Enhanced Mutation Rate, Relaxed Selection, and the “Domino Effect” are associated with Gene Loss in Blattabacterium, A Cockroach Endosymbiont

Yukihiro Kinjo,†,1 Nathan Lo,†,2 Paula Villa Martín,1 Gaku Tokuda ‡,3 Simone Pigolotti,1 and Thomas Bourguignon *,1

1Okinawa Institute of Science & Technology Graduate University, Tancha, Onna-son, Okinawa, Japan
2School of Life and Environmental Sciences, University of Sydney, Sydney, NSW, Australia
3Tropical Biosphere Research Center, University of the Ryukyus, Nishihara, Okinawa, Japan
†These authors contributed equally to this work.
*
Corresponding author: E-mail: thomas.bourguignon@oist.jp.

Abstract

Intracellular endosymbionts have reduced genomes that progressively lose genes at a timescale of tens of million years. We previously reported that gene loss rate is linked to mutation rate in Blattabacterium, however, the mechanisms causing gene loss are not yet fully understood. Here, we carried out comparative genomic analyses on the complete genome sequences of a representative set of 67 Blattabacterium strains, with sizes ranging between 511 and 645 kb. We found that 200 of the 566 analyzed protein-coding genes were lost in at least one lineage of Blattabacterium, with the most extreme case being one gene that was lost independently in 24 lineages. We found evidence for three mechanisms influencing gene loss in Blattabacterium. First, gene loss rates were found to increase exponentially with the accumulation of substitutions. Second, genes involved in vitamin and amino acid metabolism experienced relaxed selection in Cryptocercus and Mastotermes, possibly triggered by their vertically inherited gut symbionts. Third, we found evidence of epistatic interactions among genes leading to a “domino effect” of gene loss within pathways. Our results highlight the complexity of the process of genome erosion in an endosymbiont.

Key words: genetic drift, population size, Muller’s Ratchet, genome reduction, comparative genomics.

Introduction

Many bacteria have abandoned their free-living lifestyle to couple with eukaryotes as mutualistic intracellular endosymbionts, growing in specialized host cells, and being maternally transmitted through cytoplasmic heredity. In these associations, the bacteria benefit from the stable environment provided by their host and, in exchange, they provide various services, most typically the biosynthesis of nutrients that the hosts cannot produce on their own (Baumann 2005).

Living within host cells comes with consequences and profoundly affects endosymbiont genomes. The most noticeable consequence is reductive genome evolution. The genomes of anciently evolved endosymbionts, such as Blattabacterium cuenotii (hereafter, Blattabacterium) and many other insect endosymbionts, are invariably highly reduced, typically below 1,000 kb (Moran and Bennett 2014), and as small as 112 kb (Bennett and Moran 2013). This genome reduction takes place in two main phases. During the first phase, endosymbiont genomes experience massive loss of genes that are not necessary in the stable environment provided by the host (Toft and Andersson 2010; McCutcheon and Moran 2012). No selective pressure acts on the maintenance of these non-essential genes, leading to their rapid removal. During this phase, there may be selection for a streamlined genome (Koskiiniemi et al. 2012), and endosymbionts become specialized for particular functions complementing host metabolism (Moran et al. 2008; Moya et al. 2008). The second phase is a long and slow process of genome erosion, during which endosymbiont genomes largely preserve their synteny, sometimes for more than 200 My (Latorre and Manzano-Marín 2017; McCutcheon et al. 2019), but gradually lose genes.

The evolutionary processes that drive the initial genome erosion have been largely investigated (e.g., Moran and Plague 2004; Burke and Moran 2011; Oakeson et al. 2014; Clayton et al. 2016; Manzano-Marín and Latorre 2016). In contrast, the second phase, characterized by slow and gradual erosion of specialized endosymbiont genomes, like those of Blattabacterium, is not completely understood. Several mechanisms may explain this slow and gradual genome erosion. The most generally accepted mechanism is genetic drift...
Endosymbionts are generally thought to have small effective population sizes, as they go through transmission bottlenecks at each generation of their host, enhancing genetic drift. In addition, because endosymbionts strictly reproduce asexually and experience no recombination, they are thought to undergo irreversible accumulation of slightly deleterious mutations, impeding gene function, and promoting pseudogenization. This process, known as Muller’s ratchet (Muller 1964; Lynch et al. 1993), is largely believed to be the main cause of genome reduction by gene loss in endosymbionts (Moran et al. 2008; McCutcheon and Moran 2012).

A second mechanism was proposed by Marais et al. (2008), who suggested that enhanced mutation rate is the main driver of genome reduction in endosymbionts. Endosymbionts typically lack many genes involved in DNA repair mechanisms (Moran et al. 2008; Kuwahara et al. 2011), enhancing the mutation rate, and increasing the effect of mutation bias in bacteria toward adenine and thymine (A + T) and deletions (Hershberg and Petrov 2010). The enhanced mutation rate hypothesis also has the potential to explain reductive genome evolution in free-living bacteria (Marais et al. 2008), which, unlike endosymbionts, have large effective population sizes (Giovannoni et al. 2005; Boscaro et al. 2013; Batut et al. 2014; Giovannoni et al. 2014; Brewer et al. 2017). A third mechanism is the host acquisition of cosymbionts, which is thought to relax selection on many primary symbiont genes, and speed up the loss of redundant genes. This phenomenon has been described repeatedly in species of insects associated with multiple endosymbiont species (Moran and Bennett 2014; Douglas 2016; Sudakaran et al. 2017). Finally, the “domino effect” hypothesis postulates that gene loss induces the loss of other dependent genes (e.g., within the same pathway) that are rendered nonfunctional (Dagan et al. 2006; Martinez-Can et al. 2018).

Cockroaches and termites inherited Blattabacterium, an intracellular symbiotic bacterium growing in specialized fat body cells (Brooks 1970), from their common ancestor. Genome sequences have shown that Blattabacterium partakes in recycling of nitrogen wastes, and provides amino acids and vitamins to their host cockroach (López-Sánchez et al. 2009; Sabree et al. 2009). Blattabacterium has been strictly vertically transmitted across generations of their cockroach hosts since the symbiotic association was established more than 200 Ma (Lo et al. 2003; Bourguignon et al. 2018; Evangelista et al. 2019; Arab et al. 2020). Termites, except the primitive species Mastotermes darwiniensis, and cave-dwelling cockroaches of the genus Nocticola are the only two blattodean lineages that are known to have lost Blattabacterium (Bandi et al. 1995, Lo et al. 2003, 2007).

The largest genomes of Blattabacterium are 630 – 640 kb. Several lineages independently experienced further reduction and have genomes as small as 511 kb (López-Sánchez et al. 2009; Sabree et al. 2009; Neef et al. 2011; Huang et al. 2012; Sabree et al. 2012; Kambhampati et al. 2013; Patiño-Navarrete et al. 2013; Tokuda et al. 2013; Kinjo et al. 2015, 2018; Vicente et al. 2018; Bourguignon et al. 2020). For example, the sister strains of Blattabacterium associated with the woodroach Cryptocercus punctulatus and M. darwiniensis have reduced genomes of 610 and 590 kb, respectively, and have experienced independent genome reduction, as indicated by the relatively large 637 kb genome of the strain of Cryptocercus kyebangensis (Kinjo et al. 2018). Interestingly, many genes involved in amino acid biosynthesis were lost in parallel in both strains (Kinjo et al. 2018), possibly compensated by the stable association their host established with their largely vertically transmitted gut microbes (Brune 2014). How common parallel gene loss is in Blattabacterium remains to be understood.

We previously obtained 67 Blattabacterium genomes including strains associated with all cockroach families known to host Blattabacterium and found that, unexpectedly, gene loss rates do not correlate with the ratios of nonsynonymous to synonymous substitutions (dN/dS), a measure of the relative strength of selection and genetic drift (Bourguignon et al. 2020). Therefore, reduced effective population size, which is expected to strengthen genetic drift, does not appear to be the main driver of gene loss in Blattabacterium, potentially reflecting the existence of strong purifying selection acting on endosymbiont genomes at the host level (Rispe and Moran 2000; Pettersson and Berg 2007). In contrast, gene loss rates strongly correlate with time-controlled synonymous substitution rates (dS/time), a proxy for mutation rate, not only in endosymbiotic lineages, such as Blattabacterium and Buchnera, but also in free-living bacterial and archaeal lineages. This suggests that enhanced mutation rate is commonly linked to genome reduction by gene loss in prokaryotes (Bourguignon et al. 2020). However, it is unclear whether gene loss rates vary linearly with mutation rates in endosymbiont lineages, or whether the relationship between these two rates follows different rules. Furthermore, mutation rate did not fully explain the mechanism by which genes involved in amino acid biosynthesis were selectively lost from the genomes of the Blattabacterium strains associated with the woodroaches Cryptocercus spp. and the termite M. darwiniensis (Kinjo et al. 2018), suggesting that additional factors drive the genome evolution of bacterial endosymbionts. The present study employs a large-scale comparative genomic approach to unveil the comprehensive evolutionary history of gene loss in the anciently evolved endosymbionts of cockroaches, Blattabacterium. In particular, we mathematically characterize the relationship between gene loss rates and mutation rates among 67 genomes of Blattabacterium and evaluate how Blattabacterium genomes have been affected by other evolutionary driving forces, especially relaxed selection and a within-pathway “domino effect.”

Results and Discussion

Massive Parallel Gene Loss Occurred among Blattabacterium Strains

We reconstructed gene loss on the maximum likelihood phylogenetic tree and the time-calibrated Bayesian phylogenetic tree of Bourguignon et al. (2020). The two phylogenetic trees had very similar topologies, closely recapitulating the phylogenetic tree inferred from host mitochondrial genomes.
and their lifestyles (Williams and Wernegreen 2008), we used an ancestral state reconstruction model that permitted gene loss, but no gene gain. The model was an adaptation of the maximum likelihood model described by Pagel (1994). The model of gene loss, run on the time-calibrated Bayesian tree and the maximum likelihood tree, estimated a total of 938.6 and 974.1 independent protein-coding gene loss events across the 566 orthologous genes of the 67 Blattabacterium strains, respectively (fig. 1A; supplementary tables S3 and S4, data S1 and S2, Supplementary Material online). The difference in number of gene loss events was almost entirely due to differences in branch lengths, as shown by the analyses run on the two trees without branch lengths that estimated the number of gene loss events to be 896.0 for the time-calibrated Bayesian tree and 896.6 for the maximum likelihood tree. Therefore, these analyses are robust to small variations in tree topology and not sensitive to tree reconstruction method.

Using these estimations of gene loss across the Blattabacterium phylogenetic tree, we determined how gene losses are distributed among genes and strains. Interestingly, the number of gene loss events varied considerably among genes. The reconstruction of gene loss events on the maximum likelihood phylogenetic tree with branch length revealed that, out of the 566 protein-coding genes analyzed, 200 genes were lost in at least one lineage, and 25 genes were lost between 10 and 24 times independently (fig. 1B). We carried out a Kruskal—Wallis test on number of independent gene loss events across COG categories and found that some categories of genes are lost more frequently than others ($\chi^2 = 90.1, df = 20, P < 10^{-10}$). For example, five of the six genes involved in defense mechanisms (COG category V), a function often lost in endosymbionts (e.g., Shigenobu et al. 2000), were lost multiple times in parallel. Our post hoc test, carried out on five groups, genes of COG categories E, H, and J, hypothetical genes, and genes of all other categories combined, showed that genes involved in amino acid and coenzyme biosynthesis (COG categories E and H) ($P < 0.01$), as well as hypothetical genes ($P < 0.01$), were lost more frequently than other genes. This is in line with the fact that endosymbionts often contribute to the nutrition of their host, retaining biosynthetic genes over long evolutionary periods, which they can easily lose when their host acquires new sources of nutriments, including from new cosymbionts (Douglas 2016). Likewise, biosynthetic genes are rapidly lost in free-living bacteria cultivated on nutritionally rich substrate (D’Souza et al. 2014; D’Souza and Kost 2016).

Overall, these results show that parallel gene loss is not limited to Blattabacterium strains associated with cockroach lineages possessing a social lifestyle, as M. darwiniensis and Cryptocercus spp. (Kinjo et al. 2018), but that the phenomenon is widespread across the Blattabacterium strains of all cockroaches. Parallel gene loss has been described in other endosymbiont lineages, but so far not in the extent found in Blattabacterium. For example, a total of 55 genes were independently lost between two and four times among six sequenced Blochmannia genomes (Williams and Wernegreen 2015). The sequencing of additional genomes in these lineages could also reveal massive parallel gene loss, involving hundreds of genes lost in parallel up to 24 times, as is the case in Blattabacterium for the gene coding for the tryptophan synthase beta chain.

Some strains of Blattabacterium lost genes at a faster pace than others. For example, the Blattabacterium genome of the strain associated with Blattella germanica was 641 kb in size and was inferred to miss only nine genes compared with the last common Blattabacterium ancestor, whereas that of Euphyllodromia sp. was 511 kb and lacked 130 genes. We carried out ancestral reconstruction on genome sizes in the 67 Blattabacterium strains under Brownian motion with a directional trend. The estimated genome size in the last common ancestor of the 67 Blattabacterium strains was estimated to be 668 kb (supplementary fig. S3A and B, Supplementary Material online). The last common ancestor of the 67 Blattabacterium strains was inferred to have existed 282 Mya (fig. 1A), and already lacked about five genes. Therefore, in 282 My, four genes were lost and genome size experienced a 27 kb reduction in the branch leading to B. germanica, and 125 genes were lost and genome size

### Footnotes

---

(ParaFitGlobal = 1.10; $P = 0.001$) (supplementary fig. S1, Supplementary Material online). We found seven disagreements between the Blattabacterium and cockroach phylogenetic trees, of which only one featured bootstrap support values above 75%. These results are largely congruent with those previously reported by Arab et al. (2020), and support the absence of host switching in Blattabacterium, and therefore the absence of DNA recombination among strains of different cockroach species. We further investigated the rate of recombination using GENCONV (Sawyer 1989), and identified eight putative DNA recombination events among Blattabacterium strains, all of which involved DNA fragments shorter than 120 bp, and 29 putative recombination events with bacteria other than Blattabacterium, all of which involved DNA fragments shorter than 15 bp (supplementary table S1, Supplementary Material online). These results indicate that, should these putative recombination events be genuine rather than false positives, recombination only affects DNA fragments much shorter than genes, and recombination does not therefore appear to be involved in gene loss. Next, we searched for evidences of horizontal gene transfers using HGTest (Zhu et al. 2014). Three genes were identified as candidate horizontally transferred genes by HGTest, but subsequent BlastP searches against the RefSeq database indicated that they were false positives (supplementary table S2, Supplementary Material online). Overall, the congruence between the phylogenetic trees of Blattabacterium and their cockroach hosts is in agreement with the vertical mode of inheritance of Blattabacterium, and the low rates of recombination and gene acquisition detected in this study are in agreement with the almost perfectly preserved synteny observed among the 67 Blattabacterium genomes examined, with only four inversions found in three strains (see supplementary fig. S2, Supplementary Material online).

Because we found that Blattabacterium genomes do not gain foreign genes by horizontal transfers, which is consistent with their lifestyle (i.e., growing within bacteriocytes physically separated from other bacteria; Moya et al. 2008), we used an ancestral state reconstruction model that permitted gene loss, but no gene gain. The model was an adaptation of the maximum likelihood model described by Pagel (1994). The model of gene loss, run on the time-calibrated Bayesian tree and the maximum likelihood tree, estimated a total of 938.6 and 974.1 independent protein-coding gene loss events across the 566 orthologous genes of the 67 Blattabacterium strains, respectively (fig. 1A; supplementary tables S3 and S4, data S1 and S2, Supplementary Material online). The difference in number of gene loss events was almost entirely due to differences in branch lengths, as shown by the analyses run on the two trees without branch lengths that estimated the number of gene loss events to be 896.0 for the time-calibrated Bayesian tree and 896.6 for the maximum likelihood tree. Therefore, these analyses are robust to small variations in tree topology and not sensitive to tree reconstruction method.

Using these estimations of gene loss across the Blattabacterium phylogenetic tree, we determined how gene losses are distributed among genes and strains.
FIG. 1. Evolution of genome reduction by gene loss in *Blattabacterium*. (A) Time-calibrated phylogenetic tree showing the evolution of genome reduction by gene loss in *Blattabacterium*. The tree was reconstructed with BEAST, using a set of 31 marker protein-coding genes, with third codon position excluded (see Bourguignon et al. 2020). Branch color represents cumulative gene loss. Node labels represent posterior distribution. Node bars indicate 95% highest posterior density. (B) Histogram representing the frequency of independent gene loss estimated from the 200 protein-coding genes lost in at least one lineage. The numbers above each bar indicate the frequency of independent gene loss. Values were derived from the reconstruction of gene loss on the maximum likelihood phylogenetic tree with branch length and were rounded to the nearest integer.
Gene Loss Rate Increases Exponentially with Mutation Rate
The smallest genomes of Blattabacterium have lost genes that directly affect the mutation rate. For example, the gene coding for superoxide dismutase, an enzyme catalyzing the removal of reactive oxygen species, was lost in the strain harbored by Euphyllodromia sp. Similarly, the gene coding for DNA polymerase III sliding clamp, without which the accuracy of DNA replication is reduced, was lost in all Blattabacterium genomes smaller than 600 kb, except that of Anaplecta spp. and M. darwiniensis. Increased mutation rate (dS/time) is linked to genome reduction (Bourguignon et al. 2020), for the strains that evolve faster, accumulate mutations at a faster pace, and lose genes at a faster pace. Here, we built mathematical models to quantify the gene loss process as mutations are accumulated (fig. 2; supplementary fig. S4, Supplementary Material online). We calculated the loss probability of a gene at a given evolutionary time t, by estimating the maximum likelihood of instantaneous rate of gene loss μ(t) (hereafter referred to as “gene loss rate”) at time t, with the assumption that there is no gene gain. We then investigated how gene loss rate varies over evolutionary time. We used branch lengths from the maximum likelihood phylogenetic tree of Bourguignon et al. (2020) as an estimate of mutation accumulation (evolutionary time) for every Blattabacterium strain. The phylogenetic tree was inferred from 353 protein-coding genes without third codon positions. Therefore, the branch lengths were mainly estimated from nonsynonymous substitutions, which, unlike synonymous substitutions, show no sign of saturation. As the level of genetic drift was constant across Blattabacterium genomes (Bourguignon et al. 2020), branch length is a good proxy for evolutionary time. In these models, we assumed that genes are lost independently from one another.

Our first model assumes gene loss at a constant rate μ, which we fitted for each gene using maximum likelihood method. In these analyses, the 95% confidence intervals were analytically determined, and corresponded to 1.96 standard deviations. The model predicted that the total number of lost genes rapidly increased at short times and gradually slowed down as the number of remaining nonessential genes declines (supplementary fig. S4A, Supplementary Material online). This did not match the data, which suggested an increased rate of gene loss as Blattabacterium genomes accumulate mutations.

To account for this discrepancy, we built three additional models, that assumed that genes are lost at a rate dependent on the cumulative number of mutations accumulated from the root. In this way, the overall rate of gene loss is allowed to increase over evolutionary time as Blattabacterium genomes accumulate substitutions. In particular, the second model assumes that gene loss rate grows linearly with time, μ(t) = μ₀ + t; the third model assumes that gene loss rate follows a power law, μ(t) = μ₀ tⁿ; and the fourth model assumes that gene loss rate increases exponentially, μ(t) = μ₀ e^t. Each of these three models is characterized by two gene-specific parameters μ₀ and τ, which determine the increase of gene loss rate and are estimated using maximum likelihood. We fitted these additional three models for each gene in our data set. Based on likelihood ratio tests between the first (constant rate) model and the other three models, we found that the second model did not improve the likelihood when compared with the first model (χ² = 148.82, df = 155, P = 0.62; supplementary fig. S4B and C, Supplementary Material online). In contrast, the third and fourth models, the power law and exponential models, fitted the data more precisely than the linear model (χ² = 260.64, df = 155, P < 10^-6, and χ² = 290.08, df = 155, P < 10^-9, for the third and fourth models, respectively) and the best-fit model was the exponential model (fig. 2). Exponential lineage-specific gene loss over time has also been described for mitochondrial genomes (Janouskovec et al. 2017), suggesting that the relationship might be a characteristic of endosymbionts. The exact mechanism driving this relationship remains to be determined.

Amino Acid and Coenzyme Biosynthetic Genes Are Affected by Relaxed Selection
Our exponential model provides a global description of genome erosion in Blattabacterium. However, some genomes, especially those of the strains associated with Cryptocercus spp. and M. darwiniensis, lost a number of genes significantly different from that predicted by our model, which implies that additional mechanisms affect genome erosion in Blattabacterium. Notably, these genomes disproportionately lost genes involved in amino acid and coenzyme biosynthesis (COG categories E and H), and genes with unknown functions (fig. 3). These observations might be explained by the acquisition of secondary symbionts inducing relaxed selection on some categories of genes, as is common in many insects associated with obligatory bacterial endosymbionts (Douglas 2016). If this is the case, our model should keep its explanatory power for gene categories that remain unaffected by relaxed selection. To test this hypothesis, we ran our exponential model without genes involved in amino acid biosynthesis (COG category E), genes involved in coenzyme biosynthesis (COG category H), and genes with unknown function (fig. 4). We found that removing COG category E improved the gene loss prediction of our exponential model for the Blattabacterium strains of Cryptocercus spp. and M. darwiniensis (F = 1.76, df = 66, P = 0.02; fig. 4A). Therefore, these strains especially lost genes involved in amino acid biosynthesis, supporting the idea that the largely vertically transmitted
gut bacteria of Cryptocercus spp. and M. darwiniensis replaced Blattabacterium for the biosynthesis of several amino acids (Brune 2014; Kinjo et al. 2018). Removing COG categories E, H, and unknown genes together also improved the gene loss prediction of the exponential model ($F = 1.71$, $df = 66$, $P = 0.03$; fig. 4B). This highlights that amino acid and coenzyme biosynthetic genes experienced relaxed selection in some lineages of Blattabacterium.

The loss of vitamin and amino acid biosynthetic genes in several Blattabacterium strains either occurred passively, by relaxed selection, or under the action of positive selection for a streamlined genome. The presence of multiple pseudogenes in the Blattabacterium genomes of the strains associated with Cryptocercus spp. indicate that the complete removal of pseudogenes takes upward of 1 My, suggesting that relaxed selection is the main mechanism (Kinjo et al. 2018). Whether or not the complete removal of pseudogenes is also a slow process in the Blattabacterium strains associated with other cockroaches, suggesting the absence or presence of positive selection for a streamlined genome, needs to be investigated by future studies examining genomes of closely related Blattabacterium strains. In the strains associated with Cryptocercus spp. and M. darwiniensis, the probable route of transmission of gut microbes, from parents to offspring, is through proctodeal trophalaxis, a phenomenon during which one individual provides a droplet of fecal fluid to a congener (Nalepa et al. 2001). No such route of transmission exists for the gut microbes of other cockroaches; however, Wolbachia and Rickettsia, which are frequently reported as secondary symbionts or parasites and are maternally inherited like Blattabacterium, were detected in several cockroach lineages harboring Blattabacterium with reduced genomes (supplementary table S5, Supplementary Material online).

This raises the possibility that the many genes of COG category H lost by the Blattabacterium strains associated with Ectobiidae are complemented by these secondary symbionts. However, the presence of Wolbachia and Rickettsia was also detected in association with other Blattabacterium strains and these secondary symbionts did not appear to experience long-term coevolution with their host cockroaches. This suggests that alternative factors triggered the loss of vitamin and amino acid biosynthetic genes in other Blattabacterium strains. One such factor is putatively the cockroach diet, which can be a source of vitamins. The presence of such vitamins in the diet might relax selection on biosynthetic genes, as has been suggested to explain the loss of vitamin biosynthetic genes in some mammal genomes (Hellwell et al. 2013). Overall, our results show that gene loss is accumulated over time in Blattabacterium, with occasional burst of gene loss for vitamin and amino acid biosynthetic genes, and genes with unknown function. As is the case for many insects, in which these genes were replaced by those encoded in secondary symbiont genomes (e.g., Koga and Moran 2014; Husnik and McCutcheon 2016; Sudakaran et al. 2017), the Blattabacterium strains associated with Cryptocercus spp. and M. darwiniensis appear to have lost genes functionally redundant with the genes of their gut microbial symbionts.

**Domino Effect: Evidence for Within-Pathways Correlated Gene Loss**

So far, our analyses have been conducted on the premise that the loss of one gene does not affect the loss of other genes. However, there is a possibility that lost genes affect the loss probability of other genes because of epistatic interactions. To test for a “domino effect” scenario, in which gene loss triggers further gene loss, we carried out Pagel tests for correlated evolution of discrete characters (Pagel 1994) (hereafter referred to as Pagel tests) using a model assuming that gene loss rate increases exponentially with evolutionary time for all possible pairwise combinations of genes lost in at least two lineages. Note that using the classical constant gene loss rate model for the Pagel tests provided identical results. A total of 146 genes were lost at least twice, for a total of 10,585 pairs. We found significant correlations (fig. 5), indicating that there is genome-scale correlated gene loss. However, these results do not necessarily imply epistatic interactions. Instead, they might be explained by the unequal rate of gene loss among Blattabacterium strains. Genes are especially lost by Blattabacterium strains with reduced genomes, and these

---

**Fig. 2.** Exponential model investigating gene loss as a function of substitution accumulation. (A) Exponential model of gene loss and observed number of gene loss in Blattabacterium genomes. The model includes two gene-specific loss parameters: an initial gene loss rate $\mu_0$ and an exponential time scale $\tau$. (B) Relationship between $\mu_0$ and $\tau$ for all genes in the data set. (C) Simplified phylogenetic tree of Blattabacterium. Branch colors correspond to symbol colors in (A).
strains tend to lose the same genes multiple times in parallel, possibly leading to correlated gene loss without epistatic interactions. To rule out this possibility, and to determine whether epistatic interactions trigger gene loss in Blattabacterium, we separated Pagel tests into three groups according to the nature of the pairs of genes compared: genes from different COG categories, genes from the same COG category but from different metabolic pathways, and genes from the same metabolic pathways (fig. 5). We compared the proportion of significant correlations in each group using chi-square tests. Pagel tests comparing genes from the same COG category but from different metabolic pathways comprised more significant correlations than Pagel test comparing genes from different COG categories ($\chi^2 = 12.21, df = 1, P < 10^{-3}$). Pagel tests comparing genes from the same metabolic pathways yielded many more significant correlations than Pagel
The clearest evidences of correlated gene loss were for genes composing metabolic pathways. Perhaps the most illustrative example is that of the most disrupted pathways (supplementary table S3, data S1, orthologous genes 55, 58, 313, 315, 318, 320, 402, were fixed at 10^-16). Therefore, our results indicate that correlated gene loss is more likely to occur among interdependent genes forming metabolic pathways.

Overall, our results provide support to the “domino-effect” theory, which posits that epistatic interactions among genes contribute to gene loss (Dagan et al. 2006; Martinez-Cano et al. 2018). The clearest evidences of correlated gene loss were for genes composing metabolic pathways.

Conclusion

We carried out comparative genomic analyses on 67 Blattabacterium genome sequences to determine the mechanisms responsible for gene loss within one endosymbiont lineage. We found that parallel gene loss has been common in Blattabacterium, with the most extreme case of one gene lost independently in at least 24 lineages. To the best of our knowledge, such a high number of parallel losses for one gene has never been reported for an endosymbiont, possibly because of the scarcity of studies comparing many genome sequences. Our results also indicate that genome reduction in Blattabacterium is a complex process under the influence of...
at least three mechanisms: enhanced mutation rate, relaxed selection, and within-metabolic-pathway domino effect. Notably, we found that gene loss rate exponentially increases with the accumulation of substitutions, as has also been described for mitochondrial genomes (Janouškovec et al. 2017).

Materials and Methods

Genome Sequences and Phylogenetic Analyses

We carried out all our analyses on the 67 Blattabacterium genome sequences used by Bourguignon et al. (2020). We also used the genome annotation and the phylogenetic trees of Bourguignon et al. (2020). The mitochondrial genome sequences of the corresponding host cockroaches, or of different cockroach specimens of the same species, were used to reconstruct the host phylogeny (supplementary table S4, Supplementary Material online). Briefly, we aligned the 22 transfer RNA genes, the 13 protein-coding genes and the two ribosomal RNA genes using MAFFT v7.305 and the command line: -maxiterate 1000 –globalpair (Katoh and Standley 2013). Protein-coding genes were aligned as amino acid and back-translated to nucleotide sequences using pal2nal v14 (Suyama et al. 2006). The 37 gene alignments were concatenated and the third codon position of protein-coding genes was excluded. The data set was divided into four partitions: one for the 22 tRNA genes, one for the two rRNA genes, one for the first codon position of protein-coding genes, and one for the second codon position of protein-coding genes. The concatenated alignment was used to reconstruct a maximum likelihood phylogenetic tree with RAxML version 8.2.4 with the GTRCAT model (Stamatakis 2014). Node supports were estimated from 100 bootstrap replicates.

Analyses of Cophylogeny, Recombination, Horizontal Gene Transfer, and Genome Synteny

We used the cophylo function of the R package phytools (Revell 2012) to investigate the congruence between the maximum likelihood phylogenetic tree of Blattabacterium inferred from 353 protein-coding genes without third codon positions (see Bourguignon et al. 2020) and the maximum likelihood phylogeny of cockroaches inferred from host cockroach mitochondrial genomes. The statistical method used to test coevolution was Parafit (Legendre et al. 2002). One cockroach species, Rhabdoblatta sp., was removed from this analysis due to the low quality of its mitochondrial genome assembly (see supplementary table S4, Supplementary Material online).

DNA recombination events among Blattabacterium, and with bacterial taxa other than Blattabacterium, were identified using GENCONV (Sawyer 1989). We carried out the analysis on all the genes of Blattabacterium, which included 564 protein-coding genes found in all strains of Blattabacterium. Closely related strains of Cryptocercus punctulatus were removed from the analysis, and only one Blattabacterium genome associated with C. punctulatus was retained. Statistical significance was assessed using the command “-numsim = 1000,” which specified that significance is estimated from 1,000 random permutations, and “-Indel_blocs,” to properly handle missing data. Other parameters were set on default values.

We used the software HGTector (Zhu et al. 2014) to identify genes that were putatively acquired through horizontal transfers. The analysis was carried out on the 566 orthologous protein-coding genes found in the 67 Blattabacterium strains used in this study. As recommended by the developer of HGTector, we used the command line options “-evalue 1e–20 –identity 50 –coverage 50” for the search stage, and “-maxhits 100 –evalue 1e–50 –identity 80 –coverage 80” for the analysis stage.

We investigated the variations in gene order across the 67 Blattabacterium strains used in this study. The gene synteny was depicted with full genome alignment using progressiveMauve (Darling et al. 2010).

Reconstruction of Gene Loss and Genome Size

We previously reconstructed gene loss for the 67 genome sequences of Blattabacterium used in this study with the “ace” function from the R package ape (Paradis et al. 2004). For each gene, we used presence/absence data to reconstruct ancestral state using maximum likelihood method (Pagel 1994). We used the option “model=matrix(c(0, 1, 0, 0, 2))” to specify no gain. The orthologous gene groups were manually curated and the orthologous groups annotated as “hypothesetical gene” and comprising less than five genes were removed from downstream analyses. The function “plotTree” of the R package phytools was used to visualize the results of each reconstruction (Revell 2012). We also used the cumulative maximum likelihood estimation of ancestral states (gene presence/absence) to calculate expected number of gene loss events across the tree. Using the estimated number of loss events for each gene, we tested whether some categories of genes (COG) are lost more frequently than others. We used a nonparametric Kruskal–Wallis test to determine differences among categories. Significant differences among categories were then determined with the kruskal.test function implemented in R. We only present the results of the analysis carried out on COG categories E, H, J and unknown, which were the most represented categories. Other categories were pooled together. Analyses with all genes assigned to their respective COG categories yielded similar results.

We reconstructed the evolution of genome sizes for the 67 Blattabacterium genomes using the anc.trend function of the R package phytools. We applied a directional Brownian motion model for the ancestral state reconstruction. We used the phenogram function implemented in the same package to plot the estimated ancestral genome sizes.

Correlation Between Gene Loss and Evolutionary Rate

We built a mathematical model to investigate the relationship between rate of gene loss and evolutionary distance from the root (evolutionary time). Our model used the branch length of the maximum likelihood phylogenetic tree of Bourguignon et al. (2020) as a measure of evolutionary time. The phylogenetic tree was inferred from the 353 orthologous protein-coding genes present across all strains of
Blattabacterium. Third codon sites were removed from the final alignment and the data set was partitioned into two subsets: one containing the first codon sites and one containing the second codon sites. We estimated gene loss probability along each branch using previously described methods (Pagel 1994; Borenstein et al. 2007). For all models, these probabilities can be expressed in an explicit form, which considerably sped up computations.

We described the presence and absence of a gene $i$ on a given branch of the phylogenetic tree by a variable $s_i$ that takes the values $s_i = 1$ and $s_i = 0$ if the gene is present or lost, respectively. The probability $P_i(t)$ that gene $i$ is present at evolutionary time $t$ follows the equation (Pagel 1994):

$$P_i(t + dt) = P_i(t)(1 - \mu_i(t)dt),$$

where $dt$ is an infinitesimal time interval and $\mu_i(t)$ is the time-dependent loss rate of gene $i$ (referred to as “gene loss rate” in this study). We assumed that, in Blattabacterium, lost genes cannot be recovered. We considered four scenarios of gene loss rate: constant $\mu_i(t) = \mu$, linear $\mu_i(t) = \mu_0 + \mu t$, power law $\mu_i(t) = \mu_0 t^{\gamma}$, and exponential $\mu_i(t) = \mu e^{-t/t_i}$. Using these probabilities, the likelihood can be efficiently computed by means of Felsenstein’s algorithm (Felsenstein 1981). For each gene, and each model, the likelihood was numerically maximized to obtain the parameters characterizing gene loss ($\mu$ for the first model, and $\mu_0$ and $\gamma$ for the other three models).

For all four models, the average number of lost genes $N$ for a root-to-tip distance $d$ was analytically obtained as:

$$N = \sum_i \left(1 - e^{-\int_0^d \mu_i(t)dt}\right),$$

where the sum $\sum_i$ runs over all genes and $\mu_i(t)$ is the loss rate with the fitted parameters for gene $i$. The 95% confidence interval was numerically obtained by simulating 10,000 random processes.

We ran the four mathematical models using the 200 protein-coding genes lost in at least one lineage. We also ran the exponential model an additional two times using a subsample of the 200 protein-coding genes: once using all genes but those of COG category E; once using all genes but those of COG category E, those involved in coenzyme biosynthesis (COG category H), and the genes with unknown function.

Correlated Gene Loss and Domino Effect

We used Pagel tests to investigate correlated gene loss, which we implemented as described by Pagel (1994).

In this case, variables ($s_i$, $s_j$) describe the presence and absence of a pair of genes $i, j$, taking values $(0,0), (0,1), (1,0)$, and $(1,1)$. We assumed that lost genes cannot be recovered. Each gene of the pair is lost at a rate that depends on the presence or absence of the other gene, so that the model includes four rates for each pair of genes. We assumed that these rates increase exponentially over evolutionary time. We computed the likelihood as in the previous section (Felsenstein 1981). In our numerical implementation, we used the explicit form of the probabilities of gene loss along branches to avoid the calculation of matrix exponential and speed up the analyses (see supplementary text, Supplementary Material online). For each pair of genes, the likelihood was numerically maximized to obtain the four gene loss rates and a rate-acceleration parameter depending on evolutionary time.

We carried out our analyses on the 146 genes that were lost twice or more resulting in 10,585 Pagel tests. We separated the 10,585 comparisons into three groups which we compared using chi-square tests. The first group comprised comparisons between genes from different COG categories. The second group included pairs of genes belonging to the same COG category but to different metabolic pathways. The third group included pairs of genes from the same metabolic pathways. We applied the local false discovery rate (FDR) method on the original 10,585 Pagel tests to correct multiplicity effect (Efron 2004) by using the fdrtool function implemented in the R package fdrtool (Strimmer 2008) with 0.05 local FDR threshold. The same significance threshold was applied to the subsets of the original comparison. For each group, we then counted the number of comparisons above and below the significance threshold and carried out one-sided chi-square tests to compare the first and second groups, and the second and third groups.

Detection of Candidate Secondary Symbiont in Sequence Data

We used MetaPhlAn2 (Truong et al. 2015) to investigate the presence of secondary symbionts in the sequence data generated by Bourguignon et al. (2020). MetaPhlAn2 detects clade-specific phylogenetic marker genes then performs taxonomic assignment and estimation of relative abundance. We defined as secondary symbionts bacterial taxa that make up more than five percent of bacterial reads in a given library.

Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgments

N.L. was supported by the Australian Research Council. G.T. was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI 17H01510. T.B. was supported by the JSPS KAKENHI 18K14767 and 21H02541. S.P. and T.B. were supported by the subsidiary funding to OIST.

Data Availability

All sequence data used in this study were previously published and are freely available on NCBI (see supplementary table S4, Supplementary Material online).

References

Arab DA, Bourguignon T, Wang Z, Ho SYW, Lo N. 2020. Evolutionary rates are correlated between cockroach symbionts and mitochondrial genomes. Biol Lett. 16(1):20190702.

Bandi C, Sironi M, Damiani C, Magrassi L, Nalepa CA, Laudani U, Sacchi L. 1995. The establishment of intracellular symbiosis in an ancestor of cockroaches and termites. Proc R Soc B Biol Sci. 259:293–299.
Batur B, Knibbe C, Marais G, Daubin V. 2014. Reductive genome evolution at both ends of the bacterial population size spectrum. *Nat Rev Microbiol*. 12(12):841–850.

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

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.

Borenstein E, Shlomi T, Ruppin E, Sharan R. 2007. Gene loss rate: a probabilistic measure for the conservation of eukaryotic genes. *Nucleic Acids Res*. 35(1):e7.

Boscaro V, Felletti M, Vannini C, Ackerman MS, Chain PSG, Malfatti S, 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.

Bourguignon T, Kinjo Y, Villa-Martín P, Coleman NV, Tang Q, Arab DA, Wang Z, Tokuda G, Hongo Y, Okhuma M, et al. 2020. Increased mutation rate is linked to genome reduction in prokaryotes. *Curr Biol*. 30(19):3848–3855.

Bourguignon T, Tang Q, Ho SYW, Juna F, Wang Z, Arab DA, Cameron SL, Walker J, Rentz D, Evans TA, et al. 2018. Transoceanic dispersal and plate tectonics shaped global cockroach biotas of cockroaches and termites (Blattodea). *Proc Natl Acad Sci U S A*. 115(37):E5416–E5424.

Janouškovec J, Gavelský GS, Burki F, Dinh D, Bachvaroff TR, Gornik SG, Bright KJ, Imanian B, Strom SL, Delwiche CF, et al. 2017. Major transitions in dinoflagellate evolution unveiled by phylogenomics. *Proc Natl Acad Sci U S A*. 114(2):E171–E180.

Kambhampti S, Alleman A, Park Y. 2013. Complete genome sequence of the endosymbiont *Blattabacterium* from the cockroach *Nauphoeta cinerea* (Blattodea: blaberidae). *Genomics*. 102(5-6):479–483.

Kato H, Standley DM. 2013. MAFFI multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780.

Kinjo Y, Bourguignon T, Tong KJ, Kuwahara H, Lim SJ, Yoon KB, Shigenobu S, Park YC, Nalapa CA, Hongo Y, et al. 2018. Parallel and gradual genome erosion in the *Blattabacterium* endosymbionts of *Mastotermes darwiniensis* and *Cryptocercus* wood roaches. *Genome Biol. Evol*. 10(6):1622–1630.

Kinjo Y, Saitoh S, Tokuda G. 2015. An efficient strategy developed for next-generation sequencing of endosymbiont genomes performed using crude DNA isolated from host tissues: a case study of *Blattabacterium cuenoti* inhabiting the fat bodies of cockroaches. *Microbes Environ*. 30(3):208–220.

Koga R, Moran NA. 2014. Swapping symbionts in siphlet bugs: evolutionary replacement of a reduced genome symbiont. *ISME J*. 8(6):1237–1246.

Koskimäki S, Sun S, Berg OG, Andersson DI. 2012. Selection-driven gene loss in bacteria. *PLOS Genet*. 8(6):e1002787.

Kuwahara H, Takaki Y, Shimamura S, Yoshida T, Maeda T, Kunieda T, Maruyama T. 2011. Loss of genes for DNA recombination and repair in the reductive genome evolution of thioautotrophic symbionts of *Cylindrolesmus* clams. *BMC Evol Biol*. 11:285.

Latorre A, Manzano-Martin A. 2017. Dissecting genome reduction and trait loss in insect endosymbionts. *Ann N Y Acad Sci*. 1389(1):52–75.

Legender P, Desseviers Y, Bazin E. 2002. A statistical test for host–parasite coevolution. *Syst Biol*. 51(2):217–234.

Lo N, Bandi C, Watanabe H, Nalapa CA, Beninati T. 2003. Evidence for cladcogenesis between diverse dicyopteran lineages and their intracellular endosymbionts. *Genome Biol. Evol*. 20(6):907–913.

Lo N, Beninati C, Stone F, Walker J, Sacchi L. 2007. Cockroaches that lack *Blattabacterium* endosymbionts: the phylogenetically divergent genus *Nocticola*. *Biol Lett*. 3(3):327–330.

López-Sánchez MJ, Neef A, Peretó J, Patiño-Navarrete R, Pignatelli M, Latorre A, Moya A. 2009. Evolutionary convergence and nitrogen metabolism in *Blattabacterium* strain Bge, primary endosymbiont of the cockroach *Blattella germanica*. *PLOS Genet*. 5(11):e1000721.

Lynch M, Bürger R, Butcher D, Gabriel W. 2013. The mutational meltdown in asexual populations. *J Hered*. 84(5):339–344.

Manzano-Marín A, Latorre A. 2016. Snapshots of a shrinking partner: genome reduction in *Serratia* symbionts. *Sci Rep*. 6:32590.

Marais GB, Calleau A, Tenailleau O. 2008. Mutation rate and genome reduction in endosymbiotic and free-living bacteria. *Genetica* 134(2):205–210.

Martínez-Can ø D, Bor G, Moya A, Delaye L. 2018. Testing the domino theory of gene loss in *Buchnera aphidicola* the relevance of epistatic interactions. *Life* 8(2):17.

McCUTCHEON JP, BOYD BM, DALE C. 2019. The life of an insect endosymbiont from the cradle to the grave. *Curr Biol*. 29(11):R485–R495.

McCUTCHEON JP, MORA N A. 2012. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol*. 10(1):13–26.
Moran NA. 1996. Accelerated evolution and Muller’s ratchet in endosymbiotic bacteria. *Proc Natl Acad Sci. 93*(7):2873–2878.

Moran NA, Bennett GM. 2014. The tiniest tiny genomes. *Annu Rev Microbiol. 68*:195–215.

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

Moran NA, Plague GR. 2004. Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev. 14*(6):627–633.

Moya A, Peretó J, Gil R, Latorre A. 2008. Learning how to live together: genomics insights into prokaryote–animal symbioses. *Nat Rev Genet. 9*(3):218–229.

Muller HJ. 1964. The relation of recombination to mutational advance. *Mutat Res Mol Mech Mutagen. 1*(1):2–9.

Nalepa CA, Bignell DE, Bandi C. 2001. Detritivory, coprophagy, and the evolution of digestive mutualisms in Dictyoptera. *Insectes Sociaux. 48*:194–201.

Nee F, Latorre A, Peretó J, Silva FJ, Pignatelli M, Moya A. 2011. Genome economization in the endosymbiont of the wood roach *Cryptocercus punctulatus* due to drastic loss of amino acid synthesis capabilities. *Genome Biol Evol. 3*:1437–1448.

Oakerson KF, Gil R, Clayton AL, Dunn DM, von Niederhausern AC, Hamil C, Aoyagi A, Duval B, Baca A, Silva FJ, et al. 2014. Genome degeneration and adaptation in a nascent stage of symbiosis. *Genome Biol Evol. 6*:351–361.

Pagel M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc R Soc Lond B Biol Sci. 255*:37–45.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics. 20*(2):289–290.

Patirio-Navarrete R, Moya A, Latorre A, Peretó J. 2013. Comparative genomics of *Blattabacterium cuenoti*: the frozen legacy of an ancient endosymbiont genome. *Genome Biol Evol. 5*(2):351–361.

Pettersson ME, Berg OG. 2007. Muller’s ratchet in symbiont populations. *Genetica. 130*(2):199–211.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol. 3*(2):217–223.

Rispe C, Moran NA. 2000. Accumulation of deleterious mutations in endosymbionts: muller’s ratchet with two levels of selection. *Ann Natl. 156*(4):425–441.

Sabree ZL, Huang CY, Arakawa G, Tokuda G, Lo N, Watanabe H, Moran NA. 2012. Genome shrinkage and loss of nutrient-providing potential in the obligate symbiont of the primitive termite *Mastotermes darwininensis*. *Appl Environ Microbiol. 78*(1):204–210.

Sabree ZL, Kambhampati S, Moran NA. 2009. Nitrogen recycling and nutritional provisioning by *Blattabacterium*, the cockroach endosymbiont. *Proc Natl Acad Sci U S A. 106*(46):19521–19526.

Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol. 6*:526–538.

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.

Sloan DB, Nakabachi A, Richards S, Qi J, Murali SC, Gibbs RA, Moran NA. 2014. Parallel histories of horizontal gene transfer facilitated extreme reduction of endosymbiont genomes in sap-feeding insects. *Mol Biol Evol. 31*(4):857–871.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics. 30*(9):1312–1313.

Strimmer K. 2008. fdttools: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics. 24*(12):1461–1462.

Sudakaran S, Kost C, Kaltenpoth M. 2017. Symbiont acquisition and replacement as a source of ecological innovation. *Trends Microbiol. 25*(5):375–390.

Suyama M, Torres D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res. 34*(Web Server Issue):W609–W612.

Tof C, Andersson SGE. 2010. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat Rev Genet. 11*(7):465–475.

Tokuda G, Elbourne LDH, Kinjo Y, Saitho S, Sabree Z, Hojo M, Yamada A, Hayashi Y, Shigenobu S, Bandi C, et al. 2013. Maintenance of essential amino acid synthesis pathways in the *Blattabacterium cuenoti* symbiont of a wood-feeding cockroach. *Biolett. 9*(3):20121153.

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Hutterenower C, Segata N. 2015. MetaPhAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods. 12*(10):902–903.

Vicente CS, Mondal SI, Akter A, Ozawa S, Kikuchi T, Hasegawa K. 2018. Genome analysis of new *Blattabacterium* spp., obligatory endosymbionts of *Periplaneta fuliginosa* and *P. japonica*. *PloS One 13*(7):e0200512.

Williams LE, Wennergren JJ. 2015. Genome evolution in an ancient bacteria-ant symbiotic parallel gene loss among *Blochmannia* spanning the origin of the ant tribe Camponotini. *PeerJ. 3*:e881.