Adaptation by Deletogenic Replication Slippage in a Nascent Symbiont

Adam L. Clayton,*[,1] D. Grant Jackson,[1] Robert B. Weiss,[2] and Colin Dale[1]

[1]Department of Biology, University of Utah

[2]Department of Human Genetics, University of Utah

***Corresponding author:** E-mail: aclayto@gmail.com.

Associate editor: James McInerney

Data deposition: The following genomes were used in this study: *Sodalis praecaptivus* str. HS1 (accession nos. CP006569 and CP006570), *Candidatus* Sodalis pierantonius str. SOPE (accession no. CP006568), *Sodalis glossinidius* (accession nos. AP008232, AP008233, AP008234, AP008235, and DQ785801), Secondary endosymbiont of *Ctenarytaina eucalypti* (accession no. CP003546), *Wigglesworthia glossinidia* (accession no. NC_016893), *Candidatus* Blochmannia pennsylvanicus str. BPEN (accession no. NC_007292).

Abstract

As a consequence of population level constraints in the obligate, host-associated lifestyle, intracellular symbiotic bacteria typically exhibit high rates of molecular sequence evolution and extensive genome degeneration over the course of their host association. While the rationale for genome degeneration is well understood, little is known about the molecular mechanisms driving this change. To understand these mechanisms we compared the genome of *Sodalis praecaptivus*, a nonhost associated bacterium that is closely related to members of the *Sodalis*-allied clade of insect endosymbionts, with the very recently derived insect symbiont *Candidatus* Sodalis pierantonius. The characterization of indel mutations in the genome of *Ca*. Sodalis pierantonius shows that the replication system in this organism is highly prone to deletions resulting from polymerase slippage events in regions encoding G+C-rich repetitive sequences. This slippage-prone phenotype is mechanistically associated with the loss of certain components of the bacterial DNA recombination machinery at an early stage in symbiotic life and is expected to facilitate rapid adaptation to the novel host environment. This is analogous to the emergence of mutator strains in both natural and laboratory populations of bacteria, which tend to reach high frequencies in clonal populations due to linkage between the mutator allele and the resulting adaptive mutations.

Key words: **symbiosis, sodalis, slippage, mutator, evolution.**

Introduction

Obligate mutualistic symbiotic bacteria are found in an estimated 10% of all insects (Buchner 1965; Douglas 1989). These symbionts perform a wide range of functions for their insect hosts, ranging from nutritional supplementation of the host diet (Douglas 1998; Baumann 2005), to host protection from parasites (Oliver et al. 2003), pathogens (Scarborough et al. 2005; Hedges et al. 2008; Teixeira et al. 2008), and environmental stress (Russell and Moran 2006). Although the functions of symbionts vary between hosts, symbiont genome evolution generally follows the same degenerative trend. Over long periods of time symbiont genomes reduce dramatically in size compared with their free-living relatives (Andersson and Kurland 1998; Dale and Moran 2006). This process of degeneration results from a gradual erosion and elimination of genes that are nonessential in the insect host, which is a nutritionally rich and static environment (Moran 1996). Gene loss is anticipated to be accelerated by the loss of recombination and repair enzymes, that presumably increases the frequency of mutations leading to gene inactivation and deletion (Moran and Wernegreen 2000; Rocha 2003). The loss of DNA recombination machinery is also expected to prevent symbionts from acquiring novel DNA via horizontal gene transfer. As such, their isolation leads to irreversible and extensive gene loss.

For example, the genomes of long established symbionts are extremely small and contain few, if any inactivated genes (pseudogenes) or mobile DNA (phage or insertion sequences; IS-elements). However, the genomes of recently derived symbionts are more similar to their free-living relatives in terms of size, but often contain large numbers of pseudogenes, phage, and IS-elements (Parkhill et al. 2003; Moran and Plague 2004; Plague et al. 2008). *Candidatus* Sodalis pierantonius str. SOPE is the primary endosymbiont of the rice weevil, *Sitophilus oryzae*. Its genome is close to the size of free-living relatives (4.5 Mb), but it maintains an abundance of pseudogenes and mobile DNA in the form of phage and IS-elements, such that the functional component of its genome is actually reduced to less than half of its original coding capacity (Oakeson et al. 2014). The role of *Ca.* Sodalis pierantonius involves nutritional supplementation, with recent studies showing that it provides the aromatic amino acid precursors of DOPA (tyrosine and phenylalanine) that the insect incorporates into its cuticle during development (Vigneron et al. 2014).

Article

Open Access

In order to understand the state of genome degeneration in *Ca.* Sodalis pierantonius, we previously compared its genome to that of *Sodalis praecaptivus* str. HS1, which is a very close (but nondegenerate) relative, originally isolated from a human wound (Clayton et al. 2012). Since *Ca.* Sodalis pierantonius and *S. praecaptivus* differ by only 2% genome-wide nucleotide substitutions, it is possible to align their DNA sequences and determine the molecular basis of all gene-inactivating mutations that have taken place since the onset of symbiosis. Previous work showed that the most abundant mutations in the *Ca.* Sodalis pierantonius genome are insertions and deletions (indels; Clayton et al. 2012). In this study, we show that a large proportion of mutations in the *Ca.* Sodalis pierantonius genome arise from deletogenic DNA polymerase slippage events involving closely spaced, G + C-rich repeat sequences. While the mechanism of replication slippage is well understood, and slippage has been identified as a source of frameshifting indels in symbionts (Williams and Wernegreen 2013), its impact on the early stages of symbiont genome evolution has not been previously described. We propose that a large number of polymerase slippage events occur due to the loss of the RecFOR and SbcCD repair enzymes that participate in repairing stalled DNA replication forks.

Results and Discussion

In this study, we compared whole-genome sequences of a recently acquired insect symbiont (*Ca.* Sodalis pierantonius) with a free-living progenitor (*S. praecaptivus*) to determine the types of mutations that result in gene inactivation and loss in the process of symbiosis. Our analysis focused primarily on coding sequences shared between these organisms, because this allows us to differentiate between frameshifting and nonframeshifting indels and to understand the effect of such events on gene function and the trajectory of degenerative evolution in the symbiosis.

The Nature of Indel Mutations in *Ca.* Sodalis pierantonius

Alignments were generated for 1,601 putative coding sequences shared between *S. praecaptivus* and *Ca.* Sodalis pierantonius (Oakeson et al. 2014). Alignments that were free of 3′- and 5′-terminal deletions were then examined for the presence of internal indels. The mutations were then categorized according to their context (fig. 1), as indels of <3 bases (either uncategorized in context or associated with homopolymers), or as polymerase slippage events occurring in repeat sequences (either direct or gapped) or "quasi-repeats," which are imperfect repeats (defined here as indel termini of length ≥5 bases with a Hamming distance of < 0.3/base). Repeat sequences are known to facilitate indel mutations as a consequence of misalignment and slippage during replication (Lovett 2004). Numerous laboratory studies indicate that replication slippage occurs as a consequence of DNA polymerase stalling and dissociating from DNA, leading to the potential mispairing of template and nascent strands (Lovett et al. 1993). Once replication continues, either an insertion or deletion of repeat subunits can occur depending on the strand that is mispaired (Fresco and Alberts 1960; Levinson and Gutman 1987). These mutations are known to occur spontaneously at low frequency in wild type strains of *Escherichia coli*, however their frequency can be increased substantially by introducing mutations in numerous components of the DNA replication machinery (Saveson and Lovett 1997; Feschenko and Lovett 1998).

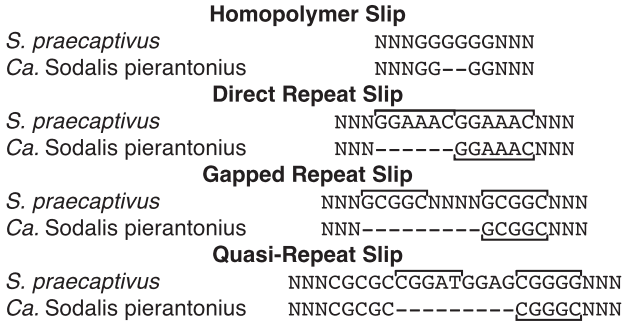

Fig. 1. Examples of intragenic indels identified from an alignment of all coding sequences shared between *Sodalis praecaptivus* and *Candidatus* Sodalis pierantonius. Most indels are the result of DNA polymerase slippage events associated with one of four repeat classes: homopolymer, direct, gapped, and quasi-repeats.

DNA polymerase slippage events in homopolymeric regions of the genome are known to be a common source of small (typically frameshifting) indels, especially in organisms with compositionally biased (A + T-rich) genomes (Moran et al. 2009; Williams and Wernegreen 2012, 2013). However, the data obtained from analyses of the alignments generated from the *S. praecaptivus* and *Ca.* Sodalis pierantonius orthologs (fig. 2) shows that the majority of intragenic indel mutations in *Ca.* Sodalis pierantonius have occurred in regions that contain direct, gapped, and quasi-repeat sequences, resulting in deletions ranging from 2 to 100 base pairs in size (fig. 3). While repeat sequences have long been known to catalyze polymerase slippage through slip-strand mispairing, they have not previously been reported to be a strong driver of gene inactivation and genome erosion in natural populations. Alongside homopolymers, short simple sequence repeats (<6 bases) have been identified as a source of indel mutations in the carpenter ant symbiont, *Candidatus* Blochmannia spp. However, very few of these indels were found to inactivate genic reading frames (Williams and Wernegreen 2012, 2013).

In natural populations of bacteria, the expansion and contraction of short direct repeat motifs in genic and promoter sequences facilitate a phenomenon known as phase variation in which rapid and reversible genetic changes optimize survival in a constantly changing environment (van der Woude and Bäumler 2004; Bayliss and Palmer 2012). Repeat sequences involved in phase variation are typically short (<10 bases) and exist in tandem arrays (Saunders et al. 1998; Maxon et al. 2006; Bayliss and Palmer 2012). The expansion or contraction of these repeats catalyzed by polymerase slippage, provides a means to modulate transcription or change the protein-coding sequence of a gene

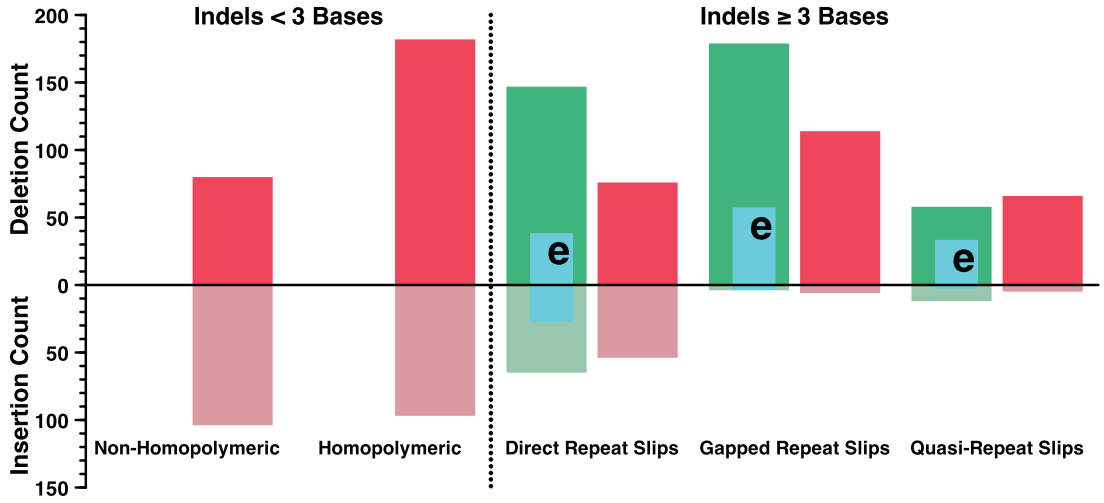

Fig. 2. Counts of intragenic indels in *Candidatus* Sodalis pierantonius. Deletions and insertions are shown above and below the *x* axis, respectively. Frameshifting indels are shown in red and those that preserve the integrity of the reading frame are shown in green. Note that there is a bias toward deletions among indels of size ≥3 bases. The inner blue bars show the expected numbers of nonframeshifting indels based on the assumption of a 2:1 ratio of frameshifting to nonframeshifting indels under neutrality in natural selection.

(Hallet 2001). For example, in *Haemophilus influenzae*, a 5′ variable length tetramer of CAAT nucleotides is responsible for maintaining or shifting the reading frame of the *licA* gene, involved in modifying LPS with phosphorylcholine allowing *H. influenza* to evade immune detection (Weiser et al. 1989; Humphries and High 2002). In *Bordetella pertussis* a variable length homopolymeric tract of cytosine bases upstream of the fimbrial protein-coding gene, *fim3*, regulates the level of expression by modulating the distance between RNA polymerase and a transcriptional activator. Altered levels of *fim3* expression also provide a means of immune evasion (Willems et al. 1990; Chen et al. 2010). It is reasonable to assume that the switch to a static, obligate symbiotic lifestyle would obviate the need for a bacterium to engage in phase variation, at least at the level of providing antigenic or metabolic diversity to cope with a changing environment or host. However, polymerase slippage could be useful to introduce mutations that inactivate genes that have no benefit in the symbiotic lifestyle but might prove costly in terms of their expression. In this sense, the switch to a symbiotic lifestyle could be considered an extreme case of phase variation, albeit one in which the adaptive benefit of slippage would be to inactivate gene functions. Indeed, the high frequency of repeat-associated deletions in *Ca.* Sodalis pierantonius (which exceeds the frequency of nonsense mutations) is consistent with the notion that the replication complex has evolved an increased propensity to undergo deletogenic slippage in the proximity of direct, gapped, and quasi-repeats.

The Role of Polymerase Slippage in Gene Inactivation

Inspection of our data set revealed that slips in homopolymeric sequences was constrained to 1–2 bases in size and therefore always result in the disruption of the genic reading frame. We found no examples in which two or more homopolymeric slips were located in close proximity (within 100 bases) such that they restored the genic reading frame in a compensatory fashion. Larger indels, mediated by polymerase slippage between direct repeats, gapped repeats, and quasi-repeats only cause frameshifts when the size of the indel is not divisible by three. Thus, under a scenario of neutrality (in which a protein-coding gene has no selective benefit), the expected ratio of frameshifting to nonframeshifting indels is 2:1. However, since some proportion of the genes in the genome are beneficial or necessary for survival, cells that obtain frameshifting mutations in those genes should be removed from the population by natural selection, increasing the relative ratio of nonframeshifting to frameshifting indels. In the case of direct repeats this is difficult to test, because it is likely that some proportion of these repeats function in phase variation and are therefore "programmed" to induce indels whose size is divisible by three. However, the gapped and quasi-repeats detected in our analysis are not anticipated to function in phase variation and therefore should provide a more reliable marker for determining the action of selection on slippage. Notably, all three classes of repeat-derived slips detected in our study have a higher than expected ratio of nonframeshifting indels (fig. 2). This highlights the trade-off between adaptive benefit and deleterious consequences that is often observed with an elevated mutation rate (Giraud et al. 2001).

Deletogenic Bias of Replication Slippage Events

Our data show that polymerase slippage events in *Ca.* Sodalis pierantonius have a strong deletional bias (fig. 2). Notably, deletions occur almost exclusively in the case of gapped-repeat and quasi-repeats slips, which makes sense because these slips are not anticipated to play a role in adaptive phase variation. At a mechanistic level, deletions are predicted to occur when DNA lesions or secondary structure forms in the template strand causing replication fork stalling (Levinson and Gutman 1987; Cox et al. 2000; Bikard et al. 2010). If the source of the stalling event cannot be resolved via recombination and repair pathways, then replication can only be resumed if the nascent strand pairs with a complementary

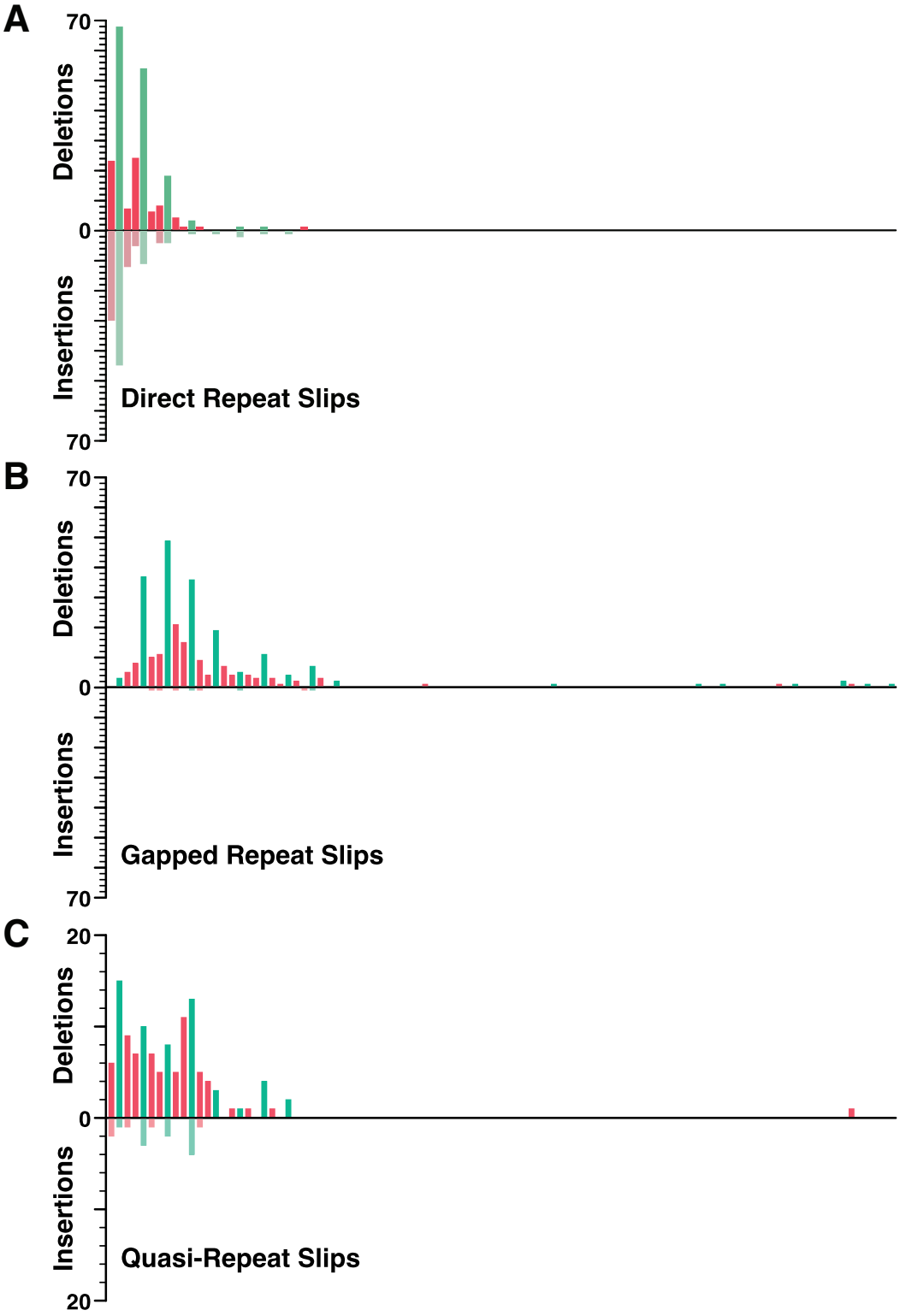

FIG. 3. Distribution of indel sizes resulting from intragenic direct (*A*), gapped (*B*), and quasi-repeat (*C*) slippage events. Deletions and insertions are shown above and below the *x* axis, respectively. The first column in each plot represents indels of 2 bp in length, with each subsequent column representing a 1 bp increase in indel size. Frameshifting indels are shown in red and those that preserve the integrity of the reading frame are shown in green.

sequence downstream. In the case of *Ca.* Sodalis pierantonius, it appears that those sequences do not need to be exactly complementary, given that many quasi-repeat slips were observed in our analysis. With respect to the ability to repair stalled replication forks, it is notable that *Ca.* Sodalis pierantonius lacks functional copies of a number of genes that are anticipated to play a role in recombination and repair, including the *recFOR* recombination pathway (fig. 4). This pathway

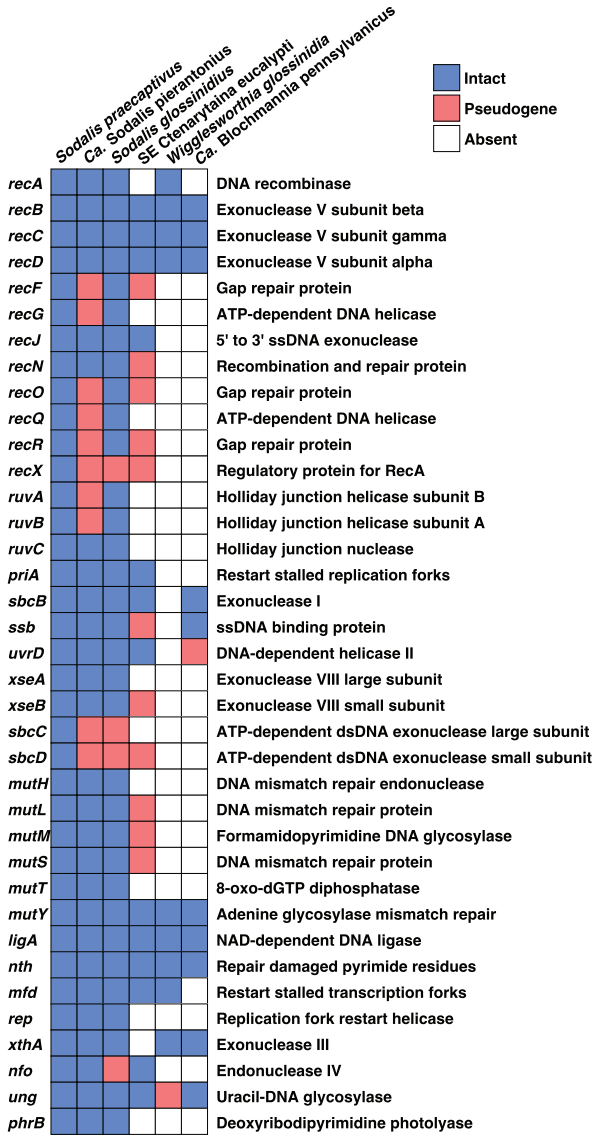

Fig. 4. Genes in recombination and repair pathways have been inactivated and lost in many symbionts.

is involved in the repair of stalled replication forks by recruiting RecA for recombination between the nascent DNA and a homologous chromosome (Kowalczykowski 2000; Morimatsu et al. 2012; Lenhart et al. 2014). The loss of the *recFOR* pathway in *Ca.* Sodalis pierantonius implies that the cell no longer retains the ability to repair a stalled replication fork by homologous recombination, forcing an outcome in which replication must be reinitiated by the nonconservative single strand annealing process.

The SbcCD nuclease (fig. 5) functions to repair replications forks that stall as a result of hairpins and cruciform formation in the DNA. SbcCD cleaves the hairpin or cruciform allowing resection (deletion) of the sequence involved in the secondary structure followed by annealing and ligation of the free ends (Bzymek and Lovett 2001). Curiously, genes encoding *sbcC* and *sbcD* are duplicated in the chromosome of *Ca.* Sodalis pierantonius, however analysis of the sequence reveals that *sbcC* was inactivated prior to its duplication, based on the fact that both copies of this gene maintain several shared inactivating mutations (fig. 5). The genomic region encompassing the duplication contains 29 genes (22 excluding IS-elements), and only two of these remain intact in both copies, one of which is *aroL* (fig. 5). The *aroL* gene encodes shikimate kinase II that functions as a rate-limiting step in aromatic amino acid biosynthesis (Dell and Frost 1993; Krämer et al. 2003). The duplication of *aroL* is therefore expected to be of key importance in the symbiosis, since *Ca.* Sodalis pierantonius was recently shown to provide its host with tyrosine and phenylalanine (Vigneron et al. 2014). Based on this evidence it seems likely that the adaptive value of this duplication lies in the amplification of *aroL* and that *sbcC* and *sbcD* were simply neighboring genes that were hitchhiked in the duplication event. With respect to the mutation phenotype described in this study, it should be noted that the loss of the SbcC nuclease is expected to prevent the breakage of DNA at stalled replication forks that occur during replication (Leach 1994), thereby increasing the number of slip-strand mispairing opportunities associated with polymerase stalling. It has also been hypothesized that mutations in components of the recombination machinery (including RecF) lead to an increased frequency of chromosome breaks that can catalyze duplication of chromosomal regions (Reams and Roth 2015). In addition to the example highlighted in this study, *Ca.* Sodalis pierantonius is known to maintain several other duplicated genomic regions including one that harbors the genes encoding the GroEL chaperone, which is thought to also provide an adaptive benefit (Oakeson et al. 2014).

DNA Polymerase Slippage Occurs Preferentially at G + C-Rich Sites

Further analysis of the sequence context of intragenic repeat associated deletions revealed that both the repeats and their intervening sequences (in the case of gapped and quasi-repeats) are substantially biased in composition toward increased G + C, relative to a chromosomal mean of 57.1% G + C (fig. 6). Indeed, the so-called "quasi-repeats" are classified as such by virtue of the fact that they are G + C-rich. In order to determine if G + C richness is correlated with deletions of larger sizes (>100 bases) in the genome of *Ca.* Sodalis pierantonius, we analyzed the G + C-content of all gaps in a global alignment of the *Ca.* Sodalis pierantonius and *S. praecaptivus* genome sequences generated using the NUCmer algorithm (Delcher et al. 2002). Analysis of the base composition of these sequences revealed that only fragments of < 600 bases in length are significantly biased towards increased G + C-content (fig. 7 and table 1). This fits with the notion that the size of the replication fork constrains the amount of sequence that is available for slippage to several hundred bases in length (Lovett 2004).

It is interesting to note that it is not just the termini of deleted fragments that are G + C-rich, but often the entire deleted fragment (fig. 7). This suggests that G + C-rich DNA between the gapped and quasi-repeat sequences plays a role in facilitating replication slippage events. This is expected based on the fact that G + C-rich sequences are known to induce polymerase stalling and slippage, specifically when they are located in secondary structures in the template

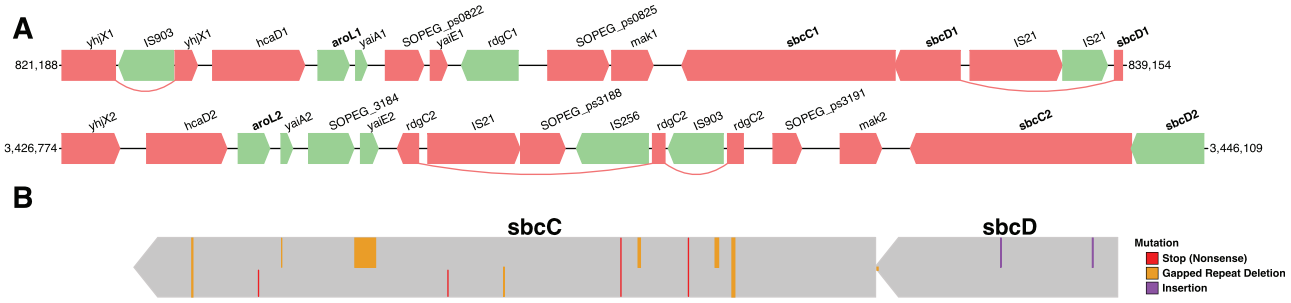

Fig. 5. Duplication of the chromosomal region encoding *aroL* and *sbcCD* in *Candidatus* Sodalis pierantonius. Intact genes and those containing inactivating mutations are shown in green and red respectively. The duplication includes *aroL*, the rate-limiting step in aromatic amino acid biosynthesis (*A*). Mutations present in each of the two copies of *sbcC* and *sbcD* in *Ca.* Sodalis pierantonius are depicted in the diagram (*B*) as tick marks in the top and bottom sections of the image. Connected tick marks reveal shared mutations that likely arose in *sbcC* prior to duplication.

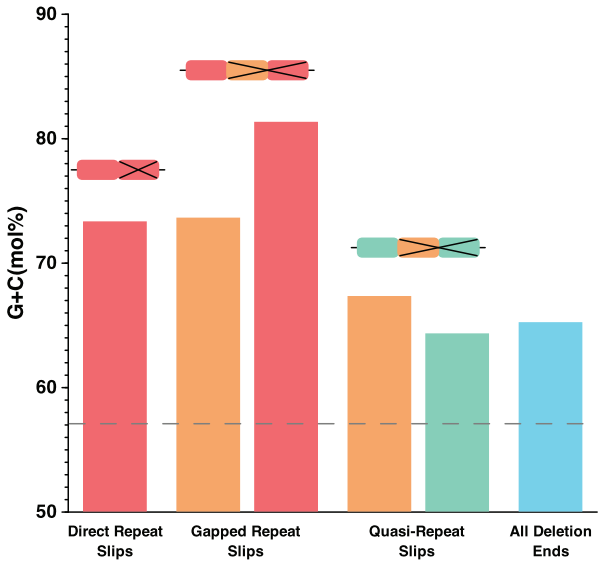

Fig. 6. G + C-content of deletions in *Candidatus* Sodalis pierantonius. The mean G + C-content of the *Sodalis praecaptivus* genome is represented by the gray dashed line. The diagrams above the bars show the slip prone structures with repeats shown in red, quasi-repeats shown in green and intervening DNA shown in orange. The "X" represents the region deleted in each of these slip-prone structures. All of the slip-prone structures have a G + C-content that is substantially elevated above the genomic mean. Furthermore, the G + C-content of the terminal 10 bases of larger deletions (>100 bases; identified using NUCmer) also shows increased G + C-content.

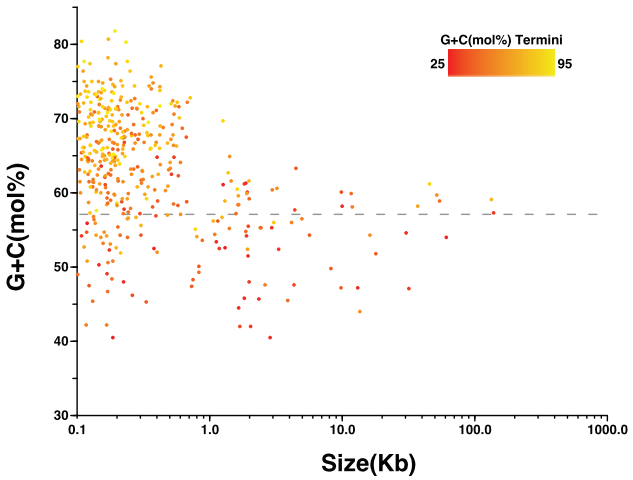

Fig. 7. G + C-content of NUCmer predicted deletion fragments and their 10 base termini. Position on the *y* axis shows the G + C-content of the complete deleted sequence. The color of each point shows the G + C-content of the terminal 10 bases of the deleted sequence. Note that a large proportion of the smaller deletion fragments (<600 bases) are G + C-rich and have G + C-rich termini. The dotted line illustrates the chromosomal mean G + C-content of *S. praecaptivus* (57.1%).

Table 1. G + C-Content of Terminal Sequences of NUCmer Predicted Deletion Fragments.

Terminal Bases	10	15	20	25
100–1,000 bases	67.6%	67.1%	67.2%	67.1%
>1,000 bases	53.6%	53.6%	54.0%	54.6%

DNA (Viguera et al. 2001). In laboratory based studies involving *E. coli* it has been observed that inverted repeats promote slips by inducing hairpin and cruciform structures that cause the replication machinery of the cell to stall, dissociate, and slip inside the replication fork (Bzymek and Lovett 2001). It is also likely that the interstrand pairing events that facilitate the reinitiation of replication are more stable for G + C-rich sequences.

Is Replication Slippage Adaptive at the Onset of Symbiosis?

The genomes of endosymbiotic bacteria are known to undergo severe degeneration, primarily as a consequence of a low effective population size that results from frequent population bottlenecks during host reproduction (Wernegreen 2002; Kuo et al. 2009; Batut et al. 2014). In the absence of recombination, low effective population size exacerbates the effect of Muller's ratchet, in which irreversible deleterious mutations accumulate over time, leading ultimately to the extinction of the bacterial population as a consequence of mutational meltdown (Lynch et al. 1993). However, many extant endosymbiotic associations are known to be ancient in origin and the bacterial partners in these associations have attained extremely small genome sizes (Tamas et al. 2002; Nakabachi et al. 2006; Bennett and Moran 2013), indicating that they are functionally stable in

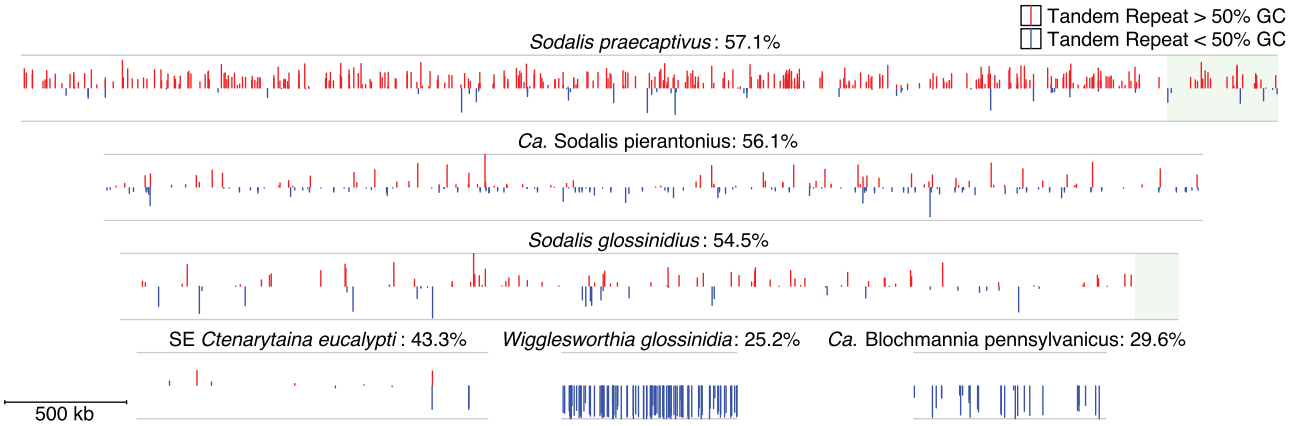

Fig. 8. G + C-content of direct repeats in *Sodalis*-allied symbionts. Genomes are linearized and scalar relative to the *S. praecaptivus* genome with tick marks indicating the location of direct repeats. Plasmids are highlighted in light green. G + C-content of each direct repeat is also shown above (red) or below (blue) 50% G + C. The genome average G + C-content is also shown.

spite of such a high genetic load. This leads to the hypothesis that there is a second level of selection, imposed by hosts, validating the functionality of individual populations of symbionts (Pettersson and Berg 2007). This is anticipated to prevent the accumulation of symbiont mutations that negatively impact host fitness (O'Fallon 2008), facilitating the functional stability of symbiotic associations that is apparent in nature. This selective force should also be capable of increasing the efficiency of the symbiotic association by selecting for mutations in the symbiont genome that improve mutualistic functions. An example of this is described earlier in the current study and involves the duplication of the chromosomal region encoding shikimate kinase II (fig. 5). This amplification is expected to have occurred after the onset of genome degeneration in the symbiont because there are a number of pseudogenes in this duplicated region with shared degenerative mutations that most likely occurred prior to the duplication event.

Gene inactivation and deletion have the potential to increase fitness (Hottes et al. 2013) by streamlining the gene inventory and removing functions that have a high cost to benefit ratio (Koskiniemi et al. 2012). In a symbiotic relationship it is conceivable that hosts might select for populations of symbionts with elevated mutation rates that undergo more rapid streamlining at the onset of the association, when the symbiont has the largest number of dispensable genes. This could be achieved by selecting mutant strains of bacteria that have elevated mutation rates (designated as mutators), as a consequence of the loss of genes involved in DNA repair. While these mutator strains have no intrinsic adaptive value, they can fix or reach high frequencies in populations as a consequence of the adaptive mutations that they generate, which become linked to the mutator allele (Wielgoss et al. 2013). Notably, mutator phenotypes often arise when bacteria encounter a new environment in which the potential for adaptive mutations is high and mutation rate is a limiting factor in adaptation (Arjan et al. 1999). The loss of genes encoding DNA recombination and repair machinery often occurs at an early stage following the onset of a symbiotic relationship (Moran et al. 2008), as evidenced in this study with the loss of the RecFOR recombination system (Dale et al. 2003). Such loss is anticipated to be irreversible in the asexual lifestyle in which symbionts are isolated from opportunities to engage in lateral gene transfer.

The Fate of G + C-Rich DNA in Insect Endosymbionts

Our data clearly show that polymerase slippage events in G + C-rich repeat regions of the genome are responsible for numerous deletions and gene inactivating mutations in *Ca.* Sodalis pierantonius. Left unchecked, this process would be expected to progressively erode these repeats until only those genes that are essential remain. In order to determine the fate of such repeats in the genomes of *Sodalis*-allied symbionts, we analyzed the frequencies of direct repeats in symbionts that are at different stages in the process of genome degeneration (fig. 8). As expected, both *Ca.* Sodalis pierantonius and *Sodalis glossinidius* (another recently derived symbiont) have substantially fewer G + C-rich repeats than *S. praecaptivus*, despite having very similar genomic G + C-content (fig. 8 and table 2). In addition, the mean copy number of direct repeat units that remain in each genome is reduced in *Ca.* Sodalis pierantonius (3.8 copies/kb) relative to *S. praecaptivus* (5.02 copies/kb), supporting the notion that slippage has a deletogenic bias in *Ca.* Sodalis pierantonius. This is consistent with the results of our study, indicating that G + C-rich repeats are a hotspot for deletions. A + T-rich repeats were found to be much less frequent in all three genomes and notably, the frequencies of these repeats are very similar between *S. praecaptivus*, *Ca.* Sodalis pierantonius, and *S. glossinidius* (Table 2), with the caveat that the number of A + T-rich repeats in *Ca.* Sodalis pierantonius is elevated by the presence of an A + T-rich repeat in an IS-element that is only abundant in this symbiont (Oakeson et al. 2014).

The genome of the secondary symbiont of the psyllid, *Ctenarytaina eucalypti*, which has a reduced G + C-content (43.3%) and genome size (1.4-Mb), contains very few G + C-rich or A + T-rich direct repeats. The highly reduced genomes of *Ca.* Blochmannia pennsylvanicus and *Wigglesworthia glossinidia*, also have no G + C-rich direct repeats, which can readily be explained by their low G + C-content (<30%).

Table 2. Density of Genome Wide G + C-rich and A + T-Rich Direct Repeats.

	G + C-rich repeats/kb	A + T-rich repeats/kb	Genomic G + C (%)
Sodalis praecaptivus	0.075	0.011	57.1
Ca. Sodalis pierantonius	0.019	0.032	56.1
Sodalis glossinidius	0.017	0.007	54.5
SE *Ctenarytaina eucalypti*	0.004	0.003	43.3
Wigglesworthia glossinidia	0.000	0.160	25.2
Ca. Blochmannia pennsylvanicus	0.000	0.035	29.6

However, both of these genomes maintain substantial numbers of A + T-rich repeats, indicating that when A + T-rich repeats arise as a consequence of the A + T-bias that is common in ancient symbionts, they appear to be stable. Taken together this data suggest that G + C-rich repeats, but not A + T-rich repeats, are essentially mutagenic because they cause slip-induced deletions in the absence of the RecFOR repair system.

The collapse of G + C-rich repeats cannot itself be responsible for the A + T-bias that often occurs in the genomes of ancient symbionts because the composition bias exists genome wide. However, the emergence of an A + T-bias is expected to ameliorate the effects of deleterious mutations arising from slippage in G + C-rich regions. The adaptive value of a mutator phenotype is predicted to be greatest at the onset of the symbiotic association when a large proportion of genes are evolving under relaxed selection and there is great potential for adaptive mutations (Sniegowski et al. 1997; Giraud et al. 2001; Elena and Lenski 2003; Wielgoss et al. 2013). In addition, mutators are predicted to be beneficial (and therefore most likely to achieve fixation) in environments that are static and therefore amenable to extreme specialization (Giraud et al. 2001).

While mutators are often selected in response to a change in environment, their fitness advantage is expected to decline over time as the population adapts to its new environment. At some point, presumably when the ratio of dispensable to adaptive genes drops below some critical threshold, it becomes selectively advantageous to reduce the deleterious burden of mutation. In simple terms this could be achieved by restoring the status of a dysfunctional gene by mutational reversion or lateral gene transfer (Denamur and Matic 2006). However, in practice, a complete reversion is not likely to be selectively advantageous given that a partial reduction in the mutation rate might suffice to generate a more favorable equilibrium between adaptive potential and deleterious consequences. In this case, a mutator lineage might be partially "rescued" by compensatory mutation(s) that reduce the severity of the mutator phenotype. For example, Wielgoss et al. (2013) documented the fixation of a *mutT* mutator strain in an experimental system that had a point mutation rate approximately 150-fold higher than wild type. This strain was later complemented by a mutation in *mutY* that reduced the mutation rate by 40–60%.

One of the most striking degenerative changes in the genomes of insect symbionts involves the loss of many genes encoding DNA repair and recombination functions (Moran et al. 2008). These genes appear to be lost progressively over the course of the symbiotic association (fig. 4), and this may be a consequence of epistatic (complementary) effects. It is also conceivable that the effects of a mutator phenotype could be mitigated by changes in the biology of the symbiosis. One possibility is a change in the transmission dynamics of the symbiont to lessen the potency of Muller's ratchet. Another is a reduction in the frequency of symbiont replication and an increase in cellular longevity, thereby reducing the number of mutations that arise during replication. Yet another option is for the host to provide novel replication, recombination, and repair functions, potentially rescuing the hapless situation that ensues in ancient symbionts, which, notably, also lack many vital components of the DNA polymerase III machinery (Moran et al. 2008).

Materials and Methods

Ortholog Identification and Analysis

Sequence alignments of all orthologous genes in *S. praecaptivus* and *Ca.* Sodalis pierantonius were generated using a Smith–Waterman alignment algorithm implemented in cross_match (Gordon et al. 1998). Alignments were analyzed both computationally (using custom processing scripts) and manually to identify indels. Each indel was classified according to its sequence context and analyzed using custom Java scripts designed to detect direct and gapped repeat sequences.

The coordinates of duplicated regions of the *Ca.* Sodalis pierantonius genome were identified in a previous study (Oakeson et al. 2014). Mutations were identified in each copy of the *Ca.* Sodalis pierantonius *sbcC* and *sbcD* genes by alignment with the intact *S. praecaptivus* orthologs using Geneious version 8.1.7 (Kearse et al. 2012).

G + C-Content of Large Deletions

Large deletions were identified in the *Ca.* Sodalis pierantonius genome as "GAP" mutations using the NUCmer algorithm of the MUMmer software package (Delcher et al. 2002). The resulting output file was then curated to include only deletions > 100 bases in size. The G + C-content of each deletion and its 10 base termini was calculated using custom perl scripts.

Direct Repeat Comparisons

Genome-wide direct repeats were identified in each genome using Tandem Repeats Finder version 4.07b (Benson 1999) using the following settings: Alignment parameters (match, mismatch, indels) = (2, 7, 7), minimum alignment score = 50, maximum period size = 500. The resulting output was manually curated to remove nested repeats. The complete data set is available in supplementary table S1, Supplementary Material online. The genomic context and G + C-content of each repeat were determined using custom perl scripts. Direct repeats from *S. praecaptivus* were used to query the *Ca.* Sodalis pierantonius genome using a combination of custom perl scripts, BLASTn (Camacho et al. 2008), and the Mauve

plugin (Darling et al. 2010) as implemented in Geneious version 8.1.7 (Kearse et al. 2012).

Supplementary Material

Supplementary table S1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the National Institutes of Health (grant number 1R01AI095736 awarded to C.D.).

References

Andersson SG, Kurland CG. 1998. Reductive evolution of resident genomes. *Trends Microbiol*. 6:263–268.

Arjan JA, Visser M, Zeyl CW, Gerrish PJ, Blanchard JL, Lenski RE. 1999. Diminishing returns from mutation supply rate in asexual populations. *Science* 283:404–406.

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

Batut B, Knibbe C, Marais G, Daubin V. 2014. Reductive genome evolution at both ends of the bacterial population size spectrum. *Nat Rev Microbiol*. 12:841–850.

Bayliss CD, Palmer ME. 2012. Evolution of simple sequence repeat-mediated phase variation in bacterial genomes. *Ann N Y Acad Sci*. 1267:39–44.

Bennett GM, Moran NA. 2013. Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect. *Genome Biol Evol*. 5:1675–1688.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 27:573–580.

Bikard D, Loot C, Baharoglu Z, Mazel D. 2010. Folded DNA in action: hairpin formation and biological functions in prokaryotes. *Microbiol Mol Biol Rev*. 74:570–588.

Buchner P. 1965. Endosymbiosis of animals with plant microorganisms. [Revised English Translation by Bertha, Mueller with the collaboration of Frances H. Fockler, editor]. New York: Interscience Publishers.

Bzymek M, Lovett ST. 2001. Evidence for two mechanisms of palindrome-stimulated deletion in *Escherichia coli*: single-strand annealing and replication slipped mispairing. *Genetics* 158:527–540.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2008. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Chen Q, Decker KB, Boucher PE, Hinton D, Stibitz S. 2010. Novel architectural features of *Bordetella pertussis* fimbrial subunit promoters and their activation by the global virulence regulator BvgA. *Mol Microbiol*. 77:1326–1340.

Clayton AL, Oakeson KF, Gutin M, Pontes A, Dunn DM, von Niederhausern AC, Weiss RB, Fisher M, Dale C. 2012. A novel human-infection-derived bacterium provides insights into the evolutionary origins of mutualistic insect-bacterial symbioses. *PLoS Genet*. 8:e1002990.

Cox MM, Goodman MF, Kreuzer KN, Sherratt DJ, Sandler SJ, Marians KJ. 2000. The importance of repairing stalled replication forks. *Nature* 404:37–41.

Dale C, Moran NA. 2006. Molecular interactions between bacterial symbionts and their hosts. *Cell* 126:453–465.

Dale C, Wang B, Moran N, Ochman H. 2003. Loss of DNA recombinational repair enzymes in the initial stages of genome degeneration. *Mol Biol Evol*. 20:1188–1194.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5:e11147.

Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res*. 30:2478–2483.

Dell KA, Frost JW. 1993. Identification and removal of impediments to biocatalytic synthesis of aromatics from D-glucose: rate-limiting enzymes in the common pathway of aromatic amino acid biosynthesis. *J Am Chem Soc*. 115:11581–11589.

Denamur E, Matic I. 2006. Evolution of mutation rates in bacteria. *Mol Microbiol*. 60:820–827.

Douglas AE. 1989. Mycetocyte symbiosis in insects. *Biol Rev Camb Philos Soc*. 64:409–434.

Douglas AE. 1998. Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria Buchnera. *Annu Rev Entomol*. 43:17–37.

Elena SF, Lenski RE. 2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat Rev Genet*. 4:457–469.

Fresco JR, Alberts BM. 1960. The accommodation of noncomplimentary bases in helical polyribonucleotides and deoxyribonucleotides and deoxyribonucleic acids. *Proc Natl Acad Sci U S A*. 46:311–321.

Feschenko VV, Lovett ST. 1998. Slipped misalignment mechanisms of deletion formation: analysis of deletion endpoints. *J Mol Biol*. 276:559–569.

Giraud A, Matic I, Tenaillon O, Clara A, Radman M, Fons M, Taddei F. 2001. Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *Science* 291:2606–2608.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. *Genome Res*. 8:195–202.

Hallet B. 2001. Playing Dr Jekyll and Mr Hyde: combined mechanisms of phase variation in bacteria. *Curr Opin Microbiol*. 4:570–581.

Hedges LM, Brownlie JC, O'Neill SL, Johnson KN. 2008. *Wolbachia* and virus protection in insects. *Science* 322:702.

Hottes AK, Freddolino PL, Khare A, Donnell ZN, Liu JC, Tavazoie S. 2013. Bacterial adaptation through loss of function. *PLoS Genet*. 9:e1003617.

Humphries HE, High NJ. 2002. The role of *licA* phase variation in the pathogenesis of invasive disease by *Haemophilus influenzae* type b. *FEMS Immunol Med Microbiol*. 34:221–230.

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

Koskiniemi S, Sun S, Berg OG, Andersson DI. 2012. Selection-driven gene loss in bacteria. *PLoS Genet*. 8:e1002787.

Kowalczykowski SC. 2000. Initiation of genetic recombination and recombination-dependent replication. *Trends Biochem Sci*. 25:156–165.

Krämer M, Bongaerts J, Bovenberg R, Kremer S, Müller U, Orf S, Wubbolts M, Raeven L. 2003. Metabolic engineering for microbial production of shikimic acid. *Metab Eng*. 5:277–283.

Kuo CH, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. *Genome Res*. 19:1450–1454.

Leach DR. 1994. Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. *Bioessays* 16:893–900.

Lenhart JS, Brandes ER, Schroeder JW, Sorenson RJ, Showalter HD, Simmons LA. 2014. RecO and RecR are necessary for RecA loading in response to DNA damage and replication fork stress. *J Bacteriol*. 196:2851–2860.

Levinson G, Gutman GA. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol*. 4:203–221.

Lovett ST. 2004. Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol Microbiol*. 52:1243–1253.

Lovett ST, Drapkin PT, Sutera VA Jr, Gluckman-Peskind TJ. 1993. A sister-strand exchange mechanism for *recA*-independent deletion of repeated DNA sequences in *Escherichia coli*. *Genetics* 135:631–642.

Lynch M, Bürger R, Butcher D, Gabriel W. 1993. The mutational meltdown in asexual populations. *J Hered*. 84:339–344.

Maxon R, Bayliss C, Hood D. 2006. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu Rev Genet*. 40:307–333.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A*. 93: 2873–2878.
Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet*. 42:165–190.
Moran NA, McLaughlin HJ, Sorek R. 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* 323:379–382.
Moran NA, Plague GR. 2004. Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev*. 14:627–633.
Moran NA, Wernegreen JJ. 2000. Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends Ecol Evol*. 15:321–326.
Morimatsu K, Wu Y, Kowalczykowski SC. 2012. RecFOR proteins target RecA protein to a DNA gap with either DNA or RNA at the 5′ terminus: implication for repair of stalled replication forks. *J Biol Chem*. 287:35621–35630.
Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M. 2006. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* 314:267.
Oakeson 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*. 6:76–93.
O'Fallon B. 2008. Population structure, levels of selection, and the evolution of intracellular symbionts. *Evolution* 62:361–373.
Oliver KM, Russell JA, Moran NA, Hunter MS. 2003. Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. *Proc Natl Acad Sci U S A*. 100:1803–1807.
Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden MT, Churcher CM, Bentley SD, Mungall KL, et al. 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat Genet*. 35:32–40.
Pettersson ME, Berg OG. 2007. Muller's ratchet in symbiont populations. *Genetica* 130:199–211.
Plague GR, Dunbar HE, Tran PL, Moran NA. 2008. Extensive proliferation of transposable elements in heritable bacterial symbionts. *J Bacteriol*. 190:777–779.
Reams AB, Roth JR. 2015. Mechanisms of gene duplication and amplification. *Cold Spring Harb Perspect Biol*. 7:a016592.
Rocha EP. 2003. An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. *Genome Res*. 13:1123–1132.
Russell JA, Moran NA. 2006. Costs and benefits of symbiont infection in aphids: variation among symbionts and across temperatures. *Proc Biol Sci*. 273:603–610.
Saunders NJ, Peden JF, Hood DW, Moxon ER. 1998. Simple sequence repeats in the Helicobacter pylori genome. *Mol Microbiol*. 27:1091–1098.
Saveson CJ, Lovett ST. 1997. Enhanced deletion formation by aberrant DNA replication in *Escherichia coli*. *Genetics* 146:457–470.
Scarborough CL, Ferrari J, Godfray HC. 2005. Aphid protected from pathogen by endosymbiont. *Science* 310:1781.
Sniegowski PD, Gerrish PJ, Lenski RE. 1997. Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* 387:703–705.
Tamas I, Klasson L, Canbäck B, Näslund AK, Eriksson AS, Wernegreen JJ, Sandström JP, Moran NA, Andersson SG. 2002. 50 million years of genomic stasis in endosymbiotic bacteria. *Science* 296:2376–2379.
Teixeira L, Ferreira A, Ashburner M. 2008. The bacterial symbiont *Wolbachia* induces resistance to RNA viral infections in *Drosophila melanogaster*. *PLoS Biol*. 6:e2.
van der Woude MW, Bäumler AJ. 2004. Phase and antigenic variation in bacteria. *Clin Microbiol Rev*. 17:581–611.
Vigneron A, Masson F, Vallier A, Balmand S, Rey M, Vincent-Monégat C, Aksoy E, Aubailly-Giraud E, Zaidman-Rémy A, Heddi A. 2014. Insects recycle endosymbionts when the benefit is over. *Curr Biol*. 24:2267–2273.
Viguera E, Canceill D, Ehrlich SD. 2001. Replication slippage involves DNA polymerase pausing and dissociation. *Embo J*. 20:2587–2595.
Weiser JN, Love JM, Moxon ER. 1989. The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. *Cell* 59:657–665.
Wernegreen JJ. 2002. Genome evolution in bacterial endosymbionts of insects. *Nat Rev Genet*. 3:850–861.
Willems R, Paul A, van der Heide HG, ter Avest AR, Mooi FR. 1990. Fimbrial phase variation in *Bordetella pertussis*: a novel mechanism for transcriptional regulation. *Embo J*. 9:2803–2809.
Wielgoss S, Barrick JE, Tenaillon O, Wiser MJ, Dittmar WJ, Cruveiller S, Chane-Woon-Ming B, Médigue C, Lenski RE, Schneider D. 2013. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc Natl Acad Sci U S A*. 110:222–227.
Williams LE, Wernegreen JJ. 2012. Purifying selection, sequence composition, and context-specific indel mutations shape intraspecific variation in a bacterial endosymbiont. *Genome Biol Evol*. 4:44–51.
Williams LE, Wernegreen JJ. 2013. Sequence context of indel mutations and their effect on protein evolution in a bacterial endosymbiont. *Genome Biol Evol*. 5:599–605.