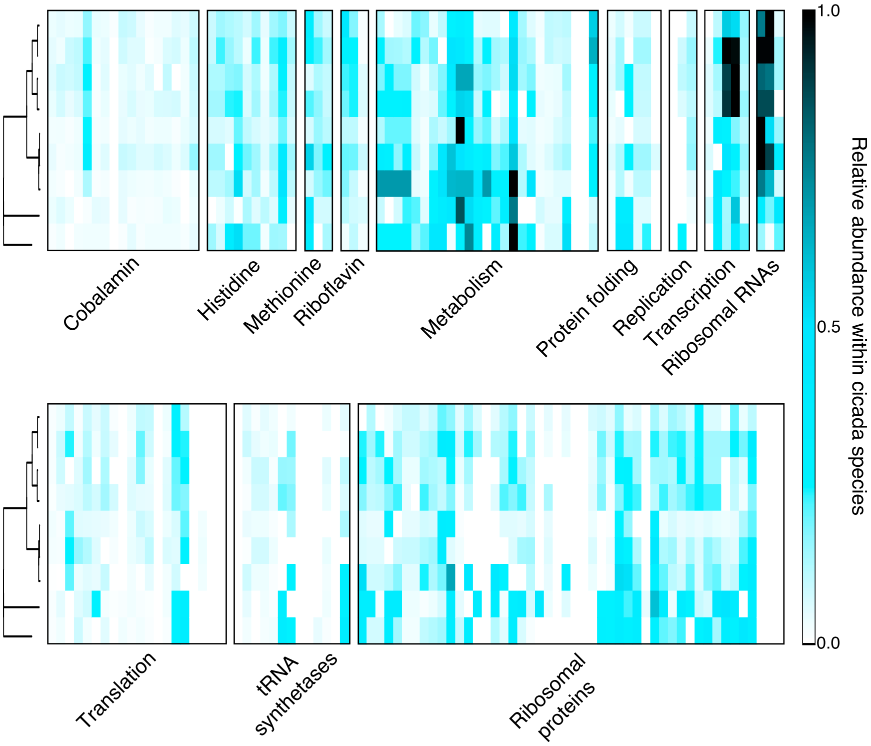


Figure 3. Relative Gene Abundance in All Study Species

Heatmaps showing the relative abundance of each gene in each species, ordered by gene category. A value of one (black) indicates the most abundant gene in that species, and zero (white) indicates that the gene is absent in that species. Columns that are completely white represent genes that were not annotated in any species and so have been lost, are present on broken contigs, or are present on contigs that did not otherwise assemble in our experiments. Trees are taken from Figure 1. See also Figure S3.

to the CSI scores calculated for these comparisons (compare Figure 2C to Figure 2D). These data show that although the patterns of *Hodgkinia* genome fragmentation are different in divergent *Magicicada* species, the overall set of retained genes is similar. For a sense of scale, the same analysis for *Hodgkinia* from cicadas diverged for dozens of millions of years [13–16], such as *Magicicada* and *D. semicincta*, *Magicicada* and *T. ulnaria*, and *D. semicincta* and *T. ulnaria* gives values of 0.82, 0.77, and 0.92, respectively. We note again that all *Hodgkinia* genes present in *Magicicada* may not have fully assembled due to the complexity of the dataset, so the true values for *Magicicada* HGCs may be higher than what we report here.

Lineage Splitting Leads to Different Gene Dosages

To estimate the similarity in gene dosages in different cicada species, we summed the average coverage of all contigs on which a given functional gene is found, scaled to the most abundant gene for each species. We find that the relative abundances of genes are similar within species groups (cicadas diverged less than 50 thousand years ago [10]), but not between species groups (Figure 3). This phylogenetic pattern is evident in a principle coordinates analysis (Figure S3A) and is clearer when only considering genes annotated in all species (Figure S3B). This grouping is qualitatively similar to the CSI results and suggests that there is not a convergent pattern of gene dosage outcomes as might be expected if the host were dictating the *Hodgkinia* splitting process or if the process were beneficial to the *Hodgkinia* community in some way. Rather, the gene dosage outcomes are stochastic and thus only similar in comparisons between very closely related cicadas.

DISCUSSION

Many endosymbioses consist of two or more partners that are strictly reliant on one another for survival. Even in symbioses that become highly genetically and cell-biologically integrated, the evolutionary trajectories of the partners are not inevitably aligned and may directly conflict because each partner can experience selection and drift independent of the other [19–28]. Although the engulfed partner is capable of exerting selfish tendencies in some cases [29–31], there are several mechanisms that the host employs to constrain the evolution of its symbionts [32–35]. In bacterial endosymbioses, this host-level constraint is often reflected in the genomic stasis of the bacterial partner. Endosymbiont genomes can remain stable in gene content and structure for tens [3], hundreds [9], or even thousands [5, 36] of millions of years.

However, secondary genome instability subsequent to this stasis is now recognized as relatively common, especially in mitochondria [37, 38]. Mitochondrial genomic instability manifests both as genome reduction [39, 40] that sometimes leads to outright genome loss [41–45] and as genome fragmentation [46–49] that sometimes leads to massive genome expansion with little obvious functional change [50–53]. We suggest that what unites these starkly different outcomes is a shift away from the host-driven constraint of the endosymbiont genome toward (sometimes temporary) symbiont-driven instability. In cases of mitochondrial reduction and loss, the host ecology changes such that the function of the organelle is no longer needed and therefore no longer under selective constraint from the host [42–44]. For example, many eukaryotes that live in anaerobic environments no longer require the oxidative respiratory function of their mitochondria, so the genes for this process are free to be lost [40]. The forces promoting mitochondrial genome fragmentation and expansion are less clear, but these expansions sometimes seem to be associated with increases in mitochondrial mutation rates [51] and have been hypothesized to result from less efficacious host-level selection against slightly deleterious symbiont mutations [53, 54].

Depending on whether one takes a *Hodgkinia*- or cicada-centric perspective, the outcomes that we report here could be interpreted either as a genome-reductive or genome-expansive

Please cite this article in press as: Campbell et al., Idiosyncratic Genome Degradation in a Bacterial Endosymbiont of Periodical Cicadas, Current Biology (2017), https://doi.org/10.1016/j.cub.2017.10.008

process [7, 8]. From *Hodgkinia*'s perspective, the splitting and deletion process leads to individual circular molecules that resemble the extremely degraded genomes of mitochondria found in some eukaryotes. The idiosyncratic nature of these circles in different cicada species (Figure 2C) is consistent with stochastic gene loss through mutation and suggests a process with no particular goal or end point. But an important difference between cases of mitochondrial genome reduction and *Hodgkinia* is that in *Magicicada*, the host ecology has not changed such that *Hodgkinia*'s functions are no longer required. The massive gene loss on individual *Hodgkinia* circles is most likely only tolerable because from the host's perspective, the combined HGCs seem to have retained *Hodgkinia*'s overall nutritional contribution to the symbiosis (Figure 2D). This splitting and genome-reductive process results in a combined *Hodgkinia* "genome" size that is an order of magnitude larger than the ancestral single genome (Table 1). How all of the numerous proteins, RNAs, and metabolites are shared between *Hodgkinia* cells is unknown, but it is likely that the host is heavily involved in this process, similar to other endosymbionts that are highly integrated with their hosts [55].

In our view, the most interesting parallel to what we report here for *Hodgkinia* can be found in the mitochondrial genomes of the angiosperm genus *Silene* [51, 56]. Like many plants, some *Silene* mitochondrial genomes consist of a single "master circle" with multiple "subcircles" that arise primarily from recombination [57]. Other *Silene* species, though, have experienced dramatic increases in mitochondrial mutation rates, which seem to be accompanied by the expansion to dozens of enormous mitochondrial chromosomes [51]. These mitochondrial chromosomes, some encoding few or no detectable genes, can be rapidly lost or gained in closely related *Silene* lineages [56]. Like *Hodgkinia*, this diversity of genome expansion outcomes in closely related plant hosts is not accompanied by any detectable increase in functional capacity. We previously hypothesized that the increased complexity of *Hodgkinia* in *Magicicada* results from an increased number of *Hodginia* genome-replication events due to the unusually long life cycle of *Magicicada* [8]. We hypothesize that an increase in *Hodgkinia* mutations per host life cycle enables lineage splitting and eventually results in stochastic differences between HGCs from different cicada species (Figure 2C). Although *Hodgkinia* genes are (mostly) maintained in all HGCs, they are now present at wildly different abundances in different cicada species groups (Figure 3). We suggest that lineage splitting and changes in gene dosages are either maladaptive or neutral for the host. There is no benefit from *Hodgkinia* degeneration, but the cicada host must tolerate it because it is wholly dependent on *Hodgkinia* for survival.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
    - ○ DNA extraction
    - ○ Library preparation and sequencing
    - ○ Genome assembly and annotation
    - ○ Phylogenetic analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures and two tables and can be found with this article online at https://doi.org/10.1016/j.cub.2017.10.008.

AUTHOR CONTRIBUTIONS

Conceptualization, M.A.C. and J.P.M.; Methodology, M.A.C. and P.Ł.; Formal Analysis, M.A.C.; Investigation, M.A.C.; Resources, C.S. and J.P.M.; Data Curation, M.A.C. and C.S.; Writing – Original Draft, M.A.C.; Writing – Review & Editing, M.A.C., P.Ł., C.S., and J.P.M.; Visualization, M.A.C.; Supervision, J.P.M.; Funding Acquisition, C.S. and J.P.M.

ACKNOWLEDGMENTS

We thank all members of the McCutcheon lab for helpful discussion and comments and Scott Miller for suggesting the use of the Jaccard index. Funding for the sequencing and analysis was supported by National Science Foundation (NSF) grants IOS-1256680 and IOS-1553529 and National Aeronautics and Space Administration Astrobiology Institute award NNA15BB04A. Funding for cicada collecting was provided by NSF DEB-09-55849.

Received: August 23, 2017
Revised: September 27, 2017
Accepted: October 3, 2017
Published: November 9, 2017

REFERENCES

1. McCutcheon, J.P., and Moran, N.A. (2011). Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. *10*, 13–26.
2. Bennett, G.M., and Moran, N.A. (2015). Heritable symbiosis: the advantages and perils of an evolutionary rabbit hole. Proc. Natl. Acad. Sci. USA *112*, 10169–10176.
3. 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.
4. McCutcheon, J.P. (2010). The bacterial essence of tiny symbiont genomes. Curr. Opin. Microbiol. *13*, 73–78.
5. Boore, J.L. (1999). Animal mitochondrial genomes. Nucleic Acids Res. *27*, 1767–1780.
6. 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.
7. Van Leuven, J.T., Meister, R.C., Simon, C., and McCutcheon, J.P. (2014). Sympatric speciation in a bacterial endosymbiont results in two genomes with the functionality of one. Cell *158*, 1270–1280.
8. Campbell, M.A., Van Leuven, J.T., Meister, R.C., Carey, K.M., Simon, C., and McCutcheon, J.P. (2015). Genome expansion via lineage splitting and genome reduction in the cicada endosymbiont *Hodgkinia*. Proc. Natl. Acad. Sci. USA *112*, 10192–10199.
9. McCutcheon, J.P., McDonald, B.R., and Moran, N.A. (2009). Convergent evolution of metabolic roles in bacterial co-symbionts of insects. Proc. Natl. Acad. Sci. USA *106*, 15394–15399.
10. Sota, T., Yamamoto, S., Cooley, J.R., Hill, K.B.R., Simon, C., and Yoshimura, J. (2013). Independent divergence of 13- and 17-y life cycles

among three periodical cicada lineages. Proc. Natl. Acad. Sci. USA *110*, 6919–6924.

11. Moulds, M.S. (2003). An appraisal of the cicadas of the genus *Abricta Stal* and allied genera (Hemiptera: Auchenorrhyncha: Cicadidae). Rec. Aust. Mus. *55*, 245–304.
12. Jaccard, P. (1901). Etude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull. Soc. Vaud. Sci. Nat. *37*, 547–579.
13. Cooper, K.W. (1941). *Davispia bearcreekensis Cooper*, a new cicada from the Paleocene, with a brief review of the fossil *Cicadidae*. Am. J. Sci. *239*, 286–304.
14. Poinar, G., Jr., and Kritsky, G. (2012). Morphological conservatism in the foreleg structure of cicada hatchlings, *Burmacicada proteran*. gen., n. sp. in Burmese amber, *Dominicicada youngin*. gen., n. sp. in Dominican amber and the extant *Magicicada septendecim*(L.) (Hemiptera: Cicadidae). Hist. Biol. *24*, 461–466.
15. Poinar, G., Jr., Kritsky, G., and Brown, A. (2012). *Minyscapheus dominicanus* n. gen., n. sp. (Hemiptera: Cicadidae), a fossil cicada in Dominican amber. Hist. Biol. *103*, 1–5.
16. Marshall, D.C., Hill, K.B.R., Moulds, M., Vanderpool, D., Cooley, J.R., Mohagan, A.B., and Simon, C. (2016). Inflation of molecular clock rates and dates: molecular phylogenetics, biogeography, and diversification of a global cicada radiation from Australasia (Hemiptera: Cicadidae: Cicadettini). Syst. Biol. *65*, 16–34.
17. Wernegreen, J.J. (2002). Genome evolution in bacterial endosymbionts of insects. Nat. Rev. Genet. *3*, 850–861.
18. Patiño-Navarrete, R., Moya, A., Latorre, A., and Peretó, J. (2013). Comparative genomics of *Blattabacterium cuenoti*: the frozen legacy of an ancient endosymbiont genome. Genome Biol. Evol. *5*, 351–361.
19. Bennett, G.M., McCutcheon, J.P., MacDonald, B.R., Romanovicz, D., and Moran, N.A. (2014). Differential genome evolution between companion symbionts in an insect-bacterial symbiosis. MBio *5*, e01697–14.
20. Keeling, P.J., and McCutcheon, J.P. (2017). Endosymbiosis: the feeling is not mutual. J. Theor. Biol. *434*, 75–79.
21. Eberhard, W.G. (1990). Evolution in bacterial plasmids and levels of selection. Q. Rev. Biol. *65*, 3–22.
22. Otto, S.P., and Orive, M.E. (1995). Evolutionary consequences of mutation and selection within an individual. Genetics *141*, 1173–1187.
23. Okasha, S. (2005). Multilevel selection and the major transitions in evolution. Philos. Sci. *72*, 1013–1025.
24. Maynard Smith, J. (1964). Group selection and kin selection. Nature *201*, 1145–1147.
25. Hurst, L.D., Atlan, A., and Bengtsson, B.O. (1996). Genetic conflicts. Q. Rev. Biol. *71*, 317–364.
26. Kiers, E.T., and West, S.A. (2015). Evolving new organisms via symbiosis. Science *348*, 392–394.
27. West, S.A., Fisher, R.M., Gardner, A., and Kiers, E.T. (2015). Major evolutionary transitions in individuality. Proc. Natl. Acad. Sci. USA *112*, 10112–10119.
28. Sachs, J.L., Skophammer, R.G., and Regus, J.U. (2011). Evolutionary transitions in bacterial symbiosis. Proc. Natl. Acad. Sci. USA *108* (*Suppl 2*), 10800–10807.
29. Taylor, D.R., Zeyl, C., and Cooke, E. (2002). Conflicting levels of selection in the accumulation of mitochondrial defects in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA *99*, 3690–3694.
30. Bastiaans, E., Aanen, D.K., Debets, A.J.M., Hoekstra, R.F., Lestrade, B., and Maas, M.F.P.M. (2014). Regular bottlenecks and restrictions to somatic fusion prevent the accumulation of mitochondrial defects in *Neurospora*. Philos. Trans. R. Soc. Lond. B Biol. Sci. *369*, 20130448.
31. Ma, H., and O'Farrell, P.H. (2016). Selfish drive can trump function when animal mitochondrial genomes compete. Nat. Genet. *48*, 798–802.
32. Bergstrom, C.T., and Pritchard, J. (1998). Germline bottlenecks and the evolutionary maintenance of mitochondrial genomes. Genetics *149*, 2135–2146.
33. Rispe, C., and Moran, N.A. (2000). Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection. Am. Nat. *156*, 425–441.
34. Leigh, E.G. (1983). When does the good of the group override the advantage of the individual? Proc. Natl. Acad. Sci. USA *80*, 2985–2989.
35. Maynard Smith, J. (1976). Group selection. Q. Rev. Biol. *51*, 277–283.
36. Adams, K.L., and Palmer, J.D. (2003). Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol. Phylogenet. Evol. *29*, 380–395.
37. Smith, D.R., and Keeling, P.J. (2015). Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. Proc. Natl. Acad. Sci. USA *112*, 10177–10184.
38. Burger, G., Gray, M.W., and Lang, B.F. (2003). Mitochondrial genomes: anything goes. Trends Genet. *19*, 709–716.
39. Turmel, M., Lemieux, C., Burger, G., Lang, B.F., Otis, C., Plante, I., and Gray, M.W. (1999). The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor*: two radically different evolutionary patterns within green algae. Plant Cell *11*, 1717–1730.
40. Conway, D.J., Fanello, C., Lloyd, J.M., Al-Joubori, B.M.A.S., Baloch, A.H., Somanath, S.D., Roper, C., Oduola, A.M.J., Mulder, B., Povoa, M.M., et al. (2000). Origin of *Plasmodium falciparum* malaria is traced by mitochondrial DNA. Mol. Biochem. Parasitol. *111*, 163–171.
41. Mai, Z., Ghosh, S., Frisardi, M., Rosenthal, B., Rogers, R., and Samuelson, J. (1999). Hsp60 is targeted to a cryptic mitochondrion-derived organelle ("crypton") in the microaerophilic protozoan parasite *Entamoeba histolytica*. Mol. Cell. Biol. *19*, 2198–2205.
42. Hackstein, J.H.P., Akhmanova, A., Boxma, B., Harhangi, H.R., and Voncken, F.G.J. (1999). Hydrogenosomes: eukaryotic adaptations to anaerobic environments. Trends Microbiol. *7*, 441–447.
43. van der Giezen, M. (2009). Hydrogenosomes and mitosomes: conservation and evolution of functions. J. Eukaryot. Microbiol. *56*, 221–231.
44. Embley, T.M., van der Giezen, M., Horner, D.S., Dyal, P.L., and Foster, P. (2003). Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. Philos. Trans. R. Soc. Lond. B Biol. Sci. *358*, 191–201, discussion 201–202.
45. Tovar, J., León-Avila, G., Sánchez, L.B., Sutak, R., Tachezy, J., van der Giezen, M., Hernández, M., Müller, M., and Lucocq, J.M. (2003). Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation. Nature *426*, 172–176.
46. Shao, R., Kirkness, E.F., and Barker, S.C. (2009). The single mitochondrial chromosome typical of animals has evolved into 18 minichromosomes in the human body louse, *Pediculus humanus*. Genome Res. *19*, 904–912.
47. Shao, R., Zhu, X.-Q., Barker, S.C., and Herd, K. (2012). Evolution of extensively fragmented mitochondrial genomes in the lice of humans. Genome Biol. Evol. *4*, 1088–1101.
48. Shao, R., Li, H., Barker, S.C., and Song, S. (2017). The mitochondrial genome of the guanaco louse, *Microthoracius praelongiceps*: insights into the ancestral mitochondrial karyotype of sucking lice (Anoplura, Insecta). Genome Biol. Evol. *9*, 431–445.
49. Vlček, C., Marande, W., Teijeiro, S., Lukeš, J., and Burger, G. (2011). Systematically fragmented genes in a multipartite mitochondrial genome. Nucleic Acids Res. *39*, 979–988.
50. Sloan, D.B. (2013). One ring to rule them all? Genome sequencing provides new insights into the 'master circle' model of plant mitochondrial DNA structure. New Phytol. *200*, 978–985.
51. Sloan, D.B., Alverson, A.J., Chuckalovcak, J.P., Wu, M., McCauley, D.E., Palmer, J.D., and Taylor, D.R. (2012). Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol. *10*, e1001241.
52. Alverson, A.J., Rice, D.W., Dickinson, S., Barry, K., and Palmer, J.D. (2011). Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. Plant Cell *23*, 2499–2513.
53. Rice, D.W., Alverson, A.J., Richardson, A.O., Young, G.J., Sanchez-Puerta, M.V., Munzinger, J., Barry, K., Boore, J.L., Zhang, Y., dePamphilis, C.W.,

et al. (2013). Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. Science *342*, 1468–1473.

54. Lynch, M., Koskella, B., and Schaack, S. (2006). Mutation pressure and the evolution of organelle genomic architecture. Science *311*, 1727–1730.
55. Singer, A., Poschmann, G., Mühlich, C., Valadez-Cano, C., Hänsch, S., Hüren, V., Rensing, S.A., Stühler, K., and Nowack, E.C.M. (2017). Massive Protein Import into the Early-Evolutionary-Stage Photosynthetic Organelle of the Amoeba Paulinella chromatophora. Curr. Biol. *27*, 2763–2773.
56. Wu, Z., Cuthbert, J.M., Taylor, D.R., and Sloan, D.B. (2015). The massive mitochondrial genome of the angiosperm *Silene noctiflora* is evolving by gain or loss of entire chromosomes. Proc. Natl. Acad. Sci. USA *112*, 10185–10191.
57. Palmer, J.D., and Shields, C.R. (1984). Tripartite structure of the *Brassica campestris* mitochondrial genome. Nature *307*, 437–440.
58. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.
59. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477.
60. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.
61. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
62. Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. *39*, W29–W37.
63. 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.
64. Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. *32*, 11–16.
65. Seemann, T. barrnap. https://github.com/tseemann/barrnap.
66. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.
67. Oksanen, J., Kindt, R., Legendre, P., and O'Hara, B. (2007). The vegan package. Community Ecology Package *10*, 631–637. https://cran.r-project.org/web/packages/vegan/index.html.
68. R Core Team. (2015). R: a language and environment for statistical computing. https://www.R-project.org.
69. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Please cite this article in press as: Campbell et al., Idiosyncratic Genome Degradation in a Bacterial Endosymbiont of Periodical Cicadas, Current Biology (2017), https://doi.org/10.1016/j.cub.2017.10.008

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
M. cassini bacteriome tissue	Jefferson Co. OK Brood IV; 34 09’ 521” N, 98 00’ 181” W; US 70 at River, W. Side of Wauhika; 2 June 2015, D. Marshall	N/A
M. tredecassini bacteriome tissue	4701 Leeb Drive, Tupelo, MS. In front of America’s Best Value Inn; 34 19’ 5.0” N, 88 47’ 30.2” W; 30 May 2015, C. Simon	N/A
M. septendecula bacteriome tissue	St. Mary’s Co., MD Brood II; HWY 235, btw MacArthur RD. & Millstone Landing Rd.; 1 June 2013, C. Simon & S. Chiswell	N/A
M. tredecula bacteriome tissue	Marshall Co. KY Brood XXIII; Calvert City, Super 8 Motel; 37.0086 N, 88.3308 W; 4 June 2015, J. Yoshimura	N/A
M. neotredecim bacteriome tissue	Dewitt Co. IL Brood XXIII; Clinton, N 830Rd. & Iron Bridge Rd.; 40.09714 N, 88.9884 W; 1 June 2015, J. Yoshimura	N/A
M. septendecim bacteriome tissue	New Haven Co. CT, Brood II; Gwen Rd., Meriden; 41 32′ 51.42” N, 71 50’ 45.23” W; 9 June 2013, C. Simon	N/A
M. tredecim bacteriome tissue	Baton Rouge Parish, LA Brood XXII; Warren Watson Mem. Park; 9 May 2014, C. Simon	N/A
T. crassa bacteriome tissue	Australia; 21 Jan. 2004, D. Marshall and K. HIll	N/A
A. curvicosta bacteriome tissue	Sydney, Australia; Queen Elizabeth Park; 8 Feb. 1997, C. Simon	N/A
Critical Commercial Assays		
NEBNext Ultra DNA Library Prep Kit	Illumina	Cat#E7370S
DNeasy Blood and Tissue kit	QIAGEN	Cat#69506
TruSeq PCR-free kit	Illumina	Cat# FC-121-9006DOC
Deposited Data		
Raw sequencing reads	This paper	SRA: SRR5753865, SRR5753866, SRR5753867, SRR5753868, SRR5753869, SRR5753870, SRR5753871, SRR5753872, SRR5753873, SRR5753874, SRR5753875, SRR5753876, SRR5753877, SRR5753878, SRR5753879, SRR5753880
HGC sequences and annotations	This paper	GenBank: NXGL00000000, NXGM00000000, NXGN00000000, NXGO00000000, NXGP00000000, NXGQ00000000, NXGR00000000, NXGS00000000, NXGT00000000
Software and Algorithms		
FASTX version 0.0.13	N/A	http://hannonlab.cshl.edu/fastx_toolkit/
Trimmomatic version 0.35	[58]	http://www.usadellab.org/cms/?page=trimmomatic
Spades version 3.6.2	[59]	http://cab.spbu.ru/software/spades/
TBLASTN 2.2.31+	[60]	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
BWA version 0.7.12-r1039	[61]	https://sourceforge.net/projects/bio-bwa/files/
HMMER v. 3.1b2	[62]	http://hmmer.org/download.html

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
RNAmmer 1.2	[63]	http://www.cbs.dtu.dk/cgi-bin/sw_request?rnammer
Aragorn v1.2.36	[64]	http://mbio-serv2.mbioekol.lu.se/ARAGORN/Downloads/
barrnap 0.6	[65]	https://github.com/tseemann/barrnap
RAxML v. 8.2.0	[66]	https://github.com/stamatak/standard-RAxML
R package Vegan 2.4-3	[67]	https://cran.r-project.org/web/packages/vegan/index.html
R	[68]	https://www.r-project.org/
mafft v. 7.221	[69]	https://mafft.cbrc.jp/alignment/software/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, John McCutcheon (john.mccutcheon@umontana.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Bacteriomes were dissected from a single male of *T. crassa*, a single female of *A. curvicosta* and *M. tredecim*, and two females of the remaining species. All specimens were collected in the wild between 1997 and 2015.

METHOD DETAILS

DNA extraction

DNA was extracted from all dissected bacteriomes using a DNeasy Blood and Tissue kit (QIAGEN catalog #69506). Extracted DNA was stored at −20C.

Library preparation and sequencing

Genomic DNA from *M. tredecim* was sheared to an average fragment size of 550 base pairs using a Covaris E220. Sheared DNA was made into a sequencing library using the NEBNext Ultra DNA Library Prep Kit for Illumina (catalog #E7370S), according to the standard protocol. The library was sequenced at the University of Montana Genomics Core on a MiSeq benchtop sequencer with a v3 600 cycle kit.

Genomic DNA from *A. curvicosta* was sheared to an average size of 480 base pairs using a Covaris E220. Sheared DNA was made into a sequencing library using a TruSeq PCR-free kit (Illumina) and sequenced as ∼1/12 of a multiplexed lane at NGX Bio in San Francisco, CA using a HiSeq 2500 Rapid SBS kit (Illumina).

Genomic DNA from *T. crassa* was sheared to an average of 570 base pairs using a Covaris E220. Sheared DNA was made into a sequencing library using the NEBNext Ultra DNA Library Prep Kit for Illumina (catalog #E7370S), according to the standard protocol. The library was sequenced as ∼1/4 of a multiplexed lane at the University of Montana Genomics Core on a MiSeq benchtop sequencer with a v3 600 cycle kit.

Genomic DNA from *M. neotredecim*, *septendecim*, *tredecassini*, *cassini*, *tredecula*, and *septendecula* was sheared to an average of 500 base pairs using a Covaris E220. Sheared DNA was made into a sequencing library using the NEBNext Ultra DNA Library Prep Kit for Illumina (catalog #E7370S), according to the standard protocol. Libraries were sequenced on two lanes on a HiSeq 2500 with 250 cycles at the Johns Hopkins School of Medicine Genetic Resources Core Facility.

Genomic DNA from *M. neotredecim*, *M. septendecim*, *M. tredecassini*, *M. cassini*, *M. tredecula*, and *M. septendecula* was used for making libraries with a Nextera Mate Pair Sample Prep Kit (Illumina), according to the standard protocol. These libraries were sequenced on a single lane on a HiSeq 2500 with 100 cycles at the Case Western Reserve University Genomics Core Facility.

Genome assembly and annotation

Raw reads were quality filtered using Trimmomatic version 0.35 [58]. Remaining reads were further filtered using fastq_quality_filter from FASTX version 0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/).

Assembly of the filtered reads was done using Spades version 3.6.2 [59], using kmer sizes 127, 151, 191, and 291, both individually as well as combined together. Putative *Hodgkinia* contigs were identified with TBLASTN 2.2.31+ [60] with an E-value cutoff of 10e-5 using previously annotated *Hodgkinia* genes as the query. To remove redundant contigs, all putatively *Hodgkinia* contigs were queried against themselves, and any contig > = 97% identical to another over > = 80% of its length was considered redundant and removed. Any contigs with BLASTN E-values less than 10e-10 to the mitochondrial genome were also removed. Coverage of individual contigs was calculated by the total coverage at each base, divided by the length of the contig. Completely assembled *Hodgkinia* circles were identified based on sequence overlap on both ends of the contig. To identify putative circular contigs, filtered paired end and mate pair reads were mapped back to the assembly using BWA version 0.7.12-r1039 [61] with default parameters. Contigs were considered putatively circular if more than five read pairs mapped with one mate mapping in the first 10% of the contig, while its mate mapped in the last 10% of the contig. Putatively circular contigs were then closed when possible by PCR and Sanger sequencing.

Annotation of the *Hodgkinia* circles was done using a custom Python pipeline based around the Jackhmmer module of HMMER v. 3.1b2 [62], RNAmmer 1.2 [63], and Aragorn v1.2.36 [64]. Occasionally RNAmmer misannotated the 23S rRNA gene, so barrnap 0.6 [65] was used for corrections. The completely closed *Hodgkinia* circles were then checked manually for any long open reading frames that could contain missing genes.

Phylogenetic analysis

Host phylogeny was reconstructed using RAxML v. 8.2.0 [66] based on manually inspected alignments of 15 mitochondrial genes (13 protein-coding and two rRNA) of the total length of 12744 bp, divided into four partitions corresponding to three codon positions and to rRNA genes. Rapid bootstraping (100 replicates) was used to estimate node support.

To construct individual gene phylogenies, homologous nucleotide sequences were translated into amino acids and aligned using mafft v. 7.221 [69]. Visually inspected alignments were analyzed using RAxML v. 8.2.4 [66] using a PROTGAMMAWAG model of amino acid substitution and 100 bootstrap replicates. Trees were rooted using *Aleeta-Tryella* as outgroups (whenever they formed a single monophyletic clade), or alternatively on the longest branch separating well-supported clades that included species from all or most hosts in a comparison.

QUANTIFICATION AND STATISTICAL ANALYSIS

To compare the similarity in gene content and size of HGC circles between cicada species, a Circle Similarity Index (CSI) score was calculated for all pairwise comparisons of all circles, as explained in Results. The pair with the highest CSI score was kept for each circle.

To determine relative coverage of all *Hodgkinia* genes, the coverage of all *Hodgkinia* genes was summed based on the coverage of the contig on which it was annotated. These abundance values were then normalized based on the most abundant gene. Principal coordinates analysis was done using the R package Vegan 2.4-3 [67].

DATA AND SOFTWARE AVAILABILITY

Raw reads from this project are available at the Sequence Read Archive (SRA) with accession numbers SRA: SRR5753865, SRR5753866, SRR5753867, SRR5753868, SRR5753869, SRR5753870, SRR5753871, SRR5753872, SRR5753873, SRR5753874, SRR5753875, SRR5753876, SRR5753877, SRR5753878, SRR5753879, SRR5753880. Assembled sequences and annotations are available from NCBI BioProject PRJNA390936 with accession numbers GenBank: NXGL00000000, NXGM00000000, NXGN00000000, NXGO00000000, NXGP00000000, NXGQ00000000, NXGR00000000, NXGS00000000, NXGT00000000.