Idiosyncratic Genome Degradation in a Bacterial Endosymbiont of Periodical Cicadas

Matthew A. Campbell, Piotr Łukasik, Chris Simon, and John P. McCutcheon
Division of Biological Sciences, University of Montana, 32 Campus Drive, Missoula, MT 59812, USA
Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N Eagleville Road Unit 3043, Storrs, CT 06269, USA
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-map-
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 the origin of the genus *Magicicada*.

---

**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 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:

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

(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

---

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

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

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 CSI between the single Hodgkinia genomes of Dicero-procta 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 (500 thousand years diverged [10]), M. -cassini species with M. -decula species (4 million years diverged [10]) give average CSI scores of 0.43, 0.46, and 0.29, respectively.

![Figure 2. CSI Scores for Individual Circles and HGCs](image)

(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. decim 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 (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.
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
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 *Hodgkinia* genome-replication events due to the unusually long life cycle of *Magicicada* [5]. 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:

- 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.T.; 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.T., 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**


