Intrahost Genetic Diversity of Bacterial Symbionts Exhibits Evidence of Mixed Infections and Recombinant Haplotypes

Shelbi L. Russell*[,1,2] and Colleen M. Cavanaugh*[,1]

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA

[2]Department of Molecular Cellular and Developmental Biology, University of California Santa Cruz, Santa Cruz, CA

Associate editor: Miriam Barlow

***Corresponding authors:** E-mails: shelbilrussell@gmail.com; cavanaug@fas.harvard.edu.

Abstract

Even the simplest microbial-eukaryotic mutualisms are comprised of entire populations of symbionts at the level of the host individual. Early work suggested that these intrahost populations maintain low genetic diversity as a result of transmission bottlenecks or to avoid competition between symbiont genotypes. However, the amount of genetic diversity among symbionts within a single host remains largely unexplored. To address this, we investigated the chemosynthetic symbiosis between the bivalve *Solemya velum* and its intracellular bacterial symbionts, which exhibits evidence of both vertical and horizontal transmission. Intrahost symbiont populations were sequenced to high coverage (200–1,000×). Analyses of nucleotide diversity revealed that the symbiont genome sequences were largely homogeneous within individual host specimens, consistent with vertical transmission, except for particular regions that were polymorphic in ~20% of host specimens. These variant sites were also found segregating in other host individuals from the same population, colocalized to several regions of the genome, and consistently co-occurred on the same short read pairs (derived from the same chromosome). These results strongly suggest that these variant haplotypes originated through recombination events, potentially during prior mixed infections or in the external environment, rather than as novel mutations within symbiont populations. This abundant genetic diversity could have a profound influence on symbiont evolution as it provides the opportunity for selection to limit the extent of reductive genome evolution commonly seen in obligate intracellular bacteria and to enable the evolution of adaptive genotypes.

Key words: **symbiosis, intrahost evolution, transmission mode.**

Introduction

Mutualistic symbiotic interactions have enabled a large number of taxa to occupy novel niches through partnering eukaryotic structural complexity with bacterial physiological capacities (Moya et al. 2008; Cavanaugh et al. 2013; Flórez et al. 2015). These associations range from single bacterial taxa to complex communities living with eukaryotic hosts (McFall-Ngai et al. 2013). Symbionts and hosts cooperate to grow and reproduce, however large disparities exist between their population structures and generation times. Even associations between just one host and one bacterial strain can reach densities of 10^6 to 10^{12} symbiont cells within a single host individual (Komaki and Ishikawa 2000; Wollenberg and Ruby 2009; Cavanaugh et al. 2013; Klose et al. 2015; Duperron et al. 2016; Sender et al. 2016). Further distinguishing symbiont life histories from that of their hosts, bacteria undergo multiple generations within each host generation to divide and populate tissues (McFall-Ngai 2014; Zhang et al. 2016). Thus, symbionts experience evolutionary pressures on different time scales than their hosts.

Symbiont populations within host tissues have the potential to exhibit high diversity due to their large sizes and the acquisition of mutations during reproduction within the host. This genetic variation provides the material for selection to potentially act during the host's lifespan. Two processes can generate genetic diversity in these intrahost symbiont populations: de novo mutation and mixed infections. De novo mutations in an originally clonal population may be lost, retained, or fixed in the population (e.g., clonal interference [Lang et al. 2013]). Infections involving more than one "species" of bacteria are termed coinfections, whereas mixed infections involve multiple genotypes of one symbiont "species." Numerous symbiotic systems exhibit coinfections. These range from two-member associations, such as *Bathymodiolus* mussels, in which two different gammaproteobacteria, a methanotroph and a hydrogen/sulfide-oxidizer, coexist within gill cells (Distel et al. 1995; Petersen et al. 2011), and *Osedax* worms, in which two dominant symbiont types inhabit the trophosome (Goffredi et al. 2014), to microbiomes, e.g., the human gut microbiome, which consist of tens to hundreds of bacterial taxa fulfilling particular niches (Costello et al. 2012; Rey et al. 2013; Coyte et al. 2015). These infections may be stable over time, or may be dynamic with new migrants from other symbiont populations introduced via horizontal transmission events. Recombination among different symbiont genomes and de novo mutation both have the potential to generate genetic diversity in symbiont populations contained in host cells and tissues, as illustrated in figure 1.

Article

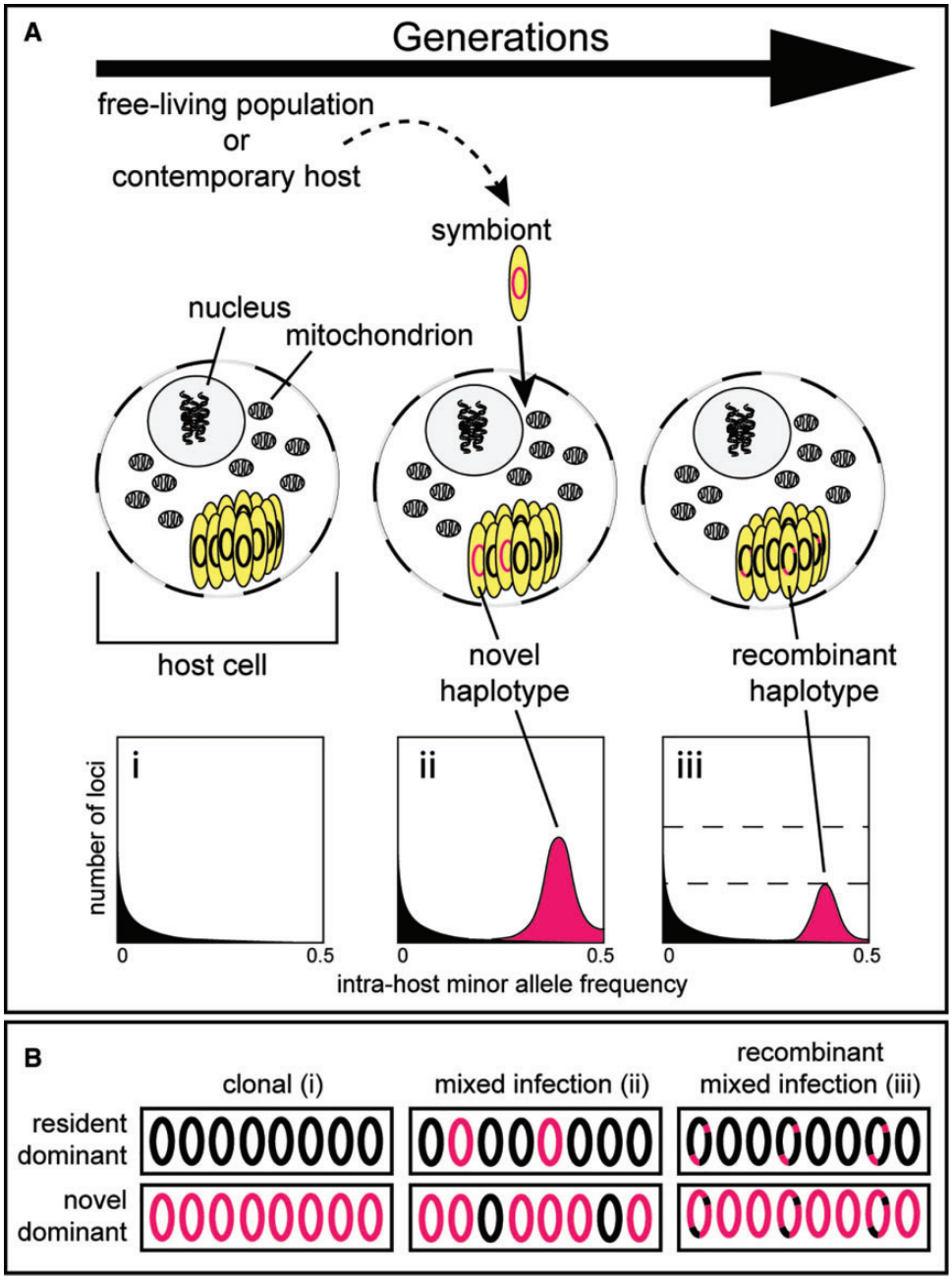

Fig. 1. Model of intrahost symbiont population admixture via horizontal transmission. (*A*) Columns from left to right represent host generations and each circle depicts a host containing symbionts (yellow) with one of two haplotypes, shown in black or pink. (i) Allele frequency spectrum (AFS) for intrahost populations under solely vertical transmission. Note that only low frequency variants produced by mutation in this host or a recent ancestor are present. (ii) When a horizontal transmission event occurs, all variant sites in the novel symbiont (pink) are at the frequency of that haplotype in the population. (iii) If recombination occurs between the two genotypes in the mixed infection, the recombinant genotype (pink tract in black haplotype) may be inherited by the next generation of host, and will have a number of variant sites proportional to the length of the recombinant tracts (dashed lines) at a frequency equal to that of the recombinant haplotype in the intrahost population. (*B*) Horizontal transmission followed by vertical inheritance of the resulting intrahost symbiont population can result in any of the six shown population compositions, with AFS similar to i–iii in A as indicated.

Despite these sources of variation, symbiont populations within a single host are often thought to be clonal or have extremely low genetic diversity (e.g., Woyke et al. 2010). This hypothesis is theoretically plausible for two reasons. First, given that symbionts must colonize hosts via horizontal transmission through the environment or vertical transmission from a parent, low symbiont diversity may be a byproduct of a transmission bottleneck in which only a few symbionts colonize the host (Wollenberg and Ruby 2009; Kaltenpoth et al. 2010; Stephens et al. 2015; Didelot et al. 2016; Grubaugh et al. 2016). However, some associations are inoculated by a relatively large number of symbionts,

yet result in low within-host diversity due to the low diversity of the source population. This is often the case in vertical transmission, where hundreds or thousands of symbionts are inherited, however diversity is low because population sizes are restricted to the capacity of host tissues and undergo a bottleneck at every host generation (see Mira and Moran 2002; Kaltenpoth et al. 2010). Second, low intrahost diversity may be maintained by the host to prevent competition among symbiont genotypes, as has been proposed for mitochondria (see Greiner et al. 2014). Genetic diversity is potentially damaging to a host because selection can act on standing genetic variation to produce phenotypes with higher reproduction rates, generating competition among symbionts and resulting in virulence via redirection of energy from host-beneficial functions or direct damage to the cells or tissues (e.g., as in malaria (de Roode et al. 2005) and see Frank 1996; Vautrin et al. 2008; Bennett and Moran 2015). Thus, both bottleneck and competition mechanisms may be important in controlling the amount of genetic diversity in microbial mutualisms.

Genetic diversity within bacterial populations has been historically difficult to analyze because individual symbionts are pooled together within and among host tissues, making it difficult to isolate and identify individual genotypes. The first investigations of bacterial diversity were reliant on microscopy, which is severely limited due to the low morphological diversity exhibited among bacteria (van Leeuwenhoek 1800; Siefert and Fox 1998; Young 2007). Later attempts to characterize bacterial communities relied on culture-dependent techniques that are now well known for grossly underestimating bacterial diversity, as typically $<1\%$ of bacteria are culturable (Amann et al. 1995; Pham and Kim 2012). With the development of the 16S rRNA gene as a marker for microbial diversity (Lane et al. 1985) and the invention of Sanger DNA sequencing, bacterial communities could be characterized based on marker loci amplified by PCR and cloned to isolate sequences from individual bacterial chromosomes (Hugenholtz et al. 1998). However, these investigations were limited in their ability to detect genetic diversity across the genome (limited number of loci amplified) and within bacterial communities and populations (limited number of clones sequenced), thus also underestimating bacterial genetic diversity (e.g., Reuter and Keller 2003; Luyten et al. 2006; Stewart and Cavanaugh 2009; Fay and Weber 2012; Weigel and Erwin 2016). Next generation sequencing (NGS, e.g., Illumina) now provides the sequencing depth and breadth, i.e., across the whole genome, needed at a relatively low cost to overcome these historical barriers to characterizing bacterial genetic diversity (e.g., Worby et al. 2014; Sim et al. 2015; Walter et al. 2016).

The obligate symbiosis between the protobranch bivalve *Solemya velum* and its chemosynthetic gammaproteobacterial gill symbionts presents an excellent system to assess the diversity of bacterial symbionts within host individuals (Stewart and Cavanaugh 2006). *S. velum* occurs in reducing mudflats along the eastern coast of North America, from the intertidal to subtidal zone, where it digs Y-shaped burrows that allow access to sulfide from pore water below and oxygenated seawater from above (Stanley 1970). The symbionts reside within gill epithelial cells (bacteriocytes, fig. 2A), where they oxidize sulfide to synthesize ATP and fix carbon dioxide. The host is reliant upon symbiont autotrophy, having a highly degenerated gut and acquiring the majority of its nitrogen and carbon from its symbiont metabolism (Stewart and Cavanaugh 2006). Within *S. velum*, symbiont populations reach $1.2 \pm 0.4 \times 10^9$ symbiont cells per gram of wet gill tissue (Cavanaugh 1983) and are comprised of a single 16S rRNA phylotype (Stewart et al. 2009). This symbiosis is vertically transmitted (Krueger et al. 1996) with occasional horizontal transmission events (Russell et al. 2017), providing the opportunity for both de novo mutation and symbiont admixture to generate intrahost diversity.

S. velum symbiont populations are nested: populations of symbionts occur within host individuals (intrahost), between hosts within a geographic locality (interhost symbiont subpopulation), and between hosts from different localities (interhost symbiont metapopulation; see fig. 2A). Furthermore, symbionts exhibit highly structured diversity between hosts from different subpopulations (Russell et al. 2017). The complicated structure of these symbiont populations allows the information contained in the genetic variation to be leveraged across these scales. Additionally, host mitochondria have been shown to be clonal due to maternal transmission (Russell et al. 2017), providing a homogeneous internal control for intrahost symbiont genetic diversity.

To analyze the genetic diversity of *S. velum* symbionts within host tissues, we employed a population genomics approach. We obtained high coverage whole genome data for symbiont populations from gill tissues. Mapping Illumina reads to the symbiont reference genome (Dmytrenko et al. 2014) allowed us to analyze nucleotide diversity across the genome and to calculate allele frequency spectra (AFS) for each population of symbionts from individual hosts. These analyses revealed striking evidence of genetic diversity that appears to be due to mutation as well as mixed infections with recombination.

Results

Deep-coverage whole genome shotgun Illumina sequencing of the gill tissue of individual *S. velum* revealed genetic variation in symbiont genomes, but not in mitochondrial genomes (supplementary table S2, Supplementary Material online). Intrahost genetic variation was localized to specific regions of the symbiont genome and colocalized to the same bacterial chromosomes in these populations, suggesting the presence of recombinant chromosomes that may have functional significance.

Intrahost Diversity Was Adequately Sampled

Rarefaction curves plotted for intrahost symbiont diversity at randomly subsampled sequencing coverage indicate that common as well as rare genotypes were detected (supplementary fig. S1, Supplementary Material online). Specimens with intermediate-frequency variants ($>10\%$ allele frequency; e.g., MA16, MA18, and RI53) exhibited detection plateaus around $50\times$ coverage and detection of low frequency

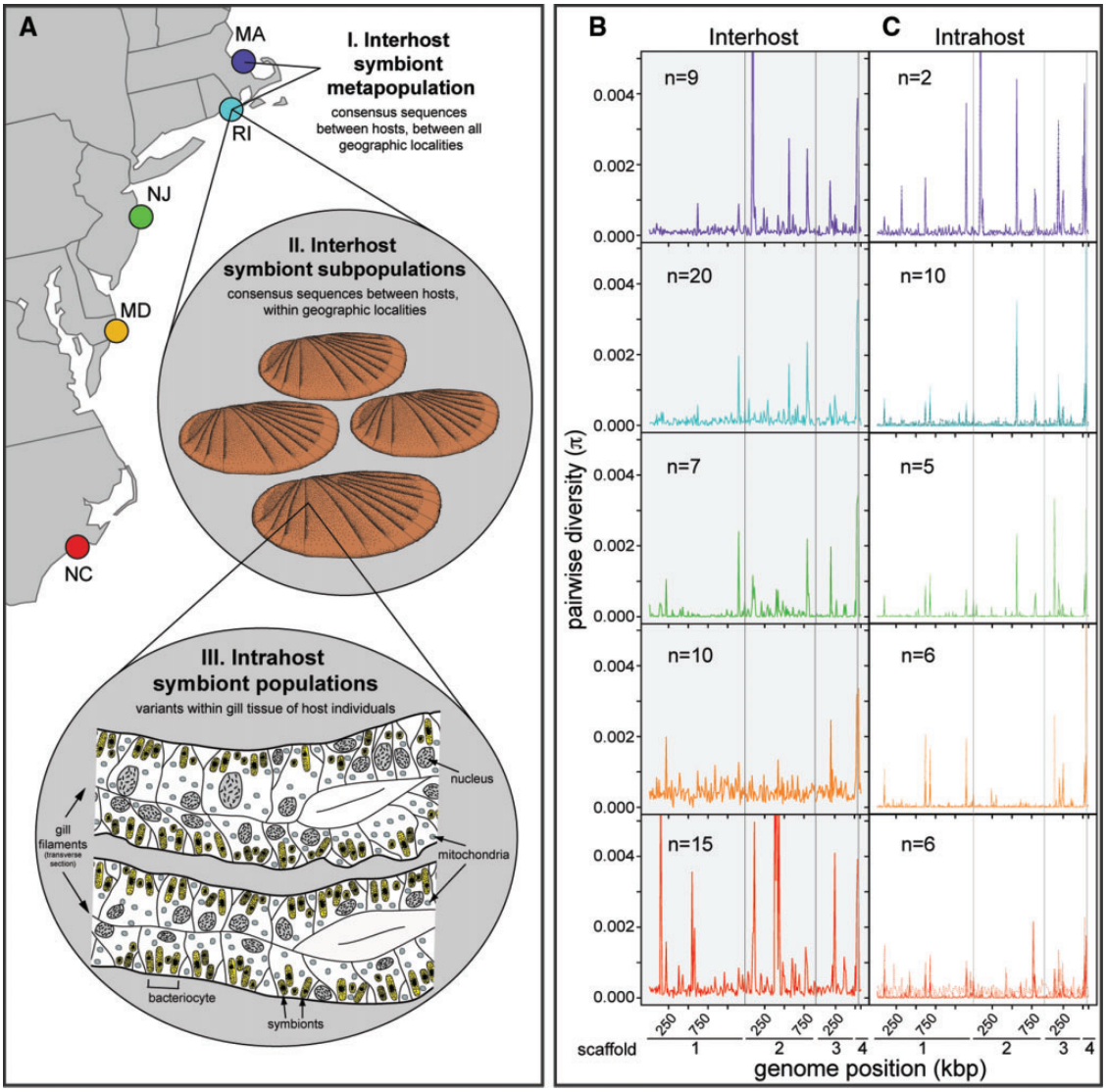

Fig. 2. Pairwise diversity along the genome of the *S. velum* bacterial symbiont at inter- and intrahost population levels. *S. velum* specimens were collected from five geographic localities, marked by colored dots, and sequenced in Russell et al. (2017). (*A*) Schematic of the nested population structure of the *S. velum* symbionts. I. Interhost metapopulation: *S. velum* occurs as a metapopulation, with subpopulations located along the coast that are genetically clustered by habitat (Russell et al. 2017). Colored dots mark the collection localities. II. Interhost populations: Within each subpopulation, symbionts are genetically diverse among hosts. III. Intrahost populations: Within each host individual, an entire population of symbionts ($\sim 10^9$ cells/gram) resides intracellularly within gill epithelial tissue. (*B*, *C*) Pairwise nucleotide diversity (p) in 10 kb nonoverlapping windows along the symbiont reference genome for interhost consensus sequence comparisons (Russell et al. 2017) within a subpopulation (B) and intrahost comparisons (C). Colors indicate locality as in (A). Gray vertical lines mark breaks between symbiont reference genome scaffolds. Number of samples used in each calculation (*n*) are given for each plot.

variants plateaued around 200× coverage (supplementary fig. S1*A* and *B*, Supplementary Material online). Plotting the AFS (described below) for each subsample revealed that intermediate-frequency variants are detected at lower coverages (supplementary fig. S1*C*, Supplementary Material online). Given this finding, additional specimens that were sequenced at 50× or higher coverage in Russell et al. 2017 were analyzed to better estimate the rate of intermediate frequency variation in intrahost symbiont populations (supplementary fig. S2, Supplementary Material online). The results of those analyses are included below, however it should be noted that the lower bound of the AFS is truncated in these samples because the lower sequencing coverage prevented confidently calling alleles with frequencies much lower than 10% (i.e., 5 out of 50 reads).

Intrahost Symbiont Diversity Is Common

In the majority of intrahost populations, average pairwise nucleotide diversity was an order of magnitude lower than that found in interhost symbiont subpopulations. However specific regions of the genome exhibited elevated levels of diversity (fig. 2*B* and *C* and table 1). Many of these variable sites correspond to segregating sites detected between hosts within a geographic subpopulation, suggesting that the diversity within a host may be influenced by the diversity found between hosts, e.g., via horizontal transmission.

Table 1. Pairwise Genetic Diversity (π) and Number of Segregating Sites (S) in Interhost and Intrahost Symbiont Populations.

S. velum Subpopulation Locality	Interhost[a]		Sample	Intrahost[b]										Functional Diversity[e]	
				All Sites			S Interhost Metapopulation		S Interhost Subpopulations		AFS Modes[c]				
	Average π	Pairwise S (average: range)		Intrahost verage π	SNPs	Indels	SNPs	Indels	SNPs	Indels	Intermediate AF Peak Median	S in Peak	S Supporting Haplotype[d]	Intrahost GO Enriched Categories	Interhost GO Enriched Categories
Massachusetts	2.81E-04	1776: 1386–2197	MA16[f]	2.25E-04	2,052	284	842	67	727	63	35%	522	206	20	4
			MA18[f]	2.17E-04	2,211	295	838	64	723	57	22%	736	305	24	2
Rhode Island	2.14E-04	1807: 1549–2369	RI36[f]	5.59E-05	1,083	173	411	36	373	36	4%	799	35	0	12
			RI37[f]	4.58E-05	765	162	138	38	104	37	na		105	na	4
			RI38[f]	4.44E-05	736	139	144	32	114	30	na		114	na	4
			RI39[f]	3.69E-05	673	123	110	21	84	21	na		79	na	4
			RI44	5.68E-05	584	195	100	29	70	28	28%	101	11	0	0
			RI47	7.18E-05	691	188	229	25	203	24	9%	307	6	0	0
			RI51	6.87E-05	589	166	287	25	246	23	15%	274	7	0	0
			RI53[f]	9.64E-05	1,043	184	390	40	363	37	31%	307	11	0	24
			RI61	5.64E-05	516	160	158	21	139	21	na		22	na	0
			RI62	6.59E-05	503	144	116	22	86	20	na		62	na	0
New Jersey	1.54E-04	1039:672–1344	NJ2[f]	6.50E-05	907	180	184	51	58	48	na		55	na	4
			NJ3[f]	4.61E-05	720	185	138	47	51	45	na		52	na	5
			NJ4	3.64E-05	325	187	63	33	33	30	na		38	na	0
			NJ6	4.56E-05	469	229	87	36	28	35	na		36	na	0
			NJ9	3.92E-05	302	204	66	29	15	26	na		24	na	0
Maryland	4.84E-04	3020:1490–4210	MD7[f]	4.41E-05	750	215	110	54	60	49	7%	489	12	0	7
			MD8[f]	4.34E-05	681	206	71	51	32	48	na		34	na	4
			MD13	5.62E-05	452	209	109	40	64	35	na		24	na	4
			MD16	6.32E-05	498	179	128	27	107	26	na		60	na	4
			MD27	5.46E-05	440	208	95	34	37	28	na		41	na	4
			MD28	3.39E-05	286	164	47	26	17	24	na		27	na	4
North Carolina	5.78E-04	6714:1508–8221	NC2[f]	6.38E-05	1,182	240	428	48	380	44	3%	829	42	0	4
											45%	56	33	0	
			NC9	6.14E-05	550	236	97	45	36	42	na		36	na	4
			NC10	2.12E-04	1860	289	1078	55	991	51	17%	975	91	1	6
			NC12[f]	7.73E-05	1,108	190	275	15	230	14	5%	650	69	1	0
											16%	188	2	0	
											33%	90	4	0	
			NC16	5.83E-05	430	237	92	37	46	35	na		45	na	4
			NC18	2.54E-04	1521	253	618	47	559	44	46%	635	20	0	9

[a]Columns describe genetic diversity of symbiont populations between hosts (interhost populations), calculated from consensus sequences called from host individuals.
[b]Columns describe intrahost genetic diveristy for entire intrahost populations (all sites), describe the subset of intrahost segregating sites found in symbionts from other host individuals (S interhost metapopulation), and the subset of these interhost sites specific to the sample's subpopulation (S interhost subpopulation). For allele frequency spectra (AFS) containing modes at intermediate allele frequencies (AF), the number of sites contained within each mode are shown (AFS Modes).
[c]For specimens with intermediate AF modes (light gray boxes), columns describe medians of intermediate frequency modes, number of sites contained within the median allele frequency $\pm$ 4%, and the number of sites found to be on the same read pairs indicating linkage on a haplotype. For specimens with only low frequency variants, all intrahost segregating sites were tested for linkage ("na", not applicable). See supplementary table S3, supplementary material online for haplotype testing results.
[d]Number of variant sites cooccurring on the same reads or read pairs, which support haplotypes within at least one Illumina insert length (~350 bp).
[e]Potential functional changes encoded by symbiont genetic variation detected by enrichment in gene ontology (GO) categories.
[f]specimens sequenced at a high depth-of-coverage (204–1,115×). Other specimens (unmarked) at 52–118×.

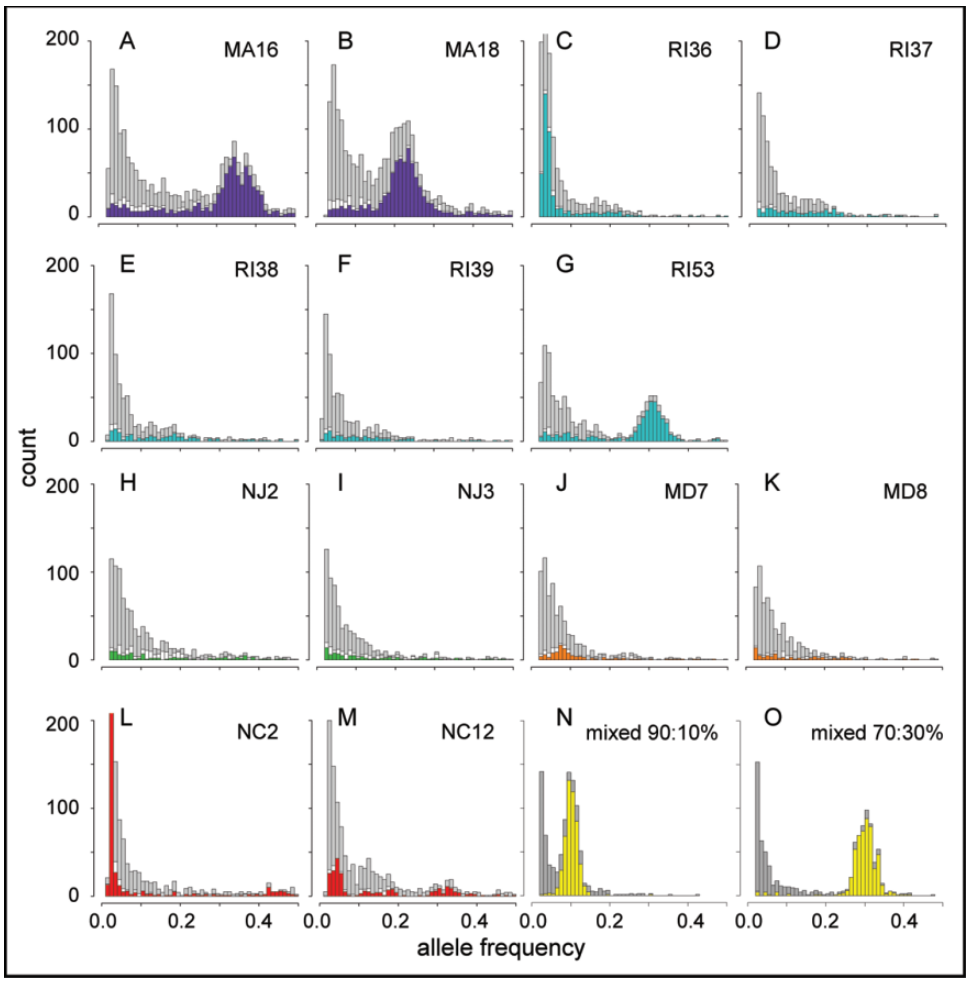

Fig. 3. Folded allele frequency spectra for *S. velum* intrahost symbiont populations and simulated mixed infections. (*A*–*M*) All intrahost variant sites detected in each host specimen are plotted as gray bars. The subset of these sites segregating in the *S. velum* interhost symbiont population (within or between subpopulations) are replotted in white and only the subset of sites segregating between hosts within a given subpopulation are replotted in color by collection locality as in figure 2. (*N*, *O*) Simulated mixtures of two symbiont haplotypes sharing 99.97% identity at 10:90% and 30:70% ratios, respectively.

Intrahost Allele Frequency Spectra Are Multimodal, Revealing Intermediate-Frequency Variants

An allele frequency spectrum (AFS) is a population genetic summary statistic that reveals information about the demographic processes that have occurred in a population, such as population size changes, migration, substructure, and selection (Galtier et al. 2000; Nielsen 2005). Variants that arose by mutation while populating the gill will be at very low frequency, and should not be segregating in the interhost symbiont population under both infinite- and finite-sites models of mutation (Hudson 1983; Yang 1996). In contrast, genetic variation that arose by mixed infection would be segregating in symbiont populations from different hosts (esp. in the same geographic locality) and may be at intermediate frequencies in the host.

Visual inspection of the symbiont AFS for each host specimen revealed that some intrahost populations exhibit multimodal frequency distributions (e.g., fig. 3*A*, *B*, and *G*). In all deeply sequenced specimens, low frequency alleles were at highest abundance and were not segregating between hosts (gray bars in fig. 3 and supplementary fig. S2 plots, Supplementary Material online), although we cannot rule out that some low frequency alleles are segregating at low frequency between hosts, preventing their detection. Some specimens also exhibited an abundance of alleles at intermediate frequencies, which were comprised almost entirely of interhost segregating sites from the specimen's subpopulation (colored bars in fig. 3 and supplementary fig. S2 plots, Supplementary Material online). For the purposes of discussion, we define intermediate frequency allele modes to have a mean of 10–50% with a distribution distinct from the low frequency variants. In particular, the two specimens from Massachusetts, MA16 and MA18, three specimens from Rhode Island, RI53, RI47, and RI51, and two specimens from North Carolina, NC10 and NC18, exhibited intermediate frequency modes. Some specimens exhibited a potentially lower-frequency mode consisting of interhost segregating sites overlapping the low frequency tail of the distribution (e.g., fig. 3; RI36, MD7, NC2, and NC12). Across all coverage depths, ~20% of *S. velum* exhibited a high allele frequency mode (204–1115× coverage: 3/13 specimens [fig. 3] and 52–118× coverage: 4/16 specimens [supplementary fig. S2,

Supplementary Material online]). Most of these frequency peaks contained fewer segregating sites than occur on average between symbiont interhost haplotypes (table 1 and fig. 2C).

Simulated Data

We simulated mixed infections to ground our expectations for what the shape of the within-host symbiont AFS should look like under this demographic scenario. A mixed infection should manifest as multimodality in the AFS, reflecting the number of variant sites between the resident and new genome as well as the relative frequencies of each genome in the intrahost population (as depicted in fig. 1). This is indeed what we observed in the simulated mixtures (fig. 3N and O and supplementary fig. S3, Supplementary Material online). The mean allele frequency for each frequency mode was equal to the smaller proportion of the mixture, i.e., a 30/70 mixture had a mean of 0.3. The height of each frequency mode was proportional to the divergence between the genomes, with more divergent genomes exhibiting larger numbers of variant sites in the mode. Highlighting variant sites known to be segregating between the mixed genomes showed that the modes were comprised of these alleles (yellow bars in fig. 3N and O and supplementary fig. S3, Supplementary Material online). Comparing simulated datasets generated with and without sequencing errors revealed that some, but not all, of the low frequency tails in the AFS were due to sequencing or alignment errors (supplementary fig. S3, top two panels, Supplementary Material online). Seeing this tail in unmixed, sequencing error-free alignments indicated that alignment errors also contributed to this low frequency peak (supplementary fig. S3, bottom panel, Supplementary Material online). The fact that these errors made it through the stringent filters and were called as variants suggests that systematic errors (e.g., repetitive regions of the genome that are difficult to align to) are difficult to distinguish from low frequency allelic variants in these intrahost populations.

It is important to note that the allele frequency modes were more completely comprised of sites known to be segregating between hosts in simulated data than in specimen data because known haplotypes were combined in the simulated mixtures. This can be seen as a difference in the proportion of colored bars in the frequency modes of the AFS. In the real data, some variant sites in the frequency modes are likely segregating in other host individuals, but have not yet been sampled. For example, only nine host individuals were sequenced from MA in Russell et al. 2017, making it unlikely that all variation found in MA was detected (see also supplementary fig. S1, Supplementary Material online).

Alleles at Similar Frequency Are Present on the Same Chromosomes

The variant sites contained in the allele frequency modes that segregate in other hosts are not randomly distributed along the genome, but group in haplotype blocks and colocalize to the same read/read pair (i.e., chromosome) more than would be expected by chance (table 1,supplementary fig. S4, and table S3, Supplementary Material online). Given that these different haplotypes both occur in otherwise extremely homogeneous genomes, indicates that one arose by recombination. We define haplotype blocks to be sets of variant sites correlated in allele frequency and position along the genome that are present (linked) on the same chromosome. Two measures of haplotype linkage on reads were calculated from the data: 1) the proportion of reads with a significantly associated haplotype, in which the alleles cooccur more/less often than expected by chance (chi-squared P value < 0.05; labeled "significant" in fig. 4B and supplementary table S3, Supplementary Material online) and 2) the proportion of reads with all possible variant sites within the aligned region (labeled "full support" in fig. 4B).

There are four possible pairwise outcomes of these metrics. If both 1 and 2 are high (near 100%), then support is high for all variant sites tested being in the same haplotype. If 2 is high and 1 is low (<90%), most sites tested are likely on the same chromosome, but the associations are insignificant because the minor alleles are near 50% frequency, making it challenging to estimate the expected allele phase based on allele frequencies (e.g., NC18 in fig. 4). If 1 is high and 2 is low, then some, but not all, of the tested sites are likely linked on a chromosome (e.g., 5% allele frequency in NC12 in fig. 4). Lastly, if both 1 and 2 are low, then the extracted alleles are likely not from the same chromosome (e.g., 45% allele frequency in NC2 in fig. 4). The majority of frequency classes exhibited high values for both measures 1 and 2, indicating that these alleles likely localize to the same chromosome. Variant sites extracted from specimens without multiple allele frequency modes exhibited a wider range of patterns, suggesting that only some of these sets represent complete haplotypes in the host (supplementary table S3, Supplementary Material online).

Some Variants May Be Functionally Significant

Identifying the genes in which variant sites occur suggested that some polymorphisms might confer a functional change in symbionts within host gill tissue. Across 29 intrahost symbiont populations, 1,875 genes exhibited variation within a host. Of these, 20 had variant gene sets that were enriched for GO terms (table 1 and supplementary table S4, Supplementary Material online), and these sets were partially overlapping among individuals (supplementary fig. S4A, Supplementary Material online). In contrast, of the 12 populations that exhibited multimodal allele frequencies, only four had modes with variant gene sets enriched for GO terms (table 1 and supplementary table S4, Supplementary Material online). Genes exhibiting variability in interhost symbiont subpopulations were enriched for several GO terms, and many were consistently observed between different localities (supplementary table S4 and fig. S4B, Supplementary Material online).

Functional enrichment at each of these nested levels (i.e., distinct haplotypes [frequency mode] within a host individual [intrahost population], living in a geographic subpopulation of hosts [interhost subpopulation]) was likely independent of the level above it because different sets of terms were enriched at each level. For example, the most enriched functions within intrahost populations were transporter, channel, and porin activities (fig. S4C, Supplementary Material online), whereas the most abundant functions between-host

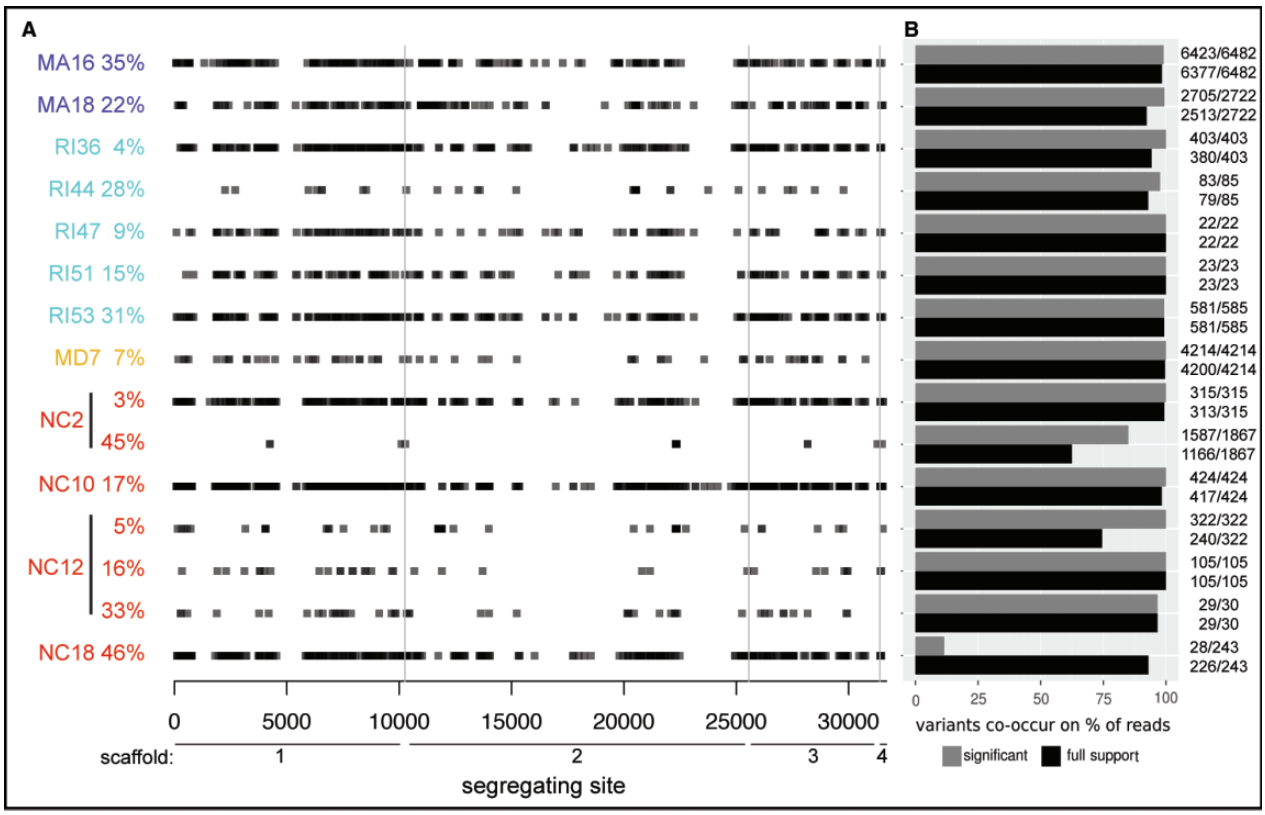

Fig. 4. Intrahost symbiont variant sites comprising AFS modes described in figure 3 are plotted by segregating site along the genome and tested for linkage on Illumina reads. (*A*) Variant sites from each sample's AFS mode with the indicated mean % (colored as in fig. 2) are plotted transparently, showing that sites with similar allele frequencies correlate in position along the genome. Vertical gray lines in plots mark breaks between scaffolds 1–4 in the symbiont reference genome. (*B*) Number of reads in support for variant sites in close proximity for each sample in A being linked on the same chromosome divided by the total number of reads covering the loci. Gray bars: Percent of read pairs with significantly more variant sites cooccurring on the same read pairs than would be expected by chance based on their frequency in the population ($P < 0.05$). Black bars: Percent of read pairs with all variant sites cooccurring on the same pair. Raw read counts are displayed to the right of the bar plot.

populations were transposition and DNA recombination and integration (see supplementary table S4, Supplementary Material online for *P* values). In contrast, there was no consistent trend in GO terms among intrahost frequency modes.

The average d*N*/d*S* for coding sequences in the *S. velum* symbiont reference genome relative to its closest relative, the *S. elarraichensis* symbiont, was 0.1118. Of the 318 stringently called orthologs between the *S. velum* and *S. elarraichensis* genomes, 305 had d*N*/d*S* values significantly different from one, which is indicative of selection at these loci (supplementary table S5, Supplementary Material online). 200 of these genes were also found variant in symbiont populations within host individuals, all of which exhibited negative d*N*/d*S* values, suggesting that these may be experiencing purifying selection to maintain gene function, e.g., as has been hypothesized for earthworms (Kjeldsen et al. 2012).

Discussion

The relative abundance of genetic diversity in intrahost symbiont populations has been a topic of much speculation (e.g., Frank 1996; Luyten et al. 2006; Vautrin et al. 2008; Wollenberg and Ruby 2009; Woyke et al. 2010; Li et al. 2013), but has been infeasible to examine robustly until very recently. Symbiont cells are pooled together within or among host cells, making these populations technically difficult to sample. Furthermore, low genetic variation within populations of bacteria makes selecting informative genetic markers difficult. The recent advent of low-cost short-read sequencing technologies, such as the Illumina HiSeq platform used in this study, have provided a means to sample intrahost populations deeply across symbiont genomes. Our study clearly demonstrates the utility of these methods for examining such, revealing evidence of intrahost symbiont diversity driven by both de novo mutation and mixed infections. With NGS, studies such as ours are now possible and will provide rapid advances in our understanding of the dynamics and evolution of intracellular symbionts, ranging from pathogens to mutualists.

The Shape of Intrahost Genetic Diversity

Populations of *S. velum* symbionts within single host specimens were found to contain substantial genetic diversity through high depth-of-coverage whole genome sequencing. The amount and patterns of diversity varied widely between intrahost populations, from few low frequency variants to numerous intermediate frequency variants (fig. 2*C*). Closer examination of the AFS for intrahost populations revealed multimodal distributions enriched in variant sites segregating

between hosts (fig. 3*A*, *B*, and *G* and supplementary fig. S2*B*, *C*, *N*, and *P*, Supplementary Material online), consistent with patterns observed in simulated mixed infections (fig. 3*N* and *O* and supplementary fig. S3, Supplementary Material online), as well as in mixed infections of *Borrelia burgdorferi* (Walter et al. 2016). Alleles from these distributions were found on the same reads far more often than expected by chance, indicating that they form distinct haplotypes. While distantly located variant sites more than an Illumina library insert length apart (ca. 350 bp, see table 1) could not be tested, the correlation between position and allele frequency over short distances detected in the read data suggests that most variant sites in a frequency mode are linked. While the majority of low-frequency variation is likely due to de novo mutation, some variants may have arisen via mixed infection with alleles at low frequency in the interhost population. These lines of evidence strongly suggest that multiple distinct symbiont haplotypes coexist within a host individual. Thus, either the intrahost symbiont populations are not yet at equilibrium or the different haplotypes do coexist and have found a way to not compete, or perhaps even cooperate (Mouton et al. 2003; Vautrin et al. 2008; Gold et al. 2009; Abkallo et al. 2015).

Collectively, these data suggest that horizontal transmission events have shaped the genetic structure of *S. velum* symbiont intrahost populations. It is extremely unlikely that the alleles at intermediate frequencies in *S. velum* intrahost populations could have arisen by de novo mutation for three reasons: 1) The overall number of variant sites found should be extremely low if all or most are attributable to mutations that occurred while dividing during host development (Lynch 2010). 2) Random mutations would occur in different genomes, so the majority of variant sites would not be expected to occur on the same chromosome. 3) In general, if variant sites were generated by mutation, they should rarely occupy the same position as sites found in other hosts in the population (Hudson 1983). In contrast, variant sites acquired from mixed infections will be comprised of sites segregating between host populations, as we observed in many of the *S. velum* specimens and in the simulated data. The intrahost populations also exhibited an abundance of mutations at the second lowest frequency class (1% above lowest detected frequency bin), compatible with the demographic model of a single recent bottleneck event (Luikart et al. 1998), indicating that additional demographic or selective effects are also needed to explain these data. Interestingly, the intrahost AFS resemble allele frequencies shaped by admixture between symbiont populations (Falush et al. 2003), illustrating the similarity between migration and mixed infection processes.

The population genetic structure seen in intrahost symbiont populations of *S. velum* appears to be a mixture of vertically transmitted and recombinant haplotypes. While a process such as diversifying selection could produce discrete haplotypes in an intrahost population, horizontal transmission with recombination is supported by finding alleles in intrahost populations 1) at intermediate frequency that can be found in other host individuals from the same locality (fig. 3), 2) that colocalize along the symbiont genome (fig. 4*A*), and 3) are present on the same chromosomes (haplotypes, fig. 4*B*), which are linked to the remainder of the invariant symbiont genome. That only fragments of the introduced symbiont genome are left in the intrahost population suggests that horizontal transmission and recombination events occurred in previous host generations, and subsequently vertical transmission homogenized genetic diversity in the population. As modeled in figure 1, the haplotype block genetic structure shown in figure 4*A* may have been produced by mixed infection events followed by recombination and inheritance. The recombinant chromosomes may ultimately go to fixation, become lost, or be maintained by diversifying selection on the haplotypes.

The variability of intrahost population diversity among specimens collected from the same locality suggests that these mixed infections are more recent than the migration events that brought the ancestors of these hosts to their habitats. It is also possible that recombinant symbiont genotypes may occur in the environment, in which case these could be examples of mixed infections in the sampled generation. It is not entirely clear whether the *S. velum* symbionts exist outside the host, as preliminary studies found only extremely low numbers in sediment and seawater from their habitat (Russell 2016). However, the fact that they undergo frequent horizontal transmission events (Russell et al. 2017) suggests that opportunities for encounters in the environment exist. More information is needed about the life history of the *S. velum* symbionts to distinguish these alternative explanations.

Sources and Sinks of Intrahost Variation

The questions of how intracellular symbionts are capable of horizontal transmission, recombination, and inheritance of variant genotypes are important to understanding the mechanistic basis of this interaction. Novel symbionts may enter adult gills via phagocytosis, as has been reported for the horizontally transmitted chemosynthetic symbionts of the bivalve *Codakia* (Gros et al. 1998) and *Wolbachia* endosymbionts of *Drosophila melanogaster*. (White et al. 2017). Alternatively, horizontal acquisition may occur earlier in development. For example, hydrothermal vent tubeworms have been shown to acquire their symbionts through their tegument after settling on the seafloor and undergoing metamorphosis (Nussbaumer et al. 2006). Following horizontal transmission, homologous recombination could take place within host cells via uptake of DNA from lysed symbiont cells or conjugation (Halkett et al. 2005). Recombination within a host has been reported for a range of bacterial pathogens (e.g., *E. coli*, *Streptococcus pneumoniae* [Didelot et al. 2016]). Alternatively, recombination could occur between symbiont lineages out in the environment prior to colonization of host tissues, potentially mediated by phages in addition to DNA uptake or conjugation (Frost et al. 2005; Rosen et al. 2015; Russell et al. 2017).

Genetic variation acquired or generated in the gill could be passed on to new generations of *S. velum* if symbionts migrate from the gill to colonize developing oocytes. This is similar in principle to how the bacterial symbionts of lice migrate from

the adult mycetome to colonize oocytes (Perotti et al. 2007) and how *Buchnera*, the primary symbionts of aphids, migrate to brooded embryos (Braendle et al. 2003). Movement between tissues could induce a reduction in genetic diversity independent of the transmission bottleneck, as has been reported for intrahost populations of the Dengue virus transferred between hosts and host tissues (Sim et al. 2015). These are just some of the possible mechanisms generating and maintaining genetic variation in symbiont populations within host individuals. More information is needed about the cellular and developmental routes of symbiont transfer in marine invertebrates in general, and *S. velum* in particular, before any more can be known about how genetic variation is generated and perpetuated in individual hosts in these associations.

Functional Consequences of Intrahost Variation

Variant haplotypes may persist in intrahost populations due to selective advantage or via neutral processes. The symbiont pathways enriched for variant sites within *S. velum* exhibited limited overlap with those enriched between hosts, suggesting that different evolutionary pressures act at these different scales. It is possible that some genetic variants provide functional benefit within the intrahost environment. For example, microscopic evidence suggests that the gill environment is quite heterogeneous. Cells at the base of gill filaments are much closer to the mucus-producing hypobranchial gland than cells towards the tips (personal observation, SLR), which could result in different accessible concentrations of O_2 and H_2S along the filaments.

We hypothesize that recombination randomly breaks up an introduced symbiont genome into fragments of the original haplotype which can become integrated in the recipient symbiont genome, producing novel haplotypes. A small subset of these different haplotypes could subsequently become adapted to particular regions of the gill, enabling their persistence in the population. In support of this idea, two distinct symbiont 16S rRNA genotypes have been observed to exhibit specific localization patterns in the gills of thyasirid clams (Fujiwara et al. 2001). However, in *S. velum* intrahost populations, finding no GO term enrichment in some of the variant gene sets, and the limited overlap in GO terms among sets suggests that these variant haplotypes may have arisen by chance, are selectively neutral (see also Renzette et al. 2016), and have not yet been lost from or fixed in the population. However, selection acting below the unit of the functional category cannot be ruled out with these analyses alone.

The persistence of multiple symbiont haplotypes within a host could allow selection to act within host tissues to purge deleterious mutations and select for advantageous ones. Numerous intracellular obligate symbionts exhibit evidence of weak purifying selection, producing an excess nonsynonymous mutations, resulting in pseudogenes in the short term (Oakeson et al. 2014), and genome erosion in the long term (Bennett et al. 2016; Moran et al. 2009). However, not all obligate intracellular symbionts, the *S. velum* symbiont included (Dmytrenko et al. 2014), exhibit these signs of erosion. One potential explanation is the accumulation of deleterious mutations may be prevented by the influx of novel genetic diversity introduced via horizontal transmission followed by recombination. In addition to being an outcome of deleterious mutation accumulation, genome erosion in bacteria can also be an adaptive process. Bacterial genomes are prone to streamlining by losing genes and becoming auxotrophic for processes supplemented by other sources (biological or inorganic) in the ecosystem (Giovannoni et al. 2014; D'Souza et al. 2014). Symbiotic bacteria often lose genes for functions supplied by the host (Bennett et al. 2014), but evolving complementary gene sets with other symbionts coinhabiting the host (López-Madrigal et al. 2014; Rao et al. 2015). For example, in the aphid *Cinara cedri* the primary symbiont, *Buchnera*, has lost some functionality that is now fulfilled by a former secondary symbiont, *Seratia symbiotica* (Lamelas et al. 2011; Manzano-Marín and Latorre 2014). Complementation can also occur from one ancestral genome, producing two, nonidentical but complementary genomes, as has occurred in bacterial symbionts of cicadas (Van Leuven et al. 2014). The introgression of novel symbiont haplotypes into intrahost populations has the potential to produce variants adapted to different conditions, which could lead to complementation. Thus, recombination between symbionts in mixed infections could provide the raw material for purifying selection, intrahost adaptation, and potentially symbiont complementation.

Materials and Methods

Collections, Sequencing, and Genome Assembly

A subset of *S. velum* gill specimens from Russell et al. (2017) (NCBI BioProject PRJNA353273) were sequenced to high-depth of coverage with Illumina sequencing in the original study (total host specimens: $n = 29$; number of specimens from each subpopulation: MA = 2, RI = 10, NJ = 5, MD = 6, NC = 6; see supplementary table S1, Supplementary Material online for specimen designations). Briefly, adult *S. velum* were collected from intertidal-subtidal sediments along the east coast of North America (fig. 2*A* and supplementary table S1, Supplementary Material online), rinsed with 0.2 μm filtered seawater, sterilely dissected, placed in 100% ethanol, flash frozen, and stored at −80 °C. Gill DNA was extracted with the DNeasy Blood and Tissue kit (Qiagen). Genomic DNA was sheared to 350 bp (Covaris S220) and Illumina paired-end libraries were made using NEXTflex adapters (Bioo) either on the Apollo 324 System (Wafergen) using the PrepX ILM kit (IntergenX) or using a custom protocol (Russell et al. 2017). Libraries were pooled and paired-end sequenced on the Illumina HiSeq2000 or HiSeq2500 platform (Bauer Core Facility, Harvard University)(see supplementary table S1, Supplementary Material online).

Reads were mapped to symbiont (Dmytrenko et al. 2014) and mitochondrial (Plazzi et al. 2013) reference genomes with Stampy 1.0.18 (Lunter and Goodson 2011). Alignments were processed with SAMTools 1.0 (Li et al. 2009), optical duplicates were removed with Picardtools (http://broadinstitute.github.io/picard/) and coverage was calculated with Bedtools 2.18.1 (Quinlan and Hall 2010). Indel realignment was

performed with the Genome Analysis Toolkit (GATK, version 3.2-2). Pileup files, reporting alignment information by genome position, were generated with mpileup in Samtools 1.0 (with parameters: –count-orphans and –max-depth 1000).

Variant Calling

Approach

As every read at a given position in each sequenced library originated from a different symbiont genome, variants were called from the pileup files directly to leverage this information. Custom perl scripts were used to perform the following filtering and output SNP allele read counts and pairwise nucleotide diversity (see below) by position: Positions within 5 bp of an indel were removed to avoid erroneous calls generated by incorrect indel alignment. To exclude regions containing potential duplications or repetitive regions relative to the reference, sites were filtered within the average genome-wide coverage $\pm$ one standard deviation. Lastly, to exclude sequencing errors, the 99% confident read coverage depth for an accurate variant call was calculated for each variant site from the cumulative binomial distribution:

$$Pr(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$

With p equal to a per-site error rate of 1% (Illumina error rate [Quail et al. 2008; Reumers et al. 2011]), n equal to the coverage at the site, and i equal to the minor allele count starting with 0 and adding one with each summation until $Pr \geq 0.99$. The final value of i (i.e., k) was used as the minimum read depth required for an alternate allele call at that site, with an absolute minimum of five at any site.

Variant sites for symbionts from different host individuals were identified from consensus sequences called with the Unified Genotyper in the Genome Analysis Toolkit for *S. velum* individuals (from Russell et al. 2017) using a custom perl script.

Validation of Methodology

Variant calls were confirmed with additional software. Both Population (Kofler et al. 2011) and the GATK's Unified Genotyper (DePristo et al. 2011) require the ploidy of the specimen to be known. Symbiont ploidy is equal to the number of symbiont lineages present per host, but is difficult to measure and introduces computational challenges because it is potentially very large. A lower bound estimate can be obtained by using the number of symbionts per host cell, as this is a first-order organizational constraint on intracellular symbiont reproduction and dispersal. This number was estimated from the ratio of host nuclear coverage to symbiont genome coverage. Given that symbionts occur in roughly half of gill cells (personal observation, SLR) and there are two copies of the nuclear genome per host cell, the average ratio of symbiont coverage to nuclear coverage was 100:1 (symbiont copies/cell/(2 nuclear copies/cell * 0.5 cells) = coverage ratio), indicating that there are roughly 100 symbionts per cell. This number is consistent with estimates made from solemyid cell morphology (Cavanaugh 1983; Krueger et al. 1996; Taylor et al. 2008) and for other chemosynthetic symbioses (Halary et al. 2008; Brissac et al. 2009; Decker et al. 2013). Population diversity statistics and variants were called with ploidy of 100 using Population and GATK, obtaining similar results to the pileup method described above, except that these methods did not exclude regions of anomalous coverage and GATK did not exclude sites around indels. In addition, EVORhA (Pulido-Tamayo et al. 2015) detected multiple haplotypes, corroborating evidence of mixed infections.

False-Positive Test for Diversity Detection

The level of mitochondrial diversity within each host specimen was investigated to confirm that the diversity of this uniparentally transmitted genome is low (Russell et al. 2017). Mitochondrial alignments were processed as described for symbionts.

Evaluation of Diversity Detection

To evaluate the extent of intrahost diversity observed by the sequencing coverage of each specimen, rarefaction curves were computed. Alignment coverage was subsampled randomly at 1, 5, 10, 20, 30, 50, 70, and 90% coverage with the view command in SAMTools 1.0 (Li et al. 2009). Alignments were converted to pileup files for variant calling. Variants were tallied for each subsample and plotted against subsample size by fraction and absolute coverage in R (R Core Development Team 2012). Allele frequency spectra were also computed for each subsample.

Analysis

Genome-Wide Pairwise Nucleotide Diversity

Genetic diversity was measured by calculating pairwise nucleotide diversity (π) (Nei and Li 1979) by site and averaging across 10 kb nonoverlapping windows:

$$\pi = \frac{\sum_{j=1}^{l} \frac{\sum_{i=1}^{a} x_i(n-x_i)}{n(n-1)}}{L}$$

Where n is the total number of individuals sampled, x_i is the number of individuals with allele i, a is the total number of sampled alleles at site j of l total sites, and L is the sequence length. Diversity was calculated within host specimens from the Illumina read alignments and between hosts from alignments of whole-genome consensus genotypes.

Allele Frequency Spectra

Folded allele frequency spectra (AFS) were calculated from allele counts and the distributions were plotted using the hist function in R (R Core Development Team 2012). For each set of intrahost segregating sites, three sets of AFS were plotted and overlain in the following order: 1) the full set of segregating sites (gray), 2) the subset of sites segregating in the entire between-host population (in white), and 3) the subset segregating in different hosts from the same geographic locality as the sample (in color according to fig. 2).

Simulated Mixed Infection

Mixed infections were simulated by mixing symbiont consensus sequences called from different host individuals. Mixtures were generated using reads simulated from consensus sequences rather than reads from the samples themselves to produce samples consisting of exactly two genotypes, opposed to an unknown mixture. Symbiont consensus sequences for MA16, MA36, RI38, RI39, RI53, NC9, NC12, and NC22 (from Russell et al. 2017), which range from 99.916% to 99.987% identical, were used to generate simulated Illumina read sets with wgsim (150 bp reads, 2×10^6 reads/specimen, 350 bp mean insert length, 0.01 error rate; Li et al. [2009]) with sequencing errors, as well as without. These reads were mapped to the reference symbiont genome with Stampy as described above. These pseudo-monoclonal samples were mixed in 10/90, 20/80, and 30/70 ratios, and variants were called and analyzed as described above. Variants were also called from alignments of mixed and unmixed simulated reads with no errors to assess the efficacy of variant call filtering on alignment error.

Variant Linkage

To test whether the alleles contained in the observed allele frequency modes were derived from the same chromosomes, and thus constituted novel haplotypes, read linkage patterns were investigated. Alleles within frequency modes (mean of mode $\pm$ 4%) were extracted, plotted by relative position along the genome and tested for presence on the same read or read pair with custom perl scripts. Each alignment file was parsed read-by-read, recording reads aligning to variant site positions. For each read pair, the number of variant sites contained in the pair was divided by the number of variant sites within the read pair range to compute the fraction of variant sites found on the same chromosome. The significance of the variant distribution among reads was calculated with the chi-squared test using the Statistics::Distributions perl module, with degrees of freedom equal to the number of variants in range minus 1, and a 5% significance cutoff. Expected values (chance of observing all variants sites on one read pair by chance) were equal to the product of variant allele frequencies within the read range. Recombination events can be inferred from finding sections of alleles in linkage, as they all have much of the same sequence across the rest of the genome and therefore require a recombination event into the background genotype to produce the observed haplotype.

Haplotype Functional Analysis

Gene ontology (GO) terms were annotated in the *S. velum* symbiont reference genome using Blast2GO (Conesa et al. 2005). Blast results for amino acid sequences for each gene were obtained with command line blast (blastp, *e*-value 1e-6). Genes with high numbers of variant sites were identified for two partitions of the data: variant sites in intrahost allele frequency modes and intrahost populations. Gene sets exhibiting high amounts of variation were tested for over enrichment in specific functions relative to the reference genome by gene ontology analysis in Blast2Go (implementing the protocol in Al-Shahrour et al. [2004]). Both *P* values and false discovery rates (FDR) were calculated, and results were filtered by FDR < 0.05. Enrichment results were compared between gene sets variant in intrahost populations and between host-populations to assess whether these results are largely independent, or if intrahost enrichment is a byproduct of between-host enrichment. The abundance of variant gene functional types was also investigated by comparing gene product abundances and visualized with word clouds (woordle.net).

Substitution effects on coding sequences were predicted for the *S. velum* symbiont (dN/dS). To calculate dN/dS, coding sequences from the *S. velum* symbiont reference sequence were aligned to their homologs in its closest known relative, the *Solemya elarraichensis* symbiont (Russell et al. 2017) with the codon-aware aligner, MACSE (Ranwez et al. 2011). Coding sequences were included only if they consistently mapped with average coverage in the majority of samples. Homologs in the *S. elarraichensis* symbiont genome were identified by best-reciprocal-blast-hits (Moreno-Hagelsieb and Latimer 2008). dN/dS values were calculated for the *S. velum* symbiont coding sequences in codeml in PAML version 4.8 (Yang 2007), using default parameters to estimate omega and calculate the log likelihood of omega for the estimated as well as a fixed omega of 1.

All perl scripts used in these analyses are available at https://github.com/shelbirussell/Russell-and-Cavanaugh-2017.

Supplementary Material

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

Acknowledgments

We thank Russ Corbett-Detig, Gonzalo Giribet, Peter Girguis, Cassandra Extavour, members of the Cavanaugh Lab, and two anonymous reviewers for valuable suggestions and comments. This work was supported by Harvard University's William F. Milton Fund, Department of Organismic and Evolutionary Biology, and Microbial Sciences Initiative.

References

Abkallo HM, Tangena J-A, Tang J, Kobayashi N, Inoue M, Zoungrana A, Colegrave N, Culleton R. 2015. Within-host competition does not select for virulence in malaria parasites; studies with *Plasmodium yoelii*. *PLoS Pathog.* 11:e1004628.

Al-Shahrour F, Diaz-Uriarte R, Dopazo J. 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20:578–580.

Amann R, Ludwig W, Schleifer K. 1995. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microb Mol Biol Rev.* 59:143–169.

Bennett GM, McCutcheon JP, MacDonald BR, Romanovicz D, Moran NA. 2014. Differential genome evolution between companion symbionts in an insect-bacterial symbiosis. *mBio.* 5:e01697–14.

Bennett GM, Moran NA. 2015. Heritable symbiosis: the advantages and perils of an evolutionary rabbit hole. *Proc Natl Acad Sci U S A.* 112:10169–10176.

Bennett GM, McCutcheon JP, McDonald BR, Moran NA. 2016. Lineage-specific patterns of genome deterioration in obligate symbionts of sharpshooter leafhoppers. *GBE* 8:296–301.

Braendle C, Miura T, Bickel R, Shingleton AW, Kambhampati S, Stern DL. 2003. Developmental origin and evolution of bacteriocytes in the aphid: *Buchnera* symbiosis. *PLoS Biol.* 1:e21.

Brissac T, Gros O, Merçot H. 2009. Lack of endosymbiont release by two Lucinidae (Bivalvia) of the genus *Codakia*: consequences for symbiotic relationships. *FEMS Microb Ecol*. 67:261–267.

Cavanaugh CM. 1983. Symbiotic chemoautotrophic bacteria in marine invertebrates from sulphide-rich habitats. *Nature* 302:58–61.

Cavanaugh CM, McKiness ZP, Newton I, Stewart FJ. 2013. Marine chemosynthetic symbioses. In: Rosenberg, E, editor. The prokaryotes: prokaryotic biology and symbiotic associations. Berlin, Heidelberg: Springer-Verlag. p. 579–607.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674–3676.

Costello EK, Stagaman K, Dethlefsen L, Bohannan BJM, Relman DA. 2012. The application of ecological theory toward an understanding of the human microbiome. *Science* 336:1255–1262.

Coyte KZ, Schluter J, Foster KR. 2015. The ecology of the microbiome: networks, competition, and stability. *Science* 350:663–666.

D'Souza G, Waschina S, Pande S, Bohl K, Kaleta C, Kost C. 2014. Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria. *Evolution* 68:2559–2570.

de Roode JC, Pansini R, Cheesman SJ, Helinski ME, Huijben S, Wargo AR, Bell AS, Chan BH, Walliker D, Read AF. 2005. Virulence and competitive ability in genetically diverse malaria infections. *Proc Natl Acad Sci U S A*. 102:7624–7628.

Decker C, Olu K, Arnaud-Haond S, Duperron S. 2013. Physical proximity may promote lateral acquisition of bacterial symbionts in vesicomyid clams. *PLoS One* 8:e64830.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 43:491–498.

Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. 2016. Within-host evolution of bacterial pathogens. *Nat Rev Micro*. 14:150–162.

Distel DL, Lee HK, Cavanaugh CM. 1995. Intracellular coexistence of methano- and thioautotrophic bacteria in a hydrothermal vent mussel. *Proc Natl Acad Sci U S A*. 92:9598–9602.

Dmytrenko O, Russell SL, Loo WT, Fontanez KM, Liao L, Roeselers G, Sharma R, Stewart FJ, Newton IL, Woyke T, et al. 2014. The genome of the intracellular bacterium of the coastal bivalve, *Solemya velum*: a blueprint for thriving in and out of symbiosis. *BMC Genomics* 15:924.

Duperron S, Quiles A, Szafranski KM, Léger N. 2016. Estimating symbiont abundances and gill surface areas in specimens of the hydrothermal vent mussel *Bathymodiolus puteoserpentis* maintained in pressure vessels. *Front Mar Sci*. 3:16.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–1587.

Fay SA, Weber MX. 2012. The occurrence of mixed infections of *Symbiodinium* (Dinoflagellata) within individual hosts. *J Phycol*. 48:1306–1316.

Flórez LV, Biedermann PHW, Engl T, Kaltenpoth M. 2015. Defensive symbioses of animals with prokaryotic and eukaryotic microorganisms. *Nat Prod Rep*. 32:904–936.

Frank S. 1996. Host-symbiont conflict over the mixing of symbiotic lineages. *Proc R Soc Lond*. 263:339–344.

Frost LS, Leplae R, Summers AO, Toussaint A. 2005. Mobile genetic elements: the agents of open source evolution. *Nat Rev Micro*. 3:722–732.

Fujiwara Y, Kato C, Masui N, Fujikura K, Kojima S. 2001. Dual symbiosis in the cold-seep thyasirid clam *Maorithyas hadalis* from the hadal zone in the Japan Trench, western Pacific. *Mar Ecol*. 214:151–159.

Galtier N, Depaulis F, Barton NH. 2000. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics* 155:981–987.

Giovannoni SJ, Cameron Thrash J, Temperton B. 2014. Implications of streamlining theory for microbial ecology. *ISME J.* 8: 1553–1565.

Goffredi SK, Yi H, Zhang Q, Klann JE, Struve IA, Vrijenhoek RC, Brown CT. 2014. Genomic versatility and functional variation between two dominant heterotrophic symbionts of deep-sea *Osedax* worms. *ISME J.* 8:908–924.

Gold A, Giraud T, Hood ME. 2009. Within-host competitive exclusion among species of the anther smut pathogen. *BMC Ecol*. 9:11.

Greiner S, Sobanski J, Bock R. 2014. Why are most organelle genomes transmitted maternally?. *Bioessays* 37:80–94.

Gros O, Frenkiel L, Moueeza M. 1998. Gill filament differentiation and experimental colonization by symbiotic bacteria in aposymbiotic juveniles of *Codakia orbicularis* (Bivalvia: Lucinidae). *Invert Repro Dev*. 34:219–232.

Grubaugh ND, Weger-Lucarelli J, Murrieta RA, Fauver JR, Garcia-Luna SM, Prasad AN, Black WC, Ebel GD. 2016. Genetic drift during systemic arbovirus infection of mosquito vectors leads to decreased relative fitness during host switching. *Cell Host Microbe* 19:481–492.

Halary S, Riou V, Gaill F, Boudier T, Duperron S. 2008. 3D FISH for the quantification of methane- and sulphur-oxidizing endosymbionts in bacteriocytes of the hydrothermal vent mussel *Bathymodiolus azoricus*. *ISME J.* 2:284–292.

Halkett F, Simon J-C, Balloux F. 2005. Tackling the population genetics of clonal and partially clonal organisms. *Trends Ecol Evol*. 20:194–201.

Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. *Theor Pop Biol*. 23:183–201.

Hugenholtz P, Goebel B, Pace N. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol*. 180:4765–4774.

Kaltenpoth M, Goettler W, Koehler S, Strohm E. 2010. Life cycle and population dynamics of a protective insect symbiont reveal severe bottlenecks during vertical transmission. *Evol Ecol*. 24:463–477.

Kjeldsen KU, Bataillon T, Pinel N, De Mita S, Lund MB, Panitz F, Bendixen C, Stahl DA, Schramm A. 2012. Purifying selection and molecular adaptation in the genome of *Verminephrobacter*, the heritable symbiotic bacteria of earthworms. *Genome Biol Evol*. 4:307–315.

Klose J, Polz MF, Wagner M, Schimak MP, Gollner S, Bright M. 2015. Endosymbionts escape dead hydrothermal vent tubeworms to enrich the free-living population. *Proc Natl Acad Sci U S A*. 112:11300–11305.

Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, Kosiol C, Schlötterer C. 2011. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One* 6:e15925.

Komaki K, Ishikawa H. 2000. Genomic copy number of intracellular bacterial symbionts of aphids varies in response to developmental stage and morph of their host. *Insect Biochem Mol Biol.* 30:253–258.

Krueger DM, Gustafson RG, Cavanaugh CM. 1996. Vertical transmission of chemoautotrophic symbionts in the bivalve *Solemya velum* (Bivalvia: Protobranchia). *Biol Bull*. 190:195–202.

Lamelas A, Gosalbes MJ, Manzano-Marín A, Peretó J, Moya A, Latorre A. 2011. *Serratia symbiotica* from the aphid *Cinara cedri*: a missing link from facultative to obligate insect endosymbiont. *PLoS Genet.* 7:e1002357.

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A*. 82:6955–6959.

Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, Botstein D, Desai MM. 2013. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* 500:571–574.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079.

Li Y, Thompson CM, Trzciński K, Lipsitch M. 2013. Within-host selection is limited by an effective population of *Streptococcus pneumoniae* during nasopharyngeal colonization. *Infect Immun*. 81:4534–4543.

López-Madrigal S, Beltrà A, Resurrección S, Soto A, Latorre A, Moya A, Gil R. 2014. Molecular evidence for ongoing complementarity and horizontal gene transfer in endosymbiotic systems of mealybugs. *Front Microbiol.* 5:449.

Luikart G, Allendorf FW, Cornuet JM, Sherwin WB. 1998. Distortion of allele frequency distributions provides a test for recent population bottlenecks. *J Hered*. 89:238–247.

Lunter G, Goodson M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res*. 21:936–939.

Luyten YA, Thompson JR, Morrill W, Polz MF, Distel DL. 2006. Extensive variation in intracellular symbiont community composition among members of a single population of the wood-boring bivalve *Lyrodus pedicellatus* (Bivalvia: Teredinidae). *Appl Envir Microb*. 72:412–417.

Lynch M. 2010. Evolution of the mutation rate. *Trends Genet*. 26:345–352.

Manzano-Marín A, Latorre A. 2014. Settling down: the genome of *Serratia symbiotica* from the aphid *Cinara tujafilina* zooms in on the process of accommodation to a cooperative intracellular life. *Genome Biol Evol*. 6:1683–1698.

McFall-Ngai MJ. 2014. Divining the essence of symbiosis: insights from the squid-*Vibrio* model. *PLoS Biol*. 12:

McFall-Ngai M, Hadfield MG, Bosch TC, Carey HV, Domazet-Lošo T, Douglas AE, Dubilier N, Eberl G, Fukami T, Gilbert SF, et al. 2013. Animals in a bacterial world, a new imperative for the life sciences. *Proc Natl Acad Sci U S A*. 110:3229–3236.

Mira A, Moran NA. 2002. Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. *Microb Ecol*. 44:137–143.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A*. 93:2873–2878.

Moran NA, McLaughlin HJ, Sorek R. 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* 323:379–82.

Moreno-Hagelsieb G, Latimer K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* 24:319–324.

Mouton L, Henri H, Bouletreau M, Vavre F. 2003. Strain-specific regulation of intracellular *Wolbachia* density in multiply infected insects. *Mol Ecol*. 12:3459–3465.

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

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A*. 76:5269–5273.

Nielsen R. 2005. Molecular signatures of natural selection. *Annu Rev Genet*. 39:197–218.

Nussbaumer A, Fisher C, Bright M. 2006. Horizontal endosymbiont transmission in hydrothermal vent tubeworms. *Nature* 441:345–348.

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.

Perotti MA, Allen JM, Reed DL, Braig HR. 2007. Host-symbiont interactions of the primary endosymbiont of human head and body lice. *FASEB J*. 21:1058–1066.

Petersen JM, Zielinski FU, Pape T, Seifert R, Moraru C, Amann R, Hourdez S, Girguis PR, Wankel SD, Barbe V, et al. 2011. Hydrogen is an energy source for hydrothermal vent symbioses. *Nature* 476:176–180.

Pham VHT, Kim J. 2012. Cultivation of unculturable soil bacteria. *Trends Biotechnol*. 30:475–484.

Plazzi F, Ribani A, Passamonti M. 2013. The complete mitochondrial genome of *Solemya velum* (Mollusca: Bivalvia) and its relationships with Conchifera. *BMC Genomics* 14:409.

Pulido-Tamayo S, Sánchez-Rodríguez A, Swings T, Van den Bergh B, Dubey A, Steenackers H, Michiels J, Fostier J, Marchal K. 2015. Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations. *Nucleic Acids Res*. 43:e105–e105.

Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 5:1005–1010.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

R Core Development Team. 2012. R: a language and environment for statistical computing. Vienna, Austria: the R Foundation for Statistical Computing.

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS One* 6:e22594.

Rao Q, Rollat-Farnier P-A, Zhu D-T, Santos-Garcia D, Silva FJ, Moya A, Latorre A, Klein CC, Vavre F, Sagot M-F, et al. 2015. Genome reduction and potential metabolic complementation of the dual endosymbionts in the whitefly *Bemisia tabaci*. *BMC Genomics*. 16:226.

Renzette N, Pfeifer SP, Matuszewski S, Kowalik TF, Jensen JD. 2016. On the analysis of intra-host and inter-host viral populations: human cytomegalovirus as a case study of pitfalls and expectations. *J Virol*. 91. pii:e01976-16.

Reumers J, De Rijk P, Zhao H, Liekens A, Smeets D, Cleary J, Van Loo P, Van Den Bossche M, Catthoor K, Sabbe B, et al. 2011. Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nat Biotechnol*. 30:61–68.

Reuter M, Keller L. 2003. High levels of multiple *Wolbachia* infection and recombination in the ant *Formica exsecta*. *Mol Biol Evol*. 20:748–753.

Rey FE, et al. 2013. Metabolic niche of a prominent sulfate-reducing human gut bacterium. *Proc Natl Acad Sci U S A*. 110:13582–13587.

Rosen MJ, Davison M, Bhaya D, Fisher DS. 2015. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science* 348:1019–1023.

Russell SL. 2016. Mode and fidelity of bacterial symbiont transmission and its impact on symbiont genome evolution. PhD thesis. Cambridge, MA: Harvard University, p. 123–139.

Russell SL, Corbett-Detig R, Cavanaugh CM. 2017. Mixed transmission modes and dynamic genome evolution in an obligate animal-bacterial symbiosis. *ISME J*. 11:1359–1371.

Sender R, Fuchs S, Milo R. 2016. Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol*. 14:e1002533.

Siefert JL, Fox GE. 1998. Phylogenetic mapping of bacterial morphology. *Microbiology* 144:2803–2808.

Sim S, Aw P, Wilm A, Teoh G, Hue K. 2015. Tracking dengue virus intra-host genetic diversity during human-to-mosquito transmission. *PLoS Negl Trop Dis*. 9:e0004052.

Stanley SM. 1970. *Relation of shell form to life habits of the Bivalvia (Mollusca). Boulder, Colo: Geological Society of America Memoir* 125. p. 119–121, 132–133.

Stephens WZ, Wiles TJ, Martinez ES, Jemielita M, Burns AR, Parthasarathy R, Bohannan BJM, Guillemin K. 2015. Identification of population bottlenecks and colonization factors during assembly of bacterial communities within the zebrafish intestine. *mBio* 6:e01163-15.

Stewart FJ, Baik A, Cavanaugh CM. 2009. Genetic subdivision of chemosynthetic endosymbionts of *Solemya velum* along the southern New England coast. *Appl Envir Microb*. 75:6005–6007.

Stewart FJ, Cavanaugh CM. 2006. Bacterial endosymbioses in *Solemya* (Mollusca: Bivalvia): model systems for studies of symbiont–host adaptation. *Antonie Van Leeuwenhoek* 90:343–360.

Stewart FJ, Cavanaugh CM. 2009. Pyrosequencing analysis of endosymbiont population structure: co-occurrence of divergent symbiont lineages in a single vesicomyid host clam. *Envir Microb*. 11:2136–2147.

Taylor JD, Glover EA, Williams ST. 2008. Ancient chemosynthetic bivalves: systematics of Solemyidae from eastern and southern Australia (Mollusca: Bivalvia). *Mem Queensl Mus*. 54:75–104.

van Leeuwenhoek A. 1800. The select works of Anthony Van Leeuwenhoek. London: G. Sidney.

Van Leuven JT, Meister RC, Simon C, McCutcheon JP. 2014. Sympatric speciation in a bacterial endosymbiont results in two genomes with the functionality of one. *Cell*. 158:1270–1280.

Vautrin E, Genieys S, Charles S, Vavre F. 2008. Do vertically transmitted symbionts co-existing in a single host compete or cooperate? A modelling approach. *J Evol Biol*. 21:145–161.

Walter KS, Carpi G, Evans BR, Caccone A, Diuk-Wasser MA. 2016. Vectors as epidemiological sentinels: patterns of within-tick *Borrelia burgdorferi* diversity. *PLoS Pathog*. 12:e1005759.

Weigel BL, Erwin PM. 2016. Intraspecific variation in microbial symbiont communities of the sun sponge, *Hymeniacidon*

heliophila, from intertidal and subtidal habitats. *Appl Environ Microb*. 82:650–658.
White PM, Pietri JE, Debec A, Russell SL, Patel B, Sullivan W. 2017. Mechanisms of horizontal cell-to-cell transfer of *Wolbachia* in *Drosophila melanogaster*. *Appl Environ Microbiol*. 83. pii: e03425-16.
Wollenberg MS, Ruby EG. 2009. Population structure of *Vibrio fischeri* within the light organs of *Euprymna scolopes* squid from Two Oahu (Hawaii) populations. *Appl Envir Microb*. 75:193–202.
Worby CJ, Lipsitch M, Hanage WP. 2014. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS Comp Biol*. 10:e1003549.
Woyke T, Tighe D, Mavromatis K, Clum A, Copeland A, Schackwitz W, Lapidus A, Wu D, McCutcheon JP, McDonald BR, et al. 2010. One bacterial cell, one complete genome. *PLoS One* 5:e10314.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.
Yang Z. 1996. Statistical properties of a DNA sample under the finite-sites model. *Genetics* 144:1941–1950.
Young KD. 2007. Bacterial morphology: why have different shapes?. *Curr Opin Microb*. 10:596–600.
Zhang Y-C, Cao W-J, Zhong L-R, Godfray HCJ, Liu X-D. 2016. Host plant determines the population size of an obligate symbiont (*Buchnera aphidicola*) in aphids. *Appl Envir Microb*. 82:2336–2346.