2020

Horizontal Transmission and Recombination Maintain forever Young Bacterial Symbiont Genomes

Shelbi Russell

Evan Pepper-Tunick

Jesper Svedberg Svedberg

Ashley Byrne

Jennie Ruelas Castillo

See next page for additional authors

Citation/Publisher Attribution

Russell SL, Pepper-Tunick E, Svedberg J, Byrne A, Ruelas Castillo J, Vollmers C, et al. (2020) Horizontal transmission and recombination maintain forever young bacterial symbiont genomes. *PLoS Genet* 16(8): e1008935. https://doi.org/10.1371/journal.pgen.1008935

PLOS GENETICS

RESEARCH ARTICLE

Horizontal transmission and recombination maintain forever young bacterial symbiont genomes

Shelbi L. Russell[1,2]*, **Evan Pepper-Tunick**[2,3], **Jesper Svedberg**[2,3], **Ashley Byrne**[1], **Jennie Ruelas Castillo**[1], **Christopher Vollmers**[2,3], **Roxanne A. Beinart**[4], **Russell Corbett-Detig**[2,3]*

1 Department of Molecular Cellular and Developmental Biology. University of California Santa Cruz, Santa Cruz, California, United States of America, **2** Department of Biomolecular Engineering. University of California Santa Cruz, Santa Cruz, California, United States of America, **3** Genomics Institute, University of California, Santa Cruz, California, United States of America, **4** Graduate School of Oceanography. University of Rhode Island, Narragansett, Rhode Island, United States of America

* shelbilrussell@gmail.com (SLR); russcd@gmail.com (RCD)

Citation: Russell SL, Pepper-Tunick E, Svedberg J, Byrne A, Ruelas Castillo J, Vollmers C, et al. (2020) Horizontal transmission and recombination maintain forever young bacterial symbiont genomes. PLoS Genet 16(8): e1008935. https://doi.org/10.1371/journal.pgen.1008935

Editor: Xavier Didelot, University of Warwick, UNITED KINGDOM

Received: November 10, 2019

Accepted: June 16, 2020

Published: August 25, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1008935

Data Availability Statement: Data and genome assemblies generated in this study are available through NCBI BioProject number PRJNA562081 (BioSample numbers listed in S1 Table). Code

Abstract

Bacterial symbionts bring a wealth of functions to the associations they participate in, but by doing so, they endanger the genes and genomes underlying these abilities. When bacterial symbionts become obligately associated with their hosts, their genomes are thought to decay towards an organelle-like fate due to decreased homologous recombination and inefficient selection. However, numerous associations exist that counter these expectations, especially in marine environments, possibly due to ongoing horizontal gene flow. Despite extensive theoretical treatment, no empirical study thus far has connected these underlying population genetic processes with long-term evolutionary outcomes. By sampling marine chemosynthetic bacterial-bivalve endosymbioses that range from primarily vertical to strictly horizontal transmission, we tested this canonical theory. We found that transmission mode strongly predicts homologous recombination rates, and that exceedingly low recombination rates are associated with moderate genome degradation in the marine symbionts with nearly strict vertical transmission. Nonetheless, even the most degraded marine endosymbiont genomes are occasionally horizontally transmitted and are much larger than their terrestrial insect symbiont counterparts. Therefore, horizontal transmission and recombination enable efficient natural selection to maintain intermediate symbiont genome sizes and substantial functional genetic variation.

Author summary

Symbiotic associations between bacteria and eukaryotes are ubiquitous in nature and have contributed to the evolution of radically novel phenotypes and niches for the involved partners. New metabolic or physiological capacities that arise in these associations are typically encoded by the bacterial symbiont genomes. However, the association itself endangers the retention of bacterial genomic coding capacity. Endosymbiont genome evolution

written and used in our analyses is available from https://github.com/shelbirussell/ForeverYoungGenomes_Russell-et-al. Underlying numerical data for all graphs and summary statistics are available as Supporting Information.

Funding: This work was supported by UC Santa Cruz, Harvard University, the Alfred P. Sloan Foundation (to RCD; sloan.org), and the NIH (R35GM128932 to RCD; nih.gov). Funding for Lau Basin collections was provided by the Schmidt Ocean Institute and NSF (OCE-1819530 to RB; nsf.gov), Funding for the Veatch Canyon collection was provided via a UNOLS Early Career Training Cruise Program funded by the NSF (OCE-1641453, OCE-1638805, OCE-1214335, OCE- 1655587, and OCE-1649756; nsf.gov) and the ONR (N00014–15-1–2583; onr.navy.mil). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

theory predicts that when bacterial symbionts become restricted to host tissues, their populations cannot remove deleterious mutations efficiently. This ultimately results in their genomes degrading to small, function-poor states, reminiscent of organellar genomes. However, many ancient marine endosymbionts do not fit this prediction, but instead retain relatively large, gene-rich genomes, indicating that the evolutionary dynamics of this process need more thorough characterization. Here we show that on-going symbiont gene flow via horizontal transmission between bivalve hosts and recombination among divergent gammaproteobacterial symbiont lineages are sufficient to maintain large and dynamic bacterial symbiont genomes. These findings indicate that many obligately associated symbiont genomes may not be as isolated from one another as previously assumed and are not on a one way path to degradation.

Introduction

Bacterial genomes encode an enormous diversity of functions, which enable them to create radically novel phenotypes when they associate with eukaryotic hosts. However, they are at risk of genome degradation and function loss within these associations. Among the diversity of eukaryotic hosts they inhabit, symbiont genome sizes range from nearly unreduced genomes that are similar to their free-living relatives (~3–5 Mb) to highly reduced genomes that are less than 10% of the size of their free-living ancestors (~0.2–0.6 Mb) [1,2]. The degradation process often leaves genes required for the bacterium's role in the symbiosis, and removes seemingly vital genes, such as those involved in DNA repair and replication [1]. Although genome erosion may be enabled in some cases due to "streamlining" benefits [3], it is clear that the process can become problematic and can ultimately result in symbiont replacement or supplementation [1,4] to accomplish the full repertoire of functions needed in the combined organism. Thus, to fully understand how symbioses evolve, we must first understand the pressures their genomes experience.

Symbiont genome evolution theory predicts that upon host restriction, bacterial genomes begin the steady and inexorable process of decay due to decreased population sizes and homologous recombination rates, which result in inefficacious natural selection [5]. The transmission bottleneck that occurs when symbionts are passed on to host offspring further exacerbates these dynamics by making deleterious mutations more likely to drift to fixation in the next generation [5,6]. Indeed, many endosymbiont taxa exhibit weak purifying selection at the gene level [7–10]. In the early stages of bacterial symbiont-host associations, deleterious mutations arise on each symbiont chromosome and a portion drift or hitchhike with adaptive mutations to fixation. Subsequently, pseudogenes are lost entirely via deletion [5]. Ultimately, this process is thought to result in an organelle-like genome that is a fraction of the size of its free-living ancestors and has relegated many of its core cellular functions (*e.g.*, DNA replication and repair, cell wall synthesis, etc.) to the host or lost them entirely.

While the degraded genomes of many endosymbionts are consistent with this general theory, such as symbionts of terrestrial sap and xylem-feeding insects [1], a diversity of associations present discrepancies. In particular, most known marine obligate endosymbionts' genomes are at least one megabase in size and contain diverse coding content [11–17], although more reduced representatives have recently been found [18]. Large symbiont genomes in obligate associations are usually interpreted as reflecting the earliest stages of genome degradation [19,20], which is surprising considering the antiquity of many of these associations (*e.g.*, [21,22]). Furthermore, even the vertically transmitted marine symbionts

whose phylogenies mirror those of their hosts have only partially degraded genomes [23]. While important genes have been lost in some lineages, such as the recombination and repair gene *recA* and transversion mismatch-repair gene *mutY* [24], their loss did not portend an organelle-like fate.

Symbioses in marine environments exhibit significantly more horizontal transmission between hosts than those in terrestrial environments [25], suggesting that symbiont gene flow between hosts may prevent genome decay by enabling high rates of homologous recombination, efficient natural selection, and the maintenance of highly diverse genome contents. This hypothesis has so far not been tested and it is not known whether marine endosymbionts represent the early stages of genome degradation, as implied by the canonical endosymbiont genome theory, or if on-going gene flow and recombination can stall such genome decay over evolutionary time. Furthermore, although the general process of symbiont genome degradation is well-understood in theory (*e.g.*, [1,2,26]), no empirical study has directly evaluated the role of horizontal transmission and recombination in facilitating efficient natural selection and, thereby, the suppression of degradation. Interspecific-population level comparisons are essential for testing these important and long-standing questions [26].

To determine which evolutionary forces prevent marine endosymbiont genomes from degrading despite host restriction, we leveraged both population and comparative genomics of six marine bacterial-animal symbioses from three host taxa that exhibit modes of transmission across the spectrum from strict horizontal transmission (*Bathymodiolus* mytilid mussels) [27–29], to mixed mode transmission (solemyid bivalves) [15,30–32], to nearly strict vertical transmission (vesicomyid clams) [23,33–35] (see Fig 1A and S1–S3 Tables). Each of these three groups evolved independently (see Fig 2) to obligately host either a single intracellular gammaproteobacterial symbiont 16S rRNA phylotype within their gill cells or two or more phylotypes, in the case of some mussels [36,37]. These symbionts provide chemosynthetic carbon fixation, either through sulfide or methane-oxidation, to nutritionally support the association [25]. Hosts are nearly completely dependent on symbiont metabolism [38], and the solemyids have lost the majority of their digestive tracts in response [39]. These associations appear to be obligate for the symbionts as well as the hosts because either the symbionts have never been found living independently, *e.g.*, solemyid [31] and vesicomyid symbionts [40], or they have only been found in the host and surrounding environment, *e.g.*, bathymodiolin symbionts [27,41,42]. Both the vesicomyids and solemyids transmit symbionts to their offspring through allocating tens to hundreds of symbiont cells to their broadcast spawned oocytes [31,33]. While the mechanism of horizontal, host-to-host transfer has not been identified in the *Bathymodiolus* or solemyid symbionts, signatures of rampant horizontal transmission are evident in the population genetics of both of these groups [15,27,28] and the developmental biology of *Bathymodiolus* [29].

This group of marine symbioses presents an ideal system in which to test the impact of transmission mode and homologous recombination rate on bacterial symbiont genome evolution. The wealth of information known about how the vesicomyid, solemyid, and bathymodiolin symbioses function and evolve makes this a powerful evolutionary model. Furthermore, the similarity of these associations to other invertebrate-bacterial associations in the marine environment, *e.g.*, in mode of reproduction (broadcast spawning), phylogeny (Mollusca-Gammaproteobacteria), immunology (innate), ancestral feeding type (filter or deposit feeders), symbiosis function (nutritional), makes this an ideal microcosm for understanding how symbiont population genetics impact the process of symbiont genome degradation. Here, we use this model study system and a powerful population genomic approach (described in S1 Fig and Sections 1-3 of S1 Text), to show that homologous recombination is ongoing even in the most strictly vertically transmitted associations and may enable the maintenance of large and intermediate genome sizes indefinitely.

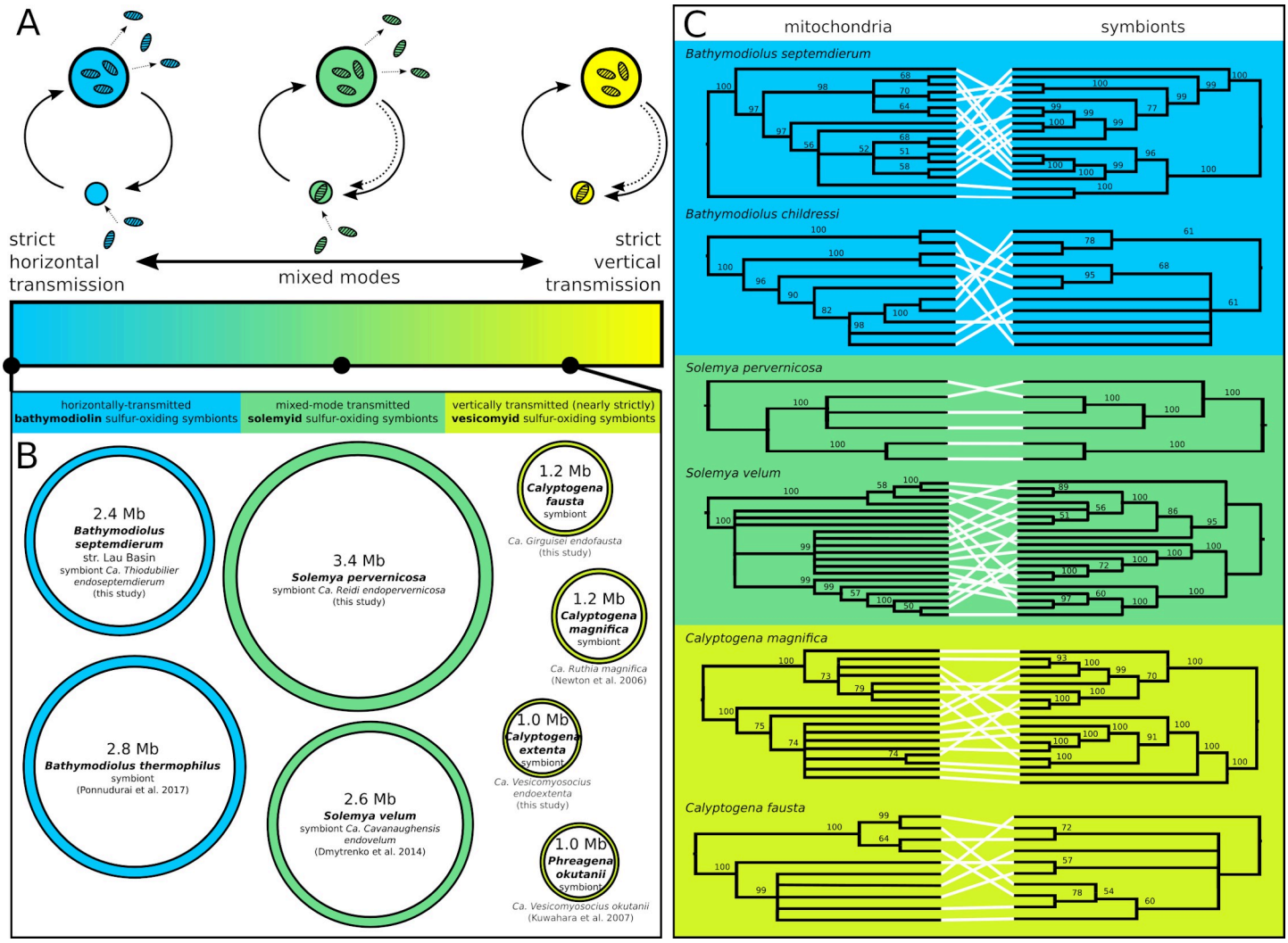

Fig 1. Nearly-strictly vertically transmitted chemosynthetic endosymbionts exhibit genome erosion despite ongoing horizontal transmission events in their populations. A) Transmission mode spectrum from strict horizontal transmission to strict vertical transmission, with a diversity of mixed modes, incorporating both strategies, in between. B) Genome sizes from this and previous studies [11–13,92] reveal consistent patterns of moderate genome erosion among the vesicomyid symbionts, but not in the other groups with higher rates of horizontal transmission. C) Mitochondrial and symbiont whole genome genealogies are discordant for all groups, indicating that sufficient amounts of horizontal transmission occur in vertically transmitted vesicomyid populations to erode the association between these cytoplasmic genomes. Maximum likelihood cladograms are midpoint rooted, and nodes below 50% bootstrap support are collapsed. Species are color coded by their symbiont transmission mode as in A).

https://doi.org/10.1371/journal.pgen.1008935.g001

Results and discussion

Comparative analyses of host mitochondrial and endosymbiont genome genealogies show strong evidence for horizontal transmission in all six populations. The genealogical discordance shown in Fig 1C indicates that horizontal transmission has occurred in the histories of all six of these populations, however, it does not suggest how much because concordance is eroded even by exceedingly low rates of horizontal transmission [43], saturating the signal genealogies can provide. Despite this apparent similarity in transmission mode, the vesicomyid endosymbiont genomes are approximately one half the size of the solemyid or mytilid endosymbiont genomes (Fig 1B), which themselves are approximately consistent with their free-living ancestors. Nonetheless, at 1–1.2 Mb, the partially degraded vesicomyid endosymbiont genomes are still ten times larger than the smallest terrestrial endosymbionts [1]. While

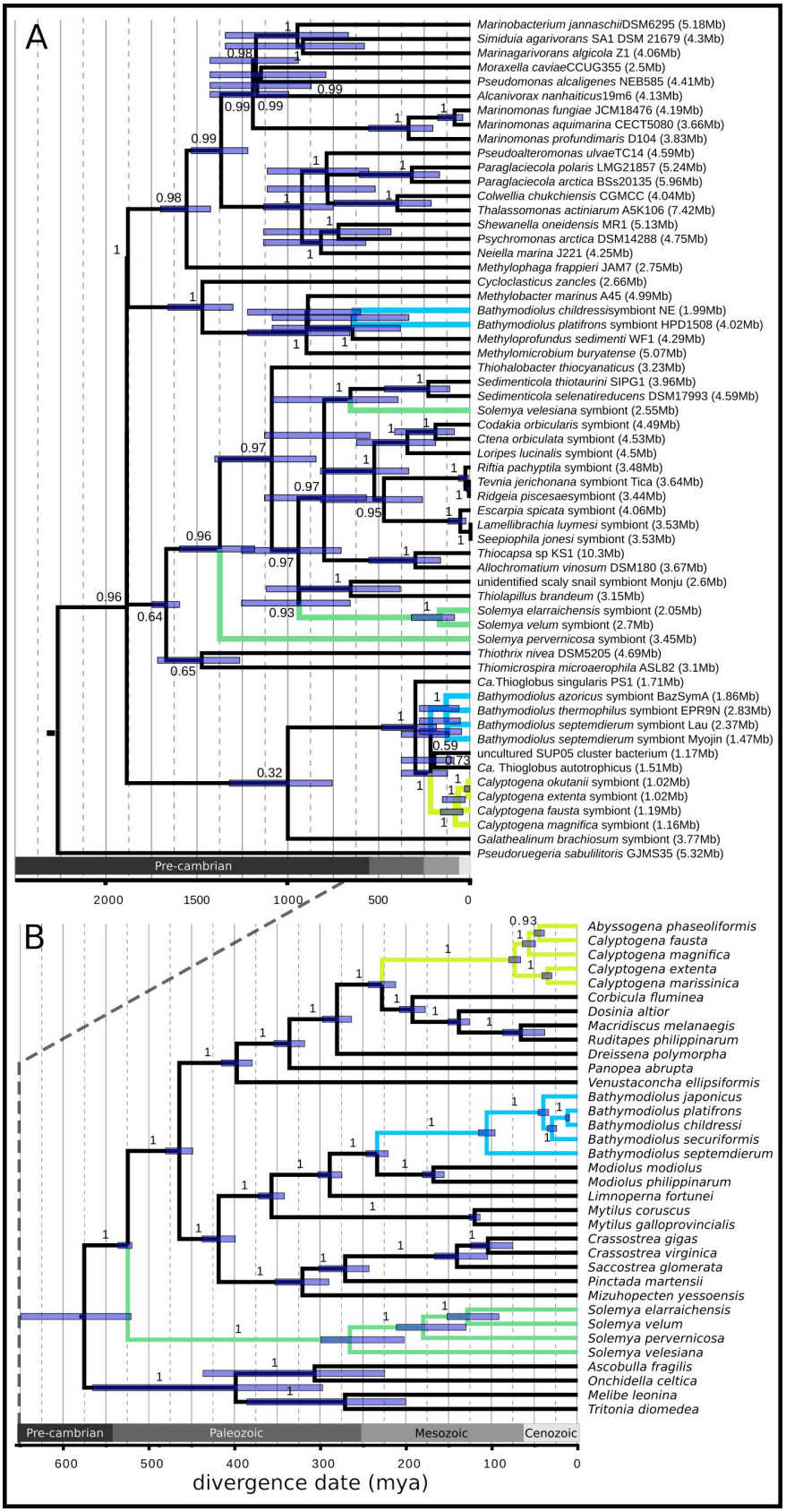
A
Marinobacterium jannaschiiDSM6295 (5.18Mb)
Simiduia agarivorans SA1 DSM 21679 (4.3Mb)
Marinagarivorans algicola Z1 (4.06Mb)
Moraxella caviaeCCUG355 (2.5Mb)
Pseudomonas alcaligenes NEB585 (4.41Mb)
Alcanivorax nanhaiticus19m6 (4.13Mb)
Marinomonas fungiae JCM18476 (4.19Mb)
Marinomonas aquimarina CECT5080 (3.66Mb)
Marinomonas profundimaris D104 (3.83Mb)
Pseudoalteromonas ulvaeTC14 (4.59Mb)
Paraglaciecola polaris LMG21857 (5.24Mb)
Paraglaciecola arctica BSs20135 (5.96Mb)
Colwellia chukchiensis CGMCC (4.04Mb)
Thalassomonas actiniarum A5K106 (7.42Mb)
Shewanella oneidensis MR1 (5.13Mb)
Psychromonas arctica DSM14288 (4.75Mb)
Neiella marina J221 (4.25Mb)
Methylophaga frappieri JAM7 (2.75Mb)
Cycloclasticus zancles (2.66Mb)
Methylobacter marinus A45 (4.99Mb)
Bathymodiolus childressisymbiont NE (1.99Mb)
Bathymodiolus platifrons symbiont HPD1508 (4.02Mb)
Methyloprofundus sedimenti WF1 (4.29Mb)
Methylomicrobium buryatense (5.07Mb)
Thiohalobacter thiocyanaticus (3.23Mb)
Sedimenticola thiotaurini SIPG1 (3.96Mb)
Sedimenticola selenatireducens DSM17993 (4.59Mb)
Solemya velesiana symbiont (2.55Mb)
Codakia orbicularis symbiont (4.49Mb)
Ctena orbiculata symbiont (4.53Mb)
Loripes lucinalis symbiont (4.5Mb)
Riftia pachyptila symbiont (3.48Mb)
Tevnia jerichonana symbiont Tica (3.64Mb)
Ridgeia piscesaesymbiont (3.44Mb)
Escarpia spicata symbiont (4.06Mb)
Lamellibrachia luymesi symbiont (3.53Mb)
Seepiophila jonesi symbiont (3.53Mb)
Thiocapsa sp KS1 (10.3Mb)
Allochromatium vinosum DSM180 (3.67Mb)
unidentified scaly snail symbiont Monju (2.6Mb)
Thiolapillus brandeum (3.15Mb)
Solemya elarraichensis symbiont (2.05Mb)
Solemya velum symbiont (2.7Mb)
Solemya pervernicosa symbiont (3.45Mb)
Thiothrix nivea DSM5205 (4.69Mb)
Thiomicrospira microaerophila ASL82 (3.1Mb)
Ca.Thioglobus singularis PS1 (1.71Mb)
Bathymodiolus azoricus symbiont BazSymA (1.86Mb)
Bathymodiolus thermophilus symbiont EPR9N (2.83Mb)
Bathymodiolus septemdierum symbiont Lau (2.37Mb)
Bathymodiolus septemdierum symbiont Myojin (1.47Mb)
uncultured SUP05 cluster bacterium (1.17Mb)
Ca. Thioglobus autotrophicus (1.51Mb)
Calyptogena okutanii symbiont (1.02Mb)
Calyptogena extenta symbiont (1.02Mb)
Calyptogena fausta symbiont (1.19Mb)
Calyptogena magnifica symbiont (1.16Mb)
Galathealinum brachiosum symbiont (3.77Mb)
Pseudoruegeria sabulilitoris GJMS35 (5.32Mb)
Pre-cambrian
2000
1500
1000
500
0
B
Abyssogena phaseoliformis
Calyptogena fausta
Calyptogena magnifica
Calyptogena extenta
Calyptogena marissinica
Corbicula fluminea
Dosinia altior
Macridiscus melanaegis
Ruditapes philippinarum
Dreissena polymorpha
Panopea abrupta
Venustaconcha ellipsiformis
Bathymodiolus japonicus
Bathymodiolus platifrons
Bathymodiolus childressi
Bathymodiolus securiformis
Bathymodiolus septemdierum
Modiolus modiolus
Modiolus philippinarum
Limnoperna fortunei
Mytilus coruscus
Mytilus galloprovincialis
Crassostrea gigas
Crassostrea virginica
Saccostrea glomerata
Pinctada martensii
Mizuhopecten yessoensis
Solemya elarraichensis
Solemya velum
Solemya pervernicosa
Solemya velesiana
Ascobulla fragilis
Onchidella celtica
Melibe leonina
Tritonia diomedea
Pre-cambrian
Paleozoic
Mesozoic
Cenozoic
600
500
400
300
200
100
0
divergence date (mya)

Fig 2. Chemosynthetic bacterial symbionts and their bivalve hosts exhibit ancient divergence times. A) Maximum likelihood phylogeny inferred from 108 orthologous protein coding genes and the 16S and 23S rRNA genes (outgroup = Alphaproteobacteria; branch labels = bootstrap support fraction) with RelTime divergence date estimates (node bars = 95% confidence intervals). Host-associated bacteria are listed as symbionts of their host species. Bacterial genome sizes are written to the right of the taxon names in the tip labels to highlight trends in genome size across clades. B) Whole mitochondrial Bayesian phylogeny for bivalves (outgroup = Gastropoda; branch labels = posterior probabilities) with divergence dates co-inferred in Beast2 (node bars = 95% highest posterior densities). In both phylogenies, members of vesicomyid, solemyid, and bathymodiolin (both thioautotrophic and methanotrophic) associations are colored yellow, green, and blue, respectively.

https://doi.org/10.1371/journal.pgen.1008935.g002

the fossil record and previous phylogenetic analyses indicate that the vesicomyid symbionts and clades of solemyid symbionts have been in continuous host association for long periods of time [21,22,44], and thus genome erosion has been prevented, precise divergence dates were needed to confirm this.

Divergence date estimates for hosts and symbionts indicate that the observed patterns in symbiont genome size have been maintained over many millions of years (Fig 2, S2 Fig, and S4–S6 Tables). Similar to prior work [21], we estimated that the vesicomyid bivalves evolved from their non-symbiotic ancestors around 73 million years ago (mya) (95% highest posterior density (HPD) = 62.59–76.70 mya; Fig 2B). We estimated a similar divergence date for the vesicomyids' monophyletic symbionts of around 84 mya (95% CI = 42.37–165.09 mya; Fig 2A), which is remarkable given that their loss of DNA repair genes and reduced selection efficacy has likely increased their substitution rate (and may have been accounted for by using a relaxed local molecular clock). The clade of gammaproteobacteria that contains the vesicomyid symbionts, termed the SUP05 clade, also contains the thiotrophic *Bathymodiolus* symbionts and free-living bacteria with genomes in the range of 1.17–1.71 Mb (Fig 2A), from which the vesicomyid symbionts are approximately 220 mya (95% CI = 127 - 381mya) diverged. This indicates that the vesicomyid symbiont genomes have eroded 58% at most, depending on the symbiont lineage and the ancestral state. Thus, over tens of millions of years of host association, the vesicomyid symbiont genomes have exhibited high degrees of stasis and have only degraded moderately (*e.g.*, in vesicomyid symbionts with 1 Mb vs. 1.2 Mb genomes).

The solemyids present a similar, but more complicated situation, likely owing to their antiquity, as the hosts first appeared in the fossil record more than 400 mya [22], when the ocean basins had much different connectivity [45]. While host-switching and novel (free-living) symbiont acquisition have certainly occurred in Solemyidae, and our data indicates such an event may have happened after *S. velum* and *S. pervernicosa* diverged (Fig 2A), it may be relatively rare across geological time. Whole genome mitochondrial and symbiont phylogenies indicate that the North Atlantic *Solemya* species *S. velum* and *S. elarraichensis* are sisters that diverged around 129 mya (95% HPD = 124–152 mya; Fig 2B), and their symbionts likely co-speciated with them around the same time (170 mya, 95% CI = 89–325 mya; Fig 2A). Divergence and subsequent speciation may have been due to the opening of the Atlantic, which occurred contemporaneously (180–200 mya; [46]). Thus, the vertically transmitted *S. velum* and *S. elarriachensis* symbionts have maintained genomes similar in size to their free-living relatives over hundreds of millions of years of host association.

Given the ancient ages of these associations (Fig 2) and their non-negligible rates of horizontal transmission (Fig 1C), a finer scaled exploration of symbiont genetic diversity was necessary to characterize the population-level processes influencing genome erosion. Homologous recombination is an important driver of genetic diversity in many bacterial populations [47,48], and could impact host-associated symbiont populations if diverse genotypes co-occur due to horizontal transmission. Novel genotypes are necessary because recombination among clonal strains, *e.g.*, within a host-restricted clonal population of vertically

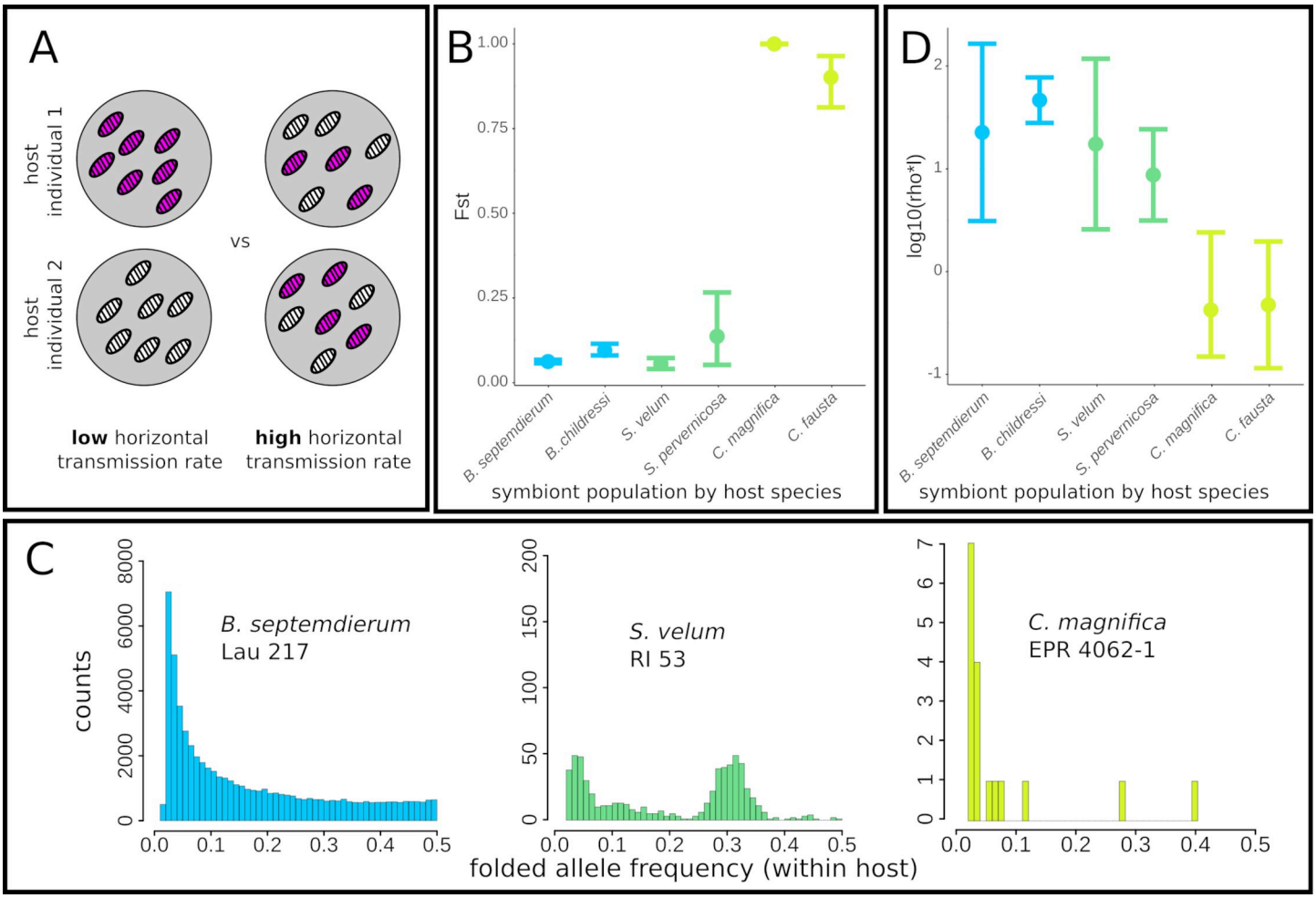

Fig 3. Horizontal transmission and recombination introduce novel alleles into symbiont populations. A) Model of endosymbiont genotype (pink vs. white) distributions under well-mixed high horizontal transmission rates and differentiated, low horizontal transmission rates. B) Horizontally transmitted mytilid (blue) and mixed mode transmitted solemyid (green) symbionts are well mixed among hosts, whereas the nearly strictly vertically transmitted vesicomyid symbionts (yellow) are highly differentiated among hosts. Error bars = 95% confidence intervals from non-parametric bootstrapping. C) Intrahost population folded allele frequency spectra (AFS) are shaped by access to gene flow, which is enabled by horizontal transmission and recombination. D) Recombination rates are significantly higher in the mytilid (blue) and solemyid (green) symbiont genomes compared to the vesicomyid symbiont genomes (yellow). Error bars = 95% confidence intervals.

https://doi.org/10.1371/journal.pgen.1008935.g003

transmitted endosymbionts, has little impact on haplotypic diversity [49]. To explore the opportunity for recombination among divergent clades of chemosynthetic symbionts, we partitioned symbiont genetic variation to between and within-host variation (*e.g.*, Fig 3A and S7 and S8 Tables). Within vesicomyid symbionts, genetic diversity is strongly subdivided by hosts, and nearly all variation distinguishes host populations (Fig 3B). Conversely, for mytilid and solemyid endosymbionts, hosts have little impact, and two endosymbionts within a host are almost as divergent as two from different hosts (Fig 3B).

The distribution of genetic diversity within a single host is even more striking than the pattern between hosts. Within host individuals, *Bathymodiolus* endosymbionts are exceptionally genetically diverse and the allele frequency spectra are qualitatively similar to expectations for an equilibrium neutrally-evolving population (Fig 3C left and S3 Fig), consistent with the high genetic diversities reported for other bathymodiolin symbionts [50]. Conversely, solemyid endosymbionts maintain intermediate and more variable within-host genetic diversity, consistent with a mixture of vertical and horizontal transmission (Fig 3C middle and S4 Fig). Finally,

vesicomyid endosymbiont populations within hosts are virtually devoid of genetic variability (Fig 3C right and S5 Fig). Therefore, despite their literal encapsulation within host cells, mytilid and solemyid symbionts have abundant opportunities to recombine and create fitter chromosomes whereas vesicomyid symbionts must only rarely encounter genetically differentiated individuals.

Although opportunities are limited for vesicomyid endosymbionts, even relatively infrequent homologous recombination events might drive patterns of genome evolution. We therefore developed a theoretical framework of symbiont evolution during mixed transmission modes. Importantly, our model demonstrates that in some conditions these populations can be approximated using a standard Kingman-coalescent and that horizontal transmission is mechanistically linked to observable recombination events between genetically diverse symbiont genomes (Section 2 of S1 Text and S6 and S7 Figs). We then performed extensive coalescent simulations and used a Random Forest-based regression framework to estimate the effective recombination rates for each population (estimated as the population-scaled recombination rate (rho) per site (l); see S9 Table and Materials and Methods). Although our model is clearly an approximation, the results are generally consistent with our prior expectations. The resulting estimated recombination rates are substantially higher in mytilid and solemyid symbionts than in vesicomyid symbionts (Fig 3D and S7 and S10 Tables), indicating that the potential for recombination within hosts is realized in these species. Given that these recombination events are occurring within symbiont populations ($\theta_{recombinant} = \theta_{genome}$), our estimate of rho*l is equivalent to r/m from previous studies of bacterial recombination rates ($r/m = rho^{*}l^{*}\theta_{recombinant} / \theta_{genome}$; see [51]). Comparing to r/m values across bacteria and archaea, which range from 0.02 to 63.6 [52], reveals that these symbiont populations have some of the largest effective recombination rates ever reported for bacteria (rho*l from the *B. septemdierum* symbionts equals 46.3, which is in the 96th percentile of previously measured rates). Despite their lack of many genes normally required for recombination [24], we found evidence for modest rates of recombination within both partially degenerate vesicomyid genomes. This capability may be enabled via "illegitimate" mechanisms, *e.g.*, RecA-independent recombination via slipped-mispairing or single strand annealing [24,53]. For all three symbiont taxa, recombination has a larger impact on genome evolution than mutation (estimated by rho*l/theta in S7 Table), in part due to the relatively low estimates of theta in the symbiont populations. Thus, these bacterial symbionts comprise what might be described as quasi-sexual, rather than clonal, populations.

A fundamental consequence of decreased homologous recombination rates for endosymbionts is that selected mutations cannot be shuffled to form higher fitness chromosomes and remain linked to neutral mutations for longer times. Ultimately, this competition among selected mutations on different haplotypes, termed clonal interference, can drive the fixation of deleterious mutations and reshape genealogies towards long terminal branches and excesses of rare alleles [54] (illustrated in Fig 4A). Similarly, even in the absence of competition among selected haplotypes, recently completed selected sweeps can reshape linked neutral genealogies if recombination is infrequent [55]. Consistent with this theory, we find abundant rare alleles in the vesicomyid symbiont genomes (Fig 4F and 4G and S7 Table; Tajima's D = -1.98 and -2.03 for *C. magnifica* and *Calyptogena fausta*, respectively), but little skew in the allele frequencies of other endosymbiont populations (Fig 4B–4E and S7 Table; D ranges from -2.03 to 1.01). Importantly, it is unlikely that differences in host species demography have driven these differences, *e.g.*, recent population expansions specifically in vesicomyid clams. In fact, we found less allele frequency skew in the mitochondrial genomes than in the vesicomyid symbionts for all species considered, and the strongest negative skew in the allele frequencies in the mitochondrial genome of *B. septemdierum* from the Lau Basin (D = -1.9, S7 Table).

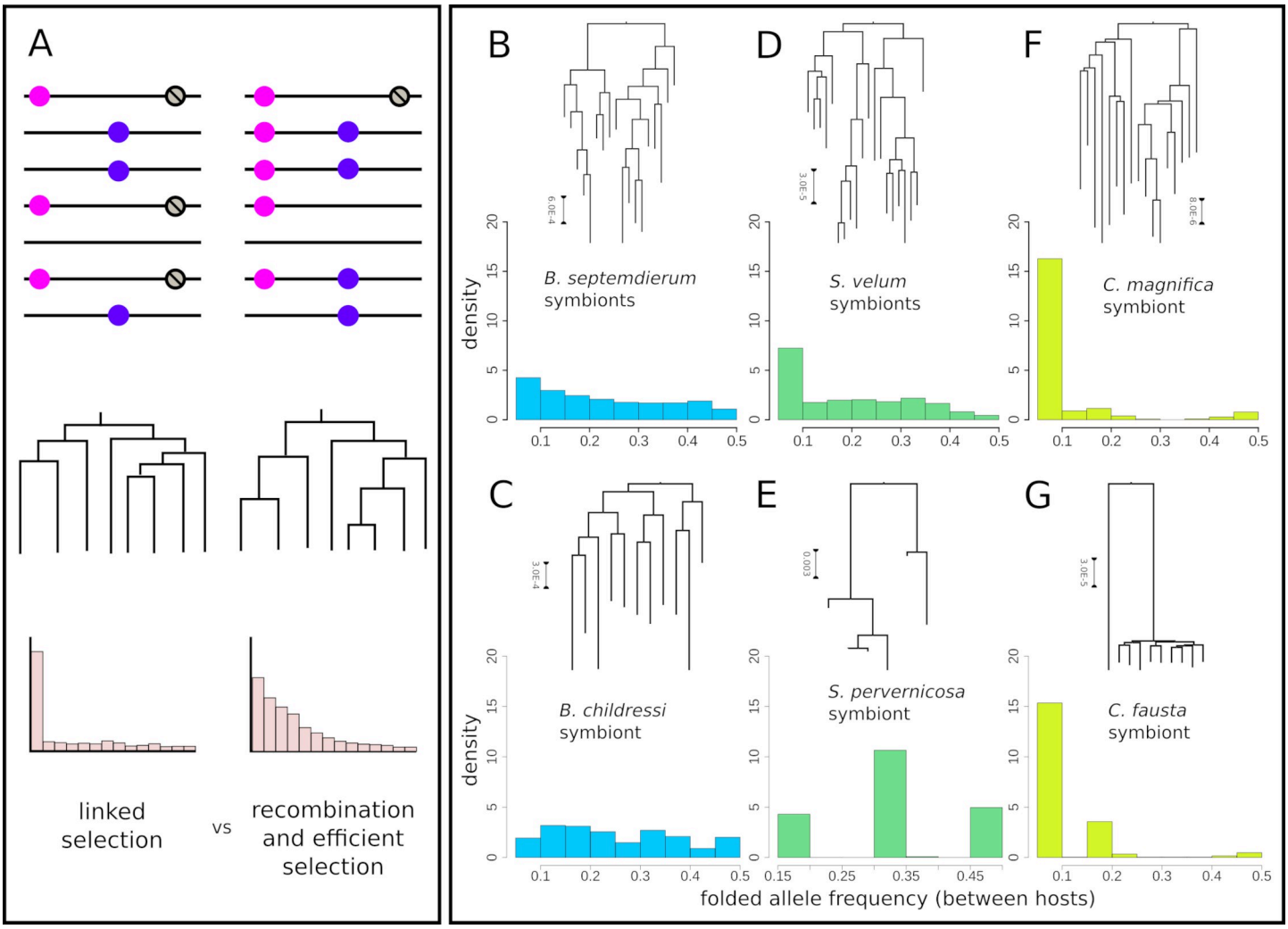

Fig 4. Consequences of access to gene flow via horizontal transmission and recombination on the distribution of symbiont genetic diversity between hosts. A) Diagram showing how beneficial alleles (pink) are linked to deleterious alleles (grey) in populations experiencing strong selection on linked sites versus free recombination, and how these processes are reflected in the underlying population genealogies and allele frequency spectra. B-G) Symbiont genealogies and between host allele frequency spectra (AFS) for each host/symbiont species.

https://doi.org/10.1371/journal.pgen.1008935.g004

Additionally, relative rates of molecular evolution between symbiont populations follow the expected trend, with dN/dS values of 0.14, 0.096, 0.083 for vesicomyid, solemyid and bathymodiolin genomes, respectively (pairwise Wilcoxon test p-values: bathymodiolin-vesicomyid p = 4.60e-14, solemyid-vesicomyid p = 1.02e-11, and bathymodiolin-solemyid p = 0.0365), consistent with recombination enabling more efficacious purifying selection for sustained periods of time.

Genome structure stasis is thought to be another hallmark of canonical endosymbiont genome evolution, and many terrestrial endosymbioses have reported static, degenerate genomes [1]. Whole genome alignments reveal that the vesicomyid symbiont genomes are highly syntenic, with few rearrangements, insertions, or deletions; whereas the other symbiont genomes are far more structurally dynamic (Fig 5 and S8 Fig). Given the role of recombination in altering bacterial genome structure ([56]; illustrated in Fig 5A), and the signals of recombination and linked selected sites in the vesicomyid symbionts (Figs 3D, 4F and 4G, respectively), recombinational processes may partially underlie genome erosion, in addition to

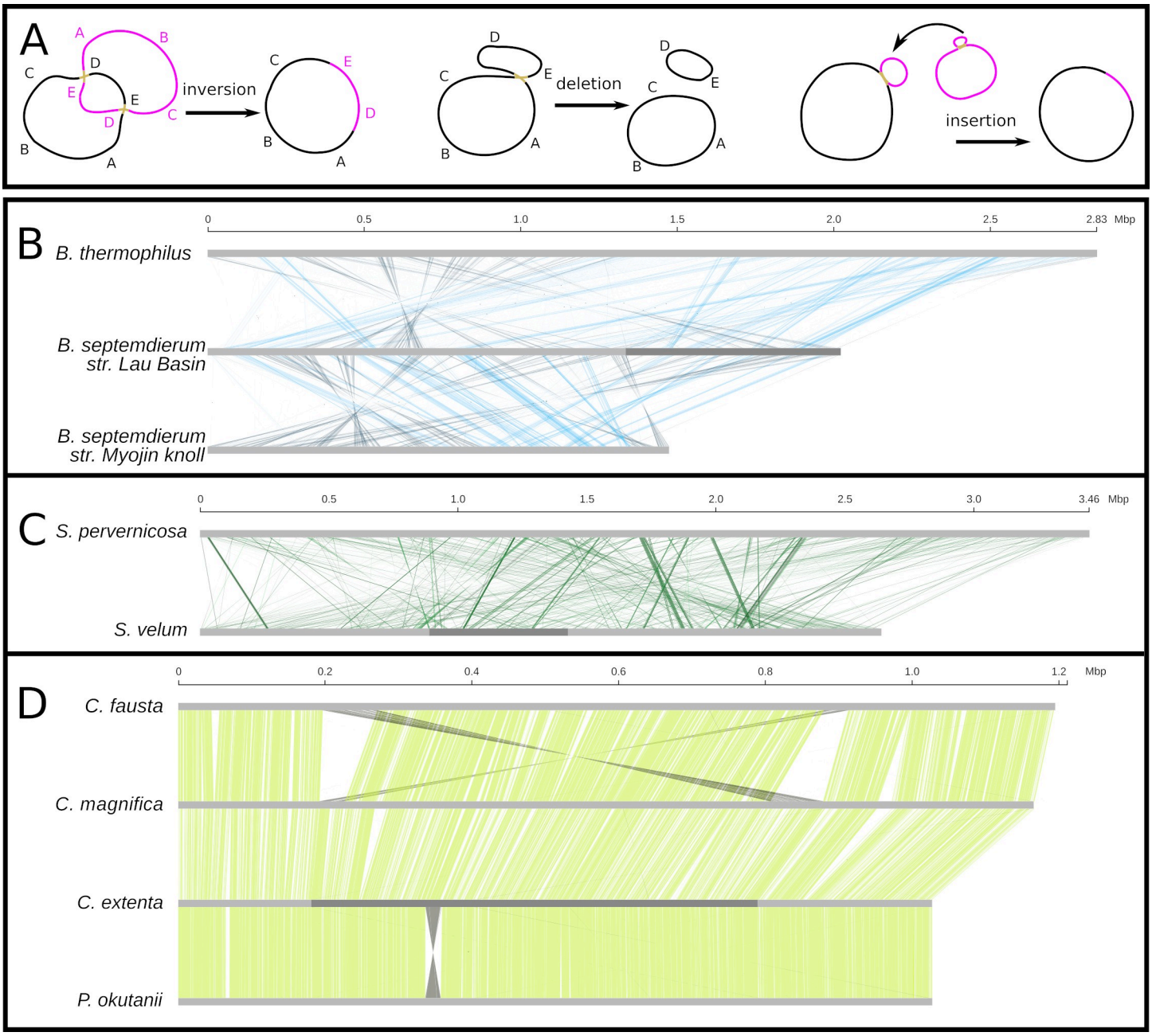

Fig 5. Genome structure is shaped by horizontal transmission and recombination. A) Models of recombination-based structural mutation mechanisms. B-D) Whole genome alignments for B) sulfur-oxidizing mytilid, C) solemyid, and D) vesicomyid symbiont genome assemblies with >1 Mb scaffolds.

https://doi.org/10.1371/journal.pgen.1008935.g005

preventing it. This could proceed in the following way: first, a rare recombination event induces a deletion that drifts to high frequency in the within-host population (other events can also induce deletions, such as strand slippage [57]). With homologous recombination events occurring so rarely in these endosymbiont genomes, natural selection would be unable to efficiently purge deleterious deletions. Although, inversions, which can be highly mutagenic by inverting the translated strand, inducing replication-transcription machinery collisions [58], are nearly absent potentially due to their high fitness costs. If symbionts with chromosomes bearing the deletion are exclusively transmitted to offspring during vertical transmission, then

the deletion would be fixed in all subsequent symbionts in that host lineage. Multiple instances of this process would incrementally reduce the size of the symbiont genome.

In contrast, the solemyid symbiont genomes exhibit increasing degrees of structural dynamics with increasing divergence time. Highly divergent solemyid symbionts, such as the *S. velum* and *S. pervernicosa* symbionts, exhibit genomes that are as structurally dynamic as strictly horizontally transmitted associations (Fig 5B vs 5C). However, over shorter time scales, such as the duration of time since the *S. velum* and *S. elarraichensis* symbionts diverged, structural changes appear to be dominated by insertions and deletions (indels; confirmed for 30 Kb segments of the *S. elarrachensis* symbiont draft genome in S7 Fig). Rapidly evolving indels have also been reported at the species level for *S. velum* [15]. Potentially underlying indel dynamics, we found that the solemyid symbionts have far more mobile genetic elements than any of the other symbiont genomes (S11 Table). High mobile element content is consistent with a combination of environmentally-exposed and host-associated periods [59], and mirrors the early stages of symbiosis [19,20], as well as the early stages of eukaryotic asexuality [60]. The mobile elements exhibit homology to different environmental bacteria, implying many independent insertion events (S12 Table).

Given the extreme ages of the solemyid and vesicomyid associations (Fig 2), our data suggest that moderate rates of recombination have allowed their symbiont genomes to maintain functional diversity characteristic of a free-living or moderately reduced genome, respectively. Evidence from the insect endosymbionts experiencing extreme genome size reduction indicates that genome erosion is possible over time frames as short as 5–20 mya (*e.g.*, [61–68]). It is plausible that vesicomyid symbionts' horizontal transmission and recombination rates are at the beginning of the range of values that permit genome decay. Our data indicates that they still undergo horizontal transmission (Fig 1C) and recombine (Fig 3D and S7 Table). Furthermore, signatures of genetic variation consistent with linked selection and sustained intermediate genome sizes indicate that selection is sufficiently efficacious to maintain some functional diversity in these populations, counter to the expectations for the original theory on endosymbiont genome evolution [5,6]. Thus, ample time has passed for these symbiont genomes to erode, but horizontal transmission and recombination have likely prevented it (as depicted in Fig 6).

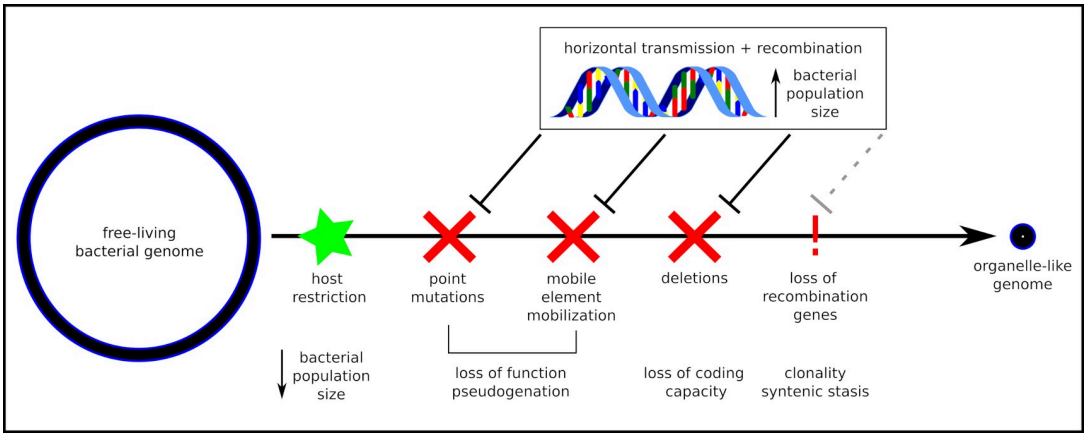

Fig 6. A conceptual model of the prevention of endosymbiont genome degradation through horizontal transmission and recombination. Sufficient levels of genetic diversity, which can be introduced via horizontal transmission of symbiont genotypes between hosts and recombination between genotypes in mixed infections, prevents or delays genome degradation by restoring functional versions of mutated or deleted regions. Prevention can continue until recombination capabilities (RecA-dependent and independent) are completely lost, at which point, genetic rescue is no longer possible without wholesale symbiont replacement.

https://doi.org/10.1371/journal.pgen.1008935.g006

Conclusion

Here we empirically show that symbiont genome sizes and functional diversity are predicted by the rate of gene flow into and among symbiont populations via horizontal transmission and homologous recombination. Although we have only investigated three independently evolved associations, we see this system serving as a microcosm for marine associations more generally, as many other associations exhibit similar biologies (*e.g.*, intracellular, autotrophic, broadcast-spawning, etc. [36]) and all are governed by the same population genetic principles. Amazingly, we found that symbiont gene flow between hosts is ongoing in one of the most intimate marine associations, the vertically transmitted vesicomyids. These results suggest that there is a range of possible intermediate genome degradation states that can be maintained over millions of years with sufficient recombination. Therefore, symbiont genome evolution following host restriction is not a one-way, inescapable process that ends in an organelle-like state as it is commonly presented [2,5,6]. These results validate long-standing but untested theory and suggest that the diversity of symbioses found to exhibit intermediate rates of horizontal transmission and incomplete genome degradation may be undergoing similar population-level processes.

Materials and methods

Samples and genomic data production

Sample collection. We obtained chemosynthetic bivalve samples from hydrothermal vents, cold seeps, and reducing coastal sediments from around the world (S1 Table). *Calyptogena magnifica* samples were collected from the East Pacific Rise (EPR) hydrothermal vent fields between 1998 and 2004. *Calyptogena fausta* were collected from the Juan de Fuca (JDF) Ridge hydrothermal vent system in 2004. The single *Calyptogena extenta* specimen was collected from Monterey Canyon in 1995. *Solemya pervernicosa* samples were obtained from the Santa Monica sewage outfall in 1992 (as in [15]). *Solemya velum* were collected from Point Judith, RI, as described in [15]. *B. childressi* were sampled from the Veatch Canyon cold seep off of New England. *B. septemdierum* was sampled from the ABE and Tu'i Malila hydrothermal vent sites in Lau Basin. All tissue samples were stored at -80˚C until sterile dissection or subsection of previously sterile-dissected gill tissue as described in [15].

DNA extraction and Illumina sequencing. We extracted DNA from gill samples for each host individual sampled using Qiagen DNeasy kits following the manufacturer's instructions. We quantified DNA concentrations using a Qubit dsDNA kit and normalized each sample to 10 ng/ul for Illumina library preparations. We produced the majority of our Illumina sequencing libraries for each sample using a Tn5-based protocol for tagmentation followed by dual-indexing PCR using HiFi DNA polymerase (Kappa Bioscience) and custom primer sequences (IDT) designed to uniquely label both i5 and i7 indexes for each sample (Tn5 enzyme was expressed and purified in-house). Indexed samples were pooled and sequenced on single lanes of a Hiseq4000. We sequenced a total of four lanes of Hiseq4000 paired-end 150 bp sequencing across the entire study. Additionally, we obtained a subset of samples for *C. magnifica* using genomic methods from our previous work ([15], S1 Table). We also used a dataset we have previously collected for *S. velum*, specifically the population from Point Judith, Rhode Island ([15], S1 Table). The specific library preparation methods and number of read pairs obtained for each sample included in this work are listed in S1 Table.

Nanopore sequencing. For each *de novo* genome assembly in this work, we selected a representative from each host population based on DNA quality as determined using an Agilent Tapestation and DNA concentration based on qubit readings. We sequenced each sample on a

single minion flow cell using the ligation-based 1D chemistry, SQK-LSK109 kit per ONT instructions with minor adjustments. The end-repair reaction was incubated for 30 minutes each at 20˚C and 65˚C and the ligation reaction was performed for 30 minutes instead of the recommended 10 minutes. Read counts obtained and mean read lengths are available in S1 Table. We performed basecalling using the Albacore basecaller v2.0.1 and we discarded the subset of reads whose mean quality score was less than 7. These are the set marked nominally as "failed" by the basecaller software.

***De novo* symbiont genome assembly.** Reference genomes for the *B. septemdierum*, *B. childressi*, *S. pervernicosa*, and *C. fausta* symbionts were assembled using combined Nanopore and Illumina reads. First, we assembled the Nanopore reads using the long-read assembly program wtdbg2 [69] using the "ont" presets option and setting the parameter -k to 15. Then, we performed two rounds of reference genome improvement by aligning Illumina sequencing reads from the same individual to the resulting unfiltered assembly and polishing with the Pilon software package [70]. We used BWA-MEM [71] to align Illumina reads in each subsequent polishing round.

We assembled the *C. extenta* symbiont genome from an Illumina library prepared for the single sampled individual using IDBA [72] and SPAdes [73]. While both assemblies were highly contiguous (N50 = 596007 and 604961 bp, respectively), the SPAdes assembly was able to merge two contigs that were split in the IDBA assembly, so the two contig SPAdes assembly was used for downstream analyses. Comparisons of synteny demonstrated that this join was found in other Vesicomyid endosymbiont genomes suggesting it is correct (see below).

Because read mixtures include host genomic DNA, mitochondrial DNA, and genomic DNA from other bacterial species, we then rigorously filtered the resulting contigs to extract only high confidence contigs contributed by bacteria of the study species. To identify symbiont contigs, we called ORFs with Prodigal v2.60 [74] and annotated coding sequences with BLAST [75] as described in [15] (NCBI nr, TrEMBL, and UniProt database accessed on April 7, 2019). We annotated ribosomal RNAs with RNAmmer [76] and transfer RNAs with tRNAscan [77]. Using the taxonomic information encoded in the annotation, we identified contigs that were confidently of symbiont origin. Then, using these contigs, we filtered the remaining contigs by GC content, read coverage, and coding density using custom scripts. Finally, we evaluated the quality of the assemblies with CheckM [78] and by testing for the presence of core bacterial phylogenetic markers [79] (see S2 Table).

Host mitochondrial genome assembly. Mitochondrial genomes were assembled from Nanopore reads, which were subsequently corrected with Illumina data, as described above, or they were assembled from Illumina reads directly. As the different samples contained different mitochondrial coverage, higher short read coverage was often better than lower long read coverage for recovering these genomes. We assembled mitochondrial genomes for *C. fausta and B. septemdierum* with IDBA [72] using Illumina data from two of the highest depth-of-coverage samples (*C. fausta* 31 and *B. septemdierum* 231, respectively). The complexity of the *B. septemdierum* data prevented IDBA from finishing within a week, so we first removed low coverage nuclear kmers (<10x) with Quake [80]. The *C. extenta* mitochondrial genome was assembled along with the symbiont genome using SPAdes [73], as described above. We were able to use the Nanopore-based assembly from *Bathymodiolus childressi* for the mitochondrial genome. Lastly, the complete mitochondrial genome for *S. pervernicosa* was available from [15], so we did not reassemble it here.

After assembly, we identified the mitochondrial scaffold by blasting the full set of scaffolds against a database containing the currently available set of 19 bivalve mitochondrial genomes. Then, we annotated the mitochondrial genome with MITOS [81]. For mitochondrial genomes lacking conserved genes, we repeated genome assembly and mitochondrial genome

identification and annotation with a different sample to verify we obtained the full sequence. See Section 1 of S1 Text for a description of the host species identification verification process.

Short read alignment. After producing endosymbiont and host mitochondrial genome assemblies for each host/endosymbiont species, we aligned short read Illumina data from each individual to a reference genome consisting of both of these genomes. Genomes assembled previously for the *C. magnifica* symbiont ([11], accession NC_008610.1) and mitochondrian ([82], accession NC_028724.1), and the *S. velum* symbiont ([13], accession NZ_JRAA00000000.1) and mitochondrian ([83], accession NC_017612.1) were used as references for these populations. We used the BWA mem software package [71], and we then sorted and removed duplicate reads using the samtools software package [84]. After this, we performed indel realignment for each sample separately using the "IndelRealigner" function within the Genome Analysis Toolkit (GATK) software package [85].

Genotyping and variant filtration for each host individual. We called consensus genotypes for each individual jointly using the GATK "UnifiedGenotyper" option and we ran the program with otherwise default parameters except we required that it output all sites rather than just all variable positions. We filtered variant sites using the vcftools software package [86] largely following the GATK best practices as we have done in our previous work [15]. Briefly, we required that each site have a minimum quality/depth ratio of 2, a maximum Fisher's strand value of 60, a minimum nominal genotype quality of 20 and a maximum number of reads with mapping quality zero at a putatively variant site of 5. For analyses of within host individual variation for fixation index (Fst) calculations, we also obtained a multiple pileup file using samtools and filtered sites that were not retained after applying these filters. See Section 1 of S1 Text for an estimate of the consensus genotyping error rate.

Within-host diversity analysis. We called within-host SNP and indel variants for endosymbionts and mitochondria using the method from [87]. Briefly, we created mpileup files from BWA bam alignment files for all individuals from each host species using SAMTools [84]. Then, we called variants and calculated pairwise diversity using the perl script from [87], which only considers sites within one standard deviation of the average genome coverage, filters SNPs around indels, and requires an alternate allele count in excess of the cumulative binomial probability of sequencing error at that site. As very closely related sister taxa have not been sampled for most of these bacterial genomes, ancestral/derived alleles could not be identified and we could not plot unfolded allele frequency spectra. Instead, folded allele frequency spectra were calculated for minor alleles and plotted in R.

No heteroplasmy was detected within the mitochondrial within-host populations (see S7 Table), suggesting that these bivalves do not experience double uniparental mitochondrial inheritance. This is important given our expectations regarding mitochondrial-symbiont co-divergence under strict vertical transmission.

Genome analyses

Population genealogy inference. We produced multiple fasta sequence files for each population for the host mitochondrial genomes and for the concatenated symbiont genomes from the set of filtered consensus genotype calls. We then used the phylogenetic software package RAxML [88] using a GTR+G model and 1,000 bootstrap replicates to estimate the phylogenetic relationships among samples and to quantify uncertainty in our phylogenetic relationships. Using FigTree, we rooted the trees by their midpoints and created cladograms for topological comparisons.

Analysis of polymorphic and recombinant sites. Using the fasta files described above, we filtered sites to only retain biallelic SNPs with a minimum genotype quality of 10. Without

indels, these resequenced genomes were already aligned. Then we used the aligned SNP data to calculate Waterson's theta [89], pi [90], and the proportion of pairwise sites where all 4-gametes, *i.e.*, all pairwise combinations of alleles, are represented. We then binned the 4-gamete sites by the distances between alleles, with bins at 1e1, 1e2, 1e3, 1e4, 1e5, and 1e6 bp, for model fitting (described below).

Whole genome structural alignment. We generated whole genome alignment plots by first aligning bacterial genome assemblies with MUMmer 3.23 [91], using the nucmer algorithm and default parameters. In addition to the mb-scale genomes we assembled and referenced above, we obtained mb-scale genomes for *Bathymodiolus septemdierum* str. Myojin knoll ([42], accession GCA_001547755.1), *Bathymodiolus thermophilus* str. EPR9N ([92], accession GCF_003711265.1), and *Vesicomyosicus okutanii* ([12], accession NC_009465.1) from NCBI for alignment. The nucmer output was converted into the BTAB format with the MUMmer tool show-coords and was then visualized using a custom Python script. When necessary, some scaffolds in the bacterial assemblies were split into two parts in order to convert a circular genome into a linear alignment plot.

We used the whole genome aligner progressiveMauve [93] to compare genome synteny on the 10s of Kb scale between the *S. velum* and *S. elarraichensis* symbionts. First, we reordered the contigs comprising the *S. elarraichensis* symbiont draft genome [15] by the *S. velum* symbiont assembly [13] with the reorder contig function in progressiveMauve. Then, we aligned the *S. velum* symbiont genome pairwise against the *S. elarraichensis* symbiont's reordered contigs and the *S. pervernicosa* symbiont genome we assembled. We plotted the alignment backbone files in the R package genoPlotR [94].

Mobile element analysis. We identified mobile elements in the endosymbiont genome sequences by BLAST. First, we generated BLAST database files with the makeblastdb command from the ACLAME [95] and ICEberg [96] nucleotide and amino acid databases of transposable, viral, and conjugative elements. Next, we used blastp and blastn to compare endosymbiont amino acid sequences and full genome sequences, respectively, to these databases (cutoff values: minimum alignment length of 50 nucleotides or 50% amino acid query coverage, 90% identity, and e-value 1e-6). Overlapping hits were consolidated into a single mobile element-containing region (S12 Table).

Ortholog identification. We identified putative orthologous sequences among sets of bacterial genomes by a reciprocal best BLAST approach. To do this, we first performed pairwise blasts between each pair of genomes' coding sequences with blastn (-best_hit_overhang 0.1 -best_hit_score_edge 0.1 -evalue 1e-6), alternating each sequence as the query/subject. We parsed these results to only retain the best hits with >50% identity and >100 bp alignment lengths. Then, using a custom perl script, we compared hits between all pairs to identify genes with identical reciprocal best hits among all taxa each homologous gene was detected in. We used the resulting matrix of these reciprocal best hits to extract the coding sequences for each ortholog for each species from the genome fasta files for downstream analysis.

dN/dS analysis. To evaluate the impact of homologous recombination on patterns of natural selection at the molecular level over long periods of time we computed the average fixation rate among endosymbiont lineages at nonsynonymous and synonymous sites (dN/dS). This ratio of values is an approximate measure of the strength of purifying selection under the assumption that most nonsynonymous substitutions are deleterious. For all genes where a single ortholog was found to be shared among all symbiont lineages we began by producing codon aware alignments using MASCE [97]. Then, we compared each orthologous alignment for pairs of symbiont lineages within each group (solemyid, vesicomyid, and bathymodiolin) to estimate dN and dS using the codeml package in the PAML v4.9 framework [98]. We excluded all comparisons for which $dS < 0.05$ or $dS > 2$, as values that exceed this range are

often thought to yield unreliable estimates of rates of molecular evolution due to low statistical power and saturated substitutions, respectively. We then compared the distributions of dN/dS for each symbiont group comparison using a Wilcoxon test.

Divergence dating

Taxon selection. To construct dated phylogenies for hosts and symbionts, we downloaded related genomes from NCBI. For the host divergence analysis, all of the bivalve mitochondrial genomes available as of early 2020 and four gastropod mitochondrial genomes were downloaded to serve as ingroups and outgroups, respectively (35 total taxa: 31 bivalves and four gastropods; see S4 Table). For the symbiont divergence analysis, bacterial genomes were identified for inclusion in the analysis by BLAST [75]. While residing in a relatively constrained clade of proteobacteria, these chemosynthetic symbionts do not form a monophyletic clade, have free-living relatives, are basal to more derived groups in Gammaproteobacteria, and are currently taxonomically unclassified, so it was necessary to fish out related genomes by identifying sequence homology. To do this, we aligned the nucleotide coding sequences from each one of the seven symbiont genomes we sequenced and/or analyzed against NCBI's Prokaryotic RefSeq Genomes database with blastn (-best_hit_overhang 0.1 -best_hit_score_edge 0.1 -evalue 1e-6). Based upon the diversity of hits across symbionts and genomes, we selected the top three best hits to each gene as taxa to include in the full genome divergence analysis (59 total taxa: 58 gammaproteobacteria and one alphaproteobacterium outgroup; see S5 Table).

Multiple sequence alignment. As bivalve mitochondrial genomes exhibit notoriously diverse structural arrangements [99], we used the whole rearrangement-aware genome aligner progressiveMauve [93] to align the molluscan mitochondrial genomes. These alignments were manually inspected and converted to fasta format in Geneious Prime (version 11.0.6+10) [100].

We identified and aligned orthologous proteins among these diverse bacterial genomes with bcgTree [101], then we back-translated the resulting amino acid alignments to nucleic acids with RevTrans [102]. As bcgTree only includes protein-coding genes, we manually extracted and aligned 16S and 23S ribosomal RNA sequences for these taxa with Mafft (using the accurate mafft-linsi setting) [103]. Although recombination clearly occurs frequently in the solemyid and mytilid symbiont populations, we decided against removing recombinant sites because doing so may exacerbate recombination-induced artifacts [104]. Finally, we concatenated these alignments with the nucleotide alignments from bgcTree/RevTrans with a custom perl script and inspected them in Geneious Prime.

Phylogenetic inference and divergence dating. We first inferred maximum likelihood phylogenies with RAxML (version 8.2.1, with parameters: f a -m GTRGAMMA -N 1000) [88] to verify that the taxa selected were able to resolve the relationships among hosts and among symbionts and free-living bacteria. Both mitochondrial and symbiont phylogenies were well-resolved (S2 Fig).

We inferred Bayesian phylogenies and dated node divergences for host mitochondria in Beast2 [105]. After several rounds of parameter testing to ascertain the speciation model and calibration date distribution that best fit the data (see Section 1 of S1 Text), we selected the Yule model of speciation, with a gamma distributed Hasegawa, Kishino, and Yano (HKY) model of substitution and a relaxed local molecular clock, and we calibrated dates to the base of the ingroup, Bivalvia. We used the fossil-based minimum appearance date for bivalves of 520 million years (first appearance estimated in Fossilworks [106] from fossil data in [107,108]). MCMC chains were run in duplicate until posterior probability convergence, around 8e8 steps for mitochondria. We also performed independent divergence date

estimations in RelTime [109,110] and PATHd8 [111] to compare to the Beast2 results (see Section 1 of S1 Text and S6 Table).

Given the consistency in RelTime and Beast2 estimates for mitochondria (S6 Table) and the unreasonably long run times necessary (several weeks) for symbiont dated phylogenies to reach posterior convergence in Beast2, we estimated symbiont divergence times in RelTime. We used the previously estimated divergence date for Gammaproteobacteria of 1.89 billion years (based on calibration to the cyanobacterial-caused atmospheric oxygenation event [112]) with a log-normal distribution to calibrate a relaxed local clock, using a gamma-distributed Tamura-Nei model of substitution, and allowing for invariant sites. All trees were plotted in FigTree.

Symbiont species descriptions

Using the genomic, phylogenetic, and divergence data we generated above, we diagnosed and described the six symbiont species sequenced in this study. These classifications will be helpful in future investigations and discussions of symbiont function and diversity. Diagnoses of symbiont genera and descriptions of symbiont species are described in the Section 3 of S1 Text and listed in S3 Table.

Parameter estimation via approximate Bayesian computation

Simulation setup. To simultaneously estimate the rates of horizontal transmission, effective homologous recombination rates, and the recombinant tract length, we used an approximate Bayesian computation approach. Here, we define the population-scaled per-base pair recombination rate, rho, to be equal to two times the effective population size times the per-base pair rate of gene conversion ($rho = 2^*N_e r$). The recombination tract length, l, is defined as the length of the gene conversion segment in base pairs. We used the bacterial sequential Markovian coalescent simulation framework, FastSimBac [113], to simulate neutral coalescence across a range of input parameters (see Section 2 of S1 Text for model proof). We drew the effective mutation rate from a log-uniform (3e-5,1e-2) distribution, the effective recombination rate from a log-uniform (1e-6,1e-2) distribution and the recombinant tract length from a uniform (1,1e5) distribution. Because the clonal frame cannot be inferred for the *Bathymodiolus* and *Solemya* symbiont populations, presumably due to the high levels of recombination relative to other bacteria, we did not supply the program with a fixed precomputed clonal frame for any simulations. In total we performed 100,000 simulations and we used subsets from each simulation to obtain summary statistics to train the variable sample size simulations.

Summary statistics. We selected a set of summary statistics that each incorporate some feature of the overall diversity (relevant for theta), and the overall effective recombination rate per site (rho^*l). Specifically, we computed two estimators of theta, Waterson's theta [89], and pi [90], and we included the proportion of pairs of non-singleton SNPs at various genomic distances where all four possible combinations of alleles are observed in our sample. We placed the divisions between distance bins for pairwise comparisons of sites at 1e1, 1e2, 1e3, 1e4, 1e5, 1e6 base pairs.

Model fitting via random forest regression. We use the scikit-learn package to perform random forest regression to obtain estimates of each parameter for each endosymbiont population using the 4-gamete sites identified above and a custom Python script available on Github. First, we confirmed that our summary statistics are sufficient to accurately fit our desired population parameters using out-of-bag score during model training and for each population sample size (S9 Table). Although the score is often slightly lower for smaller sample sizes, this approach performs sufficiently well and consistently across samples for our

applications here. We additionally obtained confidence intervals for each parameter estimate using the forestsci package [114]. It should be noted that our simulations assume an equilibrium population. If this assumption is violated in a subset of the taxa that we examined, it might affect our parameter estimates. Nonetheless, it is unlikely that the large-scale differences in estimated parameters that we observe among groups, which are consistent with prior expectations, are entirely attributable to this potential bias.

Method validation on existing datasets. In light of the relatively high recombination rates that we estimate, it is valuable to confirm that our approach for estimating rho and theta performs as expected. We therefore applied our method to the *Bacillus cereus* dataset [115] that has been studied in similar contexts using several related approaches [51,113,115]. In prior work with this dataset, estimates of the total impact of recombination, rho*l/theta, have varied somewhat, from approximately 3.7 [113] to 35.9 [115] and 229 [51]. Using our method, we obtain a value intermediate to these at 17.1, which indicates a moderate impact of recombination on genome evolution. This result suggests that our method is reliable and does not substantially inflate recombination rate estimates.

Read-backed phasing to confirm recombination estimates. Because we used consensus symbiont genomes during model fitting, it could be possible that the high estimated rates of recombination in solemyid and bathymodiolin samples are an artefact of differential haplotype coverage across the genomes of genetically diverse symbiont chromosomes. We therefore sought to confirm our recombination rate estimates via comparing the rate of occurrence of all four possible configurations of two proximal alleles using read-backed phasing. Because each read-pair must ultimately derive from a single DNA fragment, when two alleles are observed on the same read or read pair, in the absence of errors, they must reflect an allelic combination that's observed within a single bacterial chromosome.

We therefore analyzed in aggregate all reads that overlapped two or more consensus alleles for each population and computed the fraction of all four possible sampling configurations. More specifically, we computed the proportion of reads sampled from across all reads in all individuals that contained alleles, AB, Ab, aB, and ab, for two adjacent biallelic sites with alleles A/a and B/b. To reduce the impact of sequencing errors, we recorded a site as containing all four possible configurations when all configurations were present at proportion greater than 0.05 in the total set of read pairs. We further limited comparisons to alleles where we could obtain at least 100 observations of both sites on single read pairs. We excluded the populations of vesicomyid endosymbionts from this analysis because too few polymorphic sites were present within the distances spanned by individual read pairs to confidently infer the frequencies of 4D sites. However, because samples from these populations contain virtually no within-host variation, we have little reason to doubt the accuracy of the consensus genotypes resulting from differential coverage of genetically diverse bacterial haplotypes.

Because we are sampling a much larger number of lineages than we did in analyzing the consensus genome sequences, we would expect if anything is different that we observe higher rates of sites where all four possible allele configurations are present. This is precisely what we found. Specifically, we observe higher proportions of pairs of sites where all four possible allelic combinations are represented at each distance considered and in each population in the read-backed dataset than in the consensus chromosomes (S10 Table). Furthermore, because we have placed conservative cutoffs on the proportions of read pairs required to consider a pair of sites as containing all four alleles, these values are likely underestimates of the true rates. Our observed rates of four-gamete test failures is therefore consistent with our analyses of consensus genome sequences and confirms that recombination must be common within these endosymbiont populations.

Supporting information

S1 Fig. Overview of the genomic data production and analysis steps used to study the population genomic processes influencing endosymbiont genome evolution.
(TIF)

S2 Fig. Maximum likelihood phylogenies for host mitochondria (top) and symbionts (bottom). Groups of chemosynthetic associations are colored as in Fig 1: yellow = vesicomyids, green = solemyids, and blue = bathymodiolids. Mitochondrial and symbiont trees are rooted by gastropod and alphaproteobacterial outgroups, respectively. Scale bar = substitutions per site. Bootstrap support values indicated at nodes.
(TIF)

S3 Fig. Within-host symbiont folded allele frequency spectra for all *B. septemdierum* and *B. childressi* intrahost samples with more than 50x and 45x Illumina sequencing coverage, respectively (see S1 Table for coverages and S8 Table for diversity statistics).
(TIF)

S4 Fig. Within-host symbiont folded allele frequency spectra for all *Solemya velum* and *Solemya pervernicosa* intrahost samples with more than 50x Illumina sequencing coverage (see S1 Table for coverages and S8 Table for diversity statistics).
(TIF)

S5 Fig. Within-host folded allele frequency spectra for all *Calyptogena fausta* and *Calyptogena magnifica* intrahost samples with at least 50x Illumina sequencing coverage (see S1 Table for coverages and S8 Table for diversity statistics).
(TIF)

S6 Fig. Endosymbiont inheritance modes. Our generalized coalescent model of endosymbiont inheritance includes symbiont transmission modes ranging from strict horizontal transmission to strict vertical transmission, with mixed modes, exhibiting both horizontal and vertical strategies. The host populations (grey) undergo Wright-Fisher reproduction. Endosymbiont lineages (red and blue) either switch between host lineages or are inherited, depending on the transmission mode, until they coalesce in the same host lineage (purple).
(TIF)

S7 Fig. The observed number of pairwise differences across a range of parameters under the endosymbiont population model described above. Each distribution is 100 replicates with varying NH, H, and NS. The expectation following Equation 9 above is plotted as a red line and differs by less than 2 segregating sites from the observed mean for all cases investigated here.
(TIF)

S8 Fig. Local alignments suggest that few rearrangements have occurred between the *S. velum* and *S. elarraichensis* symbiont genomes. *S. elarraichensis* symbiont is the closest known relative of the *S. velum* symbiont, however material is exceedingly hard to obtain for this association, which occurs at a mud volcano at approximately 500–1000 m depth, and only a fragmented draft genome assembly was available. However, even these relatively short range segments reveal complete synteny (left). In comparison, over the same genomic distances, many rearrangements are evident between *S. velum* and *S. pervernicosa* (right), with the minority of segments retaining synteny.
(TIF)

S1 Table. Sample, sequencing library, and mapping coverage information. The second set of coverages listed for *C. fausta* apply to the libraries used for the intra-host analysis.
(XLSX)

S2 Table. *De novo* reference assemblies were assembled with Nanopore reads and polished with Illumina data. Illumina reads were used for individual sample genotype calling. There were no gaps (Ns) in any of the assemblies. The percent complete measure reflects how many of the 34 "essential genes" (see Materials and Methods) were found in the assembled genomes.
(XLSX)

S3 Table. Symbiont species named in this study and named previously. See S3 Supporting Text for full diagnoses and descriptions.
(XLSX)

S4 Table. Taxa and accession numbers used in the mitochondrial genome phylogenetic analysis and divergence dating.
(XLSX)

S5 Table. Taxa and accession numbers used in the bacterial whole genome phylogenetic analysis and divergence dating.
(XLSX)

S6 Table. Divergence date estimates from different Beast2, TimeTree, and PATHd8 runs with different parameter values.
(XLSX)

S7 Table. Between-host symbiont population statistics calculated from consensus symbiont and mitochondrial genome sequences. Random Forest (RF) theta and log10(rho*l) estimates were inferred by fitting genome-wide values of pi, Watterson's Theta, and 4-gamete sites to values generated in coalescent simulations.
(XLSX)

S8 Table. Within-host symbiont and mitochondrial genetic diversity statistics. Mapping coverages in S1 Table.
(XLSX)

S9 Table. Out-of-bag (oob) scores for random-forest models for each parameter of interest, rho*l and theta, and for each sample size of endosymbiont individuals considered. Oob scores indicate how often the trained model is able to predict known values, with perfect prediction equal to one.
(XLSX)

S10 Table. Proportion of within-host variant sites that pass the 4-gamete test for recombination based upon read and read pair data over the given genomic intervals (constrained by Illumina library insert sizes).
(XLSX)

S11 Table. Comparative symbiont genome statistics and mobile element (ME) content. MEs were identified as regions within the symbiont genomes with high sequence identity to elements in insertion sequence, phage, and integrative conjugative element databases.
(XLSX)

S12 Table. Full list of ICEberg and ACLAME database mobile element hits with > = 90% sequence identity to endosymbiont genomic regions and genes.
(XLSX)

S1 Text. Supplemental text sections 1–3 for Forever young bacterial symbiont genomes.
(PDF)

Acknowledgments

We thank two anonymous reviewers and Emma George for their helpful comments on the manuscript and Xavier Didelot for kindly providing the *Bacillus cereus* validation dataset. For *Calyptogena* freezer samples collected previously, we thank Colleen Cavanaugh and Peter Girguis. For their assistance in collecting *Bathymodiolus* samples from the Lau Basin and Veatch Canyon, respectively, we thank the crews of the R/V *Falkor* and ROV *ROPOS* and the R/V *Atlantis* and HOV *Alvin*. We thank Peter Wilton for feedback on the symbiont coalescent model derivation.

Author Contributions

Conceptualization: Shelbi L. Russell, Russell Corbett-Detig.

Data curation: Shelbi L. Russell, Evan Pepper-Tunick, Ashley Byrne, Jennie Ruelas Castillo, Christopher Vollmers, Roxanne A. Beinart, Russell Corbett-Detig.

Formal analysis: Shelbi L. Russell, Russell Corbett-Detig.

Funding acquisition: Roxanne A. Beinart, Russell Corbett-Detig.

Investigation: Shelbi L. Russell, Russell Corbett-Detig.

Methodology: Shelbi L. Russell, Russell Corbett-Detig.

Project administration: Shelbi L. Russell, Russell Corbett-Detig.

Resources: Russell Corbett-Detig.

Software: Shelbi L. Russell, Russell Corbett-Detig.

Supervision: Shelbi L. Russell, Russell Corbett-Detig.

Validation: Shelbi L. Russell, Russell Corbett-Detig.

Visualization: Shelbi L. Russell, Jesper Svedberg, Russell Corbett-Detig.

Writing – original draft: Shelbi L. Russell, Russell Corbett-Detig.

Writing – review & editing: Shelbi L. Russell, Jesper Svedberg, Christopher Vollmers, Roxanne A. Beinart, Russell Corbett-Detig.

References

1. Moran NA, Bennett GM. The Tiniest Tiny Genomes. Annu Rev Microbiol. 2014; 68: 195–215. https://doi.org/10.1146/annurev-micro-091213-112901 PMID: 24995872
2. Lo W-S, Huang Y-Y, Kuo C-H. Winding paths to simplicity: genome evolution in facultative insect symbionts. Lai E-M, editor. FEMS Microbiol Rev. 2016; 40: 855–874. https://doi.org/10.1093/femsre/fuw028 PMID: 28204477
3. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. ISME J. 2014;8.
4. Meseguer AS, Manzano-Marín A, Coeur d'Acier A, Clamens A-L, Godefroid M, Jousselin E. *Buchnera* has changed flatmate but the repeated replacement of co-obligate symbionts is not associated with

the ecological expansions of their aphid hosts. Mol Ecol. 2017; 26: 2363–2378. https://doi.org/10.1111/mec.13910 PMID: 27862540

5. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol. 2011 [cited 3 Jan 2017]. https://doi.org/10.1038/nrmicro2670 PMID: 22064560
6. Toft C, Andersson SGE. Evolutionary microbial genomics: insights into bacterial host adaptation. Nat Rev Genet. 2010; 11: 465–475. https://doi.org/10.1038/nrg2798 PMID: 20517341
7. Lambert JD, Moran NA. Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria. Proc Natl Acad Sci. 1998; 95: 4458–4462. https://doi.org/10.1073/pnas.95.8.4458 PMID: 9539759
8. Kuwahara H, Takaki Y, Yoshida T, Shimamura S, Takishita K, Reimer JD, et al. Reductive genome evolution in chemoautotrophic intracellular symbionts of deep-sea Calyptogena clams. Extremophiles. 2008; 12: 365–374. https://doi.org/10.1007/s00792-008-0141-2 PMID: 18305898
9. Herbeck JT, Funk DJ, Degnan PH, Wernegreen JJ. A Conservative Test of Genetic Drift in the Endosymbiotic Bacterium Buchnera: Slightly Deleterious Mutations in the Chaperonin groEL. Genetics. 2003; 10.
10. Shapiro BJ, Alm E. The slow:fast substitution ratio reveals changing patterns of natural selection in γ-proteobacterial genomes. ISME J. 2009; 3: 1180–1192. https://doi.org/10.1038/ismej.2009.51 PMID: 19458656
11. Newton ILG, Woyke T, Auchtung TA, Dilly GF, Dutton RJ, Fisher MC, et al. The Calyptogena magnifica Chemoautotrophic Symbiont Genome. Science. 2007; 315: 998–1000. https://doi.org/10.1126/science.1138438 PMID: 17303757
12. Kuwahara H, Yoshida T, Takaki Y, Shimamura S, Nishi S, Harada M, et al. Reduced Genome of the Thioautotrophic Intracellular Symbiont in a Deep-Sea Clam, Calyptogena okutanii. Curr Biol. 2007; 17: 881–886. https://doi.org/10.1016/j.cub.2007.04.039 PMID: 17493812
13. Dmytrenko O, Russell SL, Loo WT, Fontanez KM, Liao L, Roeselers G, et al. The genome of the intracellular bacterium of the coastal bivalve, *Solemya velum*: a blueprint for thriving in and out of symbiosis. BMC Genomics. 2014;15. https://doi.org/10.1186/1471-2164-15-15
14. Miller IJ, Vanee N, Fong SS, Lim-Fong GE, Kwan JC. Lack of Overt Genome Reduction in the Bryostatin-Producing Bryozoan Symbiont "Candidatus Endobugula sertula." Drake HL, editor. Appl Environ Microbiol. 2016; 82: 6573–6583. https://doi.org/10.1128/AEM.01800-16 PMID: 27590822
15. Russell SL, Corbett-Detig RB, Cavanaugh CM. Mixed transmission modes and dynamic genome evolution in an obligate animal–bacterial symbiosis. ISME J. 2017; 1359–1371. https://doi.org/10.1038/ismej.2017.10 PMID: 28234348
16. Hendry TA, Freed LL, Fader D, Fenolio D, Sutton TT, Lopez JV. Ongoing Transposon-Mediated Genome Reduction in the Luminous Bacterial Symbionts of Deep-Sea Ceratioid Anglerfishes. Moran NA, editor. mBio. 2018; 9: e01033–18, /mbio/9/3/mBio.01033-18.atom. https://doi.org/10.1128/mBio.01033-18 PMID: 29946051
17. Jäckle O, Seah BKB, Tietjen M, Leisch N, Liebeke M, Kleiner M, et al. Chemosynthetic symbiont with a drastically reduced genome serves as primary energy storage in the marine flatworm *Paracatenula*. Proc Natl Acad Sci. 2019; 201818995. https://doi.org/10.1073/pnas.1818995116 PMID: 30962361
18. George EE, Husnik F, Tashyreva D, Prokopchuk G, Horák A, Kwong WK, et al. Highly Reduced Genomes of Protist Endosymbionts Show Evolutionary Convergence. Curr Biol. 2020; 30: 925–933. e3. https://doi.org/10.1016/j.cub.2019.12.070 PMID: 31978335
19. Ran L, Larsson J, Vigil-Stenman T, Nylander JAA, Ininbergs K, Zheng W-W, et al. Genome Erosion in a Nitrogen-Fixing Vertically Transmitted Endosymbiotic Multicellular Cyanobacterium. Ahmed N, editor. PLoS ONE. 2010; 5: e11486. https://doi.org/10.1371/journal.pone.0011486 PMID: 20628610
20. Oakeson KF, Gil R, Clayton AL, Dunn DM, von Niederhausern AC, Hamil C, et al. Genome Degeneration and Adaptation in a Nascent Stage of Symbiosis. Genome Biol Evol. 2014; 6: 76–93. https://doi.org/10.1093/gbe/evt210 PMID: 24407854
21. Johnson SB, Krylova EM, Audzijonyte A, Sahling H, Vrijenhoek RC. Phylogeny and origins of chemosynthetic vesicomyid clams. Syst Biodivers. 2017; 15: 346–360. https://doi.org/10.1080/14772000.2016.1252438
22. Sharma PP, Zardus JD, Boyle EE, González VL, Jennings RM, McIntyre E, et al. Into the deep: A phylogenetic approach to the bivalve subclass Protobranchia. Mol Phylogenet Evol. 2013; 69: 188–204. https://doi.org/10.1016/j.ympev.2013.05.018 PMID: 23742885
23. Ozawa G, Shimamura S, Takaki Y, Takishita K, Ikuta T, Barry JP, et al. Ancient occasional host switching of maternally transmitted bacterial symbionts of chemosynthetic vesicomyid clams. Genome Biol Evol. 2017; 9: 2226–2236. https://doi.org/10.1093/gbe/evx166 PMID: 28922872
24. Kuwahara H, Takaki Y, Shimamura S, Yoshida T, Maeda T, Kunieda T, et al. Loss of genes for DNA recombination and repair in the reductive genome evolution of thioautotrophic symbionts of

Calyptogena clams. BMC Evol Biol. 2011; 11: 285. https://doi.org/10.1186/1471-2148-11-285 PMID: 21966992

25. Russell SL. Transmission mode is associated with environment type and taxa across bacteria-eukaryote symbioses: a systematic review and meta-analysis. FEMS Microbiol Lett. 2019; fnz013.

26. Wernegreen JJ. Endosymbiont evolution: predictions from theory and surprises from genomes: Endosymbiont genome evolution. Ann N Y Acad Sci. 2015; 1360: 16–35. https://doi.org/10.1111/nyas.12740 PMID: 25866055

27. Fontanez KM, Cavanaugh CM. Evidence for horizontal transmission from multilocus phylogeny of deep-sea mussel (Mytilidae) symbionts: Horizontal transmission of mussel symbionts. Environ Microbiol. 2014; 16: 3608–3621. https://doi.org/10.1111/1462-2920.12379 PMID: 24428587

28. Won Y-J, Hallam SJ, O'Mullan GD, Pan IL, Buck KR, Vrijenhoek RC. Environmental acquisition of thiotrophic endosymbionts by deep-sea mussels of the genus Bathymodiolus. Appl Environ Microbiol. 2003; 69: 6785–6792. https://doi.org/10.1128/aem.69.11.6785-6792.2003 PMID: 14602641

29. Wentrup C, Wendeberg A, Huang JY, Borowski C, Dubilier N. Shift from widespread symbiont infection of host tissues to specific colonization of gills in juvenile deep-sea mussels. ISME J. 2013; 7: 1244–1247. https://doi.org/10.1038/ismej.2013.5 PMID: 23389105

30. Gustafson RG, Reid RG. Association of bacteria with larvae of the gutless protobranch bivalve Solemya reidi (Cryptodonta: Solemyidae). Mar Biol. 1988; 97: 389–401.

31. Russell SL, McCartney E, Cavanaugh CM. Transmission strategies in a chemosynthetic symbiosis: detection and quantification of symbionts in host tissues and their environment. Proc R Soc B Biol Sci. 2018; 285: 9.

32. Krueger DM, Gustafson RG, Cavanaugh CM. Vertical transmission of chemoautotrophic symbionts in the bivalve Solemya velum (Bivalvia: Protobranchia). Biol Bull. 1996; 190: 195–202. https://doi.org/10.2307/1542539 PMID: 8652730

33. Ikuta T, Igawa K, Tame A, Kuroiwa T, Kuroiwa H, Aoki Y, et al. Surfing the vegetal pole in a small population: extracellular vertical transmission of an "intracellular" deep-sea clam symbiont. R Soc Open Sci. 2016; 3: 160130. https://doi.org/10.1098/rsos.160130 PMID: 27293794

34. Cary SC, Giovannoni SJ. Transovarial inheritance of endosymbiotic bacteria in clams inhabiting deep-sea hydrothermal vents and cold seeps. Proc Natl Acad Sci. 1993; 90: 5695–5699. https://doi.org/10.1073/pnas.90.12.5695 PMID: 8100068

35. Breusing C, Johnson SB, Vrijenhoek RC, Young CR. Host hybridization as a potential mechanism of lateral symbiont transfer in deep-sea vesicomyid clams. Mol Ecol. 2019; mec.15224. https://doi.org/10.1111/mec.15224 PMID: 31478269

36. Dubilier N, Bergin C, Lott C. Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. Nat Rev Microbiol. 2008; 6: 725–740. https://doi.org/10.1038/nrmicro1992 PMID: 18794911

37. Duperron S, Halary S, Lorion J, Sibuet M, Gaill F. Unexpected co-occurrence of six bacterial symbionts in the gills of the cold seep mussel Idas sp. (Bivalvia: Mytilidae). Environ Microbiol. 2008; 10: 433–445. https://doi.org/10.1111/j.1462-2920.2007.01465.x PMID: 18093159

38. Noellette Conway, Judith McDowell Capuzzo, Brian Fry. The Role of Endosymbiotic Bacteria in the Nutrition of Solemya velum: Evidence from a Stable Isotope Analysis of Endosymbionts and Host. Limnol Oceanogr. 1989; 34: 249–255.

39. Reid Robert G. B., Bernard Frank R. Gutless Bivalves. Sci New Ser. 1980; 208: 609–610.

40. Decker C, Olu K, Arnaud-Haond S, Duperron S. Physical proximity may promote lateral acquisition of bacterial symbionts in vesicomyid clams. López-García P, editor. PLoS ONE. 2013; 8: e64830. https://doi.org/10.1371/journal.pone.0064830 PMID: 23861734

41. Ponnudurai R, Kleiner M, Sayavedra L, Petersen JM, Moche M, Otto A, et al. Metabolic and physiological interdependencies in the Bathymodiolus azoricus symbiosis. ISME J. 2016 [cited 3 Jan 2017]. Available: http://www.nature.com/ismej/journal/vaop/ncurrent/full/ismej2016124a.html

42. Ikuta T, Takaki Y, Nagai Y, Shimamura S, Tsuda M, Kawagucci S, et al. Heterogeneous composition of key metabolic gene clusters in a vent mussel symbiont population. ISME J. 2016; 10: 990. https://doi.org/10.1038/ismej.2015.176 PMID: 26418631

43. Brandvain Y, Goodnight C, Wade MJ. Horizontal Transmission Rapidly Erodes Disequilibria Between Organelle and Symbiont Genomes. Genetics. 2011; 189: 397–404. https://doi.org/10.1534/genetics.111.130906 PMID: 21750254

44. Stewart FJ, Cavanaugh CM. Bacterial endosymbioses in Solemya (Mollusca: Bivalvia)—Model systems for studies of symbiont–host adaptation. Antonie Van Leeuwenhoek. 2006; 90: 343–360. https://doi.org/10.1007/s10482-006-9086-6 PMID: 17028934

45. Cocks LRM, Torsvik TH. Earth geography from 500 to 400 million years ago: a faunal and palaeomagnetic review. J Geol Soc. 2002; 159: 631–644. https://doi.org/10.1144/0016-764901-118

46. Biari Y, Klingelhoefer F, Sahabi M, Funck T, Benabdellouahed M, Schnabel M, et al. Opening of the central Atlantic Ocean: Implications for geometric rifting and asymmetric initial seafloor spreading after continental breakup: Opening of the Central Atlantic Ocean. Tectonics. 2017; 36: 1129–1150. https://doi.org/10.1002/2017TC004596

47. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, et al. Population Genomics of Early Events in the Ecological Differentiation of Bacteria. Science. 2012; 336: 48–51. https://doi.org/10.1126/science.1218198 PMID: 22491847

48. Rosen MJ, Davison M, Bhaya D, Fisher DS. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science. 2015; 348: 1019–1023. https://doi.org/10.1126/science.aaa4456 PMID: 26023139

49. Chong RA, Park H, Moran NA. Genome Evolution of the Obligate Endosymbiont Buchnera aphidicola. Agashe D, editor. Mol Biol Evol. 2019; 36: 1481–1489. https://doi.org/10.1093/molbev/msz082 PMID: 30989224

50. Ansorge R, Romano S, Sayavedra L, Kupczok A, Tegetmeyer HE, Dubilier N, et al. Diversity matters: Deep-sea mussels harbor multiple symbiont strains. bioRxiv. 2019 [cited 24 Jul 2019]. https://doi.org/10.1101/531459

51. Ansari MA, Didelot X. Inference of the Properties of the Recombination Process from Whole Bacterial Genomes. Genetics. 2014; 196: 253–265. https://doi.org/10.1534/genetics.113.157172 PMID: 24172133

52. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 2009; 3: 199–208. https://doi.org/10.1038/ismej.2008.93 PMID: 18830278

53. Rocha EPC. An Appraisal of the Potential for Illegitimate Recombination in Bacterial Genomes and Its Consequences: From Duplications to Genome Reduction. Genome Res. 2003; 13: 1123–1132. https://doi.org/10.1101/gr.966203 PMID: 12743022

54. Neher RA. Genetic Draft, Selective Interference, and Population Genetics of Rapid Adaptation. Annu Rev Ecol Evol Syst. 2013; 44: 195–215. https://doi.org/10.1146/annurev-ecolsys-110512-135920

55. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974; 23: 23–25. PMID: 4407212

56. Nilsson AI, Koskiniemi S, Eriksson S, Kugelberg E, Hinton JCD, Andersson DI. Bacterial genome size reduction by experimental evolution. Proc Natl Acad Sci U S A. 2005; 102: 12112–12116. https://doi.org/10.1073/pnas.0503654102 PMID: 16099836

57. Clayton AL, Jackson DG, Weiss RB, Dale C. Adaptation by Deletogenic Replication Slippage in a Nascent Symbiont. Mol Biol Evol. 2016; 33: 1957–1966. https://doi.org/10.1093/molbev/msw071 PMID: 27189544

58. Merrikh CN, Merrikh H. Gene inversion potentiates bacterial evolvability and virulence. Nat Commun. 2018; 9: 4662. https://doi.org/10.1038/s41467-018-07110-3 PMID: 30405125

59. Newton ILG, Bordenstein SR. Correlations Between Bacterial Ecology and Mobile DNA. Curr Microbiol. 2011; 62: 198–208. https://doi.org/10.1007/s00284-010-9693-3 PMID: 20577742

60. Glémin S, Galtier N. Genome Evolution in Outcrossing Versus Selfing Versus Asexual Species. In: Anisimova M, editor. Evolutionary Genomics. Totowa, NJ: Humana Press; 2012. pp. 311–335. https://doi.org/10.1007/978-1-61779-582-4_11

61. Allen JM, Reed DL, Perotti MA, Braig HR. Evolutionary Relationships of "Candidatus Riesia spp.," Endosymbiotic Enterobacteriaceae Living within Hematophagous Primate Lice. Appl Environ Microbiol. 2007; 73: 1659–1664. https://doi.org/10.1128/AEM.01877-06 PMID: 17220259

62. Lefevre C. Endosymbiont Phylogenesis in the Dryophthoridae Weevils: Evidence for Bacterial Replacement. Mol Biol Evol. 2004; 21: 965–973. https://doi.org/10.1093/molbev/msh063 PMID: 14739242

63. Degnan PH, Lazarus AB, Brock CD, Wernegreen JJ. Host–Symbiont Stability and Fast Evolutionary Rates in an Ant–Bacterium Association: Cospeciation of Camponotus Species and Their Endosymbionts, Candidatus Blochmannia. Johnson K, editor. Syst Biol. 2004; 53: 95–110. https://doi.org/10.1080/10635150490264842 PMID: 14965905

64. Moran N, Wernegreen J. Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends Ecol Evol. 2000; 15: 321–326. https://doi.org/10.1016/s0169-5347(00)01902-9 PMID: 10884696

65. Takiya DM, Tran PL, Dietrich CH, Moran NA. Co-cladogenesis spanning three phyla: leafhoppers (Insecta: Hemiptera: Cicadellidae) and their dual bacterial symbionts. Mol Ecol. 2006; 15: 4175–4191. https://doi.org/10.1111/j.1365-294X.2006.03071.x PMID: 17054511

66. Thao ML, Moran NA, Abbot P, Brennan EB, Burckhardt DH, Baumann P. Cospeciation of Psyllids and Their Primary Prokaryotic Endosymbionts. Appl Environ Microbiol. 2000; 66: 2898–2905. https://doi.org/10.1128/aem.66.7.2898-2905.2000 PMID: 10877784

67. Thao ML, Gullan PJ, Baumann P. Secondary (-Proteobacteria) Endosymbionts Infect the Primary (-Proteobacteria) Endosymbionts of Mealybugs Multiple Times and Coevolve with Their Hosts. Appl Environ Microbiol. 2002; 68: 3190–3197. https://doi.org/10.1128/aem.68.7.3190-3197.2002 PMID: 12088994

68. Santos-Garcia D, Vargas-Chavez C, Moya A, Latorre A, Silva FJ. Genome Evolution in the Primary Endosymbiont of Whiteflies Sheds Light on Their Divergence. Genome Biol Evol. 2015; 7: 873–888. https://doi.org/10.1093/gbe/evv038 PMID: 25716826

69. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Bioinformatics; 2019 Jan. https://doi.org/10.1101/530972

70. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. Wang J, editor. PLoS ONE. 2014; 9: e112963. https://doi.org/10.1371/journal.pone.0112963 PMID: 25409509

71. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013; arXiv:1303.3997.

72. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012; 28: 1420–1428. https://doi.org/10.1093/bioinformatics/bts174 PMID: 22495754

73. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012; 19: 455–477. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599

74. Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics. 2012; 28: 2223–2230. https://doi.org/10.1093/bioinformatics/bts429 PMID: 22796954

75. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10: 421. https://doi.org/10.1186/1471-2105-10-421 PMID: 20003500

76. Lagesen K, Hallin P, Rodland EA, Staerfeldt H-H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007; 35: 3100–3108. https://doi.org/10.1093/nar/gkm160 PMID: 17452365

77. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997; 25: 955–964. https://doi.org/10.1093/nar/25.5.955 PMID: 9023104

78. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015; 25: 1043–1055. https://doi.org/10.1101/gr.186072.114 PMID: 25977477

79. Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 2008; 9: 1.

80. Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. Genome Biol. 2010; 11: 1.

81. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, et al. MITOS: Improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol. 2013; 69: 313–319. https://doi.org/10.1016/j.ympev.2012.08.023 PMID: 22982435

82. Liu H, Cai S, Zhang H, Vrijenhoek RC. Complete mitochondrial genome of hydrothermal vent clam *Calyptogena magnifica*. Mitochondrial DNA Part A. 2016; 27: 4333–4335. https://doi.org/10.3109/19401736.2015.1089488 PMID: 26462964

83. Plazzi F, Ribani A, Passamonti M. The complete mitochondrial genome of Solemya velum (Mollusca: Bivalvia) and its relationships with Conchifera. BMC Genomics. 2013; 14: 1. https://doi.org/10.1186/1471-2164-14-1

84. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

85. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43: 491–498. https://doi.org/10.1038/ng.806 PMID: 21478889

86. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522

87. Russell SL, Cavanaugh CM. Intrahost Genetic Diversity of Bacterial Symbionts Exhibits Evidence of Mixed Infections and Recombinant Haplotypes. Mol Biol Evol. 2017; 34: 2747–2761. https://doi.org/10.1093/molbev/msx188 PMID: 29106592

88. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30: 1312–1313. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

89. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975; 7: 256–276. https://doi.org/10.1016/0040-5809(75)90020-9 PMID: 1145509

90. Nei M, Li W-H. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci. 1979; 76: 5269–5273. https://doi.org/10.1073/pnas.76.10.5269 PMID: 291943

91. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5: 1.

92. Ponnudurai R, Sayavedra L, Kleiner M, Heiden SE, Thürmer A, Felbeck H, et al. Genome sequence of the sulfur-oxidizing Bathymodiolus thermophilus gill endosymbiont. Stand Genomic Sci. 2017;12. https://doi.org/10.1186/s40793-017-0232-8

93. Darling AE, Mau B, Perna NT. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. Stajich JE, editor. PLoS ONE. 2010; 5: e11147. https://doi.org/10.1371/journal.pone.0011147 PMID: 20593022

94. Guy L, Roat Kultima J, Andersson SGE. genoPlotR: comparative gene and genome visualization in R. Bioinformatics. 2010; 26: 2334–2335. https://doi.org/10.1093/bioinformatics/btq413 PMID: 20624783

95. Leplae R, Lima-Mendez G, Toussaint A. ACLAME: A CLAssification of Mobile genetic Elements, update 2010. Nucleic Acids Res. 2010; 38: D57–D61. https://doi.org/10.1093/nar/gkp938 PMID: 19933762

96. Bi D, Xu Z, Harrison EM, Tai C, Wei Y, He X, et al. ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. Nucleic Acids Res. 2012; 40: D621–D626. https://doi.org/10.1093/nar/gkr846 PMID: 22009673

97. Ranwez V, Harispe S, Delsuc F, Douzery EJP. MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. Murphy WJ, editor. PLoS ONE. 2011; 6: e22594. https://doi.org/10.1371/journal.pone.0022594 PMID: 21949676

98. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007; 24: 1586–1591. https://doi.org/10.1093/molbev/msm088 PMID: 17483113

99. Plazzi F, Puccio G, Passamonti M. Comparative Large-Scale Mitogenomics Evidences Clade-Specific Evolutionary Trends in Mitochondrial DNAs of Bivalvia. Genome Biol Evol. 2016; 8: 2544–2564. https://doi.org/10.1093/gbe/evw187 PMID: 27503296

100. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012; 28: 1647–1649. https://doi.org/10.1093/bioinformatics/bts199 PMID: 22543367

101. Ankenbrand MJ, Keller A. bcgTree: automatized phylogenetic tree building from bacterial core genomes. Chain F, editor. Genome. 2016; 59: 783–791. https://doi.org/10.1139/gen-2015-0175 PMID: 27603265

102. Wernersson R. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 2003; 31: 3537–3539. https://doi.org/10.1093/nar/gkg609 PMID: 12824361

103. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013; 30: 772–780. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

104. Hedge J, Wilson DJ. Bacterial Phylogenetic Reconstruction from Whole Genomes Is Robust to Recombination but Demographic Inference Is Not. Vidaver AK, editor. mBio. 2014; 5: e02158–14. https://doi.org/10.1128/mBio.02158-14 PMID: 25425237

105. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. Prlic A, editor. PLoS Comput Biol. 2014; 10: e1003537. https://doi.org/10.1371/journal.pcbi.1003537 PMID: 24722319

106. Behrensmeyer A, Turner A. Taxonomic occurrences of Bivalvia recorded in the Paleobiology Database. In: Fossilworks [Internet]. 2013. Available: http://fossilworks.org

107. Pojeta J, Runnegar B, Kriz J. Fordilla troyensis Barrande: The Oldest Known Pelecypod. Science. 1973; 180: 866–868. https://doi.org/10.1126/science.180.4088.866 PMID: 17789257

108. Brasier MD, Hewitt RA, Brasier CJ. On the Late Precambrian–Early Cambrian Hartshill Formation of Warwickshire. Geol Mag. 1978; 115: 21–36. https://doi.org/10.1017/S0016756800040954

109. Battistuzzi FU, Tao Q, Jones L, Tamura K, Kumar S. RelTime Relaxes the Strict Molecular Clock throughout the Phylogeny. Martin B, editor. Genome Biol Evol. 2018; 10: 1631–1636. https://doi.org/10.1093/gbe/evy118 PMID: 29878203

110. Tamura K, Tao Q, Kumar S. Theoretical Foundation of the RelTime Method for Estimating Divergence Times from Variable Evolutionary Rates. Russo C, editor. Mol Biol Evol. 2018; 35: 1770–1782. https://doi.org/10.1093/molbev/msy044 PMID: 29893954

111. Britton T, Anderson CL, Jacquet D, Lundqvist S, Bremer K. Estimating Divergence Times in Large Phylogenetic Trees. Anderson F, editor. Syst Biol. 2007; 56: 741–752. https://doi.org/10.1080/10635150701613783 PMID: 17886144

112. Sheridan PP, Freeman KH, Brenchley JE. Estimated Minimal Divergence Times of the Major Bacterial and Archaeal Phyla. Geomicrobiol J. 2003; 20: 1–14. https://doi.org/10.1080/01490450303891

113. De Maio N, Wilson DJ. The Bacterial Sequential Markov Coalescent. Genetics. 2017; 206: 333–343. https://doi.org/10.1534/genetics.116.198796 PMID: 28258183

114. Polimis K, Rokem A, Hazelton B. Confidence Intervals for Random Forests in Python. J Open Source Softw. 2017; 2: 124. https://doi.org/10.21105/joss.00124

115. Didelot X, Lawson D, Darling A, Falush D. Inference of Homologous Recombination in Bacteria Using Whole-Genome Sequences. Genetics. 2010; 186: 1435–1449. https://doi.org/10.1534/genetics.110.120121 PMID: 20923983