

# Lawrence Berkeley National Laboratory

## LBL Publications

### Title

Mutational Pressure Drives Differential Genome Conservation in Two Bacterial Endosymbionts of Sap-Feeding Insects.

### Permalink

<https://escholarship.org/uc/item/26711602>

### Journal

Genome Biology and Evolution, 13(3)

### Authors

Waneka, Gus  
Vasquez, Yumary  
Bennett, Gordon  
et al.

### Publication Date

2021-03-01

### DOI


10.1093/gbe/evaa254

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial License, available at <https://creativecommons.org/licenses/by-nc/4.0/>

Peer reviewed

# Mutational Pressure Drives Differential Genome Conservation in Two Bacterial Endosymbionts of Sap-Feeding Insects

Gus Waneka <sup>1,\*</sup>, Yumary M. Vasquez<sup>2</sup>, Gordon M. Bennett<sup>2</sup>, and Daniel B. Sloan <sup>1</sup>

<sup>1</sup>Department of Biology, Colorado State University, Fort Collins, CO, USA

<sup>2</sup>Department of Life and Environmental Sciences, University of California, Merced, CA, USA

\*Corresponding author: E-mail: gus.waneka@gmail.com.

Accepted: 28 November 2020

## Abstract

Compared with free-living bacteria, endosymbionts of sap-feeding insects have tiny and rapidly evolving genomes. Increased genetic drift, high mutation rates, and relaxed selection associated with host control of key cellular functions all likely contribute to genome decay. Phylogenetic comparisons have revealed massive variation in endosymbiont evolutionary rate, but such methods make it difficult to partition the effects of mutation versus selection. For example, the ancestor of Auchenorrhynchan insects contained two obligate endosymbionts, *Sulcia* and a betaproteobacterium (*BetaSymb*; called *Nasuia* in leafhoppers) that exhibit divergent rates of sequence evolution and different propensities for loss and replacement in the ensuing ~300 Ma. Here, we use the auchenorrhynchan leafhopper *Macrostelus* sp. nr. *severini*, which retains both of the ancestral endosymbionts, to test the hypothesis that differences in evolutionary rate are driven by differential mutagenesis. We used a high-fidelity technique known as duplex sequencing to measure and compare low-frequency variants in each endosymbiont. Our direct detection of de novo mutations reveals that the rapidly evolving endosymbiont (*Nasuia*) has a much higher frequency of single-nucleotide variants than the more stable endosymbiont (*Sulcia*) and a mutation spectrum that is potentially even more AT-biased than implied by the 83.1% AT content of its genome. We show that indels are common in both endosymbionts but differ substantially in length and distribution around repetitive regions. Our results suggest that differences in long-term rates of sequence evolution in *Sulcia* versus *BetaSymb*, and perhaps the contrasting degrees of stability of their relationships with the host, are driven by differences in mutagenesis.

**Key words:** mutation spectra, insertion, deletion, variant, genome decay, extinction.

## Significance

Two ancient endosymbionts in the same host lineage display stark differences in genome conservation over phylogenetic scales. We show that the rapidly evolving endosymbiont has a higher frequency of mutations, as measured with duplex sequencing. Therefore, differential mutagenesis likely drives evolutionary rate variation in this system.

## Introduction

Sap-feeding insects in the order Hemiptera feed exclusively on nutrient deficient plant-saps (phloem and xylem) and rely on bacterial endosymbionts to provide lacking essential amino acids and vitamins (Buchner 1965; Sandström and Moran 1999). Such endosymbionts are housed intracellularly in

specialized organs called bacteriomes and are subject to strict vertical transmission from mother to offspring. This repeated transmission bottlenecking and confinement within a host lineage results in genetic drift and endosymbiont genome decay (Moran 1996; Mira and Moran 2002; McCutcheon and Moran 2012; Bennett and Moran 2015; McCutcheon et al. 2019).

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Ancient endosymbiont genomes are characterized by their extreme nucleotide composition bias and small size. In fact, many of these genomes are among the smallest known, lacking genes considered essential in free-living bacteria (e.g., cell envelope biogenesis, gene expression regulation, and DNA replication/repair; McCutcheon and Moran 2012; Moran and Bennett 2014). Ancient endosymbiont genomes generally experience rapid sequence evolution and extensive lineage-specific gene deletions, leaving only a set of “core” genes necessary to sustain a functioning nutritional symbiosis (Moran et al. 2009; Bennett, McCutcheon, et al. 2016). At the same time, these genomes exhibit remarkable structural stability, often displaying perfectly conserved synteny between lineages that diverged 10s to 100s of millions of years ago (McCutcheon et al. 2009a).

Missing DNA replication and repair genes, and conditions of relaxed selection, allow the AT mutation bias (which is common to most bacteria; Hershberg and Petrov 2010; Hildebrand et al. 2010; Long, Sung, et al. 2018) to drive AT content to levels above 75% in many endosymbionts (Moran et al. 2009; Moran and Bennett 2014; Wernegreen 2015). Notable exceptions are found in the GC rich genomes of “*Candidatus Tremblaya princeps*” (58.8% GC) and “*Candidatus Hodgkinia cicadicola*” (hereafter *Hodgkinia*; 58.4% GC; Baumann et al. 2002; McCutcheon and Von Dohlen 2011). *Hodgkinia* is an endosymbiont of cicadas with an extremely small genome (144 kb) (McCutcheon et al. 2009b). However, population level estimates of the *Hodgkinia* mutation spectrum reveal an AT mutation bias, indicating that GC prevalence in *Hodgkinia* likely results from selection on nucleotide identity (Van Leuven and McCutcheon 2012).

In extreme cases of genome decay, ancient endosymbionts can go extinct when they are outcompeted and replaced by newly colonizing bacteria or yeast (Matsuura et al. 2018; McCutcheon et al. 2019). Comparisons among ancient endosymbionts reveal massive variability in genome stability and extinction rates. Insects in the suborder Auchenorrhyncha maintain two bacterial endosymbionts: “*Candidatus Sulcia muelleri*” (hereafter *Sulcia*) and a betaproteobacterium (hereafter *BetaSymb*), which are both thought to be ancestral to the group (~300 Ma) (Moran et al. 2005; Dietrich 2009; Bennett and Moran 2013; Bennett and Mao 2018). The partner endosymbionts provision distinct and complementary sets of the ten essential amino acids that animals are unable to synthesize and are limited in the host diet. *Sulcia* has been nearly universally retained over the ~300-Myr diversification of the Auchenorrhyncha (~40,000 described species); in contrast, *BetaSymb* has been replaced in at least six major lineages, highlighting its relative instability (Dietrich 2009; Bennett and Moran 2015).

The apparent unreliability and relatively frequent loss of the *BetaSymb* lineage is mirrored by rapid rates of molecular evolution. Pairwise divergence estimates for the two ancestral

symbionts in closely related leafhopper species (family Cicadellidae) reveal that *Sulcia* genomes are highly similar (99.68% nucleotide identity), whereas *BetaSymb* genomes are 30-fold more divergent (90.47% nucleotide identity) (Bennett, Abbà, et al. 2016). The dramatic differences in *Sulcia* and *BetaSymb* rates of molecular evolution have also been documented across more divergent Auchenorrhyncha lineages (Bennett and Moran 2013; Mao et al. 2017; Bennett and Mao 2018).

Although phylogenetic comparisons have greatly improved our understanding of the changes that endosymbiotic genomes experience, they do not adequately parse the relative contributions of mutagenesis, genetic drift, and natural selection. It is possible that selection differentially constrains sequence change in *Sulcia* and *BetaSymb* (Wernegreen 2015). Alternatively, the low rate of evolution in *Sulcia* may correlate with decreased DNA replication (Silva and Santos-Garcia 2015) or low mutational input (Bennett et al. 2014). However, these hypotheses are difficult to test directly as traditional approaches for measuring bacterial mutation, such as mutation accumulation lines (Kucukyildirim et al. 2016; Long, Miller, et al. 2018), are not feasible for obligate endosymbionts. In one case, researchers were able to use the recent demographic history of the insect host to infer endosymbiont mutation rates (Moran et al. 2009), but the challenge of directly measuring mutation in endosymbionts continues to inhibit efforts to disentangle mutagenesis, drift, and selection in endosymbionts.

A direct measurement of mutation would ideally capture all variants, before some can be filtered by natural selection. Although modern DNA sequencing technologies have facilitated an enormous increase in sequenced endosymbiont genomes, they have not necessarily translated to improvements in measuring de novo mutation in endosymbiont genomes. This is because typical sequencing error rates are too high (often above  $10^{-3}$  errors per bp) to detect rare variation in DNA samples (Schirmer et al. 2016). As a result, these methods can only detect single-nucleotide variants (SNVs) that have existed long enough to rise to a substantial frequency within the sample, either via drift or positive selection.

Fortunately, the recent advent of several high-fidelity DNA sequencing techniques provides the opportunity to obtain a snapshot of extremely low-frequency SNVs in endosymbiont DNA (Sloan et al. 2018). One technique, called duplex sequencing, is particularly suited for such a measurement because it has an exceptionally low error rate of less than  $2 \times 10^{-8}$  errors per bp (Kennedy et al. 2014; Wu et al. 2020). The high accuracy of this method is achieved by tagging each original DNA fragment with random barcodes and producing multiple sequencing reads from both strands to obtain a consensus sequence. Importantly, this method is robust to single-stranded DNA damage and individual errors introduced during PCR amplification or sequencing because

it separately tracks reads originating from the two complementary strands of the original biological molecule and requires support from each.

Here, we measured low-frequency variants in the endosymbionts of a leafhopper (*Macrostes sp. nr. severini*) to test the hypothesis that differences in mutagenesis may be responsible for differential evolutionary stability between *Sulcia* and *BetaSymb* (called “*Candidatus Nasuia deltocephalinicola*” in leafhoppers; hereafter *Nasuia*). We assembled *Sulcia* and *Nasuia* reference genomes from a population of *M. sp. nr. severini* isolated from Hawaii, using standard Illumina and Oxford Nanopore (MinION) libraries. We then created duplex libraries, which were mapped to the reference genomes to detect de novo variants. We found a dramatically higher frequency of independent SNVs in *Nasuia* than in *Sulcia*, which suggests the elevated rate of evolution observed in *Nasuia* over phylogenetic scales is driven, at least in part, by large differences in mutagenesis.

## Materials and Methods

### Growth and DNA Extractions of *Macrostes sp. nr. severini* Lines

A laboratory stock population of field collected *Macrostes sp. nr. severini* from Sumida Farms, Aiea, Oahu Island, Hawaii was established on December 8, 2016. Identification of this species, which has not been formally described, followed Le Roux and Rubinoff (2009). For this experiment, eight laboratory populations were established on barley plants from a single randomly selected foundress on June 17, 2017. Populations were grown with occasional plant replacements to maintain colony health for approximately six months and harvested December 18–22, 2017. Two populations, referred to as A and B, survived to produce sufficient number of individuals for downstream genomic sequencing. Populations were further subdivided into two replicates each (A1, A2, B1, B2) for dissection and sequencing. For each replicate, bacteriomes from approximately 100 individuals were dissected out, pooled, and stored in 95% ethanol. Total DNA was extracted using a DNeasy Blood and Tissue kit (Qiagen) following the manufacturer’s protocol, with a 12-h Proteinase K digestion.

To ensure variants detected with duplex sequencing were not an artifact of endosymbiont DNA transferred to the *M. sp. nr. severini* genome nuclear genome, we sequenced DNA isolated from *M. sp. nr. severini* heads, which should contain no true endosymbiont DNA. DNA was extracted from a pool of 20 dissected heads using the Qiagen Dneasy Blood and Tissue Kit.

### Construction of Shotgun Illumina Libraries

Standard shotgun Illumina libraries were created from the most concentrated bacteriome DNA sample (B1) as well as

the head DNA sample using the NEBNext Ultra II FS DNA Library Prep Kit. We used 50 ng of input DNA, with 15- and 10-min fragmentation steps for the bacteriome and head DNA, respectively. Samples were amplified with four cycles of PCR. The libraries had average lengths of 318 and 319 bp for the bacteriome and head DNA, respectively.

### Construction of Duplex Libraries from Bacteriome DNA for Variant Detection

Separate duplex sequencing libraries were generated for each of the four replicate bacteriome DNA samples (A1, A2, B1, and B2). Duplex library preparation followed protocols described elsewhere (Wu et al. 2020). Briefly, samples were fragmented with the Covaris M220 Focused-Ultrasonicator and subsequently end-repaired (NEBNext End Repair Module), A-tailed (Klenow Fragment enzyme, 1 mM dATP), adapter ligated (NEBNext Quick Ligation Module), and treated with a cocktail of three repair enzymes (NEB CutSmart Buffer, Fpg, Uracil-DNA Glycosylase, Endonuclease III). About 50 pg of repaired and adapter-ligated products were amplified for 19 cycles with the NEBNext Ultra II Q5 Master Mix (New England Biolabs M0544) and dual-indexed with custom IDT Ultramer primers. Adapter dimers were detected when DNA was assessed on an Agilent TapeStation 2200 (High Sensitivity D1000 reagents) and subsequently removed with size selection on a 2% BluePippin gel (Sage Science), using a target range of 300–700 bp. The pooled duplex libraries had an average length of 386 bp.

### Illumina Sequencing of Shotgun and Duplex Libraries

Shotgun and duplex libraries were sequenced on a NovaSeq 6000 platform (2 × 150 bp) at the University of Colorado Cancer Center in separate sequencing runs. Sequencing resulted in 9.4 M read pairs for the bacteriome shotgun library. The shotgun library from *M. sp. nr. severini* head DNA was sequenced in two runs (the first run did not produce enough reads), which resulted in a total of 359 M read pairs. Duplex libraries were also sequenced in two runs (again, the first round of sequencing did not produce enough reads), which resulted in 37.7–44.9 M read pairs per library. The raw sequencing reads are available via the NCBI Sequence Read Archive (SRA) under accessions SRR12112868 (A1), SRR12112867 (A2), SRR12112866 (B1), SRR12112865 (B2), SRR12112862 (*M. sp. nr. severini* head tissue), and SRR12112864 (B1: shotgun) (all in BioProject PRJNA642181).

### MinION Library Construction and Sequencing

To aid in genome assembly, we also sequenced the (B1) bacteriome sample with the Oxford Nanopore MinION (Jain et al. 2016), with 150 ng of input DNA on a FLO-MINSP6 flow cell, using the Rapid Sequencing Kit (SQK-RAD004). Data were processed using the MinION software release v19.10.1 with

default MinKNOW parameters except that base calling was set to high accuracy mode. Only the first 1.1 M reads were used for genome assembly (which accounted for  $\sim 2/3$  of the data generated by the run and averaged 2,972 bp in length). The resulting reads are deposited in the NCBI SRA under accession SRR12112863.

### Genome Assembly and Characterization

Endosymbiont genomes were assembled from the bacteriome Illumina shotgun and MinION libraries with the SPAdes genome assembler (v3.11.1) using the `-nanopore` flag (Bankevich et al. 2012). Scaffolds returned by SPAdes were searched (BlastN v2.9.0+) against available *Nasuia* and *Sulcia* genomes from a *M. quadrilineatus* (GenBank accession numbers CP006059.1 and CP006060.1, respectively) (Bennett and Moran 2013), and the two longest scaffolds were identified as near complete matches to *Sulcia* and *Nasuia*.

Evidence of chromosome circularity was assessed by searching for the scaffold beginning and ending sequences in the shotgun and MinION reads. Reads that contain the sequence of both scaffold ends (and therefore span the circular gap) were identified for both genomes. In both cases, the scaffolds ended in tandem repeat regions where the number of repeat units varied in different spanning reads (repeat length heterogeneity).

In the *Sulcia* assembly, the scaffold was broken at a 6-bp microsatellite (corresponding to positions 67046–67105 in the final assembly), with anywhere from 8 to 11 repeat units in different spanning reads. The most common repeat number in spanning reads was 10, and the genome was manually adjusted accordingly. The same approach was used to confirm the most common tandem repeat number for three other regions in the *Sulcia* assembly that exhibited repeat length heterogeneity (corresponding to regions 16069–16134, 174128–174181, and 178159–178299 in the final assembly), which were identified as scaffold breakpoints or ambiguous calls (Ns) in a separate SPAdes assembly run without the MinION data.

In the *Nasuia* genome, the repeat structure that broke the assembly is far more complicated, and its resolution was only possible through the use of the long-read MinION data (supplementary fig. S1, Supplementary Material online). In spanning sequences, we found seven to 53 repeats of a  $\sim 72$ -bp sequence (which contains only one G and one C), with 22 repeats being the most common repeat number. A complicating factor is the intermittent presence of an additional 7-bp sequence, which itself is a tandem repeat of the first seven base pairs of the 72-bp repeat, resulting in a mix of 72- and 79-bp repeats. The shotgun Illumina data were used to determine that 53.3% of repeats were 72 bp and 46.7% of repeats were 79 bp. The spanning MinION read of median length, which contained 22 repeats total, was selected as

the basis for resolving the *Nasuia* genome gap (supplementary fig. S1, Supplementary Material online).

After the most common repeat numbers were determined and circularity was confirmed, the breakpoints of the chromosomes were shifted, and the *Nasuia* scaffold was reverse complemented to reflect the orientation and genome positions of *Nasuia* and *Sulcia* accessions from the closely related *M. quadrilineatus* (Bennett and Moran 2013). We then aligned our new assemblies to *M. quadrilineatus* accessions using MAFFT (v7.453) under default parameters (Katoh and Standley 2013). Gaps in the alignments were tabulated with a custom script. Pairwise distances were calculated with MEGA (v10.1.8) (Kumar et al. 2018), using the Tamura 3-parameter model and a gamma distribution. Repetitive sequences were identified with custom scripts that report the position and length of homopolymers and microsatellites of 4+ bp in length (supplementary table S1, Supplementary Material online) (Temnykh et al. 2001).

Both *Sulcia* and *Nasuia* were annotated with initial protein-coding gene predictions using Glimmer3 in Geneious v11.1.5 and were then checked against existing symbiont genomes from two other *Macrosteles* host species (Delcher et al. 2007; Bennett, Abbà, et al. 2016). All RNA annotations were done using the “annotate from reference” function in Geneious with a match threshold of  $>80\%$  similarity. *Nasuia* and *Sulcia* genomes from *M. quadrilineatus* were used as the reference genome (Bennett and Moran 2013). GC content at 4-fold degenerate sites was calculated using custom scripts.

### Variant Detection with Duplex Data

Duplex libraries were processed with our custom data analysis pipeline (Wu et al. 2020; <https://github.com/dbsloan/duplex-seq>), which uses the random barcodes on read ends to first group raw reads into single-stranded consensus sequences, which require at least three raw Illumina reads. Complementary single-stranded consensus sequences with matching barcodes are then paired to form a duplex consensus sequence (DCS). Therefore, each DCS must contain at least six Illumina reads, though the mean raw read redundancy in these libraries was higher (supplementary fig. S2, Supplementary Material online). DCSs were subsequently mapped against the newly assembled reference genomes and the resulting alignment files were parsed to identify indels, single-nucleotide variants (SNVs), and coverage per bp. Supplementary table S2, Supplementary Material online, reports total DCS coverage and percent mapping to endosymbiont genomes for the four replicates. A filtering step then checked for the presence of identified variants in a  $k$ -mer database ( $k = 39$ ) created from the *M. sp. nr. severini* head DNA shotgun library (KMC v. 3.0.0; <https://github.com/refresh-bio/KMC>) to ensure variants are not derived from bacterial DNA that has been transferred to the host genome (Hotopp et al. 2007; Nikoh et al. 2010). All variants had

counts well below those detected for *M. sp. nr. severini* nuclear and mitochondrial sequences.

## Results

### Higher Divergence in *Nasuia* than in *Sulcia* Derived from Closely Related Leafhoppers

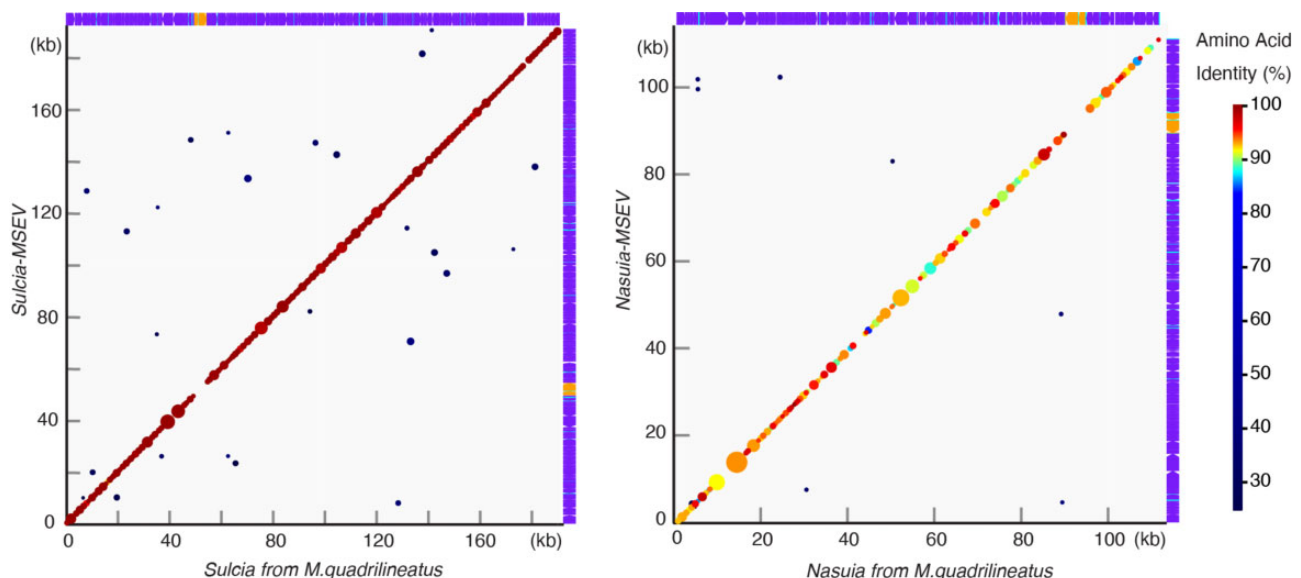
We assembled *Nasuia* and *Sulcia* genomes (GenBank accession numbers CP060019.1 and CP060020.1, respectively), using a combination of short-read (Illumina) and long-read (MinION) sequences from *M. sp. nr. severini* bacteriome DNA. The resulting genomes (referred to hereafter as *Nasuia*-MSEV and *Sulcia*-MSEV) were generally comparable to those previously sequenced from other hosts in terms of size (*Nasuia*-MSEV: 113 kb, *Sulcia*-MSEV: 190 kb), gene count (*Nasuia*-MSEV: 173, *Sulcia*-MSEV: 224), and GC content (*Nasuia*-MSEV: 16.9%, *Sulcia*-MSEV: 24.0%). We aligned our new assemblies to the previously published *M. quadrilineatus* accessions and found pairwise distances of 7.68% and 0.01% for the *Nasuia* and *Sulcia* pairs, respectively (fig. 1). The ~700-fold higher pairwise divergence in *Nasuia* than in *Sulcia* is much larger than 30-fold difference previously shown in a pairwise comparison between endosymbionts of the closely related *M. quadrilineatus* and *M. quadripunctulatus* (Bennett, Abbà, et al. 2016). This disparity is driven by the near perfect sequence conservation in *M. sp. nr. severini* and *M. quadrilineatus* *Sulcia* genomes,

which only differ at approximately one of every 10,000 positions. Transitions are responsible for 79.6% of the observed pairwise distance in *Nasuia*. The extremely small number of changes in the *Sulcia* comparison prohibits a meaningful calculation of the contribution of transitions to pairwise distance.

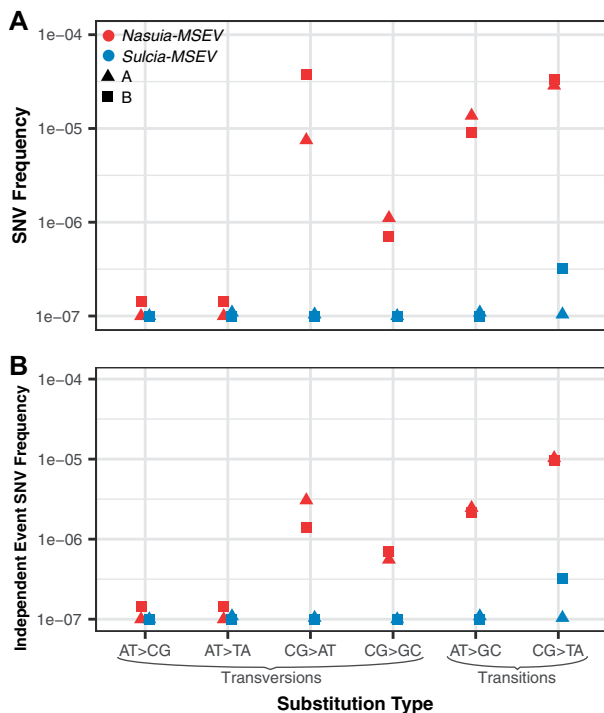
We found 1.00 and 0.28 gaps per kb in the *Nasuia* and *Sulcia* pairs, respectively (supplementary table S3, Supplementary Material online). A large (1,352 bp) gap in the *Nasuia* pairs (supplementary table S3, Supplementary Material online) corresponds to the mix of 72- and 79-bp repeats (22 total) in the *Nasuia*-MSEV assembly that we characterized with the MinION data (supplementary fig. S1, Supplementary Material online). The lack of this region in the *Nasuia* accession from *M. quadrilineatus* likely reflects differences in methods of genome assembly, rather than an actual difference in genome structure between the two *Nasuia* strains. We found complete conservation of synteny in both endosymbiont pairs, reflecting the high level of gene-order stability seen in many other ancient endosymbionts (fig. 1) (McCutcheon et al. 2019).

### Higher Frequency of SNVs in *Nasuia*-MSEV than in *Sulcia*-MSEV

We then used our new endosymbiont genome assemblies as references for mapping duplex sequence data. We detected a 106-fold higher SNV frequency in *Nasuia*-MSEV ( $2.18 \times 10^{-5}$ )



**FIG. 1.**—Synteny and sequence similarity of *Nasuia* and *Sulcia* from two closely related leafhopper species. Synteny plots reveal gene order to be perfectly conserved in both *Nasuia* and *Sulcia* of *Macrosteles quadrilineatus* and *M. sp. nr. severini*. Circles represent significant pairwise alignments between protein-coding genes in the two genomes, with circle diameter indicating alignment length. *Nasuia* exhibits a 700-fold higher sequence divergence than *Sulcia*, which is reflected in the lower amino acid identity (color scale on figure right) of *Nasuia* protein-coding sequences. The compact nature of both genomes can be observed in the gene tracks on the top (*M. quadrilineatus* associated) and right (*M. sp. nr. severini* associated) of the synteny plots, where protein-coding genes are shown in purple, rRNA genes are in yellow, tRNA genes are in light blue, and intergenic regions are white. This figure was generated with GenomeMatcher (Ohtsubo et al. 2008).



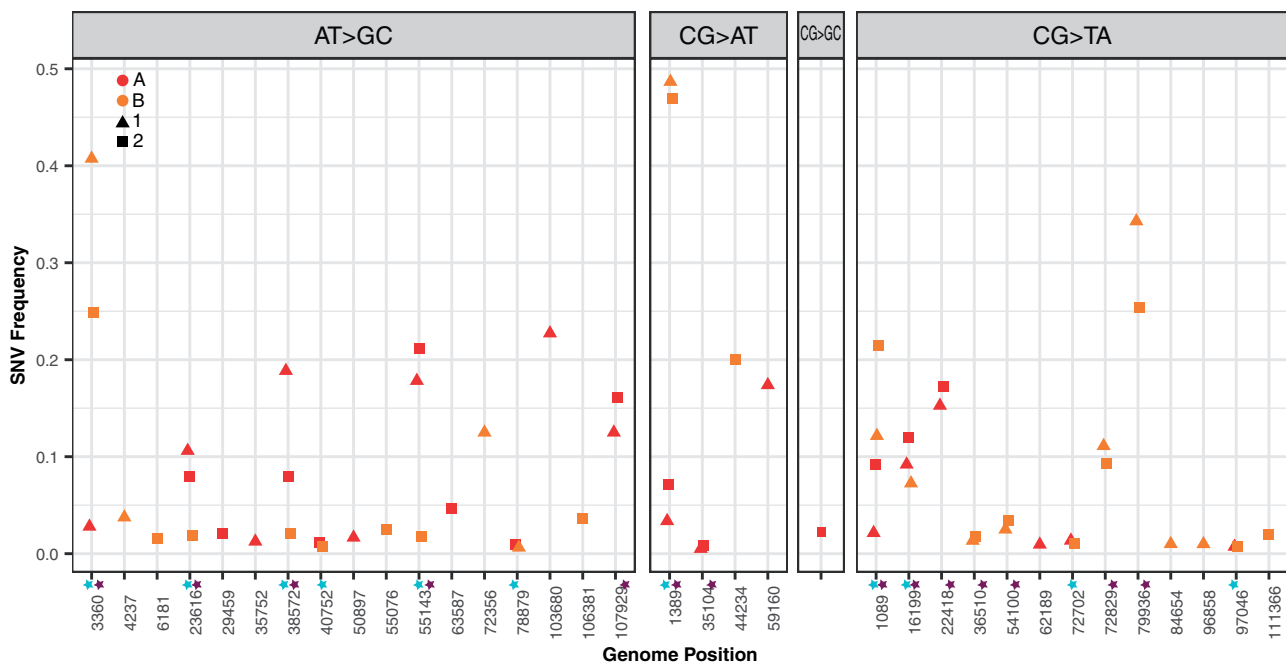
**FIG. 2.**—Duplex sequencing variant frequency in *Nasuia-MSEV* and *Sulcia-MSEV*. Variant frequencies are shown (A) in terms of total SNV frequency (DCS variant coverage/total DCS coverage) and (B) in terms of independent event frequency (sites with a variant/DCS coverage). In both, the six mutation types and are normalized by DCS coverage of the corresponding reference bases (e.g., the AT→CG changes are normalized by total AT coverage). *Nasuia-MSEV* is red and *Sulcia-MSEV* is blue. Triangles and squares are experimental populations A and B, respectively. For *Nasuia-MSEV*, the average (of the two experimental populations) independent event frequency ( $4.97 \times 10^{-6}$ ) is reduced relative to the average SNV Frequency ( $2.18 \times 10^{-5}$ ) due to the presence of 32 sites where multiple DCS reads mapped with a variant. There is no difference in *Sulcia-MSEV* variant frequencies ( $2.04 \times 10^{-7}$ ) because all *Sulcia-MSEV* SNVs are singletons (see also [supplementary fig. S1, Supplementary Material](#) online).

than in *Sulcia-MSEV* ( $2.04 \times 10^{-7}$ ) (fig. 2A and [supplementary data S1, Supplementary Material](#) online). For both endosymbionts, observed SNV frequencies are above the noise threshold of  $2 \times 10^{-8}$  errors per bp that we have previously established for our duplex protocol (Wu et al. 2020). All 17 of the SNVs identified in *Sulcia-MSEV* were “singletons” (captured by only a single DCS). In contrast, 35 of the 126 positions with SNVs in *Nasuia-MSEV* were detected at “high frequency” (captured by more than one DCS). Accordingly, the 106-fold higher SNV frequency in *Nasuia-MSEV* is partially driven by variants present at high frequency in the experimental populations. We therefore performed a comparison based on independent SNV events (sites in the genome with a SNV) and found a 18-fold higher independent event SNV frequency in *Nasuia-MSEV* than in *Sulcia-MSEV* after normalizing for sequence coverage (fig. 2B and [supplementary table S4, Supplementary Material](#) online).

For both *Sulcia-MSEV* and *Nasuia-MSEV*, the SNV type with the highest independent event frequency was CG→TA transitions, though SNV type comparisons are low powered in *Sulcia-MSEV* due to the small number of variants ([supplementary fig. S3, Supplementary Material](#) online). In *Nasuia*, AT→GC transitions and CG→AT transversions were also detected at relatively high frequencies. The proportion of transitions out of all independent SNVs was 0.82 and 0.52 for *Nasuia-MSEV* and *Sulcia-MSEV*, respectively, though the *Sulcia-MSEV* proportion is likely unreliable given that it is based on so few events. The proportion of transitions in *Nasuia-MSEV* appears to be within the range reported for other bacteria (Hershberg and Petrov 2010), and does not deviate from the proportion observed in the comparison with *Nasuia* from *M. quadrilineatus* (0.80; binomial test,  $P = 0.51$ ).

Given the extreme AT bias of the genomes (*Nasuia-MSEV*: 83.1% AT, *Sulcia-MSEV*: 76% AT), we tested if genomic AT content is at equilibrium, in which case the number of AT-increasing mutations would equal the number of AT-decreasing mutations. Of the 126 (independent event) SNVs detected in *Nasuia-MSEV*, significantly less than half (only 43) decreased AT content, whereas 75 mutations increased AT content (eight changes were AT-neutral) (binomial test,  $P = 0.0004$ ). However, the probability of detecting AT-increasing and AT-decreasing SNVs may not be equal, as relatively GC-rich regions have higher average sequence coverage due to bias against extreme AT-rich sequences during library amplification. The *Nasuia-MSEV* genome has a GC content of only 16.9%, but GC positions accrued 24.0% of the total sequencing coverage. This difference in depth of coverage may lead to the preferential detection of rare variants at GC sites. We therefore adjusted AT-decreasing SNVs to account for decreased detection probability, which we calculated as a ratio of AT coverage per bp and GC coverage per bp (AT-decreasing SNVs were 35.5% less likely to be detected). Adjusted AT-decreasing counts of 56 and AT-increasing counts of 63 were not significantly different from 50:50 equilibrium expectations (binomial test,  $P = 0.64$ ). The five AT-decreasing and six AT-increasing SNVs we detected in *Sulcia-MSEV* provide little power for testing for AT content equilibrium, but neither raw counts nor counts adjusted for the probability of detection deviate significantly from equilibrium expectations (binomial test,  $P = 0.77$ ).

Interestingly, 12 of the 35 high-frequency SNVs in *Nasuia-MSEV* were detected in both experimental populations A and B (fig. 3). Shared SNVs could have originated in the ancestor of the A and B foundresses and remained polymorphic during the six months that the experimental populations were maintained (approximately six host generations). Alternatively, the shared SNVs could have arisen independently in the two experimental lines (i.e., homoplasmy). There were no SNVs shared across any replicates in *Sulcia-MSEV*.



**Fig. 3.**—High-frequency variants in *Nasuia-MSEV*. SNV frequencies of the 35 positions in the *Nasuia-MSEV* genome captured by more than one DCS. Frequencies for experimental populations A and B are shown in red and orange, respectively. Experimental populations were subdivided into replicates 1 and 2, which are shown in triangles and squares, respectively. At 12 positions in the genome variants are shared between experimental populations A and B (teal stars) and at 14 positions in the genome variants are shared between replicates 1 and 2 within the two populations (purple stars).

In *Nasuia-MSEV* protein-coding regions, we found that the proportion of nonsynonymous changes was significantly reduced in high-frequency SNVs compared with singletons (Fisher's exact test,  $P = 0.003$ ). Reduction in the proportion of nonsynonymous SNVs in high-frequency SNVs (compared with singletons) occurs in all six substitution types, though none of the six SNV type reductions are significant on their own (Fisher's exact test,  $P > 0.079$ ; [supplementary fig. S4](#) and [table S5, Supplementary Material](#) online). The nucleotide substitution spectra also differ between singletons and high-frequency SNVs. Ignoring AT neutral substitutions (AT→TA and CG→GC), only 26 of 84 singletons decrease AT content, compared with 17 of 34 high-frequency SNVs (marginal significance with Fisher's exact test,  $P = 0.06$ ). We could not perform the same analysis for *Sulcia-MSEV* SNVs, as they were all singletons. For both endosymbionts, SNVs are not differentially distributed across intergenic, protein-coding, rRNA, or tRNA regions of the genome ([supplementary fig. S5](#) and [table S6, Supplementary Material](#) online).

Because high-frequency SNVs in *Nasuia-MSEV* show evidence of filtering by selection ([supplementary fig. S4, Supplementary Material](#) online), and differ in spectrum compared with singletons, we repeated the aforementioned AT equilibrium calculations with only singleton SNVs. AT-decreasing counts of 34 and AT-increasing counts of 50 (adjusted for probability of detection) are marginally different from equilibrium expectations (binomial test,  $P = 0.10$ ). We then used our normalized singleton counts to derive an

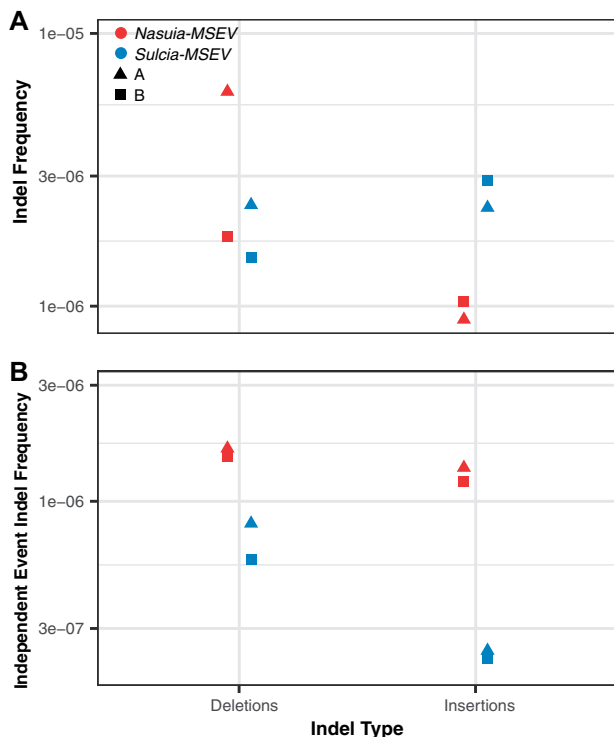
expected GC equilibrium content of 12.4%, which is lower than the actual *Nasuia-MSEV* GC content of 16.9%, but almost identical to the GC content of 12.2% at 4-fold degenerate sites, which are expected to be under minimal selection.

#### No Strand-Specific Mutation Asymmetry Detected in *Nasuia-MSEV* Protein-Coding Genes

Coding strands in CDS regions of the *Nasuia-MSEV* genome are enriched for G compared with C (positive GC skew of 0.16 or 1.4 Gs for every C). In endosymbiotic genomes, positive GC skew on the leading strand of DNA replication is thought to result from asymmetrical C→T changes on the leading strand, which is more susceptible to cytosine deamination due to its prolonged existence in a single-stranded state (Klasson and Andersson 2006). During transcription, the increased single-stranded exposure of the coding DNA strand has also been proposed to lead to asymmetrical C→T changes in bacterial genomes (Francino and Ochman 1997).

We tested if mutation drives GC skew on coding stands of *Nasuia-MSEV* by comparing reciprocal mutations (C→T vs. G→A) to expected counts given C and G coverage. We did not find evidence for asymmetry in reciprocal mutations in CG→TA transitions, or for any of other five substitution types in *Nasuia-MSEV* coding strands in CDS regions (binomial test, all  $P$  values  $\geq 0.16$ ) ([supplementary table S7, Supplementary Material](#) online). We did not attempt to perform the reciprocal





**FIG. 4.**—Indel frequencies in *Nasuia-MSEV* and *Sulcia-MSEV*. Indel frequencies shown in terms of (A) total frequency (DCS indels/DCS coverage) and (B) Independent event frequency (sites with an indel/DCS coverage). Overall average indel frequencies for *Nasuia-MSEV* and *Sulcia-MSEV* are similar—but deletions are more common in *Nasuia-MSEV* and insertions are more common in *Sulcia-MSEV*. Independent event frequencies are higher for *Nasuia-MSEV* than for *Sulcia-MSEV*. *Nasuia-MSEV* is red and *Sulcia-MSEV* is blue. Triangles and squares are experimental populations A and B, respectively.

strand analysis for *Sulcia-MSEV* given the small amount of SNVs.

#### Comparison of Indels in *Nasuia-MSEV* and *Sulcia-MSEV*

Overall, average indel frequencies for *Nasuia-MSEV* and *Sulcia-MSEV* are similar ( $4.92 \times 10^{-6}$  and  $4.51 \times 10^{-6}$  respectively). *Nasuia-MSEV* has a higher frequency of deletions than insertions, whereas *Sulcia-MSEV* has a higher frequency of insertions than deletions (fig. 4A and [supplementary data S2](#), [Supplementary Material](#) online). However, it is important to recognize that during genome assembly, determination of the dominant repeat number in regions with repeat length heterogeneity can influence whether an indel is considered a deletion or an insertion, so comparisons of the frequencies of deletions versus insertions should be cautiously interpreted. Indeed, when the three hypervariable regions that were identified during genome assembly (see Materials and Methods) are excluded from *Sulcia* indel analysis, insertions become less frequent than deletions ([supplementary table S8](#),

[Supplementary Material](#) online). Duplex sequencing did not detect any indels at the 72/79-bp tandem repeat in the *Nasuia* genome ([supplementary fig. S1](#), [Supplementary Material](#) online), despite MinION data revealing this region to be hyper-variable in terms of repeat number. Most repeats in this region are likely too long to be spanned by duplex sequences (the shortest repeat we found in the MinION data was seven repeats long) though the extreme AT content and frequent homopolymers in the region could also explain why only one DCS read from all four libraries mapped to the region (visual inspection in IGV; [Thorvaldsdóttir et al. 2013](#)). For all downstream indel analysis, no regions were excluded (i.e., the full data set was considered).

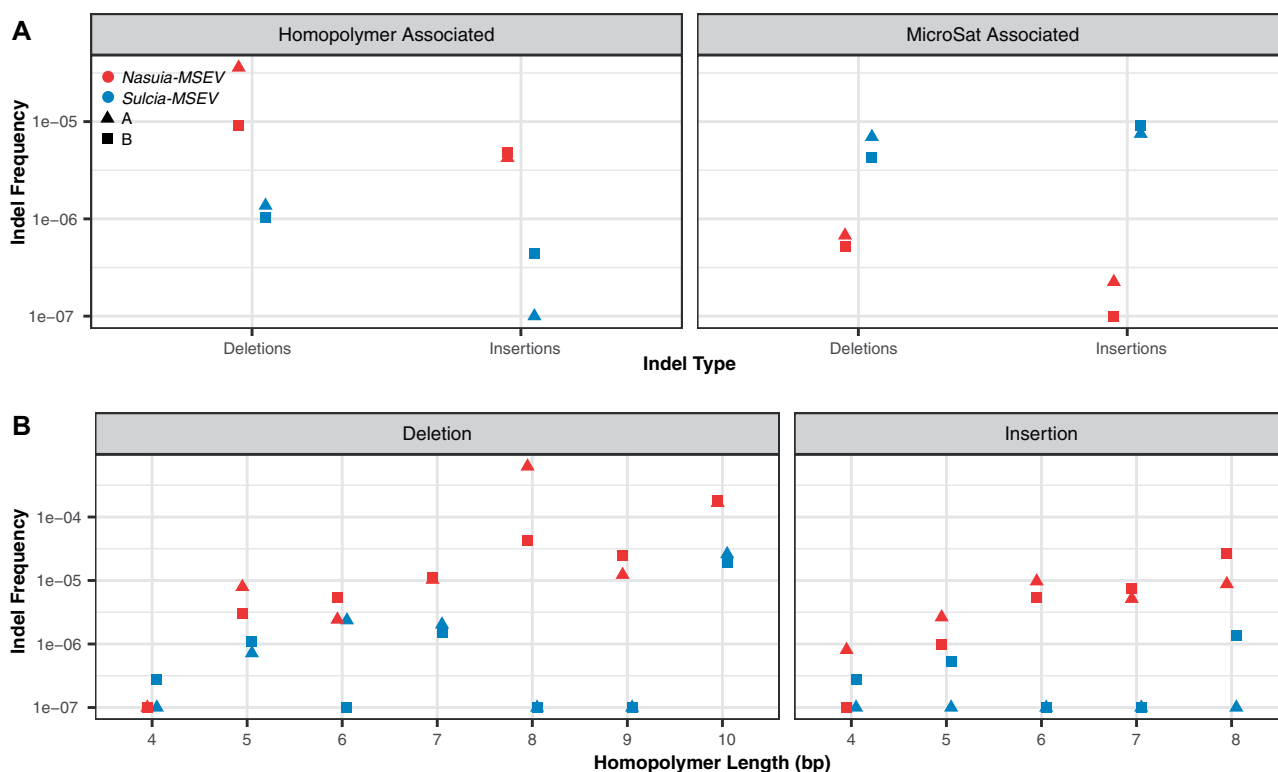
Both endosymbionts have indels that are present at high frequencies, which are sometimes shared across experimental populations A and B. Like with high-frequency SNVs in *Nasuia-MSEV*, these indels could arise from homoplasmy or from shared ancestry and persistence in a polymorphic state during the growth of experimental cultures. Local sequence context plays a large role in indel distribution across both genomes as 94.1% and 98.4% of indels overlap with repetitive genomic sequence (microsatellites and homopolymers) for *Nasuia-MSEV* and *Sulcia-MSEV*, respectively. Given the apparent influence of local sequence context on indel occurrence, and the lack of local sequence effects on SNV occurrence, we suspect that high-frequency indels often reflect multiple homoplasious events (e.g., long homopolymers can be subject to “multiple hits”). Our indel analysis therefore emphasizes total indel frequency rather than independent event indel frequency (fig. 4B).

We found that *Nasuia-MSEV* has a higher frequency of homopolymer-associated indels, whereas *Sulcia-MSEV* has a higher frequency of microsatellite-associated indels (fig. 5A). These differences do not merely reflect a difference in the repetitive landscape of the symbiont genomes, as the frequencies are normalized by the specific coverage of a given repeat type (microsatellite or homopolymer), and *Nasuia-MSEV* contains more of both repeat type per kb than *Sulcia-MSEV* ([supplementary table S1](#), [Supplementary Material](#) online). For both endosymbionts, we find that as homopolymers increase in length, indel frequency also increases (fig. 5B).

## Discussion

### Rapid Evolution of *BetaSymb* Likely Driven by Higher Mutation Rate

*BetaSymb* (represented by *Nasuia* in leafhoppers and other names depending on the insect host; *Vidania* in planthoppers and *Zinderia* in spittlebugs) evolves much more rapidly than *Sulcia*, despite the partner endosymbionts having occupied a nearly identical ecological niche (i.e., bacteriocytes in a shared host) for ~300 Myr ([Bennett, Abbà, et al. 2016](#); [Bennett and](#)



**FIG. 5.**—Frequency of repeat-associated indels in *Nasuia*-MSEV and *Sulcia*-MSEV. (A) Frequencies of indels associated with homopolymers and microsatellites and (B) frequencies of homopolymer-associated indels as function of homopolymer length. Frequencies are calculated as the total number of DCS indels in a repeat category divided by DCS coverage of that repeat category. About 94.1% of all *Nasuia*-MSEV indels and 98.4% of all *Sulcia*-MSEV indels overlap with homopolymers or microsatellites. About 90.9% and 96.1% of all indels are expansions or contractions of existing repeats for *Nasuia* and *Sulcia*, respectively. For *Nasuia*-MSEV, the majority of repeat-associated indels are located in homopolymers, and deletions are more common than insertions. For *Sulcia*-MSEV, the majority of indels are located in microsatellites. For both endosymbionts, the frequency of insertions and deletions increase as homopolymer length increases. *Nasuia*-MSEV is red, and *Sulcia*-MSEV is blue. Triangles and squares are experimental populations A and B, respectively.

Mao 2018). In addition, *BetaSymb* experiences increased evolutionary turnover and has been replaced in at least six Auchenorrhyncha lineages (Bennett and Moran 2015). This accelerated evolution with respect to *Sulcia* does not appear to be limited to *BetaSymb*, as some of its lineage-specific replacements such as *Hodgkinia* (Alphaproteobacteria) and *Baumannia* (Gammaproteobacteria) show similar asymmetries (Bennett et al. 2014; Campbell et al. 2017). Our finding of a higher frequency of de novo SNVs in *Nasuia*-MSEV (fig. 2) suggests the increased evolutionary rate in *BetaSymb* is likely driven by mutation.

Our results therefore support the hypothesis that genome decay and extinction in endosymbionts is driven by increases in mutation rate (Itoh et al. 2002). Such increases intensify Muller's ratchet, which is already prevalent in endosymbionts experiencing strict vertical transmission and host confinement (Rispe and Moran 2000). That *Sulcia* also has a tiny genome (ranging from 190 to 244 kb; Woyke et al. 2010), but exhibited a depressed mutation frequency in our study, and greatly reduced synonymous site divergence compared with multiple partner lineages (Bennett et al. 2014; Bennett and Mao

2018), illustrates that mutation is not the only force responsible for genome reduction (Bennett and Moran 2015).

Because variant frequencies are the product of both the mutation rate ( $\mu$ ) and the effective population size ( $N_e$ ), it is important to recognize that a higher  $N_e$  in *Nasuia*-MSEV could also lead to the 18- to 106-fold higher SNV frequency we observed in *Nasuia*-MSEV. However, current evidence does not support a larger  $N_e$  for *Nasuia*. Sequence coverage comparisons from our bacteriome shotgun library reveal a 1.5-fold higher per bp depth of coverage in *Sulcia*-MSEV than in *Nasuia*-MSEV (Pedersen and Quinlan 2018). This suggests the standing population size may actually be smaller in *Nasuia*-MSEV than in *Sulcia*-MSEV, though bias against AT-rich sequences during library amplification could also drive lower coverage of the *Nasuia*-MSEV genome, as coverage for GC base pairs is only 1.2-fold higher in *Sulcia*-MSEV.

The standing population size does not preclude the possibility that there may be differences in the intensity of the transmission bottlenecks for the two endosymbionts. Microscopy has revealed that, whereas *Sulcia* and *Nasuia* are housed in distinct bacteriocyte types, they both

successfully migrate to and infect terminal oocytes (Szklarzewicz et al. 2016). It is not known, however, if the number of transmitted bacteria differs for *Nasuia* versus *Sulcia* in leafhoppers. In cicadas, *Sulcia* and partner symbiont *Hodgkinia* are transmitted at similar numbers, though *Hodgkinia* transmission dramatically increases in lineages where the *Hodgkinia* genome has been fragmented into multiple chromosomes (Campbell et al. 2018). Finally, a larger  $N_e$  in *Nasuia*-MSEV should lead to more efficient purifying selection, which would be inconsistent with the fact that *Nasuia* exhibits more extensive genome decay (high AT content, reduced size, etc.) over phylogenetic scales. This final argument relies on indirect evidence, and a study of actual *Nasuia* versus *Sulcia* transmission is needed to definitively rule out the possibility that differences in  $N_e$  are driving different observed variant frequencies. Furthermore, the idea that reduced  $N_e$  drives genome size reduction has recently been challenged, as mutation rates are apparently better at predicting gene loss than  $d_N/d_S$  (a proxy for  $N_e$ ) in several independent endosymbiont lineages (Bourguignon et al. 2020).

Interestingly, 35 variants in *Nasuia*-MSEV were present in “high-frequency” (detected in more than one DCS). Compared with singletons, high-frequency SNVs in *Nasuia*-MSEV are less likely to result in nonsynonymous changes (supplementary fig. S2, Supplementary Material online), suggesting they have been filtered by selection. This idea is further supported by the observation that high-frequency SNVs differ in AT spectra compared with singletons (marginal significance in Fisher’s exact test,  $P=0.06$ ). Singletons in our data thus most accurately represent the true mutation spectra, before it can be shaped by selection. When we used singletons to predict equilibrium GC content, we got a value nearly identical to the observed GC content at 4-fold degenerate sites (12.4% and 12.2%, respectively), which are predicted to experience minimal selection. The actual GC content in *Nasuia* is higher (16.9%), suggesting that selection on nonsynonymous sites and tRNAs/rRNAs keeps genomic GC content above equilibrium values. As such, it appears that despite the already extreme AT-richness of the *Nasuia* genome, mutation bias would further erode GC content in the absence of selection (Hershberg and Petrov 2010; Hildebrand et al. 2010; Long, Sung, et al. 2018). Our observation of rapid filtering of high-frequency events in *Nasuia*-MSEV raises the possibility that a previous study finding no AT mutation bias in *Buchnera aphidicola* (an ancient endosymbiont of many aphids) (Moran et al. 2009) likely did so based on a set of mutations that had been subject to substantial effects of selection. Although our data demonstrate that selection is able to maintain some level of *Nasuia* sequence conservation, it remains possible that *Nasuia* and *Sulcia* are differentially constrained.

Twelve of the 35 *Nasuia*-MSEV high-frequency SNVs were shared between experimental populations A and B, which were reproductively isolated for six months (approximately six host generations; Capinera 2008). Although it is possible

that shared SNVs resulted from homoplasmy, some shared SNVs have risen to very high frequency (greater than 45% in one case), which demonstrates they are at least old enough to have spread to a large portion of the experimental population. We find it most likely that shared SNVs were ancestral to endosymbiont populations present in the initial insect founders and persisted as polymorphisms through the duration of insect growth, but we cannot rule out the possibility that some or all shared SNVs have arisen through homoplasmy.

### Can Differences in DNA Replication and Repair Machinery Explain Greater Genome Conservation in *Sulcia*?

Although *Nasuia*-MSEV and *Sulcia*-MSEV lack most subunits of the primary bacterial polymerase (DNA Pol III), they both retain  $\alpha$  and  $\epsilon$  subunits (*dnaE* and *dnaQ*), which in *Escherichia coli* perform polymerization and 3'  $\rightarrow$  5' exonuclease activity, respectively (Fijalkowska et al. 2012). *Nasuia*-MSEV also uniquely retains the  $\beta$  subunit (*dnaN*), which facilitates binding between subunit  $\alpha$  and the template DNA strand (Gui et al. 2011). *Nasuia*-MSEV retains two additional DNA replication and repair (RR) genes lost in *Sulcia*-MSEV: *dnaB* (helicase) and *ssb* (single-stranded binding protein) (Bennett and Moran 2013).

Only one RR gene, *mutS*, is absent in *Nasuia*-MSEV, but retained in *Sulcia*-MSEV. In other bacteria, the MutS protein recognizes and initiates repair of mismatched bases and small indels (Dettman et al. 2016; Long, Miller, et al. 2018). Although the role of *mutS* in *Sulcia* has never been experimentally validated, it contains all five of the domains present in the *E. coli mutS* (Ogata et al. 2011; Groothuizen and Sixma 2016). The extremely low frequency of SNVs we observed in *Sulcia*-MSEV (fig. 1 and supplementary table S4, Supplementary Material online) could result from *mutS*-directed mismatch repair (MMR).

Mutation accumulation experiments with *Pseudomonas aeruginosa* reveal *mutS* mutants experience a 230-fold increase in indels (average length of 1.1 bp) compared with wild-type lines, demonstrating the importance of *mutS* in short indel surveillance (Dettman et al. 2016). Although we observed a high frequency of indels in *Sulcia*-MSEV, most were expansions or contractions of repeats at existing microsatellites. In fact, only 19 of the 385 indel containing DCS reads in *Sulcia*-MSEV were of 1 bp in length (compared with 145 of 154 in *Nasuia*-MSEV) (supplementary table S9, Supplementary Material online). Even more striking is the complete lack of 1-bp indels in the alignment of *Sulcia*-MSEV with *Sulcia* from *M. quadrilineatus* (compared with forty-eight 1-bp indels in the equivalent *Nasuia* alignment) (supplementary table S3, Supplementary Material online). The rarity of 1-bp indels in *Sulcia*-MSEV thus further supports the idea that *mutS* is critical for maintaining *Sulcia* genome conservation through surveillance of homopolymer expansions/contractions.

Interestingly, the genome of *Sulcia* in leafhoppers actually has a reduced set of RR genes compared with those found in *Sulcia* genomes in other Auchenorrhynchan lineages. *Sulcia-CARI*, an endosymbiont of spittlebugs, possesses the two DNA Pol III subunits found in *Sulcia-MSEV*, but has also retained subunits *dnaB*, *dnaN*, *dnaX*, and polymerase accessory subunits *holA* and *holB*. Additional RR genes retained by *Sulcia-CARI*, but absent in leafhopper *Sulcia* genomes, include gyrase genes *gryA*, *gryB* and MMR genes *mutL*, and *mutH* (McCutcheon and Moran 2010; Bennett and Moran 2013). Despite lacking the above listed RR genes, leafhopper *Sulcia* genomes remain remarkably stable, as demonstrated by phylogenetic comparisons at multiple levels of leafhopper divergence in this and previous studies (Bennett, Abbà, et al. 2016; Mao et al. 2017).

Given that MutS works in concert with *DnaN*, *mutL*, and *mutH* products in typical bacterial MMR (Simmons et al. 2008; Jiricny 2013), it is curious that *DnaN* and *mutH* have been completely lost and *mutL* has been pseudogenized in leafhopper *Sulcia* genomes (Bennett and Moran 2013; Bennett, Abbà, et al. 2016). A recent assay of *M. quadrilineatus* gene expression revealed that in the bacteriome, leafhopper hosts often overexpresses homologs of RR genes that are absent from endosymbiont genomes, raising the possibility that leafhopper hosts may support endosymbiont genome replication and repair (Mao et al. 2018). Host support has been hypothesized to enable endosymbiont gene loss (Hansen and Moran 2011; Husnik et al. 2013; Russell et al. 2013; Sloan et al. 2014), and a recent study has demonstrated that the partner endosymbionts of cicadas likely rely of host encoded machinery to acquire correctly processed tRNAs (Van Leuven et al. 2019). However, none of the four putative *mutL*/*mutH* homologs in the *M. quadrilineatus* transcriptome showed substantial overexpression in the *Sulcia* specific bacteriocyte (compared with expression in the insect body) (supplementary table S9, Supplementary Material online). Thus, it remains unclear if *Sulcia mutS* is participating in conventional MMR or if it has acquired novel functions in support of *Sulcia* genome conservation. Functional characterization of MMR in leafhopper *Sulcia* would provide valuable insight into what appears to be an atypical role for an ancient and highly conserved mismatch recognition protein (Ogata et al. 2011; Jiricny 2013).

The missing *mutS* in *Nasuia* provides an additional opportunity for host support. Mao et al. (2018) reported strong overexpression of one host *mutS* homolog (referred to as TRINITY\_DN66078\_c1\_g1 in that study; see supplementary table S10, Supplementary Material online) in the *Nasuia* bacteriome, but it is highly unlikely that this is involved in MMR for two reasons. First, the identified homolog is related to eukaryotic *MSH4*, which is a gene family that lacks a mismatch recognition domain and is involved in recombination

rather than MMR (Ogata et al. 2011). Furthermore, it appears to encode only a partial MSH4 protein fused with a C-terminal sequence of unidentified origin. Therefore, it is likely that *Nasuia* lacks conventional MMR function.

Although *mutS* may play a role in *Sulcia* genome conservation, retention of *mutS* is not a cure-all. For example, *BetaSymb* in spittlebugs (named “*Candidatus Zinderia insecticola*”) retains the entire mismatch repair gene set (*mutS*, *mutL*, and *mutH*) but displays a rate of evolution typical of *Nasuia* and other *BetaSymb* lineages (Bennett and Moran 2013; Koga et al. 2013). We posit that the *mutS* in *Sulcia* may be particularly efficient at recognizing and removing mismatched bases and short indels. All *Sulcia* genomes analyzed to date retain *mutS*, but identification of a *Sulcia* lineage that has lost *mutS* would facilitate a test of the hypothesis that *mutS* may be the key to *Sulcia* genome conservation.

## Conclusion

The two ancestral endosymbionts in auchenorrhynchan insects vary dramatically in their propensity for extinction versus retention, thus providing a natural comparative framework for investigating what facilitates long-term, stable endosymbiotic relationships. Compared with *Sulcia*, which exhibits remarkable genome conservation for an ancient endosymbiont and has been nearly universally retained, *BetaSymb* displays substantially elevated rates of molecular evolution. We have shown that *Nasuia-MSEV*, a representative of *BetaSymb* from leafhoppers, has a substantially higher frequency of de novo mutations than *Sulcia-MSEV*. Our analysis supports the hypothesis that the evolutionary rate variation in *Sulcia* versus *BetaSymb* is driven by differences in mutational input. We posit that the low mutation rate in *Sulcia* may also explain why *Sulcia* is so consistently retained in diverse auchenorrhynchan lineages, whereas partner endosymbionts are apparently much more transient.

## Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

## Acknowledgments

We thank Meng Mao for providing Trinity assemblies and differential expression data. We also thank Kristen Poff at UH Manoa for early help with insect rearing. This work was supported by the National Institutes of Health (R01 GM118046) and a National Science Foundation graduate fellowship (DGE 1450032).

## Data Availability

The raw sequencing reads are available via the NCBI Sequence Read Archive under accessions SRR12112868, SRR12112867, SRR12112866, SRR12112865, SRR12112862, and SRR12112864. Genome annotations for *Nasuia-MSEV* and *Sulcia-MSEV* are available via GenBank under accessions CP060019.1 and CP060020.1, respectively.

## Literature Cited

- Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 19(5):455–477.
- Baumann L, Thao MLL, Hess JM, Johnson MW, Baumann P. 2002. The genetic properties of the primary endosymbionts of mealybugs differ from those of other endosymbionts of plant sap-sucking insects. *Appl Environ Microbiol.* 68(7):3198–3205.
- Bennett GM, Abbà S, Kube M, Marxachi C. 2016. Complete genome sequences of the obligate symbionts "*Candidatus Sulcia muelleri*" and "*Ca. Nasuia deltocephalinicola*". *Am Soc Microbiol.* 4:4–5.
- Bennett GM, Mao M. 2018. Comparative genomics of a quadripartite symbiosis in a planthopper host reveals the origins and rearranged nutritional responsibilities of anciently diverged bacterial lineages. *Environ Microbiol.* 20(12):4461–4472.
- Bennett GM, McCutcheon JP, MacDonald BR, Romanovic D, Moran NA. 2014. Differential genome evolution between companion symbionts in an insect-bacterial symbiosis. *MBio* 5(5):e01697–e016714.
- Bennett GM, McCutcheon JP, McDonald BR, Moran NA. 2016. Lineage-specific patterns of genome deterioration in obligate symbionts of sharpshooter leafhoppers. *Genome Biol Evol.* 8(1):296–301.
- 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.
- Bennett GM, Moran NA. 2015. Heritable symbiosis: the advantages and perils of an evolutionary rabbit hole. *Proc Natl Acad Sci U S A.* 112(33):10169–10176.
- Bourguignon T, et al. 2020. Increased mutation rate is linked to genome reduction in prokaryotes. *Curr Biol.* 30(19):3848–3855.e4.
- Buchner P. 1965. Endosymbiosis of animals with plant microorganisms. New York: Interscience.
- Campbell MA, et al. 2018. Changes in endosymbiont complexity drive host-level compensatory adaptations in cicadas. *MBio* 9(6):8–11.
- Campbell MA, Łukasik P, Simon C, McCutcheon JP. 2017. Idiosyncratic genome degradation in a bacterial endosymbiont of periodical Cicadas. *Curr Biol.* 27(22):3568–3575.
- Capinera JL. 2008. Aster Leafhopper, *Macrostelus quadrilineatus* Forbes (Hemiptera: Cicadellidae). 2nd ed. New York: Springer. p. 320–323.
- Delcher AL, Bratke KA, Powers EC, Salzberg SL. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23(6):673–679.
- Dettman JR, Sztepanacz JL, Kassen R. 2016. The properties of spontaneous mutations in the opportunistic pathogen *Pseudomonas aeruginosa*. *BMC Genomics* 17(1):27.
- Dietrich CH. 2009. Auchenorrhyncha: (Cicadas, Spittlebugs, Leafhoppers, Treehoppers, and Planthoppers). In: Resh VH, Carde RT, editors. *Encyclopedia of Insects*. New York: Academic Press. p. 56–64.
- Fijalkowska IJ, Schaaper RM, Jonczyk P. 2012. DNA replication fidelity in *Escherichia coli*: a multi-DNA polymerase affair. *FEMS Microbiol Rev.* 36(6):1105–1121.
- Francino MP, Ochman H. 1997. Strand asymmetries in DNA evolution. *Trends Genet.* 13(6):240–245.
- Groothuizen FS, Sixma TK. 2016. The conserved molecular machinery in DNA mismatch repair enzyme structures. *DNA Repair (Amst).* 38:14–23.
- Gui WJ, et al. 2011. Crystal structure of DNA polymerase III  $\beta$  sliding clamp from *Mycobacterium tuberculosis*. *Biochem Biophys Res Commun.* 405(2):272–277.
- Hansen AK, Moran NA. 2011. Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. *Proc Natl Acad Sci U S A.* 108(7):2849–2854.
- Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* 6(9):e1001115.
- Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.* 6(9):e1001107.
- Hotopp JCD, et al. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* 317(5845):1753–1756.
- Husnik F, et al. 2013. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell* 153(7):1567–1578.
- Itoh T, Martin W, Nei M. 2002. Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proc Natl Acad Sci U S A.* 99(20):12944–12948.
- Jain M, Olsen HE, Paten B, Akeson M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17(1):239.
- Jiricny J. 2013. Postreplicative mismatch repair. *Cold Spring Harb Perspect Biol.* 5(4):a012633.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30(4):772–780.
- Kennedy SR, et al. 2014. Detecting ultralow-frequency mutations by duplex sequencing. *Nat Protoc.* 9(11):2586–2606.
- Klasson L, Andersson SGE. 2006. Strong asymmetric mutation bias in endosymbiont genomes coincide with loss of genes for replication restart pathways. *Mol Biol Evol.* 23(5):1031–1039.
- Koga R, Bennett GM, Cryan JR, Moran NA. 2013. Evolutionary replacement of obligate symbionts in an ancient and diverse insect lineage. *Environ Microbiol.* 15(7):2073–2081.
- Kucukyildirim S, et al. 2016. The rate and spectrum of spontaneous mutations in *Mycobacterium smegmatis*, a bacterium naturally devoid of the postreplicative mismatch repair pathway. *G3 (Bethesda)* 6:2157–2163.
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol.* 35(6):1547–1549.
- Le Roux JJ, Rubinoff D. 2009. Molecular data reveals California as the potential source of an invasive leafhopper species, *Macrostelus* sp. nr. *severini*, transmitting the aster yellows phytoplasma in Hawaii. *Ann Appl Biol.* 154(3):419–427.
- Long H, Miller SF, Williams E, Lynch M. 2018. Specificity of the DNA mismatch repair system (MMR) and mutagenesis bias in bacteria. *Mol Biol Evol.* 35(10):2414–2421.
- Long H, Sung W, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nat Ecol Evol.* 2(2):237–240.
- Mao M, Yang X, Bennett GM. 2018. Evolution of host support for two ancient bacterial symbionts with differentially degraded genomes in a leafhopper host. *Proc Natl Acad Sci U S A.* 115(50):E11691–E11700.
- Mao M, Yang X, Poff K, Bennett G. 2017. Comparative genomics of the dual-obligate symbionts from the treehopper, *Entylia carinata* (Hemiptera: Membracidae), provide insight into the origins and evolution of an ancient symbiosis. *Genome Biol Evol.* 9(6):1803–1815.
- Matsuura Y, et al. 2018. Recurrent symbiont recruitment from fungal parasites in cicadas. *Proc Natl Acad Sci U S A.* 115(26):E5970–E5979.

- 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, McDonald BR, Moran NA. 2009a. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. *Proc Natl Acad Sci U S A.* 106(36):15394–15399.
- McCutcheon JP, McDonald BR, Moran NA. 2009b. Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *PLoS Genet.* 5(7):e1000565.
- McCutcheon JP, Moran NA. 2010. Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. *Genome Biol Evol.* 2:708–718.
- McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol.* 10(1):13–26.
- McCutcheon JP, Von Dohlen CD. 2011. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Curr Biol.* 21(16):1366–1372.
- Mira A, Moran NA. 2002. Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. *Microb Ecol.* 44(2):137–143.
- Moran NA. 1996. Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A.* 93(7):2873–2878.
- Moran NA, Bennett GM. 2014. The tiniest tiny genomes. *Annu Rev Microbiol.* 68(1):195–215.
- Moran NA, McLaughlin HJ, Sorek R. 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* 323(5912):379–382.
- Moran NA, Tran P, Gerardo NM. 2005. Symbiosis and insect diversification: an ancient symbiont of sap-feeding insects from the bacterial phylum Bacteroidetes. *Appl Environ Microbiol.* 71(12):8802–8810.
- Nikoh N, et al. 2010. Bacterial genes in the Aphid genome: absence of functional gene transfer from Buchnera to its host. *PLoS Genet.* 6(2):e1000827.
- Ogata H, et al. 2011. Two new subfamilies of DNA mismatch repair proteins (MutS) specifically abundant in the marine environment. *ISME J.* 5(7):1143–1151.
- Ohtsubo Y, Ikeda-Ohtsubo W, Nagata Y, Tsuda M. 2008. GenomeMatcher: a graphical user interface for DNA sequence comparison. *BMC Bioinformatics* 9:376.
- Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* 34(5):867–868.
- Rispe C, Moran NA. 2000. Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection. *Am Nat.* 156(4):425–441.
- Russell CW, Bouvaine S, Newell PD, Douglass AE. 2013. Shared metabolic pathways in a coevolved insect-bacterial symbiosis. *Appl Environ Microbiol.* 79(19):6117–6123.
- Sandström J, Moran N. 1999. How nutritionally imbalanced is phloem sap for aphids? In: Simpson SJ, Mordue AJ, Hardie J, editors. *Proceedings of the 10th International Symposium on Insect-Plant Relationships*. Dordrecht: Springer Netherlands. p. 203–210.
- Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* 17(1):125.
- Silva FJ, Santos-Garcia D. 2015. Slow and fast evolving endosymbiont lineages: positive correlation between the rates of synonymous and non-synonymous substitution. *Front Microbiol.* 6:1279.
- Simmons LA, Davies BW, Grossman AD, Walker GC. 2008.  $\beta$  clamp directs localization of mismatch repair in *Bacillus subtilis*. *Mol Cell.* 29(3):291–301.
- Sloan DB, Broz AK, Sharbrough J, Wu Z. 2018. Detecting rare mutations and DNA damage with sequencing-based methods. *Trends Biotechnol.* 36(7):729–740.
- Sloan DB, et al. 2014. Parallel histories of horizontal gene transfer facilitated extreme reduction of endosymbiont genomes in sap-feeding insects. *Mol Biol Evol.* 31(4):857–871.
- Szklarzewicz T, Grzywacz B, Szewdo J, Michalik A. 2016. Bacterial symbionts of the leafhopper *Evacanthus interruptus* (Linnaeus, 1758) (Insecta, Hemiptera, Cicadellidae: Evacanthinae). *Protoplasma* 253(2):379–391.
- Temnykh S, et al. 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res.* 11(8):1441–1452.
- Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 14(2):178–192.
- Van Leuven JT, Mao M, Xing DD, Bennett GM, McCutcheon JP. 2019. Cicada endosymbionts have tRNAs that are correctly processed despite having genomes that do not encode all of the tRNA processing machinery. *MBio* 10(3):e01950–e011018.
- Van Leuven JT, McCutcheon JP. 2012. An AT mutational bias in the tiny GC-rich endosymbiont genome of Hodgkinia. *Genome Biol Evol.* 4(1):24–27.
- Wernegreen JJ. 2015. Endosymbiont evolution: predictions from theory and surprises from genomes. *Ann N Y Acad Sci.* 1360(1):16–35.
- Woyke T, et al. 2010. One bacterial cell, one complete genome. *PLoS One* 5(4):e10314.
- Wu Z, Waneka G, Broz AK, King CR, Sloan DB. 2020. *MSH1* is required for maintenance of the low mutation rates in plant mitochondrial and plastid genomes. *Proc Natl Acad Sci U S A.* 117(28):16448–16455.

Associate editor: Ruth Hershberg