Hybridization breaks species barriers in long-term coevolution of a cyanobacterial population

Gabriel Birzu,[1] Harihara Subrahmaniam Muralidharan,[2] Danielle Goudeau,[3] Rex R. Malmstrom,[3] Daniel S. Fisher,[1*†] Devaki Bhaya[4*†]

[1]Department of Applied Physics, Stanford University, Stanford, CA 94305, USA
[2]Department of Computer Science, University of Maryland, College Park, MD 20742, USA
[3]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[4]Department of Plant Biology, Carnegie Institution for Science, Stanford, CA 94305, USA

*Corresponding authors. E-mail: dsfisher@stanford.edu (D.S.F.); dbhaya@carnegiescience.edu (D.B.).
†These authors contributed equally to this work.

June 6, 2023

Abstract. Bacterial species often undergo rampant recombination yet maintain cohesive genomic identity. Ecological differences can generate recombination barriers between species and sustain genomic clusters in the short term. But can these forces prevent genomic mixing during long-term coevolution? Cyanobacteria in Yellowstone hot springs comprise several diverse species that have coevolved for hundreds of thousands of years, providing a rare natural experiment. By analyzing more than 300 single-cell genomes, we show that despite each species forming a distinct genomic cluster, much of the diversity within species is the result of hybridization driven by selection, which has mixed their ancestral genotypes. This widespread mixing is contrary to the prevailing view that ecological barriers can maintain cohesive bacterial species and highlights the importance of hybridization as a source of genomic diversity.

The origin, persistence, and extinction of species in obligate sexual populations is a fundamental problem in evolutionary theory (*1*). But in bacteria, long thought to be primarily asexual, our understanding is still very limited (*2*, *3*). While it has long been known that bacteria can acquire non-essential genes from distantly-related strains (*4*), in recent years, studies across a wide range of bacteria have found that homologous recombination of core genes is also ubiquitous (*5–9*). How does the interplay between selection acting both asexually and on transferred elements, and barriers to recombination from geographical separation, mechanistic effects, and epistatic interactions lead to formation and maintenance of cohesive species?

One explanation is that ecological specialization provides barriers to recombination (*10*, *11*). This view is supported by evidence that selective sweeps of individual genes through parts of a population can lead to distinct ecological subtypes, or ecotypes, even in highly recombining bacterial populations (*12–14*). But it also raises an important open question: if species coexist and recombine with each other, do genetic differences continue to increase over evolutionary time or does recombination gradually break down barriers between species?

Answering this question is challenging because data on the underlying evolutionary dynamics is lacking. The time scales of direct observations and laboratory experiments are too short, while those corresponding to phylogenetic relationships between extant species are too long. And since spatial dynamics over these intermediate time scales are unknown in most bacteria (*8*, *15*) it is difficult to control for changes in local community structure during evolution. Therefore, microbial communities that coevolved for long time periods can act as natural experiments for studying bacterial speciation. The cyanobacterial populations found in hot springs from Yellowstone National Park form a natural incubator in which one such natural experiment is ongoing. These communities have two key features that allow us to identify and quantify the effects of coevolution on intermediate time scales.

First, like other thermophilic communities, these populations are geographically islolated (*16*, *17*). To date, the two most abundant species, *Synechococcus* A and B, have not been found outside of North America, suggesting that their global mixing time is much longer than for most other bacteria (*16*). Comparison of two isolate genomes from these species, OS-A and OS-B', suggest a divergence time considerably longer than the formation of the Yellowstone caldera around $640,000$ years ago (*18*). This would imply their ancestors colonized the caldera independently.

Second, these populations have considerable sub-species diversity, suggesting ongoing evolutionary processes generate and maintain it. Analysis of variants of 16S rRNA and marker genes was initially interpreted as evidence of asexual ecotypes occupying distinct niches (*19*, *20*). However, deep amplicon data from a diverse, but largely OS-B'-like sample revealed extensive recombination (*7*). Genetic exchange with other genomic clusters—which we refer to as hybridization—was also evident, particularly with relatives of OS-A. To date, evidence of hybridization in bacteria has mainly been from comparisons between reference genomes from large genomic databases (*21*, *22*), with only a handful of studies investigating hybridization in natural environments (*23–25*). The primary focus of the current paper is to determine the impact of hybridization on the long-term coevolution of a natural microbial community.

Characterizing hybridization requires long-distance linkage between variants that cannot be obtained from traditional metagenomics. Strain isolates are often not representative of the population diversity (*26*, *27*), while metagenome assembly destroys linkage. Recruiting metagenomic reads to reference genomes can in some cases be used to infer linkage (*8*), but the high diversity of *Synechococcus* prevents us from using such approaches. To circumvent these problems, we used single-cell genomics (*27*, *28*) to obtain over 300

genomes from the Mushroom and Octopus Springs, which contain two of the most well-studied communities from the Yellowstone caldera (*29*). This is to our knowledge the first large sample of genomes from hot spring cyanobacterial populations with minimal compositional bias.

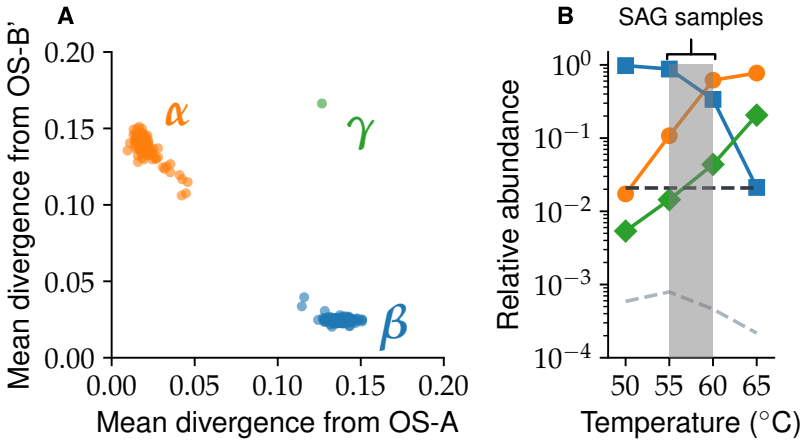

Figure 1: Synechococcus population comprises three genomic clusters. (**A**) The mean divergence across homologous genes from the two reference genomes is shown for every SAG in the data. Each dot represents a single SAG and is colored according to its assigned species. (**B**) The relative abundances of all three species across samples from different temperatures within Mushroom Spring. Abundances were estimated based on read recruitment from the metagenome to core genes from the γ SAG, as well as representative SAGs from the α and β populations. Dashed lines show the minimum detection frequencies from the SAG (black) and metagenome (gray) data, respectively (see SM, Sec. III). Shaded gray region shows temperature range of the 8 SAG samples.

Previous analysis of 16S amplicons identified two main clusters similar to OS-A and OS-B', but also several others at frequencies below 1% (*18*). But whether clusters within 16S amplicons correspond to clusters at the whole genome level in such a rapidly recombining population is not clear (*7*, *30*). Therefore, we first compared each single-amplified genome (SAG) to *Synechococcus* OS-A and OS-B'. Almost all SAGs (330 out of 331) comprised two clusters with an average nucleotide divergence of $\sim 2\%$ from one of the reference genomes and $\sim 15\%$ from the other. We labeled the cluster closer to OS-A by α (128 SAGs) and the other by β (202 SAGs), to distinguish them from the reference genomes. Surprisingly, one SAG (approximately 70% complete) was $\sim 15\%$ diverged from both OS-A and OS-B' and formed a separate cluster which we labeled γ (fig. 1A). Clustering methods based on different metrics gave identical results (figs. S2-4).

We used single-cell genomes as references and recruited reads from a large collection of metagenome samples to determine the abundance of γ across different environments (see SM, Sec. III). The abundance of γ was highly correlated with α (figs. S7 and S8) and increased rapidly with temperature in the Mushroom Spring samples (fig. 1B). One interpretation of this pattern is that α, β, and γ represent distinct species adapted to different temperatures. Alternatively, high rates of recombination could generate different genotypes within each genome cluster that are adapted to different temperatures, with the relative frequencies of each cluster being only a correlate and not a causal driver of temperature adaptation. We therefore decided to quantify the frequencies of recombination between and within species and investigate its role in maintaining genetic diversity within the population.

We measured the decrease in linkage disequilibrium along the genome in α and β (SM, Sec. IV). Within

β, the linkage disequilibrium σ_d^2 decreased by around a factor of 10 over distances of $\sim$ 1 kbp (fig. S12). This agreed quantitatively with earlier results based on deep amplicon sequencing from ref. (*7*), which showed a decrease by a factor of $\sim$ 7 over distances of up to 300 bp. Over genome-wide length scales, linkage was small ($\sigma_d^2 = 0.06$) and close to a fully-unlinked control ($\sigma_d^2 = 0.04$). We found similar results for α, but with a quantitatively slower decrease in linkage by a factor of $\sim$ 8 over $\sim$ 1 kbp. Again genome-wide linkage values were close to the fully-unlinked control ($\sigma_d^2 = 0.14$ for the data vs $\sigma_d^2 = 0.12$ for the control). However, we noticed that around 40% of genes had diversities similar to β or higher, while the remaining 60% had little to no diversity (figs. S23-25). Surprisingly, we did not find evidence of substantial differences in recombination rates between the two sets of genes (fig. S15). However, the large heterogeneity in diversity led us to hypothesize that hybridization might play an important role in the evolution of the community.

We first focused our analysis on single genes and began by constructing a set of orthologous gene groups using a standard clustering based on sequence similarity (SM, Secs. II, VI). To distinguish between hybridization and horizontal gene transfer of accessory genes, which has been documented in Synechococcus and in many other bacteria (*21*, *31–33*), we focused only on core genes that appeared to be present in all genomes, by removing those groups which had a copy number per genome that were not close to one. For each core gene, we identified distinct clusters that could be reliably separated from each other and tentatively assigned them to each of the three species. We could identify such clusters in $\sim 70\%$ of genes, including around 80 genes with more than three clusters (SM, Sec. V; fig. S17).

We searched for transfers of whole genes across species among the $\sim 70\%$ of genes that could be split into distinct clusters. We found several dozen (100 in α and 84 in β) *simple hybrid loci*, in which some but not all of the cells had undergone hybridization (fig. 2AB). However, we excluded many other candidate hybridization events based on very conservative technical criteria so the rate of hybridization in the population is likely considerably higher (SM, Sec. V). In both species, simple hybrids were scattered along the genome and contained alleles from different donor species, implying multiple hybridization events occurred independently, as expected when recombined segments are much shorter than the genome. Around 40% of α hybrid alleles were present in multiple cells, while in β the fraction was lower, at around 20%. The high frequencies of some hybrid alleles—up to $\sim 20\%$—suggest they were the result of partial sweeps. Focusing on transfers between α and β, we found that a majority of non-singleton hybrid alleles were as diverse as those in the donor population (fig. S18). This implies multiple independent hybridization events took place at these loci and suggests these hybrids could have risen to high frequencies through soft sweeps (*34*). Surprisingly, we also found that more than half of hybrid alleles were from donors not present in the population, suggesting hybridization with other species not present in our samples. However, a systematic search over a much wider range of metagenomic samples did not find evidence of other *Synechococcus* species, implying that the donors have since gone extinct (SM, Sec. III; figs. S9-11). These results suggested the current population is descended from multiple cyanobacterial species that hybridized with each other. For further evidence supporting this hypothesis, we examined the $\sim 30\%$ of core genes with a single mixed-species sequence cluster.

Closer inspection of almost 500 mixed-species sequence clusters revealed the majority were the result of extensive hybridization, with some others due to poor alignments caused by large numbers of indels (fig. S1). We therefore refer to these as *mosaic hybrid loci*. The mosaic loci were present throughout the genomes of both species, which suggests they were not the result of a small number of recent events but were instead the result of ongoing genetic exchange within the population (fig. 2AB). Within α, most

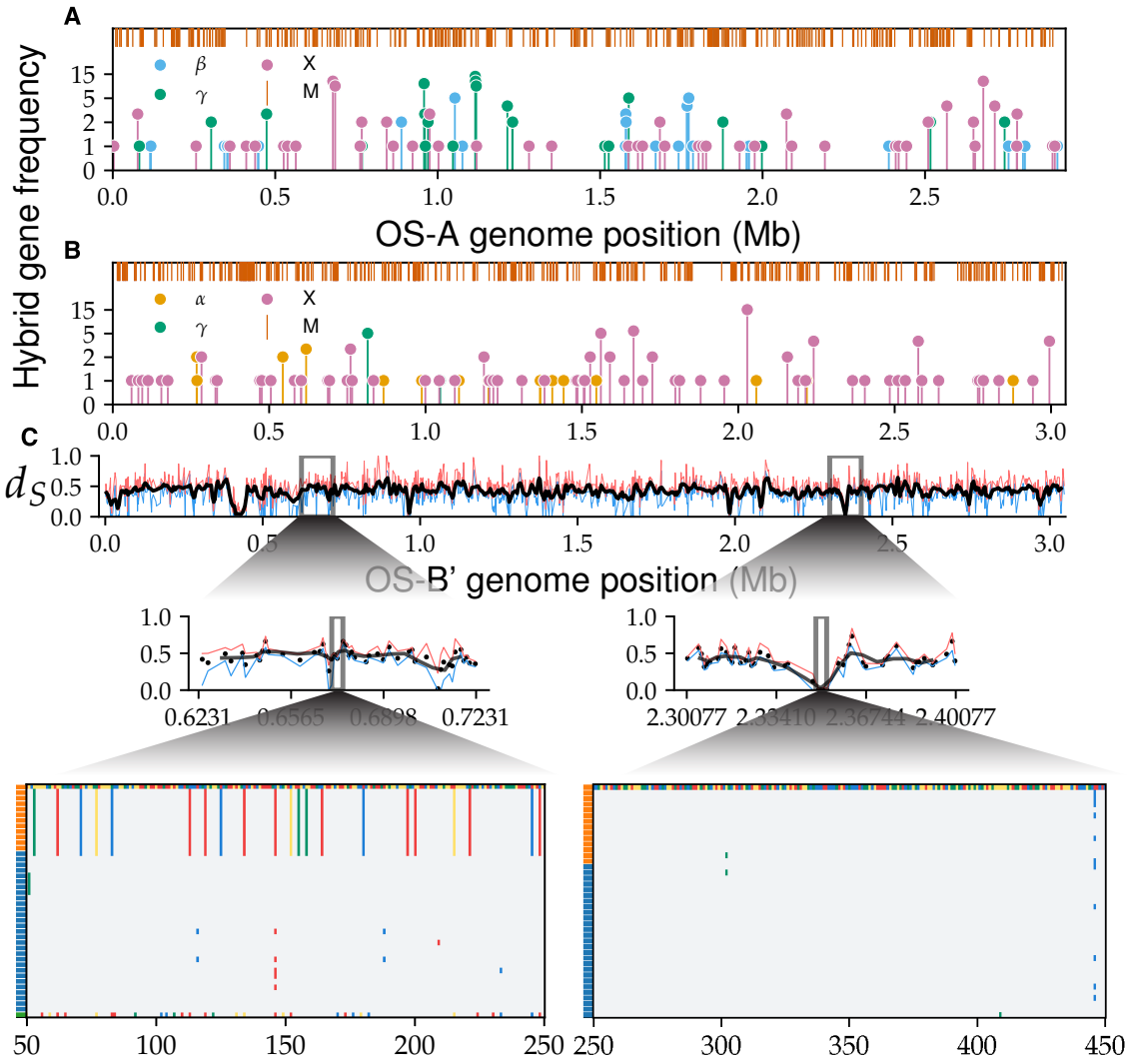

Figure 2: Transfers of whole-gene alleles between species is common. The frequency of hybrid gene alleles along the α (**A**) and β (**B**) genomes. Different colors show the donor species, with unknown species labeled X and shown in magenta. The orange lines on top show mosaic loci without distinct gene clusters. (**C**) The mean synonymous divergence between α and β at homologous loci, ordered according to OS-B'. Smoothed line over a sliding window of five genes with a step size of two genes is shown in black. Maximum and minimum divergences at each locus are shown in red and blue, respectively. (Insets, middle) A zoomed-in region from a typical genomic region (left) and a diversity trough (right), with dots representing individual genes. (Insets, bottom) Alignment segments of individual genes, subsampled to 40 sequences for clarity. The first sequence is arbitrarily chosen as the reference with each nucleotide represented by a different color. Other sites are colored gray if same as reference or using nucleotide colors if mutated. Left colored bars show the species of each genome.

of these loci were part of the $\sim$ 40% of high-diversity genes we found previously, confirming that the diversity at those loci was the result of hybridization. By contrast, the diversity of mosaic loci was typical of other loci within β, suggesting most of the β genome has undergone hybridization.

Is the genomic mixing between species the result of neutral processes or could positive selection for beneficial hybrids plan an important role? Distinguishing between these alternatives is difficult in general,

so we instead focused on cases where beneficial hybrid genes swept through the entire host species, i.e., a population-wide sweep. Such full gene sweeps would greatly reduce the synonymous divergence between α and β, making them much easier to detect.

A systematic search revealed ~ 100 very low divergence loci, spread across more than half a dozen *diversity troughs* (fig. 2C). Comparing alignments of typical genes and diversity trough genes revealed a dramatic difference in fixed substitutions between species, which can only be explained by a recent selective sweep. The largest diversity trough contained 28 genes (around 25% of diversity trough genes), most of which were part of a nitrogen-fixation pathway. The presence of the entire pathway among the diversity troughs strongly suggests at least one of the ancestors acquired the ability to fix nitrogen after it colonized the Yellowstone caldera and that this gave it a substantial selective advantage. The genetic diversity within the pathway genes gave further clues about the hybridization process. The diversity within the trough was similar for both species ($\pi_S = 0.033$ in α and $\pi_S = 0.039$ in β) and was close to the average diversity within β ($\pi_S = 0.042$), but more than an order of magnitude larger than the diversity within α at non-hybrid loci ($\pi_S = 0.002$; see SM, Sec. VII and fig. S19). This large discrepancy strongly suggests α acquired the pathway from β through multiple independent transfers, after they colonized the Yellowstone caldera.

Selective sweeps of individual genes could also lead to mixing of other genomic segments from the donor species into the host population through genetic hitchhiking. If the diversity within the two hybridizing species is low, this process would result in blocks of tightly linked SNPs consisting of primarily two haplotypes, one from the host and another from the donor (fig. 3A). We developed a systematic method to search for SNP blocks based on clustering the SNP linkage matrix (see SM, Sec. VI). Despite using conservative thresholds, we found over one thousand of short SNP blocks in each species, scattered across hundreds of genes, including in highly-conserved ribosomal genes and genes in which distinct species clusters could be assigned (figs. 3BC, S20, S21). The divergences between SNP block haplotypes was similar to that between species, consistent with hybridization (figs. S22). To confirm their hybrid origin, we mapped the haplotypes from each block to the two other species in our sample. We found $\sim 75\%$ (665 out of 877) of SNP blocks within α and $\sim 25\%$ of those within β (256 out of 1046) contained a hybrid haplotype that was identical to the consensus sequence of one of the other species (fig. 3B). In addition, we found a comparable number of blocks that were statistically similar but did not match the other species (fig. 3BC). The most likely explanation is that these blocks resulted from hybridization with other cyanobacterial species, similar to the whole-gene hybrids found previously (fig. 2). Consistent with this explanation, we found that the relative contributions from different donors to hybrid alleles and to hybrid blocks were similar in each species (Table S1).

Do genome-scale epistatic interactions, possibly associated with ecological distinctions, lead to a small number of substrains with distinct combinations of block haplotypes or are haplotypes mixed by recombination? To answer this question we calculated the linkage coefficients between haplotypes for each pair of blocks and compared their distribution to a fully unlinked control (see SM, Sec. VI for details). Within the β species the two distributions were almost identical, implying that the genomes are consistent with a random mixture of block haplotypes, as was previously proposed (7). For α the linkage coefficients were statistically larger than those predicted from the unlinked control (fig. 3C), but the overall value was still low ($\langle r^2 \rangle \approx 0.10$ for data compared to $\langle r^2 \rangle \approx 0.04$ for the unlinked model). These results are consistent with the slower rate of linkage decrease within α compared to β discussed previously. We also found similar results for the linkage between hybrid gene alleles (SM, Sec. V). Thus, rather than finding

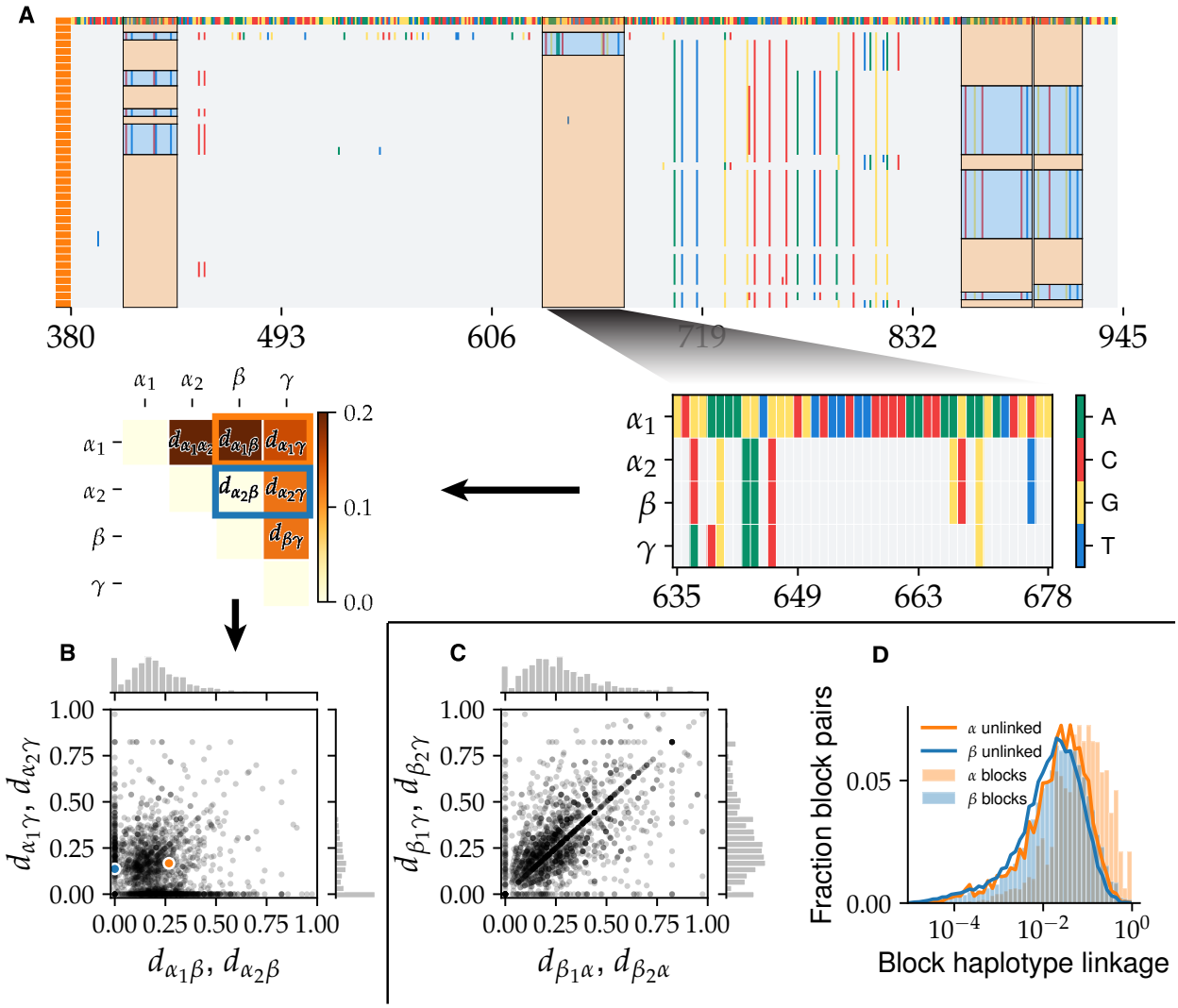

Figure 3: Hybridization between species leads to mixing of short DNA segments within single genes. (**A**, top) Alignment segment of α cells from a representative locus with four SNP blocks. The inferred ancestral α haplotype is highlighted in orange and the hybrid haplotype (all β here) is shown in blue. (Middle) Illustration of SNP block analysis pipeline. (Right) The consensus alignment of the two α haplotypes, β, and γ is constructed. (Left) Pairwise divergences between consensus sequences are calculated. Note that α_2 is identical to γ here. (**B**) The joint distribution of divergences of α SNP block haplotypes from the other species for all SNP blocks. Each point represents a haplotype with coordinates showing the divergence from β (x-axis) and γ (y-axis). The orange and blue dots show the two haplotypes from the example in (A). Marginal distributions are shown on the top and right. Note the high density of haplotypes near the axes, indicating transfer from one of the other known species. (**C**) Same as panel (B) for β block haplotypes. Note the much larger number of haplotypes far from the axes compared to (B), indicating transfers of unknown origin. (**D**) Histogram of linkage coefficients between pairs of block haplotypes at different loci for α (orange histogram) and β (blue histogram) compared to the expectation under a fully unlinked model (solid orange and blue lines, respectively).

clonal subpopulations, every one of the $\sim$ 300 cells in our sample has a unique combination of SNP block haplotypes.

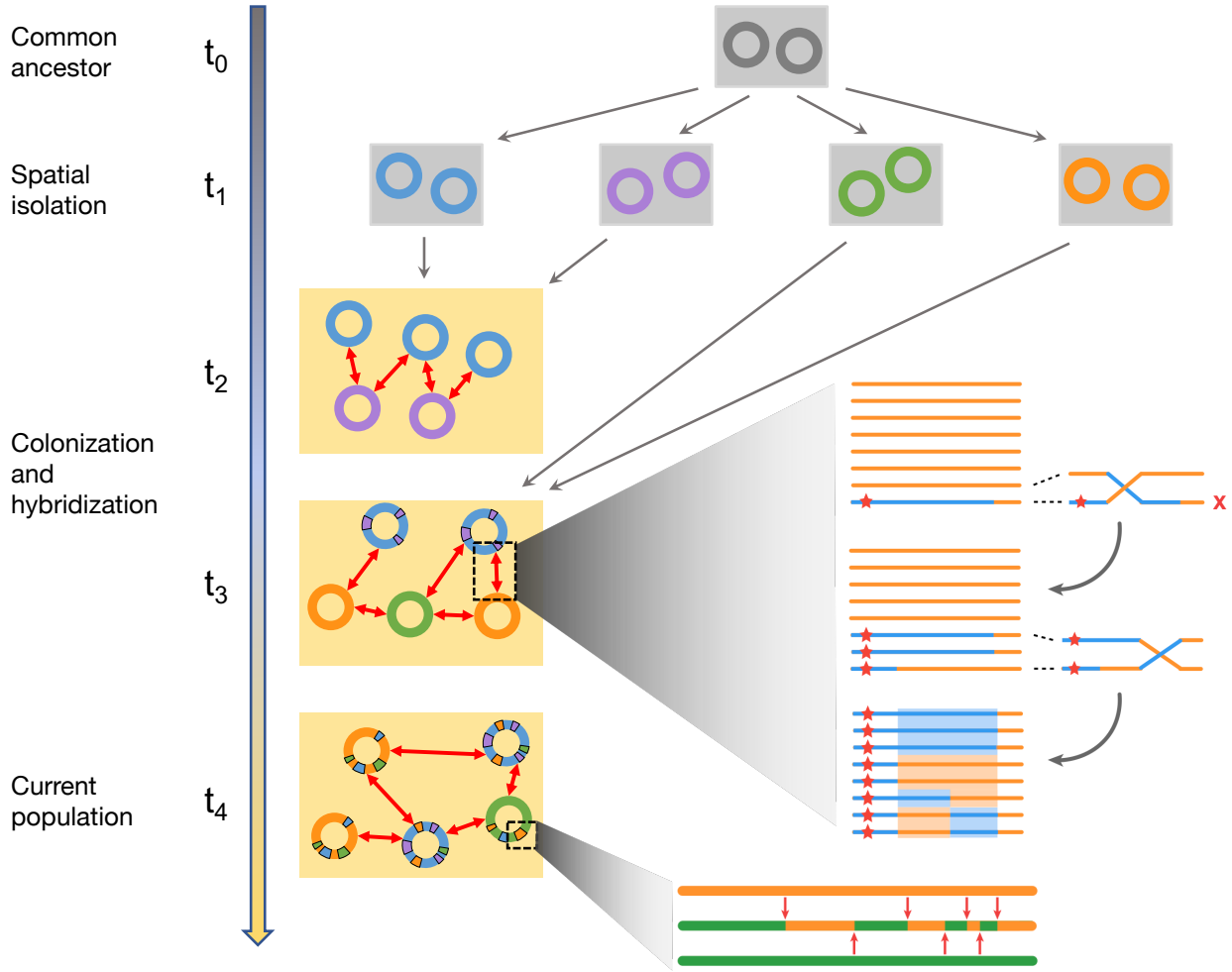

Figure 4: Hybridization between previously isolated lineages is main source of diversity within the population. (Main panel) An illustration of the evolutionary scenario consistent with the main results of this paper. A single common ancestor (gray) for the population (t_0) diverged into several species (four shown here) following spatial isolation (t_1). The species colonized the Yellowstone caldera, represented by the yellow rectangle, at different times (t_2–t_3) and began hybridizing. This process continued up to the present (t_4). (Top inset) Illustration of the mechanism for generating unlinked SNP blocks through recombination between a hybrid segment (blue) containing a beneficial region (red star) and the rest of the host species (orange). Note the four distinct haplotypes highlighted at the end. (Bottom inset) Zoom in of the ancestry of a particular genome (middle) compared with the donor (orange, top) and host (green, bottom) ancestral genomes. After the initial hybridization, recombination with other host genomes (red arrows) creates a mosaic structure across a range of genomic length scales.

By analyzing over 300 *Synechococcus* single-cell genomes, we show that hybridization plays a major role in shaping their genomic diversity. Our results suggest a simple scenario for the evolutionary history of the population (fig. 4; see also SM, Sec. VII). Multiple strains were geographically separated and evolved independently into distinct species from an ancestral population. They then colonized the Yellowstone caldera after its formation around $640,000$ years ago and have since been gradually hybridizing. The presence of genes with multiple sequence clusters suggests at least six independent colonizers, but the current population is dominated by the descendants of only three—the ancestors of α, β, and γ. The β

population has been hybridizing with others and recombining with itself long enough that its genomes are a mosaic of different ancestral segments, consistent with a quasisexual population. But α still exhibits a large fraction of low-diversity genomic segments, interlaced with clearly hybridized segments. Molecular clock estimates using the low diversity regions of the α genomes suggest it emerged around 10^{4-5} years ago and on these time scales the population is well-mixed across springs (SM, Sec. VII; fig. S26). Using the hybridized genomic segments of α as a control, we also estimated the contribution of de-novo mutations to the β diversity. This analysis suggests that the ancestor of β colonized the Yellowstone caldera soon after its formation (SM, Sec. VII).

Diversity within species is often modeled using neutral processes (*3*, *35*, *36*). But in this *Synechococcus* population, there is overwhelming evidence for the crucial role played by selection. The diversity troughs resulted from selective sweeps through the whole population that likely overcame mechanistic barriers to recombination, genetic incompatibilities, and ecological differences, evident in the apparent preferences to different temperatures of the three species. Moreover, the thousands of blocks composed of tightly linked SNPs we observe are difficult to explain through neutral drift, due to the long time it would take for them to reach observable frequencies. Instead, genetic hitchhiking during selective sweeps is by far the most likely explanation.

Selective sweeps of single genes through subpopulations have been argued to lead to the formation of separate ecotypes, recombination barriers, and eventual speciation (*12*, *13*). But we find multiple cases of genes from one species partially sweeping through another, sometimes via several independent transfer events. Thus gene sweeps do not necessarily lead to ecological specialization, but can do the opposite, acting to homogenize the population. Overall, the population shows abundant evidence of hybridization on a wide range of genomic scales leading to both homogenization between species—through the erosion of genomic clusters—and diversification within species. How extensive hybridization affects, and is affected by, ecology and how the interplay between selection at both gene and genome scales shapes the diversity of long-term coexisting populations are intriguing questions for future work.

The importance of hybridization in eukaryotic evolution is increasingly being recognized but there have been few investigations in prokaryotes (*37*, *38*). Metagenomic studies on communities in acid mine drainages revealed several hybrid bacterial and archaeal populations whose genomes were mosaics of hundreds of kbp segments from two different ancestors (*23*, *24*). Those results are consistent with very recent hybridization, after the start of mining operations (*39*). Our study shows the effects of this process over much longer time scales. At the level of individual genes, evidence for hybridization has been observed among commensal and pathogenic species of *Campylobacter* (*25*, *40*). But unlike the thermophilic *Synechococcus*, such host-associated bacteria are globally dispersed and have likely sampled many different environments, on a wide range of time scales, during their evolutionary history. It is thus difficult to identify which evolutionary forces drive their hybridization.

For the Yellowstone *Synechococcus*, a fortuitous combination of spatial separation and later mixing on the same time scales on which speciation and then hybridization can extensively shape—but not completely overwrite—their diversity, has enabled us to infer a great deal about the evolutionary history and the effects of full and partial sweeps. This provides a window into processes that shape species-level bacterial diversity. But for most bacterial populations, the effects of these processes are obscured by the broad range—and little knowledge—of the time scales on which different strains have co-occurred and are able to exchange DNA. Investigating whether hybridization plays a major role in these communities will thus require developing a predictive theory that can distinguish between possible evolutionary scenarios.

Acknowledgments

We thank Benjamin H. Good, Alana Papula, and Freddy Bunbury for helpful comments and discussions.

Funding: G.B. and D.S.F. gratefully acknowledge support from the Simons Foundation via a Postdoctoral Fellowship Award 730295 to G.B. and Sabbatical Fellowship to D.S.F. This work was also supported by NSF Grants PHY-1607606 and PHY-2210386 to D.S.F. H.S.M. was supported by the National Institutes of Health grant R01-AI-100947 to Mihai Pop. D.B. acknowledges support from the NSF/BIO-BBSRC collaborative research grant 1921429, NSF MIM grant 2125965, Joint Genome Institute Community Sequencing Project 503441 and 509352, and the Carnegie Institution for Science. Samples processed for single cell genome amplification were collected using NP Park Permits: YELL-5494 to David Ward (multi year), YELL-5660 to D.B. (2007-2008), and YELL 5694 to D.B. (2007-2009). The work (proposal: https://doi.org/10.46936/10.25585/60001132) conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

Authors contributions: G.B. designed and performed the data analysis, contributed to the design the project, and wrote the paper. H.S.M. analyzed the metagenome data. R.R.M. performed the single-cell sorting, sequencing, and assembly. D.G. performed the single-cell sorting, sequencing, and assembly. D.S.F. designed and supervised the data analysis, supervised and contributed to the design of the project, and wrote the paper. D.B. supervised and contributed to the design of the project and wrote the paper.

Competing interests: The authors declare no competing interests.

Data and materials availability: All SAG assemblies and metagenome samples used in this study can be found on the JGI website (https://genome.jgi.doe.gov/portal/) under the project ID 503441. The analyzed data and code used to produce the results are available on GitHub (https://github.com/gbirzu/yellowstone_cyanobacteria_hybridization.git).

References

1. J. A. Coyne, H. A. Orr, *Speciation* (Sinauer Associates, Sunderland, Mass., 2004), ISBN: 0-87893-091-4 978-0-87893-091-3 0-87893-089-2 978-0-87893-089-0.
2. C. Fraser, E. J. Alm, M. F. Polz, B. G. Spratt, W. P. Hanage, *Science* **323**, 741 (Feb. 2009).
3. E. P. C. Rocha, *Molecular Biology and Evolution* **35**, 1338–1347, ISSN: 0737-4038 (June 2018).
4. M. C. Rivera, R. Jain, J. E. Moore, J. A. Lake, *Proceedings of the National Academy of Sciences* **95**, 6239–6244 (May 1998).
5. X. Didelot, M. C. Maiden, *Trends in Microbiology* **18**, 315–322, ISSN: 0966-842X (July 2010).
6. K. Yahara *et al.*, *Molecular Biology and Evolution* **33**, 456–471, ISSN: 0737-4038 (Oct. 2015).
7. M. J. Rosen, M. Davison, D. Bhaya, D. S. Fisher, *Science* **348**, 1019 (May 2015).
8. N. R. Garud, B. H. Good, O. Hallatschek, K. S. Pollard, *PLOS Biology* **17**, e3000102 (Jan. 2019).
9. T. Sakoparnig, C. Field, E. van Nimwegen, *eLife* **10**, ed. by A. Nourmohammad, A. M. Walczak, e65366, ISSN: 2050-084X (Jan. 2021).
10. N. Kashtan *et al.*, *Science* **344**, 416 (Apr. 2014).
11. P. Arevalo, D. VanInsberghe, J. Elsherbini, J. Gore, M. F. Polz, *Cell* **178**, 820–834.e14, ISSN: 0092-8674 (Aug. 2019).
12. N. J. Croucher *et al.*, *Science* **331**, 430 (Jan. 2011).
13. B. J. Shapiro *et al.*, *Science* **336**, 48 (Apr. 2012).
14. Y.-J. Bao, B. J. Shapiro, S. W. Lee, V. A. Ploplis, F. J. Castellino, *Scientific Reports* **6**, 36644, ISSN: 2045-2322 (Nov. 2016).
15. C. A. Hanson, J. A. Fuhrman, M. C. Horner-Devine, J. B. H. Martiny, *Nature Reviews Microbiology* **10**, 497–506, ISSN: 1740-1534 (July 2012).
16. R. T. Papke, N. B. Ramsing, M. M. Bateson, D. M. Ward, *Environmental Microbiology* **5**, 650–659, ISSN: 1462-2912 (2003).
17. R. J. Whitaker, D. W. Grogan, J. W. Taylor, *Science* **301**, 976 (Aug. 2003).
18. M. J. Rosen, M. Davison, D. S. Fisher, D. Bhaya, *PLOS ONE* **13**, e0205396 (Nov. 2018).
19. D. M. Ward, M. J. Ferris, S. C. Nold, M. M. Bateson, *Microbiology and Molecular Biology Reviews* **62**, 1353 (Dec. 1998).
20. D. M. Ward *et al.*, *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**, 1997–2008 (Nov. 2006).
21. C. S. Smillie *et al.*, *Nature* **480**, 241–244, ISSN: 1476-4687 (Dec. 2011).
22. A. Diop, E. L. Torrance, C. M. Stott, L.-M. Bobay, *Genome Biology* **23**, 239, ISSN: 1474-760X (Nov. 2022).
23. G. W. Tyson *et al.*, *Nature* **428**, 37–43, ISSN: 1476-4687 (Mar. 2004).
24. I. Lo *et al.*, *Nature* **446**, 537–541, ISSN: 1476-4687 (Mar. 2007).
25. S. K. Sheppard, N. D. McCarthy, D. Falush, M. C. J. Maiden, *Science* **320**, 237 (Apr. 2008).

26. H. P. Browne *et al.*, *Nature* **533**, 543–546, ISSN: 1476-4687 (May 2016).

27. W. Zheng *et al.*, *Science* **376**, eabm1483, eprint: `https://www.science.org/doi/pdf/10.1126/science.abm1483` (2022).

28. C. Rinke *et al.*, *Nature* **499**, 431–437, ISSN: 1476-4687 (July 2013).

29. D. M. Ward, R. W. Castenholz, S. R. Miller, in *Ecology of Cyanobacteria II: Their Diversity in Space and Time* (Springer Dordrecht, First, 2012), pp. 39–63.

30. S. R. Miller, D. Carvey, *Microbial Ecology* **78**, 33–41, ISSN: 1432-184X (July 2019).

31. D. Bhaya *et al.*, *The ISME Journal* **1**, 703–713, ISSN: 1751-7370 (Dec. 2007).

32. M. Groussin *et al.*, *Cell* **184**, 2053–2067.e18, ISSN: 0092-8674 (Apr. 2021).

33. T. Hackl *et al.*, *Cell* **186**, 47–62.e16, ISSN: 0092-8674 (Jan. 2023).

34. P. W. Messer, D. A. Petrov, *Trends in Ecology & Evolution* **28**, 659–669, ISSN: 0169-5347 (Nov. 2013).

35. X. Didelot, D. J. Wilson, *PLOS Computational Biology* **11**, e1004041 (Feb. 2015).

36. Z. Chen *et al.*, *Nature Ecology & Evolution* **6**, 183–194, ISSN: 2397-334X (Feb. 2022).

37. S. Sankararaman *et al.*, *Nature* **507**, 354–357, ISSN: 1476-4687 (Mar. 2014).

38. B. M. Moran *et al.*, *eLife* **10**, ed. by P. J. Wittkopp, e69016, ISSN: 2050-084X (Aug. 2021).

39. V. J. Denef, J. F. Banfield, *Science* **336**, 462–466, ISSN: 00368075, 10959203 (2012).

40. E. Mourkas *et al.*, *eLife* **11**, ed. by B. S. Cooper, G. Storz, e73552, ISSN: 2050-084X (Feb. 2022).

Supplementary Materials

I Materials and methods **3**

I.1 Sample collection and sequencing . . . 3

I.2 Data processing . . . 3

I.3 Outline of data processing pipeline . . . 4

II SAG species assignment **8**

III Metagenome analysis **14**

III.1 Phylogeny of γ . . . 14

III.2 Estimating abundances of α, β, and γ . . . 16

III.3 Other *Synechococccus* species . . . 18

IV Linkage analysis **22**

IV.1 Statistical measures of linkage disequilibrium . . . 22

IV.2 Linkage decrease within vs between species . . . 23

IV.3 Comparison to theoretical predictions . . . 25

IV.4 Lack of clonal frame in α . . . 26

IV.5 Linkage on intermediate length scales . . . 27

IV.6 Linkage between multi-SNP alleles . . . 27

IV.7 Determining the main cloud cutoff for $\alpha$and β . . . 28

V Gene-level hybridization **29**

V.1 Simple hybrid genes . . . 29

V.2 Diversity within hybrid genes . . . 30

VI Linkage block detection and analysis **33**

VI.1 Processing alignments for SNP block detection . . . 33

VI.2 SNP block detection method . . . 33

VI.3 Haplotype divergences . . . 34

VII Genetic diversity analysis **37**

VII.1 Nucleotide diversity along the core genome . . . 37

VII.2 Diversity heterogeneity along core genome . . . 38

VII.3 Contribution of hybridization to genetic diversity . . . 39

VII.4 Geological constraints on microbial evolution 41
VII.5 Upper bound on mixing time between springs 41
VII.6 Proposed scenario for Synechococcus evolution 42

VIII Variation in hybridization across samples **43**
VIII.1 Variation in SNP block haplotype frequencies 43
VIII.2 Variation in hybrid gene transfers 44

IX Data quality control **45**
References 49

I. Materials and methods

I.1 Sample collection and sequencing

Cork borers (approximate diameter 0.5 to 1 mm) were used to collect mat core samples from Octopus Spring (44.53401N, 110.79781W) and Mushroom Spring (44.53861N, 110.79791W). Samples were collected in 2005, 2006, 2007 and 2009 under YNP Park Permits (YELL-5494, YELL-5660, YELL 5694). Cores were immediately frozen until later use. Frozen samples were shipped to the Joint Genome Institute (JGI) in 2019, where they were processed and sequenced.

At JGI, for disaggregation of sample, frozen mat cores were cut into pieces using a razor blade on glass over dry ice and aliquoted chunks of approximately 5 mm diameter. Next, 500 µl PBS and 50 µl of 500 mM NaEDTA were added. Samples were vortexed briefly, sonicated for 1 minute on benchtop sonicator, vortexed briefly again. Passed 25 times through a 25 gauge needle and run through a 35 µm strainer cap. Next, 100 µl were placed in 900 µl PBS and stain with SYBR Green at 1X. Samples were run on the Influx with Gate SYBR+ Cells and Gate 670/30 off of 632 autofluoresence. Next samples were sorted on a BD influx following the protocol outlined in Ref. (*1*), and single-cells genomes were amplified with WGAX protocol as described in Ref. (*2*). Certain samples from plate YuBhay.Oct006.RedA.3 received an additional treatment with lysozyme (20 min at room temp in 50 U/µl prior to alkaline lysis) because of difficulty lysing the cells. Otherwise, they followed the standard alkaline lysis described in Ref. (*2*). In brief, libraries were made with Nextera XT v2 kits and sequenced on an Illumina Novseq in 2×150 bp mode. Reads were assembled into contigs and annotated using the Integrated Microbial Genomes (IMG) platform as described in Ref. (*3*). All of the data generated for this study is available on the JGI website, under the project ID 503441.

I.2 Data processing

In this section we describe the procedure we used to process the single-amplified genomes (SAGs). Our analysis is based on grouping genes into orthologous gene clusters or *orthogroups*, which are then analyzed separately. This procedure makes it much easier to analyze the fragmented and incomplete genomes resulting from single-cell sequencing compared multiple sequence alignment (MSA) of entire genomes, which are sometimes used for analyzing genomes obtained from isolates. An alternative approach is to align the assemblies or even the raw reads to reference genomes. In our case, there are only two reference genomes available, OS-A and OS-B', so aligning to them would likely result in biases in the inferred diversity.

There are several ways of constructing orthogroups: 1. pairwise comparisons, or 2. graph methods. Graph methods have the advantage that they produce sets of mutually homologous genes so we decided to use this approach. There are several packages available for identifying homologous gene clusters among samples of different genomes. Most widely-used packages, such as OrthoDB (*4*) and OrthoFinder (*5*), were designed for finding homologs between different eukaryotic species and are inappropriate for studying large bacterial pangenomes (*6*). More recently, several groups have developed packages that allow for efficient construction of large bacterial pangenomes (*6*, *7*). However, such packages are primarily aimed at studying large collection of isolate genomes and were not designed to handle samples with low genome coverage, which are common in single-cell sequencing projects such as ours. As a result we implemented a custom pipeline for processing the single-cell genome assemblies and constructing the pangenome from these sequences. The pipeline is divided into two parts: 1. a pre-processsing step where assemblies are filtered for contamination and checked for misassemblies and misannotations, and 2. a gene clustering step, in which clusters of homologous genes are identified and paralogs are separated from true orthologs.

The second half of the pipeline closely follows the approach taken in Ref. (*6*). A more detailed outline of the pipeline is given below. For details on our definition of the core genome and the assignment of sequence clusters within orthogroups to different species see Sec. V.

I.3 Outline of data processing pipeline

1. Data pre-processing
2. Orthogroup clustering
3. Orthogroup MSA
4. Deep branch splitting
5. Orthogroup species clustering
6. Reference genome mapping

The rest of this section explains how each step in the pipeline was performed. Readers not interested in the details may wish to move on to the next section.

1. Data pre-processing. The goal of this step is to remove 1. SAGs from non-Synechococcus OS species, 2. SAGs that might have resulted from sequencing multiple cell in the same reaction well, and 3. contigs that might be the result of contamination from external DNA or other sequencing wells. We first removed all SAGs that were positive controls in which multiple cells were amplified in the same reaction well (all SAGs from the third column of the 384-well plate in our case). We then aligned all of the genes in our data to the OS-A and OS-B' reference genomes using BLAST (alignment parameters: `-evalue 1E-10 -word_size 8 -qcov_hsp_perc 75.0`). We then removed all SAGs with $> 50\%$ of genes that did not map to either reference genome.The remaining SAGs are assumed to be from Synechococcus OS cells. Next, we filtered out all contigs without any hits to OS-A and OS-B'.This was done to avoid external DNA or cross-contamination from affecting the rate of hybridization we observed.

Next, we removed any remaining SAGs that were suspected to have been the result of sequencing multiple cells in the same well. We reasoned that sequencing multiple cells would lead to an increase in the number of duplicated genes in the SAG. We used the probability of sampling a gene with more than one copy in the two reference genomes, both $\sim 10\%$, as a benchmark and removed all SAGs with a multicopy frequency $> 20\%$.

Preliminary analysis of the data following the above filtering procedure revealed the existence of a small number of palindromic contigs, in which the first half of the sequence was mirrored exactly in the second half, that were very likely the result of assembly artifacts. To remove these artifacts we labeled the genes in each contig based on the hits to the reference genomes and used these to search for palindromic sequences within each contigs. We then removed any contig that contained a palindromic subsequence that was either longer than 4 genes or contained $> 50\%$ of genes in the contig. Finally, we removed all partial genes from the remaining contigs to simplify the analysis of the multiple-sequence alignments (MSA), which we perfomed next.

2. Orthogroup clustering. The goal of this step is to obtain clusters of orthologous genes or *orthogroups*, which were used in subsequent analysis. To do this we followed Refs. (*6*, *7*) and used a Markov clustering algorithm to find groups of sequences that were similar to each other and well-separated from all other

sequences (*8*). The Markov clustering algorithm takes as input a similarity graph, where each node represents a gene sequence and the weight of each edge represents the similarity between the two sequences.

We constructed the gene similarity graph by translating all of the gene sequences in the filtered dataset and performing an all-to-all protein alignment using BLAST (alignment parameters: `-evalue 1E-3 -word_size 3 -qcov_hsp_perc 75.0`). Other approaches either perform a pre-alignment clustering (*7*) or use faster but less sensitive methods for all-to-all alignments (*6*), but we found that this was not necessary in our case. The higher sensitivity provided by BLAST could help distinguish between paralogs within large gene families and minimize gene misassignments, but we did not extensively test how the performance would be affected by using other methods. We used the bitscore from the alignments to define the weight of each edge in the similarity graph. Using the bitscore instead of the e-value is a simple way to correct for the minimum e-value of $1E-180$ in BLAST and has been shown previously to improve the clustering performance (*9*).

We used MCL (v. 14-137) to cluster the resulting similarity graph and obtain orthogroups (*8*). We used a relatively small inflation parameter (`-I 1.5`) in order to cluster genes aggressively and used subsequent steps in the pipeline to split clusters resulting from incorrect cluster assignments or the presence of paralogs (steps 5 and 6). As a first step, we did an additional clustering based on sequence length as follows. We constructed a graph in which all sequences in an orthogroup were represented as nodes and edges where attached between pairs of sequences if the difference between their lengths was less than 30% of their average length. We then identified all connected components in this graph. Because homologous genes in OS-A and OS-B' sometimes have different lengths, we hypothesized that each orthgroup should contain at most two main clusters. We therefore used the following heuristic algorithm to filter out orthogroups with unusual gene length patterns that could possibly result from overly aggressive clustering. If a single length cluster was present or the largest cluster contained $> 80\%$ of sequences, only the sequences from the largest cluster were kept and the rest were removed. If two clusters accounted for $> 80\%$ of sequences, only the sequences in those two clusters were kept and the rest removed. In all other cases we removed the orthogroup from our dataset and did not include it in subsequent analysis.

3. Orthogroup MSA. The filtered orthogroups from the previous step were aligned using a custom codon-aware alignment algorithm. Briefly, nucleotide sequences were translated into protein sequences and aligned using MAFFT (v. 7.475 with default parameters) (*10*). The alignments were then reverse-translated using nucleotide sequences as templates. After alignment, phylogenetic trees for each orthogroup were constructed using FastTree (v. 2.1.11 with parameters `-nj -noml`) (*11*).

4. Deep branch splitting. In this step we separated gene clusters within the initial orthogroups that were joined by unsually long branches in the phylogenetic tree. Specifically, we implemented the following algorithm. For each orthogroup, we checked the branch length for each node in the phylogenetic tree and grouped any sequences with a branch length greater than a cutoff value $b_c = 0.3$ into separate orthogroups. Each newly-created orthogroup was realigned as in step 3 and the procedure was repeated until no new orthogroups could be separated. We used the results from previous analysis of homologous genes in OS-A and OS-B', which showed that typical divergences were ~ 0.15 (*12*), as a benchmark to set b_c. Based on those results, we expected very few genuine homologs with divergences above $b_c = 0.3$. The clusters that resulted after this step represent the final orthogroups that were used in subsequent analysis.

5. Orthogroup species clustering. We next sought to determine which orthogroups could be reliably partitioned into distinct species clusters, which would be expected if the different species asexually diverged from a common ancestor. To do this, we implemented the following algorithm in a custom

Python script. For each orthogroup, the alignments were trimmed to remove excess gaps at the edges using a custom script. We then calculated the fraction of sites with different nucleotide between each pair of sequences and hierarchically clustered the sequences using average linkage as implemented in the Python Scipy package (v. 1.9.3) (*13*). Preliminary clusters were assigned using a distance cutoff $d_c = 0.075$. To ensure that the clusters were well-separated and robust we implemented an additional check by requiring that the minimum distance between any two sequences from different clusters be at least $c = 1.5$ times greater than the maximum distance between any two sequences within either cluster. Any clusters which did not satisfy this criterion were designated as *mosaic orthogroups*. We validated the robustness of the cluster assignments by d_c and c and found only small changes in the results.

6. Mapping to reference. For this step we use reciprocal best hit (RBH) given by BLAST to map the consensus sequences to the OS-A and OS-B' reference CDS. Since at this point in the pipeline most sequences within one OG are less that 10% diverged from each other, we use nucleotide sequences for this step to increase the accuracy of the mapping. The consensus sequence for each OG is constructed by finding at the consensus codon at each residue along the CDS after replacing all codons containing gaps by a gap codon. The resulting sequences that have RBH to OS-B' are mapped onto those respective genes. The genes without RBH to OS-B' are searched against OS-A and mapped to it if an RBH is found.

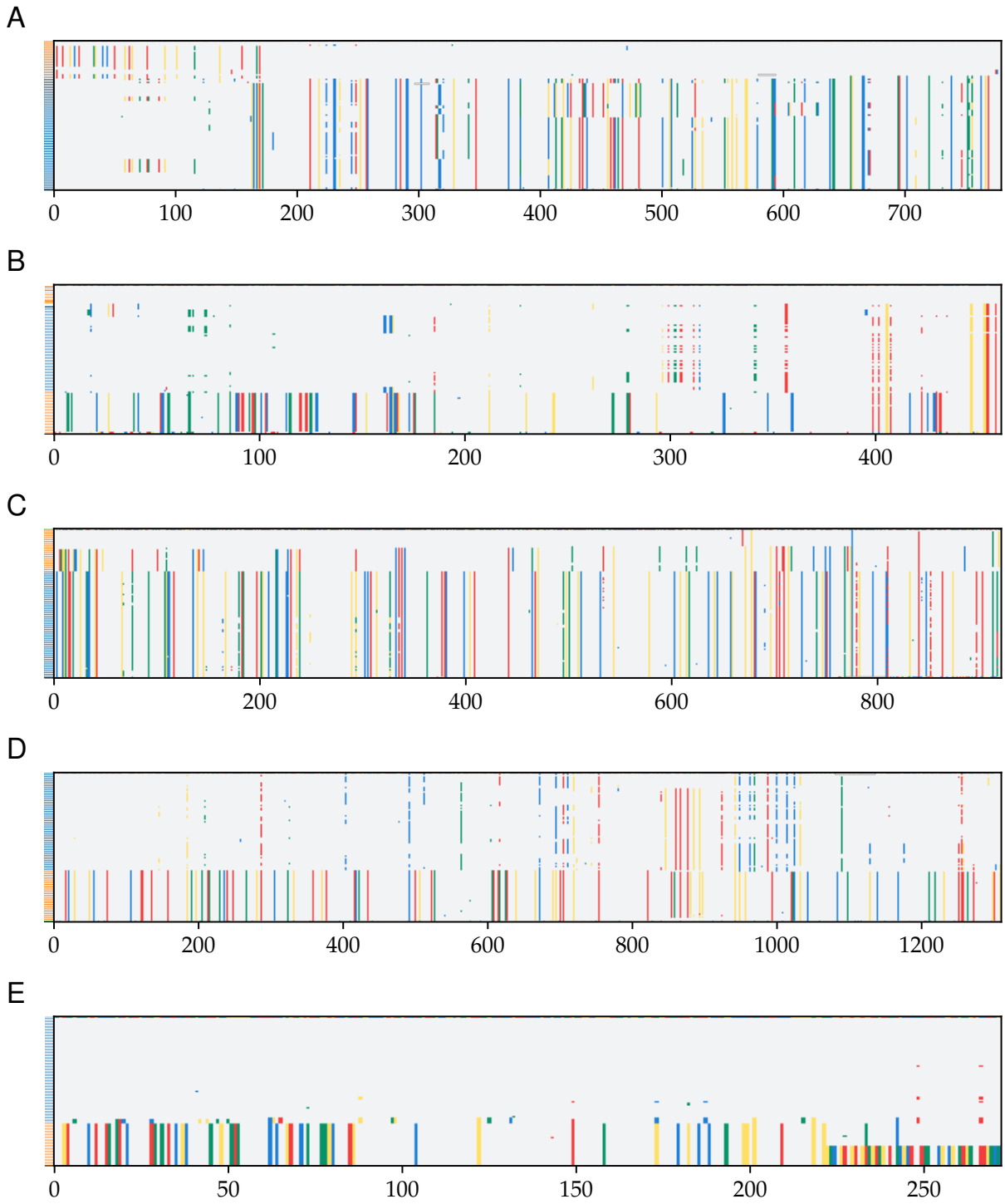

Figure S1: Examples of alignments of orthogroups with mixed-species clusters. A random sample of 5 orthogroup alignments labeled as mixed-species clusters are shown. All alignments were trimmed to remove segments containing large fraction of gaps. Panels **A**-**D** (YSG_1519, YSG_0498, YSG_0215c, and YSG_1364) show varying degrees of hybridization between α and β. Panel **E** (YSG_0291) is the result of a poor initial alignment of the orthogroup leading to a large number of indels before trimming.

II. SAG species assignment

In this section, we outline the method we used to assign genomes to each species cluster. We also verify the robustness of the species assignment by using three different methods and show that they give identical results.

Comparisons of genomes from bacterial populations often reveal clusters at different length and divergence scales (*14*). But the interpretation of such clusters is controversial (*15*, *16*). There three main possible interpretations for such patterns. One interpretation is that the clusters represent the bacterial analog of eukaryotic species, which are sexually and ecologically isolated from each other. In practice a whole-genome divergence cutoff of 5% is frequently used to define bacterial species (*14*) An alternative interpretation is that clusters are the result of clonal expansions of particular strains from a much more diverse and possibly more genetically homogeneous population. The third interpretation is that the clusters are the result of uneven sampling from a larger and more diverse population. The unique combination of having a large collection of genomes, with minimal compositional bias, from a geographically isolated population allows us to test these alternatives. We will mainly aim to distinguish between the first two interpretations and the third one by checking whether there are robust genomic clusters in our sample. More discussion on the evidence concerning the first two hypotheses can be found in Secs. III and V.

We used three different methods to check whether the genomes form robust clusters: the phylogenetic tree of the 16S rRNA sequences, average genome-wide divergence from the two reference genomes OS-A and OS-B', and patters of divergences between triplets formed by the SAG and the two reference sequences at different loci. The phylogenetic tree of the 16S is the simplest and still the most commonly used method to characterize the diversity of natural microbial communities (*17*). Because of its practical importance, we included it for comparison with the other two methods. However, in a highly recombining population like the Yellowstone Synechococcus, it is not clear to what extent the 16S divergence accurately reflects the whole genome diversity. To address this question, we used the divergences across the entire genome as a comparison. The main limitation of using summary statistics such as the average divergence across the entire genome is that it obscures heterogeneities along the genome that could have important ecological and evolutionary implications. There are different ways to address this concern. Here, we used an extension of a method first used in Ref. (*12*), which provides a coarse description of the variation in the diversity patterns across different loci. We present the results from each of these analyses next.

We first constructed the phylogenetic tree of 16S sequences. Predicted 16S rRNA sequences (117 from 331 filtered SAGs) were extracted and aligned using MAFFT (v7.453) (*10*). Two sequences (SAGs MA02M11_3 and MuA02C9) were incomplete and were removed from the analysis. The remaining (115) sequences were hierarchicaly clustered using the average linkage (UPGMA) method implemented in Scipy. Results are shown in Fig. S4, with the colored background showing the species assignment using whole genome divergences (see below) for comparison. We found perfect agreement between the whole-genome assignment and the largest three clusters in the 16S phylogenetic tree.

The average divergence within the α cluster approximately half that of β ($1 \cdot 10^{-3}$ vs. $2 \cdot 10^{-3}$). However, a significant fraction of the average divergence within β was due to two outliers (MuA1L16 and MusA1J8). After removing the outliers, the average divergence within β decreased to $1.7 \cdot 10^{-3}$. Closer investigation of the alignments revelead that both outlier β sequences contained a short segment of 3 SNPs that were identical to all α sequences and different from all other β sequences, at positions 1067–1072. In addition, MuA1L16 contained larger segment of almost 200 bp (positions 390–570) containing 10 SNPs and an

insertion that were different from all other β and were identical to the consensus α. Apart from these segments, the outlier sequences were typical of the β species. This pattern is similar to the SNP blocks discussed in the main text and in Sec. VI and therefore is likely the result of hybridization between α and β.

We compared the 16S rRNA clusters to whole-genome divergences from the two reference genomes, OS-A and OS-B'. We aligned all of the predicted genes from each SAG to all of the coding DNA sequences (CDS) from each reference genome using BLAST (v2.13.0+) (*18*). For each SAG , we calculated the divergence to each reference genome ($d = 1 - \text{pident}/100$) and calculated the average over the best hits across all genes. The results for all SAGs with more than 100 hits to both references are shown in Fig. 2A. Three clear clusters were observed, which we labeled as α, β, and γ. The species of each SAG was assigned as the label correspoding to the cluster it belonged to in Fig. 2A.

The previous two metrics do not reflect possible heterogeneity in the patterns of diversity along the genome. To characterize this variation on a gene-by-gene basis we extended a method that was previously used to analyze amplicond data from the Mushroom Spring in Ref. (*12*). We refer to this method as *gene triplet analysis*. This analysis exploits the separation of scales between and within species clusters, at most loci and in most cells, we observed previously in order to coarse grain the high-dimensional matrix of pairwise divergences into a small number of discrete patterns. Specifically, we take triplets of sequences from the SAGs , OS-A, OS-B', at loci where all three are homologous and calculate the divergences between all 3 pairs of sequences. These divergences typically form a triangle as shown in Fig. S2 [1]. To reduce the dimensionality of the data further, we define a main cloud diameter $d_M = 0.1$ for the diversity within species. We then classify the shape of the triangle into the 5 distinct patterns shown in Fig. S2, depending whether each side of the triangle is longer or shorter than d_M. For example, denote the divergence triplet by (d_{XA}, d_{XB}, d_{AB}), representing the divergences between the SAG sequence and OS-A, the SAG sequence and OS-B', and OS-A and OS-B', respectively. Then the first pattern in S2A (labeled "XA-B") contains loci where $d_{XA} < d_M$, $d_{XB} \geq d_M$, and $d_{AB} < d_M$, the second pattern (labeled "XAB") contains loci where $d_{XA} < d_M$, $d_{XB} < d_M$, and $d_{AB} < d_M$, and so on. The cutoff of $d_M = 0.1$ is the same as the one used in Ref. (*12*) and provides a good separation between species at a gene by gene basis. The results of our analysis, however, do not depend on the specific value of d_M as discussed at the end.

For each SAG we assigned all genes that were orthologous to OS-A and OS-B' to one of the 5 patterns described above. The three rows in Fig. S2 show typical examples of an α (Fig. S2A), a β (Fig. S2B), and the γ (Fig. S2C) SAG . The majority of genes across each row in Fig. S2 are found in the XA-B, XB-A, and X-A-B patterns, respectively. This is consistent with the species clusters we saw in the genome-wide average divergences from before. However, we also find a significant minority of genes forming other patterns, most notably XAB, representing genes where all three sequences are within the main cloud ($d < d_M$). Some of these genes were contained in the diversity troughs discussed in the main text and in Sec.V. In the case of the γ SAG , the fact that X-A-B is the dominant pattern rules out the possibility that γ is a hybrid resulting from the mixture of α and β. Such hybrids have been observed after recent clonal expansions in acid mine drainage microbial communities (*19*, *20*), but do not appear to be present at substantial frequency in the Yellowstone community. Instead, it appears that γ is a distinct species that

[1] Note that it is possible to have triplets of divergences that violate the triangle inequality, which cannot be represented as triangles in a two-dimensional space. In our data, such cases were rare but this may not be generally true, particularly for sequences that are not clearly separated and which are the result of extensive recombination. Note however that the classification into distinct patterns given below can still be done even in these cases.

diverged from the other two around the same time α and β diverged from each other.

Using the pattern described above allows us to compare different SAGs to each other and allows us to test the robustness of the species assignment based on the average genome-wide divergence. Specifically, we assigned to each SAG a gene triple "fingerprint" $f = (f_{XA-B}, f_{XAB}, f_{X-A-B}, f_{X-AB}, f_{XB-A})$, representing the fraction of genes with each pattern [2] We then compared the fingerprints of each SAG with the species assignment based on the average divergence from OS-A and OS-B' from Fig. 1A. We found that the fingerprints formed three clear patterns that exactly matched the assignments based on the average divergence (Fig. S3). We tested this assignment for different values of d_M ranging from 1% to 20% and found that the fingerprint patterns were robust across the entire range (data not shown). These results show that the classification into distinct species is robust and consistent across the three methods described.

[2]Note that $\sum_i f_i = 1$.

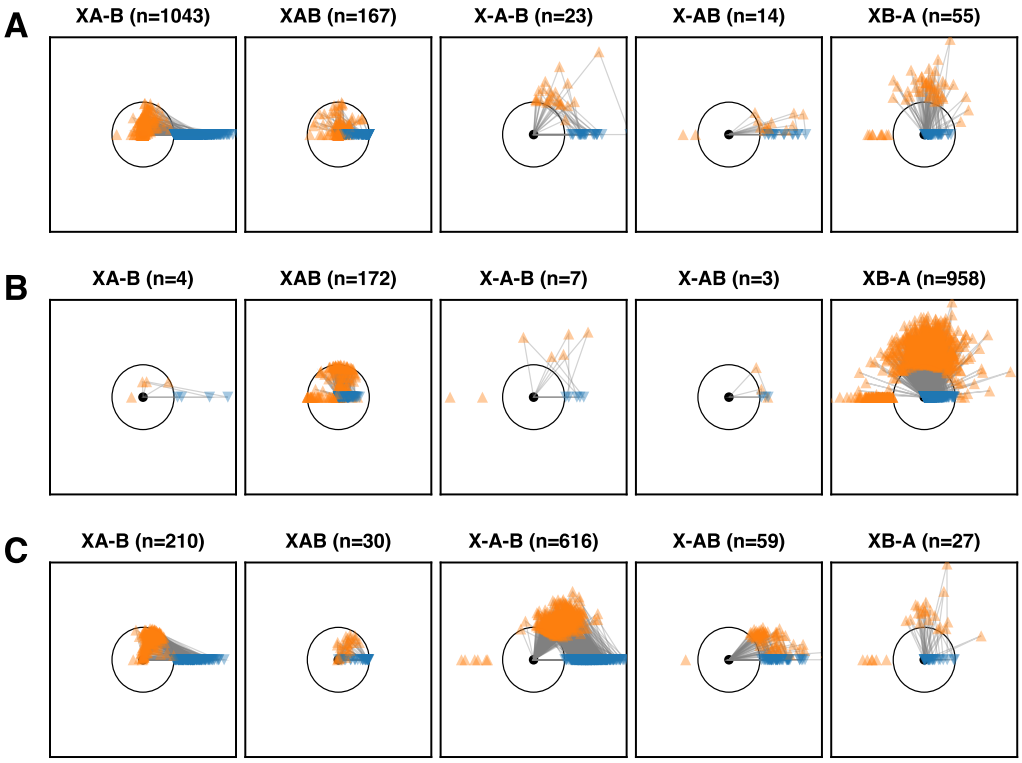

Figure S2: Species clusters at the gene level approximately match those at the genome level. Figures show gene-by-gene comparison of three cells to the OS-A and OS-B' reference genomes. The cells shown are representative of the α (**A**; OctA1J9)and β (**B**; MRedA02H14_2) clusters and the one γ cell (**C**; OcA3L13). Each core gene with homologs in both OS-A and OS-B' was compared to the references and the divergences plotted as a triangle with each side length equal to the divergence between a given pair. The cell sequence is represented by a black dot, with the OS-A and OS-B' sequences represented by the orange upward and blue downward triangles, respectively. The black circle is a 10% radius which was chosen as a cutoff to classify each gene. The different columns show genes in one of the five possible patterns labeled as follows: XA-B (the cell sequence is less than 10% divergend from OS-A and more than 10% diverged from OS-B'), XAB (all three sequences are within 10% of each other), X-A-B (all three sequences are more than 10% diverged from each other), X-AB (the cell is more than 10% diverged from both OS-A and OS-B', but both references are less than 10% diverged from each other), and XB-A (the cell is less than 10% diverged from OS-B' and more than 10% diverged from OS-A).

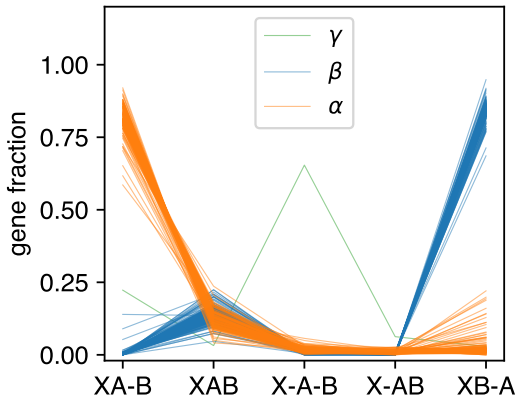

Figure S3: Proportion of gene triplet patterns shows three distinct clusters. The proportion of genes in each of the five patterns shown in S2 is show of each cell passing our quality control criteria (see SI, Sec. I.2). Lines represent different cells and are colored according to the species classification based on whole-genome divergences.

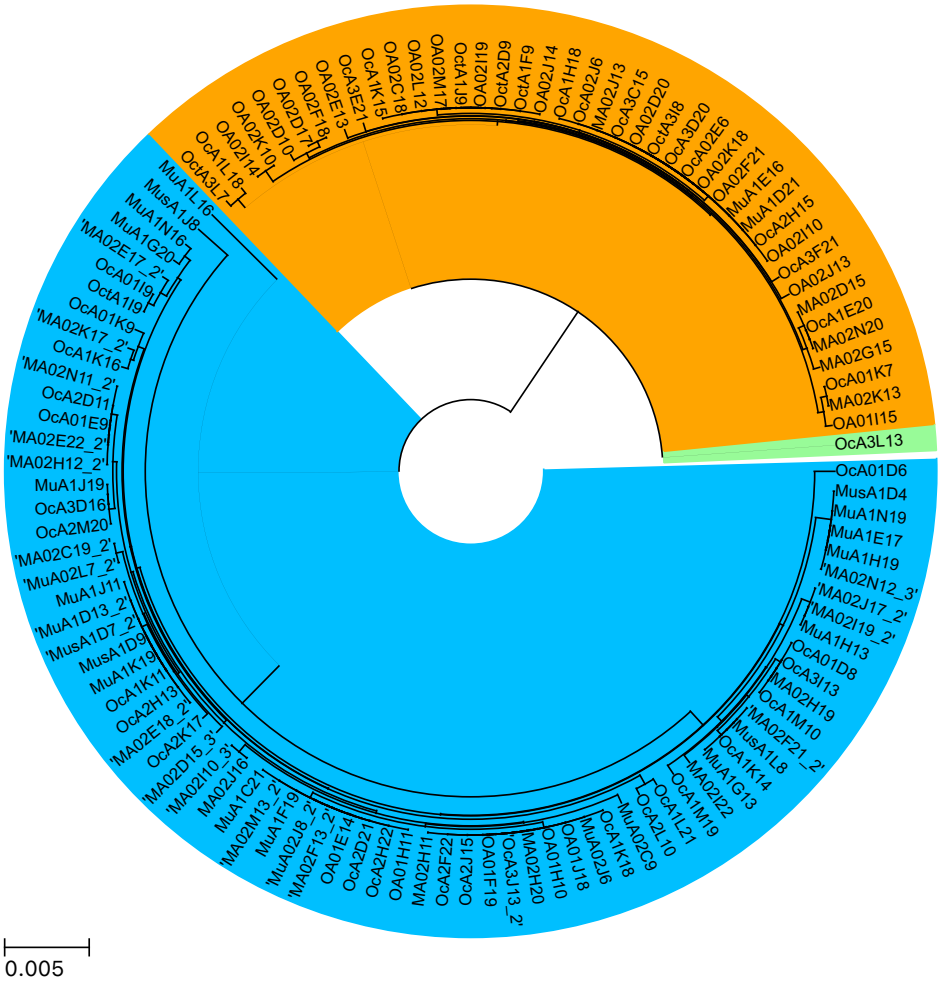

Figure S4: Phylogenetic tree inferred from 16S rRNA agrees with whole-genome species classification. The pairwise divergences between all 16S sequences passing quality control was calculated from the multiple sequence alignment performed using MAFFT (*10*). The tree was generated using a hierarchical agglomerative clustering algorithm with average linkage (UPGMA) as implemented in SciPy (*13*)

III. Metagenome analysis

III.1 Phylogeny of γ

Here we present the details of the phylogenetic analysis of the γ SAG, OA3L13 which shows that it represents a new uncharacterized species within the Yellowstone cyanobacteria population. To establish its phylogeny we used a concatenated alignment of marker genes, following Ref. (*21*). Of the 31 marker genes used in Ref. (*21*), 16 were covered in the γ SAG (*dnaG*, *frr*, *pgk*, *rplA*, *rplK*, *rplL*, *rplM*, *rplS*, *rplT*, *rpsB*, *rpsI*, *rpsJ*, *rpsK*, *rpsM*, *smpB*, and *tsf*) and were used to determine its phylogeny. We used a large collection of cyanobacterial genomes from Ref. (*22*) to check for closely related species to γ. The concatenated alignment was constructed as follows. For each of the 16 marker genes, we extracted the alignment of all species labeled "Thermal springs" that were within one of the subsections labeled I through V in (*22*) and added the γ sequence to the alignment. All "?" characters were replaced with "X". The sequences were then realigned using MAFFT (v7.453) (*10*) using default parameters and the alignments were concatenated and trimmed using trimal (v1.4.rev22) (*23*) with default parameters. Finally, the resulting sequences were realigned using MAFFT and the phylogenetic trees were inferred using FastTree (v2.1.11) (*11*) with default parameters. The resulting tree is shown in Fig. S5.

To verify that the lack of other closely-related genomes to γ was not due to the incompleteness of our reference database, we searched for homologs to the 16 γ marker genes in the NCBI database. We used the Web version of blast on NCBI with the megablast settings and word size 16. For all 16 marker genes the top two hits were from OS-A and OS-B', while all other high-scoring hits had less than 85% nucleotide identity. We also matching sequences to the 16S rRNA using the SILVA database (v138.1) (*24*). Using the non-redundant version of the database (SSU Ref NR99) only gave hits to OS-A and OS-B' sequences. Using the full database we did find several sequences that were less than 10 SNPs away from the γ sequence across the entire 16S rRNA ($> 99.7\%$ identity), including one from a previous study of Yellowstone microbiomes (*25*). The corresponding divergence of $3 \cdot 10^{-3}$ is comparable to the typical diversity within species in our SAGs of $1 - 2 \cdot 10^{-3}$ and much less than the typical divergence between species, which is approximately $2 - 4 \cdot 10^{-2}$ (Sec. II).

SILVA ID	matches/hit length	identity	notes
AFSR01000050	1424/1454	0.978	metagenome samples from MS and OS (Klatt et al., 2011)
AY884052	1424/1454	0.978	lab-grown OS-A from Allewalt et al., 2006
HM448361	1382/1386	0.997	Ref. (*25*)
HM448360	1381/1386	0.997	Ref. (*25*)
KU382141	1381/1389	0.994	
CP000239	1295/1324	0.978	Synechococcus OS-A
CP000240	1261/1324	0.952	Synechococcus OS-B'

Table S1: Best hits to OA3L13 16S rRNA in SILVA SSU Ref database.

Inferred abundances of γ from metagenomic samples are consistent with a single γ SAG in our data (see Sec. III). To further verify that no other γ SAGs were accidentally excluded during quality control (see Sec. I.2) we performed a BLAST search (alignment parameters: `-evalue 1E-10 -word_size 9 -qcov_hsp_perc 75.0`) of the γ genes against the contigs from the SAGs that were filtered out. The results were consistent with the majority of the excluded SAGs containing mostly α and β genes, with a smaller number of SAGs with divergences around twice as large (Fig. S6). This confirms that no other γ SAGs were present in our data.

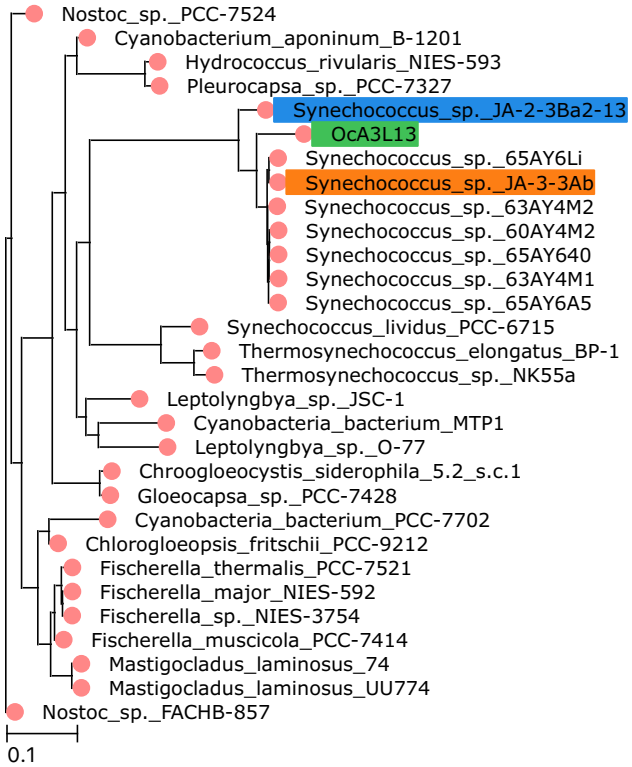

Figure S5: γ SAG represents a new distinct species within Yellowstone Synechococcus. Shows the maximum likelyhood tree obtained from an alignment of 16 marker genes present in the γ SAG together with all other thermophilic cyanobacteria from Ref. (*22*). Colored labels show the two Yellowstone Synechococcus, OS-A (orange) and OS-B' (blue), and the γ SAG OA3L13 (green).

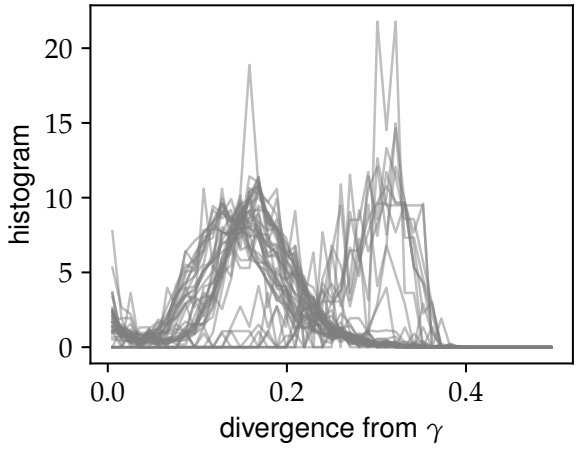

Figure S6: SAGs excluded during quality control do not include additional γ genomes. The distribution of divergences from the best BLAST hit to the γ SAG (OcA3L13) across genes for each SAG excluded during the quality control process (see Sec. I.2) is shown. Each line represents the probability density of divergences across genes for one of the 44 SAGs.

III.2 Estimating abundances of α, β, and γ

To estimate abundances of the three main species α, β, and γ in the Octopus and Mushroom Springs we recruited metagenomic reads from 34 samples taken from both springs (see Sec. I.1) to SAGs from each species. We chose the one γ SAG (OA3L13, $\approx 70\%$ coverage) along with representative SAGs from α (MA02D15, $\approx 77\%$ coverage) and β (OcA2H14, $\approx 96\%$ coverage) as reference genomes. Hybridization on length scales longer than the typical read size (150 bp in our case) can lead to errors in our estimates of species abundances. To minimize these errors we only used core genes where three distinct species clusters could be identified and there was no evidence of hybridization in the single-cell data (see Secs. I.2 and V). This resulted in set of ≈ 400 genes for each reference. We aligned all the metagenomic reads to these target gene sequences using BowTie2 (v2.3.4, default parameters) (*26*) and extracted reads aligning to each of the three reference sequences using Samtools (v1.7.0) (*27*) and bedtools bamtofastq command (v2.26.0) (*28*).

We assigned the recruited reads to each species clusters if their divergence from their respective reference was less than the "main cloud" radius d_M. We determined d_M self-consistently from the data as follows. For each reference, counted the number of reads that had divergences less than d_M from that reference sequence and more than d_M from the other two reference sequences, for d_M from 1% to 15%. If the population consists of species clusters with typical diversity π, the number of reads within the cloud will increase until $d_M \approx \pi$. As d_M approaches the typical divergence between species $d \approx 15\%$, more reads will start to map to multiple reference sequences so the number of reads uniquely assigned to each species cluster will decrease. Therefore, the number of reads uniquely assigned to each species cluster should have a maximum as a function of d_M. We indeed found this to be the case for all three species (data not shown). Empirically, we found that $d_M = 5\%$ was close to the maximum for all three species so we used this value throughout our analysis.

We found that the average depth of recruited reads across genes varied within the same sample. However the variations across samples were highly correlated, which confirmed that most reads aligning to our chosen genes came from genomes that were part of the same species cluster rather than from genes transferred across species (Fig. S7). We found that the genes that recruited the largest numbers of reads had reads depths that were much more similar within the same sample compared to other genes. Interestingly, we found much more variation in depth between genes within the γ cluster compared to α and β (Fig. S7). One likely explanation for this is that γ contains much more diversity compared to either α or β.

We estimated the abundance of each species cluster by averaging the depth of the 100 genes with the highest average depth across all samples for each cluster. We found that γ was more than an order of magnitude more abundant than our detection threshold given by the metagenomic read depth in 32 out of the 34 metagenome samples (Fig. S8). Moreover, its abundance in the two samples for which we had single-cell and metagenome data (MSe4 and OSM3) was close to the detection threshold of the single-cell samples, which was around 2%. This explains why we only obtained a single γ cell from the single-cell samples after our quality controls. Finally, we found that the abundance of γ was highly correlated with α across all samples and was only present at abundances higher than $O(1)\%$ in the Mushroom Spring sample taken from $65\,^{\circ}$C. This suggests that γ is, together with α, primarily adapted to growth at higher temperatures.

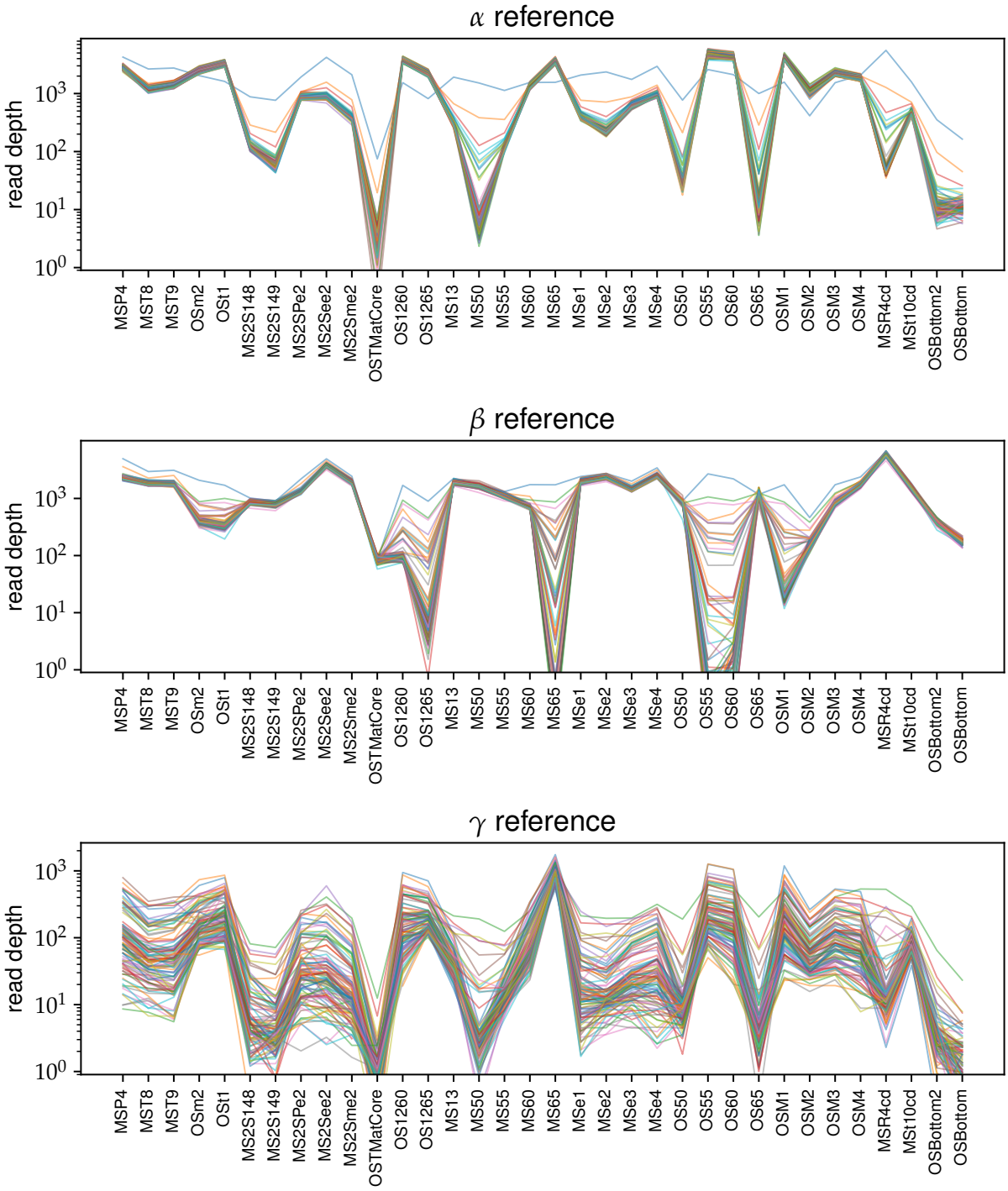

Figure S7: Number of reads recruited to different loci from the three main species are highly correlated across samples. The three panels show the average depth of reads recruited to representative SAGs from α (top), β (middle), and γ (bottom), respectively. Each panel shows 100 genes with the highest average read depth across samples, with each line representing one gene. Only reads that were $< 5\%$ diverged from the reference genomes were included.

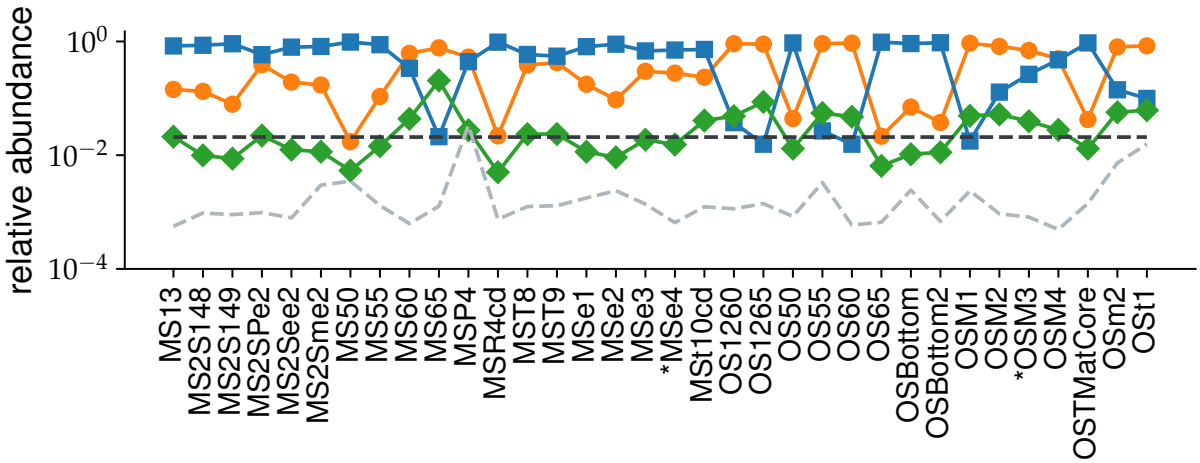

Figure S8: Abundance of γ is significantly above detection threshold in 32 of 34 metagenomic samples and is highly correlated with α. The relative abundances of α (orange circles), β (blue squares), and γ (green diamonds) across all 34 metageomic samples. The detection threshold for the single-cell ($f_t = 1/48 \approx 0.021$) and metagenomic samples are shown in black and grey dashed lines, respectively. The asterisks on the x-axis label the two samples (MSe4 and OSM3) with both single-cell and metagenomic data.

III.3 Other Synechococcus species

Are other *Synechococcus* species present in Mushroom or Octopus Spring apart from α, β, and γ? To answer this question, we again used the 34 samples from which we have high-depth metagenomes. We used BLAST with highly sensitive parameter choices to minimize the number of true *Synechococcus* reads not captured. For computational efficiency, we focused on the 100 genes with highest coverage depth in the metagenomes across all three references.

To determine how much of the population is in the α-β-γ main cloud, we used the same three SAGs as in the previous subsection as references and recruited reads using BLAST (`-task blastn -evalue 1e-3 -word_size 6 -num_threads 32`). Using these parameters, we could recruit reads belonging to the dominant *Roseiflexus* species in our samples, which is in a different phylum from *Synechococcus*. We were therefore confident that any reads from other potential *Synechococcus* species would be included in our analysis. To limit the analysis to just *Synechococcus* species, we removed all reads that were more than 20% diverged from all of our references from subsequent analysis.

Because of the large amount of data in this analysis, we focus on just the Mushroom Spring samples taken from different temperatures (MS temperature series). We recruited a total of $4,522,833$ reads with an average depth per gene of 4980. We calculated the minimum divergence from the three reference sequences for each read and found 94% within 5% divergence of one of them (Fig. S9A). For comparison, we determined the typical divergences between species by dividing the reference sequences into 150 bp segments (the same size as our reads) and calculating the divergences between all three pairs of species (Fig. S9B). The middle 95-percentile of the combined distributions extended down to a divergence of around 5% (8 SNPs), consistent with our cutoff for the intraspecies divergences at these loci. These results are consistent with all of the *Synechococcus* reads being from one of the three species and set an upper bound on the relative abundance of any other species at below 4%.

To check how sensitive our detection method is to the presence of other species, we used the Mushroom

Spring sample from 65 °C, in which γ had a relatively high abundance. To mimic the presence of another species, we compared the previous distributions to the distribution of the minimum divergence from α and β only (dashed red line in Fig. S9A). We found a 19% decrease in cumulative number of reads within 5% from the references in this case, which was very close to the 21% abundance of γ we estimated in the previous subsection based on the average depth across loci. This shows that our method is sensitive to presence of other species and illustrates the benefits of combining single-cell and metagenome analysis for identifying unknown species.

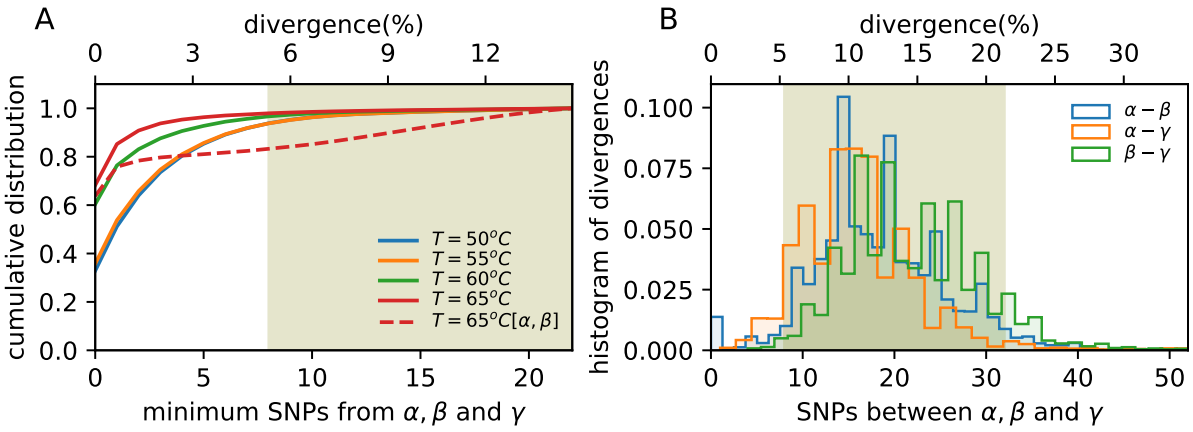

Figure S9: Vast majority of *Synechococcus* core gene sequences are within the α-β-γ main cloud. (**A**) Cumulative distribution of the minimum divergence of each read from the reference sequences across all loci used in the left panel. Different colors represent samples taken from different temperatures from Mushroom Spring. Solid lines show minimum divergence from all three references and dashed lines from α and β only. Shaded olive region shows the typical divergences between species shown in the right panel. (**B**) Histogram of pairwise divergences between three representative SAGs for α, β, and γ at the 100 core loci chosen for illustration. Distributions are across all 150 bp segments from each locus. Shaded olive region the middle 95% percentile for the combined distribution.

To further verify that no other *Synechococcus* species were present in our samples we closely examined the diversity of 16S sequences. We used only single reads in our analysis instead of trying to assemble full sequences to prevent misinference caused by hybridization. Because the diversity along the 16S sequence is highly heterogeneous, variation between read recruitment locations can dominate the variation between species. We controlled for this by limiting the analysis to reads that aligned within a given regions of the 16S. For 150 bp reads, we found that using segment sizes of 200 bp provided best compromise between increasing depth and minimizing variance due to alignment location. To find the most informative region, we calculated the nucleotide diversity at each site and averaged it over all 200 bp segments along the 16S (Fig. S10A). We found the diversity varied by more than a factor of four consistent with the presence of conserved and variable regions. The diversity was nearly identical when using all three species as reference genomes, suggesting our recruitment captured all sequences within the main cloud. The highest average diversity was close to 400-600 bp segment which we chose for further analysis.

We calculated the distribution of divergences from the γ allele for all reads recruited from this segment for MS temperature series. We found temperatures of 60 °C and below had two distinct peaks close to 4% and 8% divergences. For comparison, we calculated the range of divergences from α and β for all 150 bp subsegments within the $400 - 600$ bp segment and found these matched with the two peaks we

observed. At 65 °C we also found only two peaks, one close to zero and another close to 4%, consistent with our previous estimates showing that this sample mainly consisted of α and γ cells.

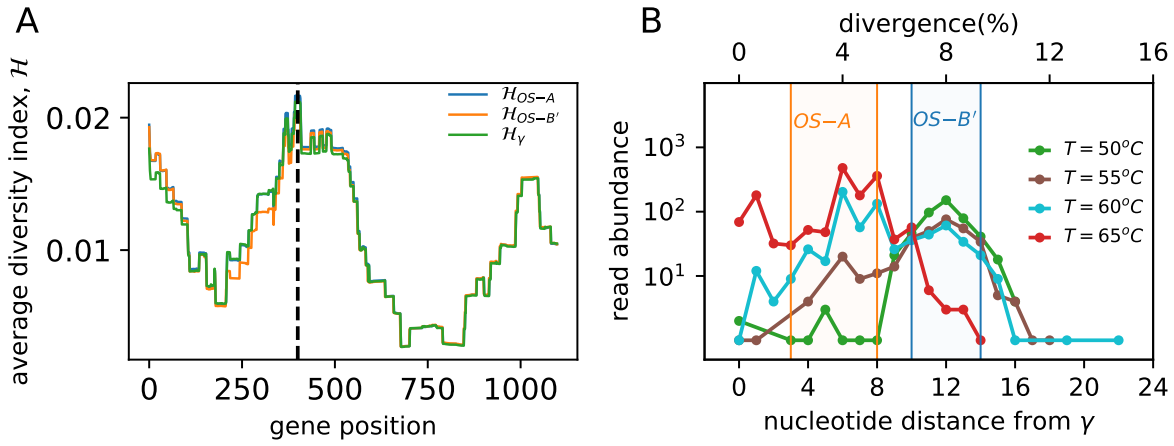

Figure S10: Diversity of *Synechococcus* 16S rRNA sequences from Mushroom Spring temperature series samples is entirely contained within the α-β-γ main cloud. (**A**) Average diversity $\mathcal{H}$ of metagenome reads from all 34 samples mapping to *Synechococcus* along the length of the 16S sequence. Averaging was done over all sites within a 200 bp window starting at the location given by the x-axis. Results for reads recruited using each of the three main species as a reference sequence are shown in different colors. Black dashed line indicates the $400 - 600$ bp segment with the highest diversity which was chosen for subsequent analysis. (**B**) Distribution of nucleotide divergences from the γ allele within the $400 - 600$ bp segment of the 16S for four different samples taken from Mushroom Spring. Shaded regions show the range of divergences of OS-A (orange) and OS-B' (blue) from γ across all 51 windows of 150 bp length, corresponding to the range of expected divergences of reads from α and β.

Do peaks correspond to single-species clusters or are they amalgamation of several distinct species with similar divergences from each other? To answer this question we constructed a multiple-sequence alignment of all of the 16S reads recruited across the 34 metagenome samples and performed a hierarchical clustering on the pairwise divergence matrix using average linkage. Pairwise divergences were calculated by padding the reads with gaps to ensure all sequences were 200 bp in length and calculating the nucleotide divergence ignoring any gaps. For the MS temperature series, we found 97% of reads were in one of the three clusters associated with α, β, or γ (Fig. S11A-D). Consisten with our previous results, samples below 65 °C had two main clusters close to OS-A and OS-B' (Fig. S11A-C), while at 65 °C they were close to γ and OS-A (Fig. S11D). Aggregating across all 34 metagenome samples, we found that α and β were the two dominant clusters, with γ being the next largest (Fig. S11EF). We also observed a large difference in the overall abundances of α and β across the two springs, but the very large heterogeneity across samples we found previously (Fig. S8) suggests this is likely due to the particular samples used in this study and may not necessarily be an accurate reflection of the overall abundances in the two springs.

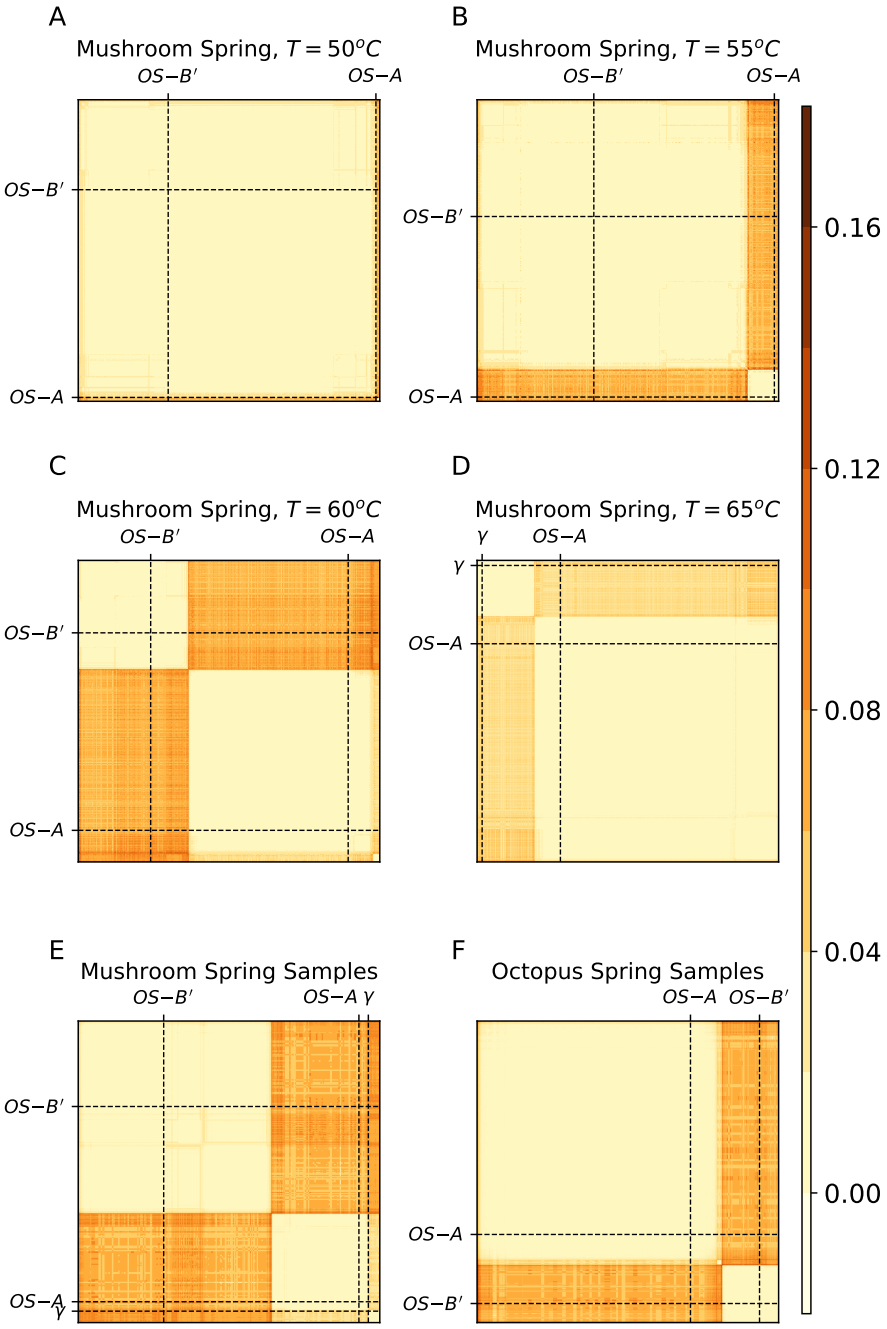

Figure S11: All *Synechococcus* 16S rRNA sequences from the metagenome samples belong to one of the three known species clusters. (**A**-**D**) Each subpanel shows the matrix of pairwise divergences between all reads recruited to the $400 - 600$ bp segment of the γ 16S rRNA sequence for the four samples from the MS temperature series. Dashed lines highlight the OS-A, OS-B', and γ reference sequences in this segment for comparison. (**E**, **F**) Same as above using a subsample of reads recruited from all 34 metagenome samples divided by spring.

IV. Linkage analysis

In this section we summarize our analysis of linkage disequilibrium within α and β. The structure of this section is as follows. We begin by reviewing definitions of two commonly-used ways to quantify linkage disequilibrium. We then compare the rate of decay of linkage disequilibrium within and between the two species. We next compare the decrease in linkage in the data to theoretical predictions from the neutral model and show that while the decay is similar to the neutral model prediction, the overall shape of the linkage decay curves cannot be fully explained by the classical neutral model. We then show that within α there is substantial heterogeneity in the rate of linkage decay in low-diversity regions compared to high-diversity ones. Surprisingly, we find that linkage decreases more rapidly in low-divergence regions, suggesting the absence of a clonal frame for α even on short time scales. The final subsection presents a technical validation of our method for processing alignments showing that our results are robust to filtering for outlier gene alleles.

IV.1 Statistical measures of linkage disequilibrium

We used two distinct but related methods for quantifying linkage disequilibrium, σ_d^2 and r^2. Both metrics are related to the correlation coefficient between alleles at two loci, but differ in their normalization. For a single pair of loci A and B with alleles A/a and B/b, we can define the correlation coefficient between their respective frequencies as

$$r_{AB}^2 = \frac{(f_{AB} - f_A f_B)^2}{f_A(1-f_A)f_B(1-f_B)}. \tag{S1}$$

Using this expression, we can define a statistical measure of linkage as a function of distance along the genome by averaging over pairs of loci with with a given separation $x = |x_A - x_B|$, where x_A and x_B are the positions of the two loci along the genome. The two most common ways to perform the average are

$$\sigma_d^2(x) = \frac{\langle (f_{AB} - f_A f_B)^2 \rangle_{|x_A - x_B| = x}}{\langle f_A(1-f_A)f_B(1-f_B) \rangle_{|x_A - x_B| = x}}, \tag{S2}$$

and

$$r^2(x) = \left\langle \frac{(f_{AB} - f_A f_B)^2}{f_A(1-f_A)f_B(1-f_B)} \right\rangle_{|x_A - x_B| = x}. \tag{S3}$$

For a neutral panmitic and well-mixed population both metrics decrease as $\sigma_d^2(x) \sim r^2(x) \sim 1/x$. For σ_d^2, there also exists an analytic expression (*29*):

$$\sigma_d^2 = \frac{10 + 2NR}{22 + 26NR + 4(NR)^2}, \tag{S4}$$

where N is the population size, and R the recombination rate per base pair per generation. When recombination is rare and occurs via crossover $R(x) \approx R_0 x$, and the expression is consistent with the scaling $\sigma_d^2(x) \sim 1/x$ mentioned previously. While numerical and theoretical studies have shown that r^2 is more sensitive to low-frequency alleles compared to σ_d^2 (*30, 31*), our analysis did not reveal significant differences between two metrics. We therefore focus here on σ_d^2 which is more commonly used and easier to compare

to theoretical predictions. Figures illustrating the analogous results for r^2 can be found at the end of this section.

We calculated $\sigma_d^2(x)$ and $r^2(x)$ for α and β separately and together. Both procedures are similar so we focus on the analysis for all α and β sequences. Unless otherwise stated, all analyses were performed on alignments of individual genes and the results were averaged across genes. Using this method, we can easily obtain linkage information over distances of up to around 1 kbp. Over longer distances the results can be dominated by specific recombination events in small number of unusually long genes and are therefore less useful for quantifying genome-wide statistical patterns. We thus perform most of our analysis up to distances of around 1 kbp, except when caculating typical genome-wide linkage as discussed below.

Alignments for each orthogroup were first trimmed to remove segments with excess numbers of gaps. We then determined a consensus sequence by taking the most common codon sequence at each location along the reading frame, excluding gaps. Each nucleotide was the labeled 0 if it was a gap, 1 if it was the same as the consensus, and 2 if it was different from the consensus. While does not distinguish between biallelic and multi-allelic sites, the later are much more less common that the former and should not affect our conclusions. Finally, we calculated the two linkage coefficients using Eqs. S2 and Eqs. S3. To estimate linkage across distances longer than a single gene, we performed the average across SNPs spanning a pair of genes and then averaged the values across a random sample of 1000 pairs of genes.

The linkage calculations within each species were performed in an analogous way, with the exception of a two-step filtering process of the alignments at the beginning. We describe this procedure in detail at the end of this section.

IV.2 Linkage decrease within vs between species

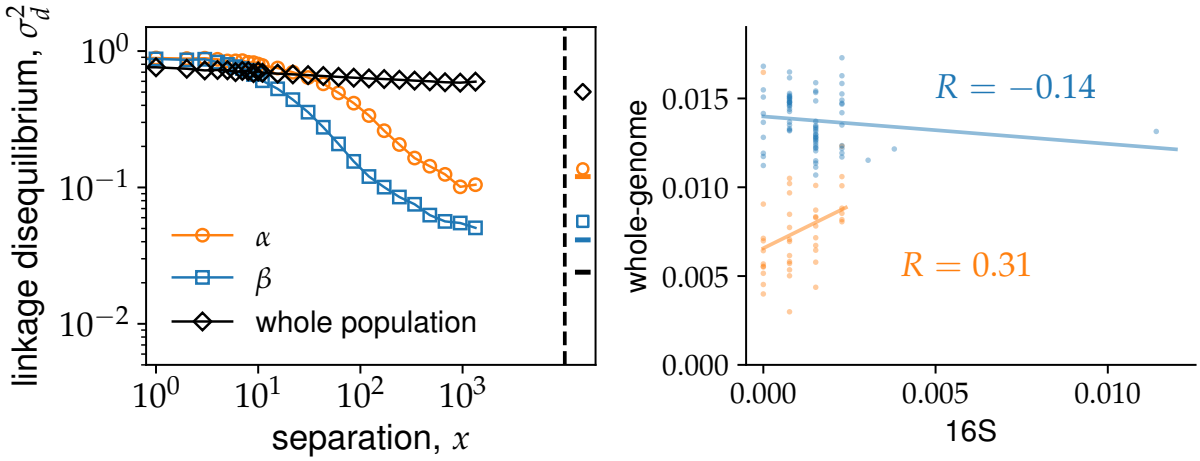

Figure S12: Genetic linkage analysis shows extensive recombination within α and β. (**A**) Linkage disequilibrium σ_d^2 as a function of separation between SNPs x averaged across all core genes among the α (orange circles), β (blue squares), or all of the cells (black diamonds). Genome-wide estimates from SNPs in different genes are shown to the right of the black dashed line using the same symbols. Fully unlinked controls for the genome-wide linkage are shown as horizontal lines of different colors. The main cloud cutoff for both α and β was set to $c = 0.05$. (**B**) Scatter plot of 16S and genome-wide divergences between all pairs of cells within the same species. Solid line shows a linear regression from which the Pearson correlation coefficient, R, within each species was calculated.

Our analysis showed that within both α and β, linkage quickly decreased by approximately an order of

magnitude over a distance of $\sim$ 1 kbp (orange circles and blue squares in Fig. S12A). However, when considering the population as a whole, we find linkage disequilibrium is maintained across the whole length of the genome (black diamonds in Fig. S12A). Within β, the linkage at distances of 1 kbp was identical to the genomewide linkage between random pairs of genes. Within α, the decrease was slower, but followed a similar pattern. The value of the genome-wide linkage was slightly higher than at 1 kbp, likely because of the reduced sample size at long genomic distances caused by poor genome coverage. Consistent with this, we found that the genomewide-linkage values in both species were close to completely unlinked controls obtained by randomly permuting each column of the alignments ($\sigma_d^2 = 0.14$ vs $\sigma_d^2 = 0.12$ in α and $\sigma_d^2 = 0.06$ vs $\sigma_d^2 = 0.04$ in β). Overall, these results suggest that the dynamics leading to the reassortment of alleles are similar in the two species and only differ in the length scale over which alleles become unlinked. We discuss this in more detail in the next section.

The fact that both species have similar levels of genomewide linkage despite the sample size of α being less than half of that of β shows that our estimates for the genomewide linkage are not significantly affected by the sample size effect. We further validated this by taking random samples from β main cloud equal to the size of the α main cloud at each locus and recalculating the linkage curves. We found the results for the full β main cloud and the subsampled main cloud were almost identical (Fig. S13). This shows that the sample size does not have a large effect on our results and cannot explain the differences in linkage decay between the two species.

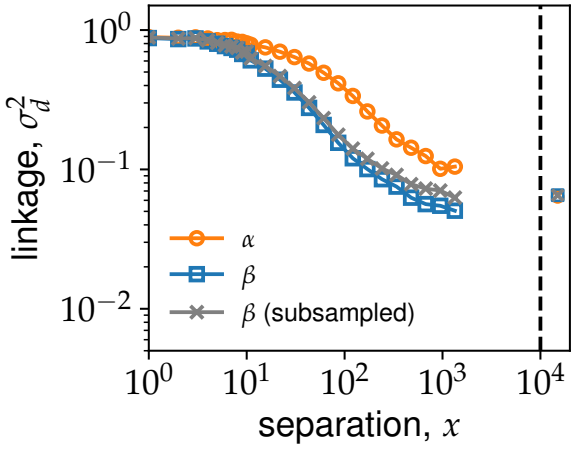

Figure S13: Samples size has small effect on linkage estimates. Linkage disequilibria curves for the whole α and β populations shown in S12 are compared to a random subsample of β sequences chosen to match the α sample size at each locus (shown in gray).

The rapid decrease in linkage sets important limits on how much information about the diversity in the population is contained in samples of marker genes. To illustrate this, we performed a linear regression between the whole-genome divergences and the 16S rRNA divergence. Consistent with our previous analysis, we found slightly stronger linkage in α compared to β, but in both cases the correlations were weak and the 16S divergences explained less than 10% of the genome-wide divergences (Fig. S12B). These results show that frequent recombination in the population makes marker genes, such as the 16S rRNA, poor predictors of the genome-wide diversity.

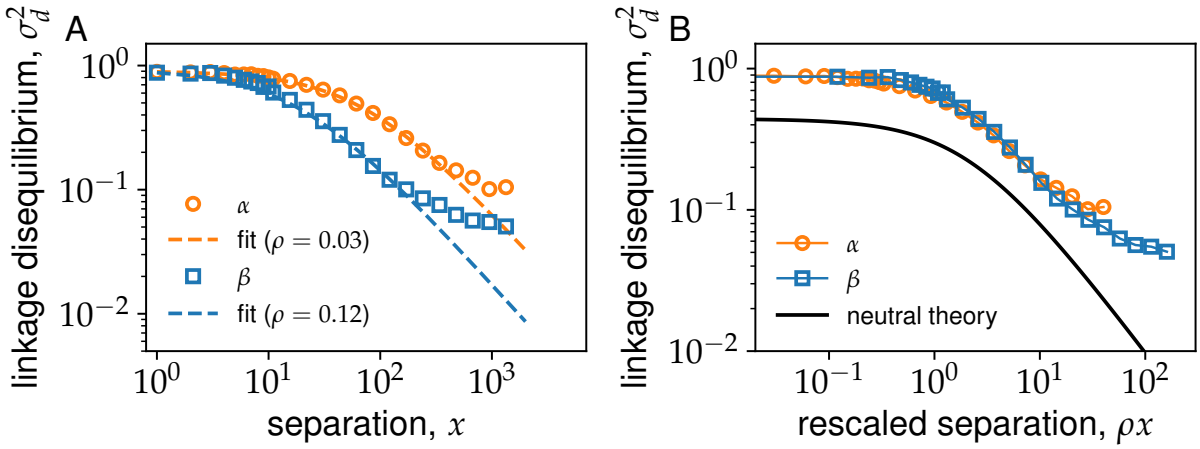

Figure S14: Linkage decay curves for α and β are collapse on to the same curve after rescaling distances by average recombination rate ρ. (**A**) Shows same data as Fig. S12A for α and β together with the fit usings Eq. (S5). (**B**) Shows the linkage decay as a function of distances between sites rescaled by the recombination rates ρ fitted from (A). The solid black line shows the theoretical prediction from the neutral model given by Eq. S4.

IV.3 Comparison to theoretical predictions

To get a more quantitative understanding of the effects of recombination on the genetic diversity in our population we compared our previous results to theoretical predictions from population genetics. At short distances, the prediction from Eq. S4 is that $\sigma_d^2 = 5/11$. However, in our data the value of σ_d^2 for nearby sites was very close to one. While the Yellostone Synechococcus have a strong mutational bias towards GC nucleotides (*12*), Eq. S4 is independent of the relative rates of mutations. As a result we conclude that the neutral model cannot explain the linkage curves we observe.

While the overall level of σ_d^2 in our data is inconsistent with the neutral prediction, we can still ask whether the rate of decay with distance has a similar functional form. To do this we followed Ref. (*32*) and fit the linkage curves to the following functional form:

$$\sigma_d^2(x) = A\frac{10 + \rho x}{22 + 26\rho x + 4(\rho x)^2}, \tag{S5}$$

where $\rho = 2NR$ is the population recombination rate. We performed the fit by choosing A such that the above expression matched the value of σ_d^2 in the data for neighboring sites ($x = 1$). We then fitted the curves by eye by adjusting the value of ρ.

We found that this functional form could fit the data very well for both species up to distances of $x \approx 300$ bp. For longer distances the linkage decrease was slightly slower than predicted by (S5). One possible interpretation of this slowdown is that longer distances have a signficant overlap with the distribution of lengths of recombined segments L_r, beyond which we expect linkage to reach the genomewide level (*32*). This suggests that recombined segments as short as 300 bp might play an important role in recombination process within the population.

We further validated that the linkage curves have the same functional form by rescaling distances between sites by the recombination rate ρ inferred previously. We found a very good collapse of the linkage curves for the two species when plotted against the rescaled distances Fig. S14B. This strongly implies that the

dynamics leading to the decay of linkage are the same in both species, with the differences between them being length scale over which sites are effectively linked.

IV.4 Lack of clonal frame in α

The lack of clonal structure implied by the rapid decrease in linkage found in our previous analysis is in line with recent studies showing recombination is more prevalent in bacteria than is traditionally thought (*33*, *34*). Nevertheless, because bacterial recombination occurs through the exchange of short segments, it is generally believed that in closely-related bacteria, corresponding to short times since their common ancestor, most of the diversity is the result of accumulation of mutations on a clonal backbone rather than reassortment of existing variants through recombination (*34*). Several studies have found mostly clonal structure for divergences up to $\sim 5 \times 10^{-3} - 10^{-2}$, with recombination playing an increasingly larger role for later times. Most of these studies have focused on human-associated bacteria, which can undergo very strong bottlenecks during transmission that may not be as common in environmental bacteria such as the Yellowstone Synechococcus. We therefore decided to take advantage of the genome heterogeneity within α to test the clonal frame hypothesis.

According to the clonal frame hypothesis, the low-diversity regions of the α genome analyzed in Sec. **??** would be the result of an earlier clonal expansion within α, with large divergence segments representing recombination with outside strains. We would therefore expect that in the low-divergence regions linkage should not decay. To test this hypothesis we averaged over the linkage curves over the low-diversity and high-diversity segments from Sec. **??** separately and compared their decay rates. Surprisingly, we found the opposite pattern—rather than staying constant, linkage in the low-diversity regions of α decreased more rapidly than high-divergence ones (Fig. S15). Moreover, within the low-divergence region there was a sudden downward shift in linkage by close to an order of magnitude at distances close around $100-300$ bp. To verify that the linkage decrease on longer distances is not due to a small number of unsual genes, we also calculated the genomewide linkage within both the low and high diversity regions as before. We found the linkage was slightly higher within the high diversity regions ($\langle\sigma_d^2\rangle = 0.15$ compared to $\langle\sigma_d^2\rangle = 0.11$ in the low diversity regions), but both values were similar to the unlinked controls ($\langle\sigma_d^2\rangle_c = 0.14$ in the high diversity regions and $\langle\sigma_d^2\rangle_c = 0.09$ in the low diversity regions).

The rapid decrease in linkage in regions of low diversity is in striking contrast to studies on human-associated bacteria and suggest that even on the much shorter time scales implied by the α low diversity backbone (see Sec. VII), recombination is the dominant evolutionary force shaping the genetic diversity of Synechococcus. To quantify the relative contributions of recombination and mutations to the diversity within α, we used the parameters inferred from fitting our data to a panmitic neutral model with recombination. The relatively poor statistics from the low-diversity regions of the α genome prevent us from obtaining the population recombination rate ρ directly from linkage decay curves. However, the comparison between the low and high diversity regions from Fig. S15 shows that the rate of linkage decay in the low diversity regions is at least comparable to the genome-wide estimate of $\rho \approx 0.03$ we inferred from Fig. S14. We therefore used this value as an estimate for the rate of recombination breakpoints. The contribution from mutations is given by the typical number of mutations that have accumulated since the common ancestor $\theta_S = 2N\mu_S$, which in the neutral model is equal to the nucleotide diversity. Using the synonymous diversity inferred in VII, we have $\theta_S \approx 2 \cdot 10^{-3}$. This gives a ratio of $\rho/\theta_S \approx 15$.

If we apply a similar analysis to the β genomes we find $\rho/\theta_S \approx 3$. The discrepancy in ρ/θ between the two species is at first sight surprising given that α and β are closely-related species with similar physiologies that have inhabited the same environment for a long time. Several factors could contribute to this discrepancy,

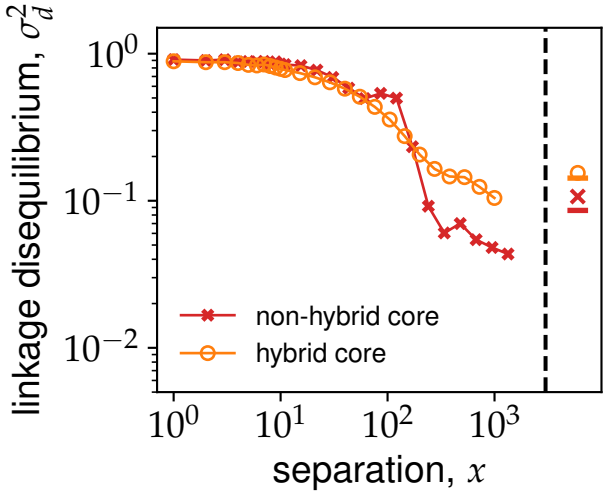

Figure S15: Hybridized and non-hybridized segments of the α genome have similar rates of linkage decrease. The decay of the linkage disequilibrium σ_d^2 averaged over α loci that have undergone hybridization (orange circles) and loci that have not undergone hybridization (red crosses) is compared. Hybridized and non-hybridized loci were determined based on a 5% cutoff in the fraction of synonymous polymorphic sites within α (see Sec. VII).

among them a decrease in the rate of recombination over evolutionary time, epistatic interactions that could suppress recombination rates at short distances, or an overestimation of the mutation rates due to a large contribution of the diversity coming from hybridization. The later would be roughly consistent with our estimates for the fraction of the diversity witin β contributed by hybridization (Sec. VI).

IV.5 Linkage on intermediate length scales

Our analysis so far was limited to either distances up to the length of typical genes ($\sim$ 1 kbp) or across the entire genome. Extending our analysis over distances in between these two extremes is difficult due to several technical challenges. The first is the fact that typical coverages for our SAGs is around 30%. Together with the a typical contig size of $\sim$ 10 kbp, this means that our sample size quickly decreases with distance. The second challenge is that at longer distances gene order might vary across genomes making it difficult to define a genomic distance at the population level.

IV.6 Linkage between multi-SNP alleles

We calculated the linkage disequilibrium between multi-SNP features, such as SNP blocks or hybrid alleles, using (S1) as in the case of biallelic SNPs. In the case of SNP blocks shown in Fig. 3E in the main text, we calculate the frequencies for the four haplotypes within SAGs that are covered at both loci. The unlinked controls (represented by solid lines in Fig. 3E) were obtained by a random permutation of the alleles at each locus independently, while keeping the SAGs that are covered at each locus fixed. This procedure preserves any correlations in genome coverage between nearby loci and provides the most accurate estimate of r^2_{AB} for a completely unlinked genome.

The linkage between hybrid gene alleles were calculated similarly as for the SNP blocks, except for the genotype assignment. At each locus, we labeled the most abundant allele using upper case letters (A, B, ...) and grouped all of the remaining ones into a single allele labeled by a lower case letter (a, b, ...). The linkage matrix r^2_{AB} and the unlinked control were calculated the same as previously described. The results for α were $\langle r^2_{AB} \rangle = 0.04 \pm 0.15$ compared to $\langle r^2_{AB} \rangle = 0.02 \pm 0.11$ for the unlinked control. For β the

results were $\langle r^2_{AB} \rangle = 0.01 \pm 0.09$ compared to $\langle r^2_{AB} \rangle = 0.009 \pm 0.08$ for the unlinked control. Note that the average linkage disequilibrium was close to the unlinked controls in both species, with α having slightly higher linkage. These agree with those from the analysis of SNP blocks shown in the main text.

IV.7 Determining the main cloud cutoff for α and β

For each species, we first selected all of the sequences that belonged to SAGs assigned to that species (see Sec. II). We then calculated the consensus sequence for this group of sequences, as described above, and removed all sequences with divergences greater that a main cloud cutoff c. This is a simple way to remove sequences that are the result of gene-level hybridization (see Sec. V and (*33*) for a similar procedure). We chose the main cloud cutoff for each species by calculating the linkage curves for different values and choosing the largest value for which the curves did not noticeably change (Fig. S16). Based on this criterion, we use $c = 0.05$ for α and $c = 0.1$ for β throughout.

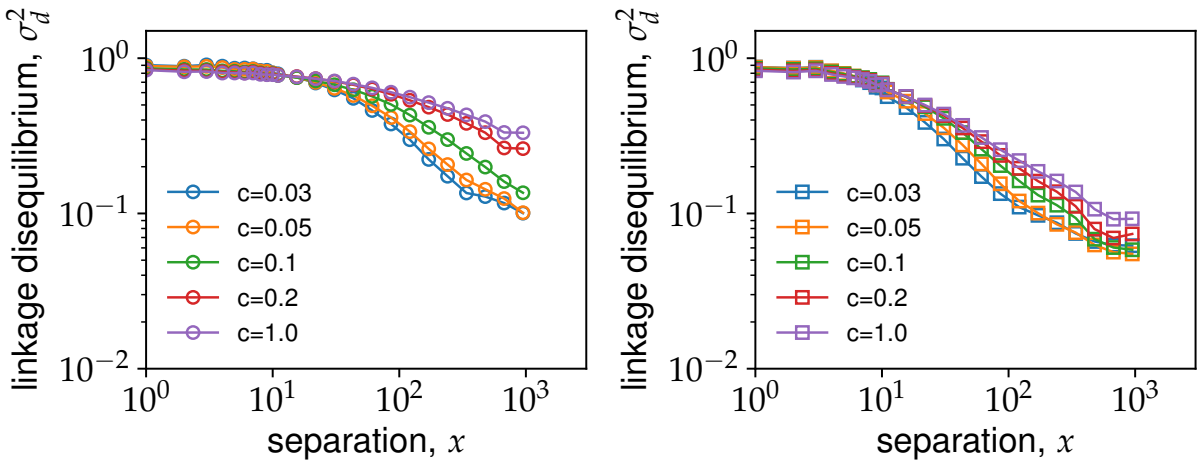

Figure S16: The effect of the main cloud diameter c on the rate of linkage decay.
The linkage decay within α (left panel) and β (right panel) is plotted for different main cloud cutoffs c (see main text). Note the convergence of the curves for $c \leq 0.05$.

V. Gene-level hybridization

This section describes the method we used to quantify hybridization at the gene level. Our method is based on hierarchical clustering of sequences based on their pairwise distances. Using the method described in Sec. I.2, we determined which loci contained distinct clusters at the divergences we expect between species ($\approx 15\%$ at the nucleotide level, as shown in Fig. 1A and Sec. II). We then assigned labels to these clusters based on the composition of species to which the sequences belonged to. Finally, we searched for discrepancies between the gene and whole-genome labels, representing transfers of whole-gene alleles from other species.

V.1 Simple hybrid genes

We first assigned labels to the species-level clusters constructed as described in Sec. I.2. Based on the gene triplet results presented in Sec. II, we expect most sequences to be contained in two main clusters, close to OS-A and OS-B', respectively, with at least one sequence forming a separate cluster whenever the orthogroup was covered in the γ SAG. Based on this expectation, we considered all of the clusters from each orthogroup in decreasing order of their sizes.

If we assume the clusters at the genome level are replicated at the gene level, we expect up to three clusters within each orthogroup, with sequences within each clusters mostly belonging to a single species. To distinguish between this and the case where rapid hybridization breaks down species clusters at the gene level we use an arbitrary threshold $S = 0.9$ for the fraction of sequences within a cluster belonging to the majority species. If the fraction of sequences in the largest cluster belonging to either α or β is less than S, we label the cluster M for *mixed-species*. Otherwise, we label the cluster A or B if the fraction of sequences from α or β genomes, respectively, is higher than S. We continue this process until both A and B labels are assigned. As a result, the labels A and B are assigned to the largest clusters that contain fractions greater than S of sequences from α and β, respectively, if these exist. If more than one A or B cluster is encountered before the other cluster is assigned, they are labeled sequentially as A1, A2, A3, etc. In cases where there are clusters still unlabeled after this process, the remaining clusters are labeled as C if they contain a sequence from the γ SAG or as X for unknown otherwise.

We next determined which orthogroups are part of the core genome. We defined a gene as a core gene if it is likely present in all of the genomes in our sample. The average coverage for each SAG in our sample was approximately 30% so it is not possible to directly determine whether some genes are always present. Instead, we use a statistical measure of the coverage to assign orthogroups to the core genome. Specifically, we used the rank-coverage plot across all orthogroups and SAGs (equivalent to the complementary cumulative distribution) to determine an appropriate cutoff for the core genome. The rank-coverage plot showed three distinct regions. First there were around a dozen orthogroups that had unsually high abundances, with the number of sequences in some cases being greater than the number of SAGs in our sample. These are likely multiple-copy genes and IS elements and were excluded from further analysis. Then there was a large plateau comprising over $1,000$ orthogroups with coverages in the range $100 - 150$, followed by a sharp decrease (on a logarithmic scale) in the coverage. We found that setting a coverage threshold for the presence at 20% of the number of SAGs from each species approximately matched the shoulder between the plateau and the sharp decrease and so we used this as a criterion for assigning core genes.

The final step in our algorithm is to catalog hybridization events, which we do by simply searching for gene sequences whose labels were different from the whole-genome species assignment. Analysis of these results showed a large number of transfers, mainly from β into α, which came from contigs that did not contain any genes labeled as the host species. One explanation for the presence of such contigs is that they

represents larger genomic segments flanked by repetitive regions (such as IS elements) that are transferred and integrated into the host genome. Long repetitive genomic segments can lead to contig breaks during assembly, which would explain the complete absence of genes belonging to the host cluster. Alternatively, such contigs could be the result of contamination with external DNA. However, the contamination hypothesis does not explain why there seem to be more transfers from β into α than vice versa. We also did not find any evidence of differences in the number of such transfers between the sample containing only α SAGs (OS2005, see Sec. IX) and samples containing mixtures of α and β. While this evidence suggests these transfers are very likely the result of hybridization, we cannot rule out the possibility of contamination. We therefore excluded all hybridization events on contigs that did not contain any genes labeled as the host genome from the analysis. The results of our analysis are summarized in Fig. 2AB in the main text.

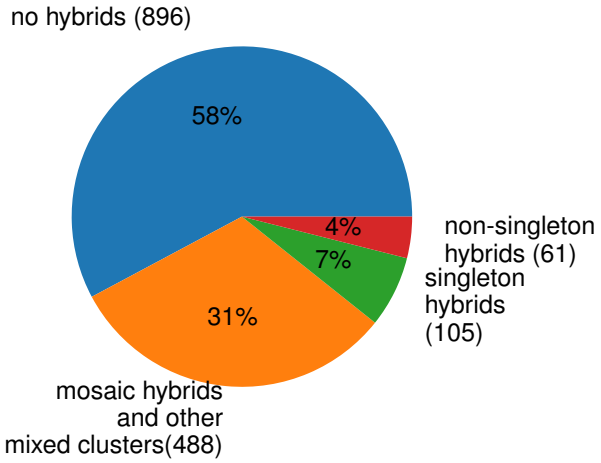

Figure S17: Majority of orthogroups contain hybrid sequences. Orthogroups were classified into four mutually exclusive groups depending on the amount of hybridization observed. The majority of mixed-species orthogroups are mosaic hybrids (see main text) and shown in orange. Orthogroups in which the local sequence clusters matched the whole genome ones were labeled as "no hybrids and shown in blue, those in which a single cell contained a sequence from a different cluster were labeled "singleton hybrids" and shown in green, and those in which multiple cells contained sequences from different clusters were labeled "non-singleton hybrids" and shown in red.

V.2 Diversity within hybrid genes

We quantified the diversity within simple hybrid genes as follows. For each non-singleton hybrid locus found in the previous step, we calculated the synonymous nucleotide diversity π_S among the hybrid sequences only, as described in VII (Fig. S18). We used two controls for comparison: the diversity within the donor population for $\alpha - \beta$ transfers, and the core genome diversity within the host population.

For transfers between α and β from either donor, we calculated the synonymous diversity in the donor species for each hybrid locus. If the hybrid gene alleles are the result of multiple independent hybridization events, we would expect the diversity among the hybrid sequences to be similar to that of the donor population. Such events would be analogous to soft sweeps (*35*). Otherwise, if the hybrid sequence was

only introduced once into the host population and then subsequently spread through recombination, we would expect much less diversity among the hybrid sequences compared to the donor population. We observed both types of transfers in our data. For transfers into α, approximately half of loci (17 out of 44) had identical hybrid sequences indicating a single introduction and subsequent spread through recombination. In the remaining cases, the typical diversity among the hybrid gene alleles was much higher than $\pi_S \approx 2 \cdot 10^{-3}$ in the core α genome (see Sec. VII), suggesting that it is the result of multiple introductions from the donor population. For transfers from β into α specifically, 4 out of the 5 loci had diversity levels in α comparable to those of β, indicating that preferential transfer of specific variants was not an important factor in the hybridization process (Fig. S18, left panel). Similar conclusions can be drawn from the analysis of hybrids transfered into β, but there there were relatively fewer loci with no diversity (5 out of 18), while the loci consistent with multiple introductions had diversity comparable to the β core genome.

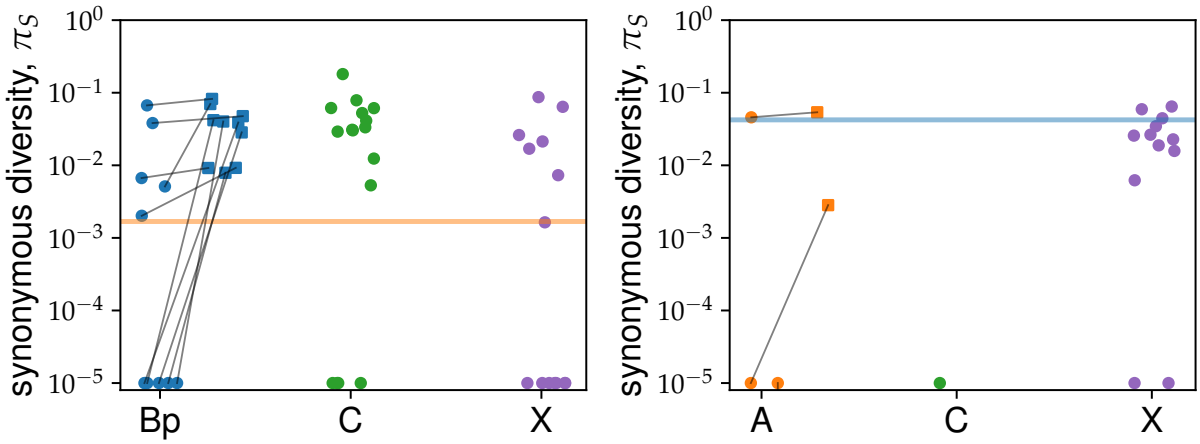

Figure S18: Simple hybrid genes have bimodal distribution of diversity across loci. Diversity among hybrid sequences at simple hybrid loci. Each circle shows one of the non-singleton loci from Fig. 2AB within the α (left) and β (right) cells. The colors represent the different donor species, as labeled on the x-axis. For transfers between α and β the diversity within the donor group is shown as squares of the same color as a control. Sequences from the same locus are connected by a black line. The horizontal lines show the nucleotide diversity within the low diversity genomic regions of α (left panel, in orange) and the core genome of β (right panel, in blue), respectively.

We performed a similar analysis for genes within diversity troughs (Fig. S19). Recall that diversity troughs were defined as regions where the synonymous divergence between α and β was much lower than typical (Fig. Fig. 2C). We found that despite the large decrease in divergence, these loci still contained considerable diversity, indicating they were either the result of soft sweeps or gene sweeps that had occurred sufficient time in the past for the present levels of the diversity to accumulate. To distinguish between these scenarios, we compared the diversity in each species at the diversity troughs to a random sample of core genes from the host species. Within β, we found that the two distributions were nearly identical (Fig. S19 right), indicating that the trough diversity within β is typical of the core genome. Within α, however, we found that core genes were on average around a factor 20 less diverse than the diversity troughs. This strongly suggest that the diversity troughs genes, including the nitrogen fixation pathway *nif*, were acquired by α from other species within the caldera after colonization. The levels of diversity within the trough genes are consistent with multiple independent transfers from β, but multiple inde-

pendent transfers from other species into both α and β could also contribute.

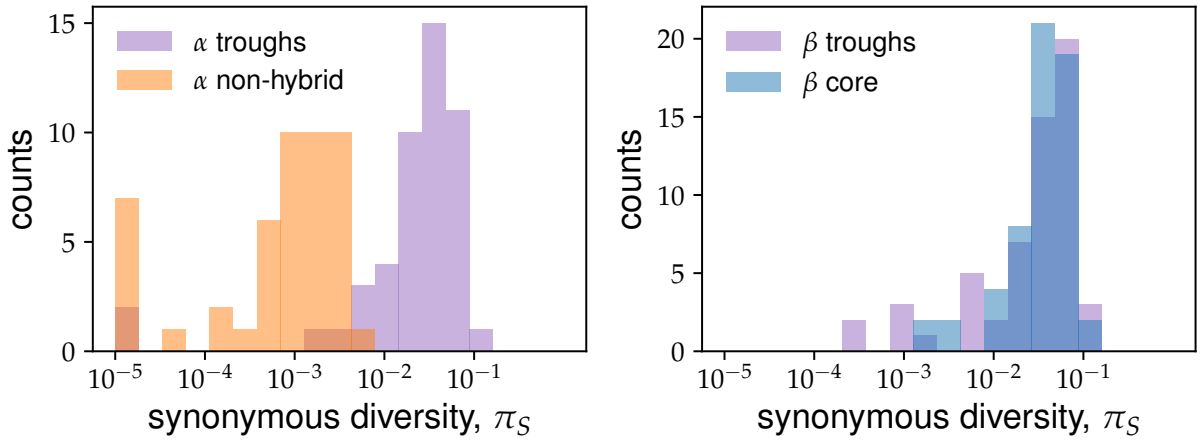

Figure S19: Typical diversity within diversity troughs is similar to β and much higher than the non-hybrid regions from α. (Left panel) The distribution of synonymous nucleotide diversity within α across all diversity trough genes is shown in purple. The distribution across a random sample of non-hybridized α genes of the same size is shown in orange for comparison. Non-hybridized α genes were determined using a 5% cutoff in the fraction of polymorphic synonymous sites (see Sec. VII). (Right panel) The distributio of synonymous nucleotide diversity with β core genes across all diversity trough genes is shown in purple. Here, the control distribution shown in blue is from a random sample of core β genes.

VI. Linkage block detection and analysis

In this section we describe the method we used to discover SNP blocks and the statistical analyses we performed.

VI.1 Processing alignments for SNP block detection

In order to identify linkage blocks originating from hybridization the raw alignments from the pangenome construction were processed as follows. We divided core OGs into distinct species clusters, as explained in Secs. I.2 and V, and removed any sequences from SAGs assigned to a different species than the cluster. This was done to avoid false positive blocks due to the presence of recently transferred whole-gene alleles from different species. For OGs with mixed-species mosaic clusters, all sequences in the cluster were initially included. Alignments were then trimmed and codons that contained a high frequency of gaps were removed. Finally, we defined a main cloud of sequences for α and β comprising sequences within a divergence diameter $d_M = 0.1$ from the consensus for each species and removed all sequences outside it.

VI.2 SNP block detection method

We used a linkage-based approach to search for evidence of hybridization within genes. Visual inspection of the alignments in the highly-diverse α genes and in most β genes revealed segments with high SNP density grouped into two haplotypes in strong linkage disequilibrium. We hypothesized that these SNP blocks were the result of intragenic hybridization and we developed an algorithm to systematically identify them.

The algorithm can be summarized by the following steps.

1. Read orthogroup alignment
2. Pre-process alignment
   (a) Trim alignment edges and remove columns with high fraction of gap codons
   (b) For each species, infer consensus sequence and remove sequences outside a predetermined diameter around the consensus
   (c) Replace all SNPs with absolute frequencies $n < n_c = 4$ with the consensus for their respective columns
3. For each species s, subsample sequences from cells assigned to s
4. Calculate the linkage matrix r^2_{ij} for all pairs of SNPs
5. Identify groups of consecutive SNPs with $r^2_{ij} = 1$, up to a rounding error of $\epsilon = 10^{-5}$, containing at least $l_{min} = 5$ SNPs

The inclusion of step 2c accounts for the possibility of more recent low-frequency SNPs emerging after the the initial hybridization took place. Note that typical alignments depths are ~ 30 for α and ~ 70 for β, and are considerably larger than n_c.

We applied the above algorithm to the alignments of all of the core genes and obtained a total of 1,442 blocks within α and 2,153 blocks within β. Most blocks were short, but those in α were on average more than twice as long as those in β ($\langle L \rangle = 62$ in α vs $\langle L \rangle = 29$ in β; Fig. S20A). Similarly, α blocks contained

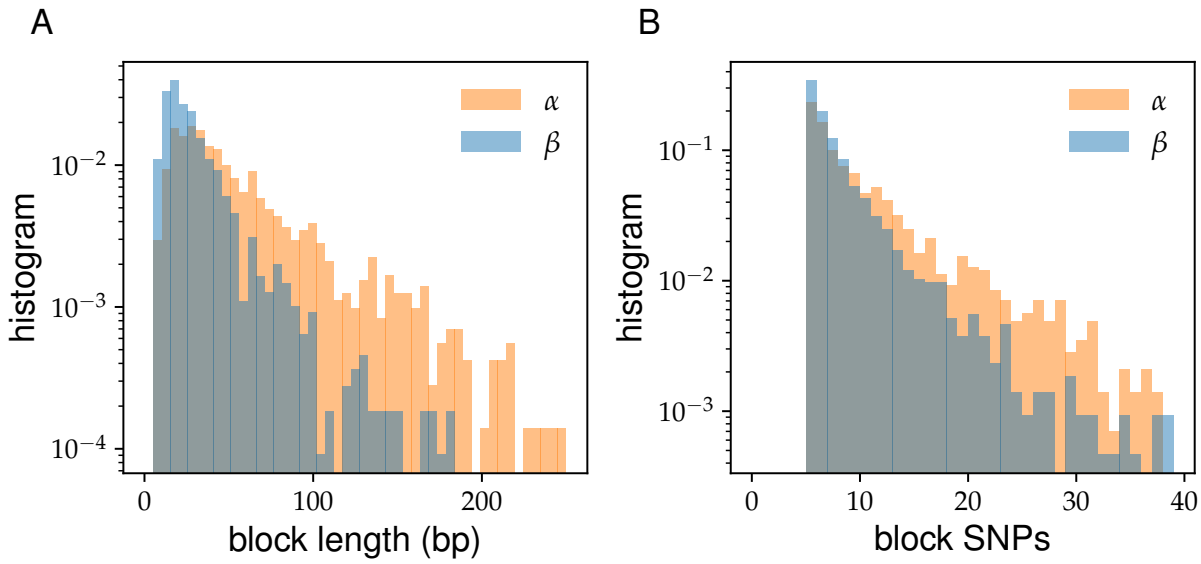

Figure S20: Typical blocks are short and contain a handful of SNPs. (**A**) The distribution of lengths of SNP blocks within α (orange) and β (blue) genomes, detected using the algorithm described in this section. (**B**) The distribution of the number of SNPs within the same blocks as in (A).

more SNPs on average compared to β, although here the difference was less pronounced ($\langle l \rangle = 10.6$ in α vs $\langle l \rangle = 7.8$ in β; Fig. S20B).

To identify hybrid haplotypes, we first determined the consensus sequences for the two block haplotypes, together with the other two species within the block segment. We then labeled any haplotypes whose sequence was identical to one of the other two species as a hybrid segment with the corresponding species as the donor. These result are shown in Table S2.

Host species	Block donor species α	β	γ	X	Total blocks	Genes with blocks	Total gene transfers	Majority mosaic genes	Total core genes
α	-	384	526	306	1442	496	100	443	1832
β	416	-	149	1397	2153	980	84	501	2172

Table S2: Number of SNP blocks within α and β containing haplotypes that mapped to different donor species.

To control for other processes that could potentially produce similar patterns of linkage disequilibrium we compared the distributions of divergences between haplotypes and block lengths for cases where a hybrid haplotype could be identified directly to cases we could not. No obvious difference in the distributions could be seen between the two cases, consistent with our interpretation that the blocks without matches to other species in our sample are also the result of hybridization.

VI.3 Haplotype divergences

We further verified that the divergences between block haplotypes were consistent with what we would expect from hybridization. We calculated the fraction of nonsynonymous and synonymous nucleotide substitutions, p_N and p_S, respectively, between the consensus sequences of each haplotype (see Sec. VII).

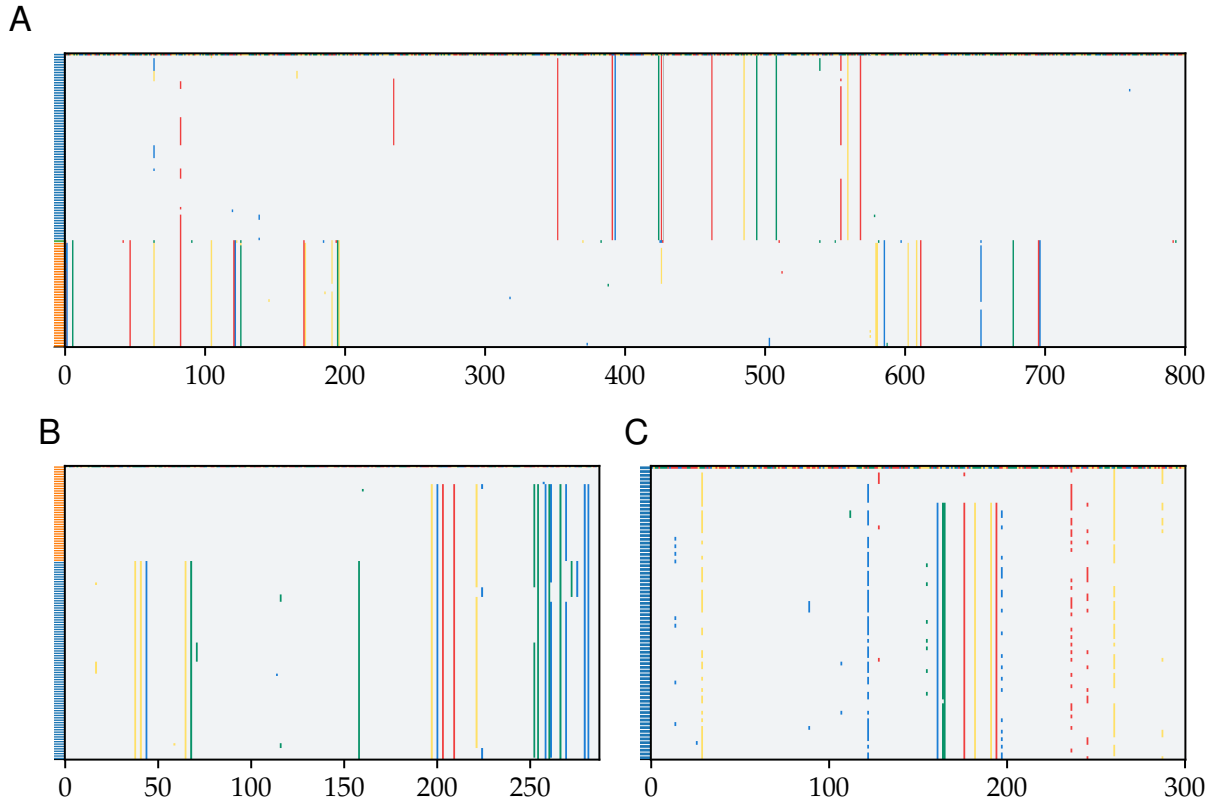

Figure S21: Examples of blocks and hybrid sequences within ribosomal genes. (**A**) Alignment of 16S sequences with a β cell containing a hybrid segment starting at around position 340 and ending around position 580 in the alignment. (**B**) Alignment of the SSU ribosomal protein rpsN with α cells ordered at the top and β cells at the bottom of the alignment. The first β cell was chosen as the reference sequence and is shown in color. (**C**) Segment of alignment of β sequences at the LSU ribosomal protein rplM. The reference sequence was chosen to illustrate short SNP block in the center of the alignment.

We then removed blocks with either p_N or p_S were above a threshold of 0.6. This was done to avoid large errors in estimating the divergence, which we calculated using the Jukes-Cantor correction $d = -(3/4)\ln(1-4p/3)$ (*36*). The overall number of such blocks was small (46 out of 1442 in α and 143 out of 2008 in β) so their removal does not affect our conclusions about the typical blocks. We found that the distributions of both the nonsynonymous and synonymous divergences were similar to those previously reported from comparisons of OS-A and OS-B' (*12*), and consistent with our hybridization hypothesis (Fig. S22).

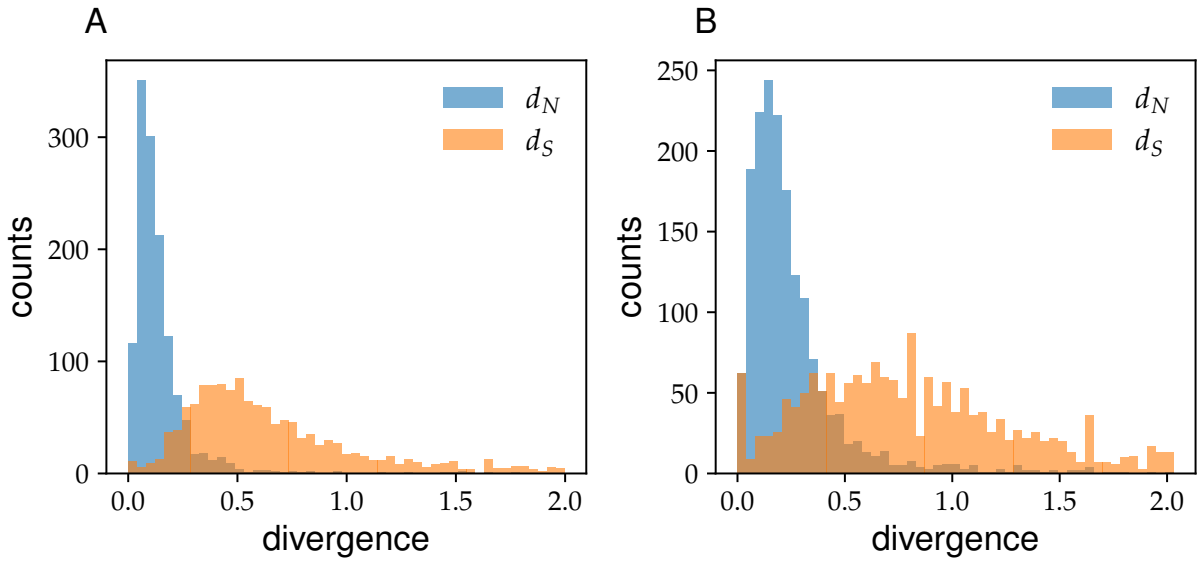

Figure S22: Block haplotypes have high synonymous and non-synonymous divergences, consistent with them being generated through hybridization. Panels show histograms of nonsynonymous (blue) and synonymous (orange) divergences between the two haplotypes of SNP blocks from α (**A**) and β (**B**), respectively.

VII. Genetic diversity analysis

In this section we present our analysis of the genetic diversity within individual genes in α and β. In the first subsection we define the different metrics we used and present the variations in diversity along the genome of both species. In the second subsection we show that a majority pf genes within α appear to have not undergone hybridization. By developing a method to differentiate between such genes and the rest we obtain an estimate for the time since the most common ancestor for α that is around an order of magnitude less than the naive estimate based on the average genomewide diversity. Finally, we compare the diversity pattern in α with β and discuss their implications.

VII.1 Nucleotide diversity along the core genome

We quantified the nucleotide diversity across all genes in both α and β using the following procedure. Alignments were trimmed for excess gaps and sequences were assigned to each species as described in Sec. II. We then define the nucleotide diversity in the usual way as

$$\pi = \frac{1}{|I|(|I|-1)} \sum_{i \neq j \in I} \pi_{ij}, \tag{S6}$$

where I is an index for sequences in the alignment, $|I|$ is the sample size, and π_{ij} is the fraction of nucleotide substitutions between i and j, excluding gaps. To calculate synonymous (π_S) and non-synonymous diversities (π_N) we replace π_{ij} with the fraction of synonymous (π_{ij}^S) and non-synonymous substitutions (π_{ij}^N), respectively. We calculated π_{ij}^S and π_{ij}^N using two different methods: the fraction of four-fold degenerate (4D) and one-fold degenerate (1D) sites, respectively, at which i and j differ and using the Nei-Gojobori method (*37*). We found that both methods gave similar results and used the one that was most convenient for each analysis.

To calculate the divergence between species, we first determined the proportion of sites that were substituted as follows:

$$p = \frac{1}{|I_\alpha||I_\beta|} \sum_{i \in I_\alpha, j \in I_\beta} \pi_{ij} \tag{S7}$$

where I_α and I_β are the indices of sequences from α and β in the alignment and $|I_\alpha|$ and $|I_\beta|$ are number of sequences from each species, respectively. The fraction of pairwise substitutions π_{ij} is the same as above. We then use the Jukes-Cantor correction (*36*) to correct for the effect of multiple substitutions at a single site

$$d = -\frac{3}{4} \ln\left(1 - \frac{4p}{3}\right). \tag{S8}$$

The synonymous (d_S) and non-synonymous divergences (d_N) were calculated in a similar way by replacing π_{ij} with π_{ij}^S and π_{ij}^N in (S7), respectively.

Finally, to calculate the diversity within each species separately we take the sum in (S6) over I_α or I_β only.

We calculated the nucleotide diversity across the core genes in α and β and compared the variation in each along the genome (Fig. S23). Within α we found π varied wildly by more than an order of magnitude over short distances of a few tens of kbp (dark orange line in Fig. S23). By constrast, the diversity within β was much more homogenous and fluctuations were rarely larger than around a factor of two (dark

blue line Fig. S23). However, within both species we found that the diversity between individual pairs of genomes was highly variable, with may pairs of genomes frequently being nearly identical at certain loci, but having otherwise typical divergences genomewide (light orange and blue lines in Fig. S23). This is consistent with our conclusion from the main text that genomewide linkage in both species is very weak (see also Sec. IV).

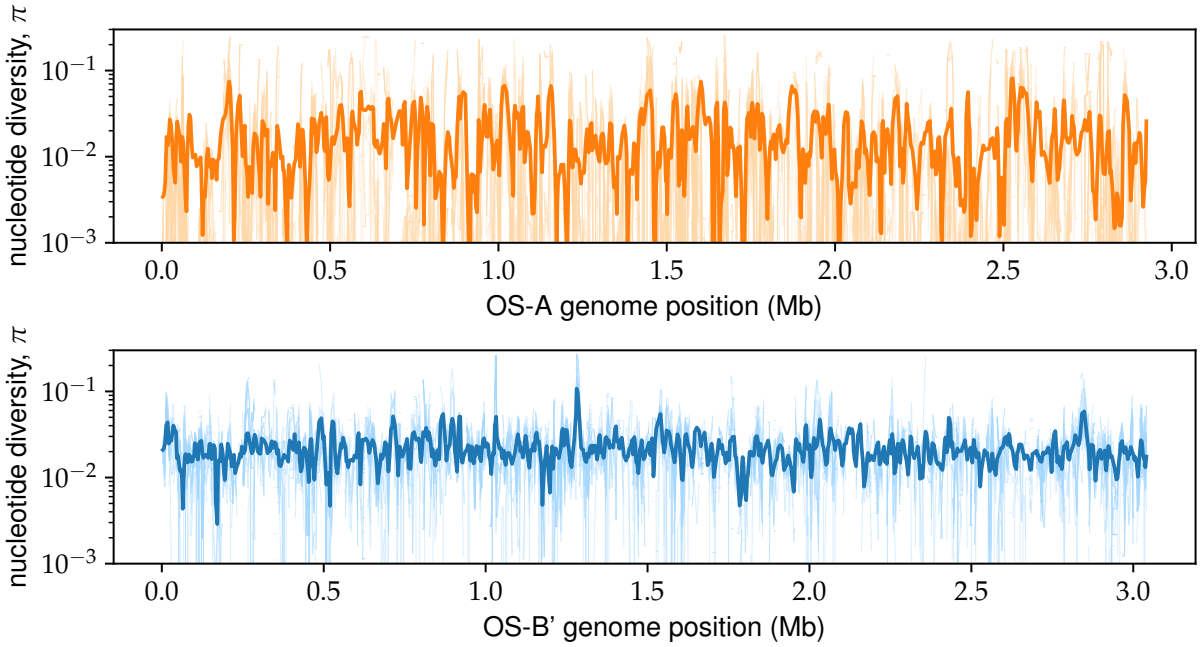

Figure S23: α and β show distinct patterns of within-species diversity. The two panels show the nucleotide diversity π within α (top) and β (bottom) at all core genes ordered according to their respective reference genomes. The solid line shows the value of π estimated from all sequences, with values calculated from a subsample of 10 randomly chosen sequences shown in lighter colors. Both lines were smoothed by averaging over a sliding window of 5 genes with a step size of 2 genes.

VII.2 Diversity heterogeneity along core genome

There are two plausible explanations for the heterogeneity within α found in S23. The low diversity genes could be due to local sweeps similar to the ones that formed the diversity troughs shown in Fig. 2 in the main text. Under this scenario, the high diversity genes would represent the typical diversity within α that has accumulated since its common ancestor and would be comparable to that of β. Alternatively, the low diversity genes could represent the typical diversity within α, with the high diversity genes being the result of hybridization with other species.

To distinguish between these two possibilities we developed a method to separate low diversity from high diversity genes within α. Specifically we calculated the fraction $f_l^S = P_l^S/L_l^S$ of 4D sites at locus l that were polymorphic within the α alignment. Here P_l^S is the number of synonymous polymorphisms and L_l^S the number synonymous sites at that gene. By using 4D sites we naturally control for differences in gene conservation. If mutation rates across loci do not vary too much, we expect the distribution of f_l^S across genes to be given by a Poisson distribution. The mean of this Poisson distribution can be fitted

from the fraction of sites that are not polymorphic

$$P(k=0|\mu) = e^{-\mu} = 1 - \frac{\sum_l P_l^S}{\sum_l L_l^S}, \tag{S9}$$

from which we get the following estimate for the mean:

$$\hat{\mu} = -\ln\left[1 - \frac{\sum_l P_l^S}{\sum_l L_l^S}\right]. \tag{S10}$$

When we compared this prediction to the data for α we found that the Poisson distribution gave an excellent fit for $f_l^S \lesssim 0.05$ (902 core orthogroups), but there was a minority of loci (654 core orthogroups) with f_l^S much larger than predicted from the Poisson (Fig. S24, left panel). Note that the agreement in the inset of Fig. S24 (left panel) is based on a single fitting parameter. Based on these results we chose a threshold of $f_l^S \leq 0.05$ to define the low diversity genes, with all other genes labeled as high diversity. To validate our partition, calculate the synonymous diversity π_S within each group. We found that our criterion produced a clear separation in the synonymous diversity, with low diversity genes having a mean $\pi_S \approx 1.7 \times 10^{-3}$, compared to $\pi_S \approx 3.4 \times 10^{-2}$ for the high diversity genes.

Within the genes classified as low diversity were many essential genes, such as genes from the ribosomal protein cluster. In addition manual inspection of the alignments of high diversity genes revealed that most were either due to gene allele transfers (Fig. 2) or long linkage blocks that were identical to the β or γ sequences (3). The simplest explanation for this pattern is then that the low diversity genes represent a genomic backbone shared by α cells from their common ancestor[3], while the high diversity genes are the result of hybridization with the other species. This result shows that molecular clock estimate of the time since the most common ancestor based on $T_{MRCA} = \pi_S/(2\mu\tau)$, where μ is the mutation rate per division and τ the generation time, should be interpreted with caution. In the case of α the estimate using the average π_S across the whole genome would be around an order of magnitude less than the true value, even without considering how mutation rates might change over such long time scales.

We performed a similar analysis for β but did not find any large peak near $f_l^S = 0$ (Fig. S25). Instead the distribution of f_l^S was close to unimodal but with significantly fatter tails compared to the Poisson. Given the excellent fit to the Poisson for α in the region where SNPs are purely due to mutations, this suggests that other mechanisms plat an important role in generating the diversity within β. The most parsimonious explanation is that the hybridization process we observed in α has covered most of the original β genome by this point, leading to a non-Poisson distribution of f_l^S similar to the high diversity regions of α. Further evidence in favour of this hypothesis is the presence of SNP blocks across a large fraction of β core genes, as discussed in the main text and VI.

VII.3 Contribution of hybridization to genetic diversity

The large number of SNP block found throughout the genomes of both α and β show that hybridization is an important source of genetic diversity for the *Synechococcus* population. However, because of the requirement for perfect linkage within SNP blocks, the total number of SNPs that are the result of hybridization is likely much higher than the one we infer from our algorithm. To obtain quantitative

[3]Note that this does not imply that the diversity within this backbone can be represented by an asexual phylogenetic tree, as is commonly assumed in the literature (*34*). Indeed, as shown in Sec. IV, we still find significant linkage decrease in the low diversity regions of the α genome.

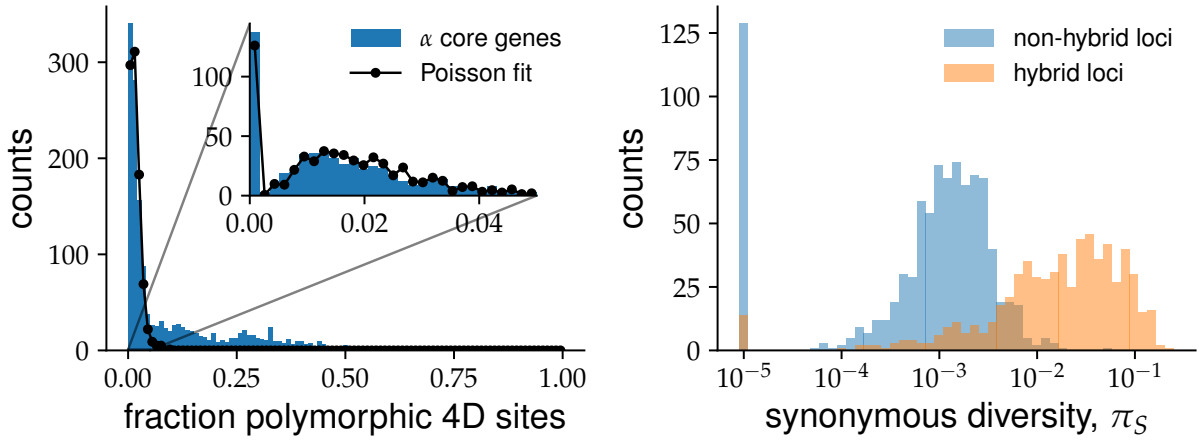

Figure S24: Core genome diversity within α has bimodal distribution. (Left) Distribution of the fraction of polymorphic four-fold degenerate (4D) sites for each gene (blue histogram) is shown together with Poisson fit based on the number of genes without any polymorphic (black line and dots). Inset shows zoomed in region of low-diversity genes ($< 5\%$ polymorphic 4D sites) together with the fit for clarity. (Right) Shows histogram of synonymous nucleotide diversity π_S for both low-diversity (blue) and high-diversity (orange) genes. Genes without polymorphic 4D sites were arbitrarily assigned a value $\pi_S = 10^{-5}$ for illustration purposes.

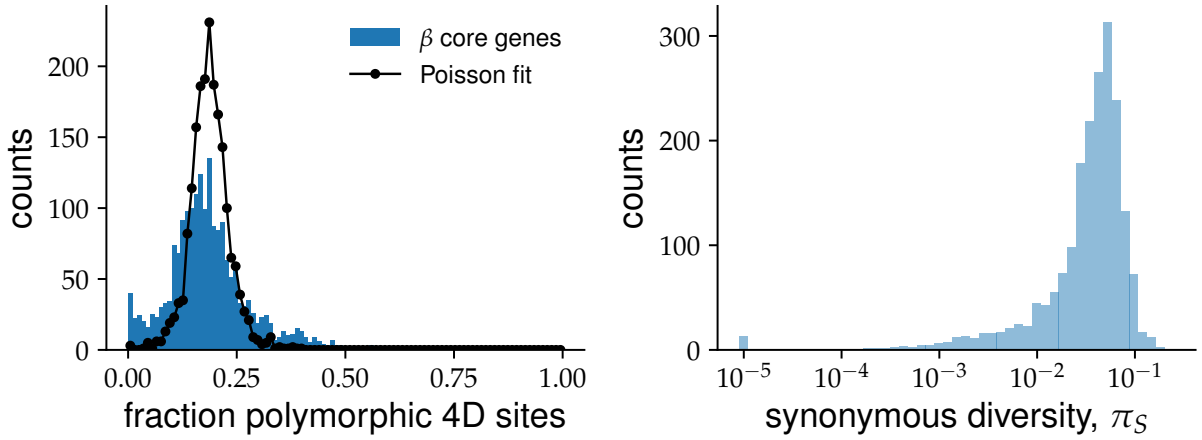

Figure S25: Core genome diversity within β is relatively homogeneous. (Left) Shows distribution of the fractions of polymorphic 4D sites for each gene (blue histogram) together with a Poisson fit based on the mean across all genes. (Right) Shows histogram of synonymous nucleotide diversity π_S. Genes without polymorphic 4D sites were arbitrarily assigned a value $\pi_S = 10^{-5}$ for illustration purposes.

estimates of the total contribution from hybridization to the diversity we again use the α as a control. As shown previously, only $\approx 5\%$ of the total diversity in the high-diversity loci from α can be attributed to mutations. These loci contained $603,588$ sites and $64,462$ SNPs, of which $10,391$ SNPs were found inside SNP blocks. Thus only around 17% of the SNPs that are likely the result of hybridization, or hybrid SNPs, are included in the SNP blocks.

Within the β core genome, we have $1,591,476$ sites and $146,906$ SNPs, of which $14,819$ were within

SNP blocks. Assuming a similar ratio of hybrid SNPs to SNPs in SNP blocks as in α, we estimate that around $90,000$ or approximately 60% of SNPs in β are the result of hybridization. This is consistent with β having evolved over longer time compared to α. However, our knowledge of how multi-locus SNP patterns change over the course of evolution is limited so the estimates for β should be interpreted with caution. In particular, if larger segments continue to be broken up due to recombination, it is likely that the fraction of hybrid SNPs within blocks would decrease over time for a constant minimum block size l_{min} (see Sec. VI). Thus the overall contribution of hybridization to the β genetic diversity may be even higher.

VII.4 Geological constraints on microbial evolution

Combining genetic diversity analysis with evidence from the geological record of the Yellowstone caldera allows us to make inferences about the evolutionary history of the Synechococcus population. If we assume that most of the fixed variants between species are the result of mutation accumulation, then we can use the typical synonymous divergence between species to infer the time since the most recent common ancestor for the whole population. Typical bacterial mutation rates for natural populations are around $\mu \sim 10^{-9.5}$ per bp per generation (*38*). Using this rates as a baseline and assuming generation times of $\tau \sim 1 - 10$ days gives an estimate of $T_{MRCA} \sim 10^{6.5}$–$10^{7.5}$ years (*12*).

For our argument, the lower bound is the more important one since it determines whether we can exclude the single-colonizer hypothesis. A T_{MRCA} below $10^{6.5}$ years would imply a shorter division time or higher mutation rate than our estimates. Our lower bound of one day for the division time is around the same as the fastest division rate observed in the laboratory. If the population is in steady state, as it appears to be, the division time sets the maximum generation time, which can be attained if the death rate is equally high. If the death rate is lower than this, other factors would limit cell division and the generation times would be longer. Taken together, we can be fairly confident that a generation time of one day is a hard lower bound.

The mutation rate is less certain since we do not have any reliable measurements and experimentally measured mutation rates might not be a good predictor of substitution rates over evolutionary times (*39*). We can instead use other bacteria to estimate a range of values. In wild-type E. coli, mutation rates of $\approx 2 \cdot 10^{-10}$ per bp per generation have been measured in laboratory conditions (*40*), but mutator strains with rates that are up to 100 times higher are also known (*41*). In natural populations, rates as high as $4 \cdot 10^{-9}$ have been inferred using longitudinal sequencing (*42*). If we use $10^{-8.5}$ mutations per bp per generation as an upper bound, we would infer a lower bound of $T_{MRCA} \sim 10^{5.5}$ years, which very narrowly overlap with the time since the foramtion of the Yellowstone caldera. Thus, while it is not inconceivable that the three species could have diverged within the caldera, such a scenario would requires all the relevant parameters to be at the limit of currently known values. In addition, such a scenario would require additional hypotheses to explain why the strains initially diverged into distinct species, with seemingly little genetic exchange between them (*12*), and then only much later started to hybridize. Taken together, current evidence suggests the most likely scenario is that α, β, and γ diverged from a common ancestor before the formation of the caldera around $640\,000$ years ago.

VII.5 Upper bound on mixing time between springs

By using the diversity within α as a proxy for time, we can obtain estimates on the time scales of spatial mixing that influences the diversity within the population. To do this, we calculated the pairwise distances in the low diversity regions of the α genome between pairs of cells in each spring separately and pairs of cells from different springs. We found that the distribution all three distributions were similar (Fig. S26).

The mean of the distribution for cells Mushroom Spring was slightly lower than those from the Octopus spring ($\pi_S = 8.6 \cdot 10^{-4}$ vs $\pi_S = 1.1 \cdot 10^{-3}$) and this difference could not be explained by difference in sample sizes between the two springs. However, both values were very close to the mean of cells from different springs ($\pi_S = 1.0 \cdot 10^{-3}$). This suggests that the time scale of spatial processes relevant for the long-term evolution of the population, τ_m is much shorter time scales than T_{MRCA} for α. Taking $T_{MRCA} \sim 3 \cdot 10^3$ years as a reference point sets an upper bound of $\tau_m \ll 3 \cdot 10^3$ years.

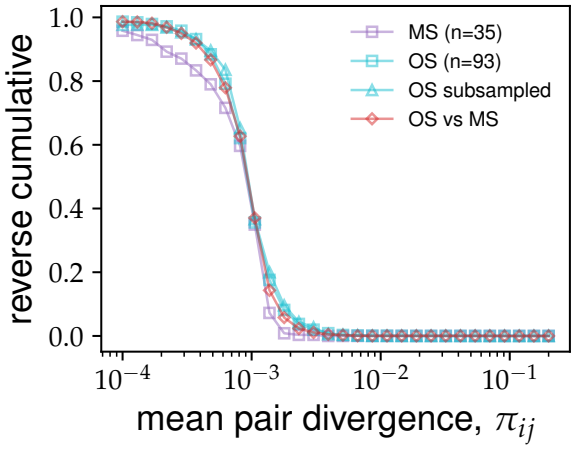

Figure S26: Non-hybrid regions of α have similar levels of diversity within and between springs. The distribution of pairwise synonymous diversity was calculated across non-hybrid genes from the alpha α population for different subgroups of cells. Square symbols show the distributions for pairs of cells from the same spring (purple for Mushroom Spring and cyan for Octopus Spring). The cyano triangles show the distribution for a random subsample of cells from the Octopus Spring of the same size as the Mushroom Spring sample ($n = 35$) for comparison. The red diamonds show the distribution for pairs of cells from different springs. Non-hybrid genes were determined using a 5% cutoff in the fraction of polymorphic synonymous sites (see text).

VII.6 Proposed scenario for Synechococcus evolution

We propose the following scenario that is consistent with our previous discussion. First we argued that a small number (3-6, but possibly more) of Synechococcus populations that had been evolving separately for $\sim 10^6 - 10^7$ years independently colonized the Yellowstone caldera after its formation 640 000 years ago. After colonization, they started to hybridize leading to a gradual mixing of their genomes (Fig. 4A). This can be most clearly seen in the α species, where the diversity in core genes that do not show signs of hybridization is more than an order of magnitude less that those that do. Moreover, genes that have undergone hybridization account for the majority of diversity within α. Within β we do not observe such heterogeneity, but we find SNP blocks are even more prevalent than in α (Table S2). The simplest explanation for this pattern is that β has been undergoing hybridization for longer and by now all of its core genome has been mixed with other species. Consistent with this explanation, we find that only a small fraction of SNP blocks within β can be traced back to α (Table S2). Most of the β SNP blocks cannot be traced to either α or γ and are likely the result of a combination of hybridization with species that are not currently within the Lower Geyser Basin, have abundances lower than our detection threshold, or have gone extinct.

VIII. Variation in hybridization across samples

In this section we describe how we quantified associations between genomic features and populations from different samples. Our primary focus was on differences in hybridization patterns so we analyzed the variations in hybrid gene and SNP block haplotype frequencies across different samples and locations. Given the small number of samples in our study, we limited the number of statistical tests we performed in order to avoid false positives. For each of the two main species, α and β, we tested for the effect of sample location and sample composition separately, as described below.

VIII.1 Variation in SNP block haplotype frequencies

We considered two distinct hypotheses that could lead to associations between block haplotypes and environmental conditions from which the sample was taken. First, we considered if the community composition was associated with specific haplotypes by comparing samples in which a single species was present to samples which contained a mixture of α and β. For α, we chose the 48 cells from pure-α sample OS2005 as the test group and the 45 α cells from MS2004, MS2006, and OS2009 as the control. For β, we chose the 44 cells from the pure-β sample MS2005 as the test group and the 46 β cells from MS2004 and MS2006 as the control. We denote these two tests as "α-composition" and "β-composition" tests, respectively. Second, we considered the hypothesis that sample location was associated with specific haplotypes. For this test we divided all cells from each species into two groups labeled MS and OS, based on the location from which the sample was taken. These tests are denoted by "α-location" and "β-location", respectively.

We used the following permutation test to find block haplotypes that were significantly associated with each environmental variable. We first removed all blocks in which the minority haplotype was below an arbitrary cutoff frequency, set at 5 sequences. This was done in order to avoid tests which were statistically underpowered. We then calculated the two-sided p-value for the association for each block individually using Fisher's exact test.

Test	Total blocks	Blocks tested	Significant ($p^{min}_{rand}/p > 1$)	Highly significant ($p^{min}_{rand}/p > 10$)	Largest p^{min}_{rand}/p
α-composition	1442	899	40	11	22.1
α-location	1442	899	3	1	188.6
β-composition	2153	1224	2	0	4.99
β-location	2153	1224	19	2	38.0

Table S3: Result of tests for associations between SNP block haplotypes and two environmental variables related to the samples. The rows show the results of the different tests described in the main text. The p-value p for each block was determined using Fisher's exact test and minimum p-value from the randomized control p^{min}_{rand} was used to correct for false-discovery rates (see main text for details). The first four columns show total number of blocks in each category. The last column shows the largest value of the p^{min}_{rand}/p^{max} (most significant association) for each test.

Because the sampling depth for each block is variable due to variations in coverage, standard methods that correct for the false-discovery rate are not well-suited in our case. We instead the following procedure to determine significant associations. We randomized the composition of the test and control cell groups for each test, while keeping the total number within each group fixed. We then recalculated the p-values using the same procedure as before and recorded the lowest p-value across all blocks p^{min}_{rand}. We used the

ratio p_{rand}^{min}/p between the minimum p-value in the randomized control and the p-value in the test group as a measure of significance of the association. We arbitrarily set a threshold value of $p_{rand}^{min}/p_t = 10$, above which associations were deemed highly significant. The results of these tests are summarized in Table S3.

Overall, we found very few blocks that were associated with either the sample composition or location. This is consistent with the results from the main text (3C), which showed that the block haplotypes are close to unlinked even on distances comparable to the whole genome. The largest number of associations was with the sample composition in the α species, consistent with the previously-mentioned results, which showed a small but statistically significant increase in linkage within the α species compared to the β. Nevertheless, the overall fraction of blocks whose haplotypes were strongly associated with the sample composition was $\approx 1.2\%$ and still small. We therefore concluded that neither the composition of the population within samples or the location have a large effect on the hybridization patterns revealed by the SNP blocks. A complete table with the results of these tests can be found in the supplementary data tables.

VIII.2 Variation in hybrid gene transfers

We also examined variations in the rate of transfers of whole gene alleles across species. The lower overall number of such events compared to the SNP blocks make statistical tests for associations much more likely to yield false positives in this case. As a result, we limited the analysis to a qualitative comparison of the number of different types of transfers across samples shown in Fig. S27. We did not observe any systematic variations with either sample location or sample composition for either species.

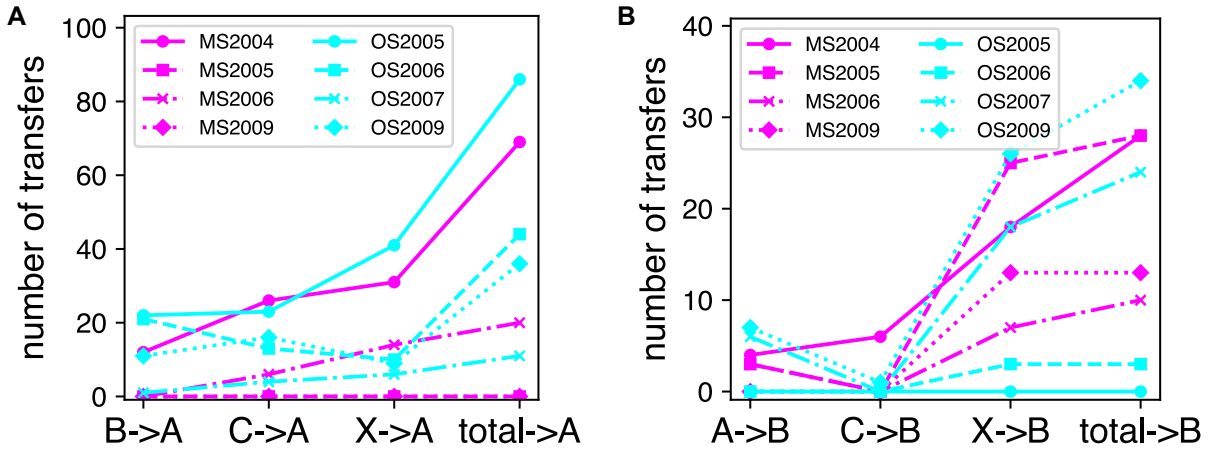

Figure S27: Numbers and types of gene transfers between species are similar across samples. The number of hybrid gene transfers from each donor found in α (**A**) and β (**B**) cells across all eight samples in the data. The labels A, B, and C, denote α, β, and γ, respectively, while X denotes other Synechococcucus clusters.

IX. Data quality control

Beyond the processing steps outlined in Sec. I.2, we also explicitly verified that the hybridization patterns we observe cannot be explained by sequencing and genome assembly artifacts. Specifically, we tested if misassembly of reads from distinct alleles of the same gene could explain the presence of hybrid sequences. We considered two main sources of such reads: cross-contamination from neighboring reaction wells during whole-genome amplification (WGA) and the presence of paralogs horizontally transferred from other species.

Distinguishing between hybridization and artifacts is hard because both can lead to similar signals. Here, structure of data provides us natural controls. First, α has genes with low diversity and little sign of hybridization and high-diversity genes with extensive hybridization. We can use former to establish baseline for genome coverage and check for deviations in the later. If the hybridized genes are caused by misassemblies of actual between α genes and contamination from β or between different alleles of the same gene within α cells, we should see an increase in the coverage at these loci. Second, the variability in species composition across samples provides controls for cross-contamination between reaction wells. Specifically, OS2005 contained 48 cells, all classified as α, while MS2004, MS2006, and OS2009 together contained 45 α cells amplified on plates with mixtures of α and β cells in similar proportions (Fig. S28). We will refer to the former as the *control SAGs sample* and the later as the *test SAGs sample*. We used these two groups as a measure of the contribution of cross-contamination from neighboring β cells to the observed hybridization in α.

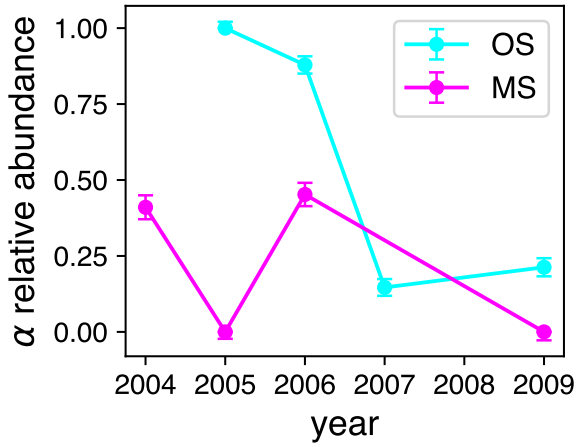

Figure S28: Species composition is highly variable across samples.

We first divided the core α OGs into test and control groups as follows. We first divided OGs into low- and high-diversity groups based on the probability that synonymous sites are segregating in the sample using a cutoff at 0.05. From the low-diversity OGs, we chose the subset that were grouped into α and β, or α, β, and γ clusters during the pagenome construction (approximately 300 each out of a total of $\sim$ 800). These OGs were nearly clonal in the α sample and showed little no evidence of hybridization in our entire data. These OGs will form the *control OG group*. From the high-diversity OGs, we selected those where all sequences were grouped together in a single mixed-species cluster ($\sim$ 300 out of a total of $\sim$ 600). These include the cases with the most extensive hybridization and thus formed the *test OG group*.

We validated that core gene coverage can be described by a random process in which each gene has a fixed probability p of being present in the assembly. We fitted the baseline probability p using the low-diversity genes from the OS2005 α cells. Many of these genes were clonal within the entire α sample and showed no evidence of hybridization with the other species in our sample. They thus provide a good baseline for the expected coverage in the absence of contamination. While the coverage among different SAGs was highly variable (Fig. S29A), we found that the distribution of coverages of different genes was well fitted by a Gaussian distribution with the mean and variance given by assuming each gene is an independent Bernoulli random variable with success probability $p \approx 0.29$. We tested the independence assumption by calculating the covariance between the presence of different genes (Fig. S29C). Under our very simple model, any correlations between the coverage at different loci would be purely due to statistical fluctuations caused by finite sample sizes. The distribution of diagonal (C_{aa}) and off-diagonal (C_{ab}) components of the covariance matrix can then be calculated exactly given p. We compared the values obtained from the data with these predictions and found a very good agreement (Fig. S29D).

Having validated our gene presence model, we can now use it to perform simple but quite powerful test for different artifacts. Consider the extreme case where all of the hybrid genes we observe are the result of misassemblies of reads from two distinct alleles, one from α and one from either β or γ, from the same reaction well. Note that for this argument it is unimportant whether the alternative allele is due to contamination or genuine horizontal transfer that resulted in two distinct, unmixed copies of the gene. In either case, the probability of observing the gene in our assembly would be $P_{obs} = p(2-p) \approx 0.5$, which is significantly higher than for a true single-copy gene. We tested for such an increase by comparing the distribution of coverages among α cells in the OS2005 sample of high-diversity genes that formed a single mixed cluster across all sequences in our data (see SI, Sec. I.2) and low-diversity genes with distinct α and β or α, β, and γ clusters. We found the two distributions were very similar to each other (Fig. S30A), with a small but statistically significant increase in the mean coverage at low-diversity genes with distinct species clusters ($p_{control} \approx 0.3$ vs $p_{test} \approx 0.28$). The difference is in the opposite direction from the one expected from the misassembly hypothesis, but it corresponds to less than an extra copy of a gene on average and is unlikely to have any consequence for our conclusions.

Note that the misassembly hypothesis above would also predict a specific pattern for the the sequences within α, with a large fraction $2(1-p)/(2-p) \approx 0.83$ having either the pure α or the pure β alleles, and $p/(2-p) \approx 0.17$ containing a mixture. We do not find evidence of such patterns in our data. More importantly, the 0.17 fraction is an upper bound on the maximum frequency of mixtures that can be explained by through misassembly as any lower amounts of contamination would lead to fewer chimeric sequences. Note that both of the above tests take advantage of the fact that the typical genome coverage is low.

Using a similar method to the one described above we can also check for contributions from cross-contamination between reaction wells on the formation of chimeric sequences. Here, we compared the coverage distribution of mixed-species orthogroups in the mixed $\alpha - \beta$ samples described previously with the control group from the all α sample OS2005. If cross contamination with β was a significant contributor to the presence of mixed-species orthogroups one would expect a large difference between the two distributions. We did observe a large systematic difference between the two groups in the overall coverage p across all genes (data not shown), which was likely due to batch effects in the amplification and sequencing. We therefore compared the mean-centered distributions between the two groups and found no difference between them and the fit to the independent presence-absence model from before (Fig. S30B).

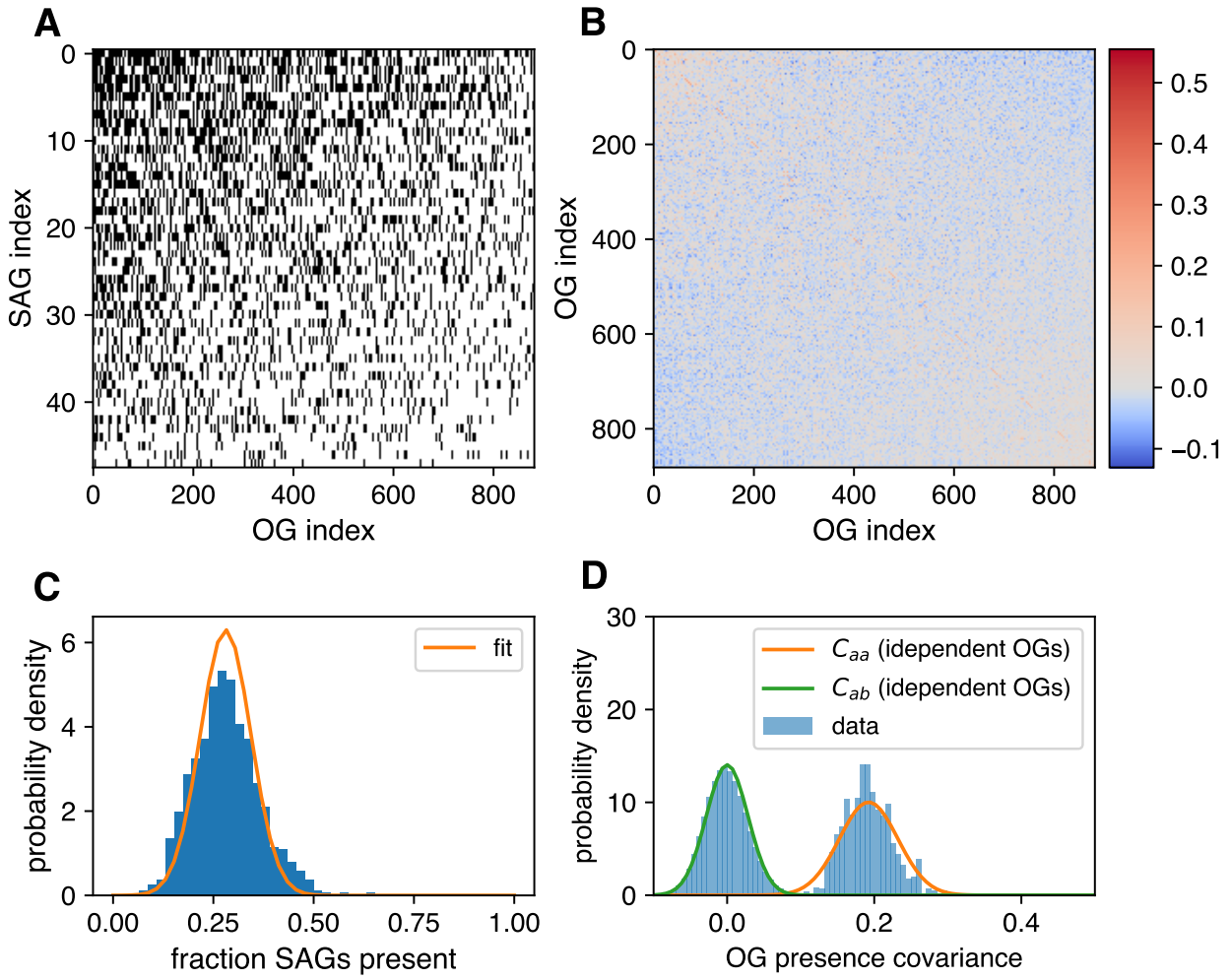

Figure S29: Genome coverage across core genes is well-explained by independent presence-absence model. **(A)** shows the presence-absence matrix P_{ia} of low-diversity OGs in the control SAG sample OS2005 (see main text for definitions). Rows and columns were ordered in decreasing order of the sums across the columns and rows, respectively. **(B)** shows the correlation matrix between OG presences defined as $C_{ab} = N^{-1}\sum_{i=1}^{N}\langle(P_{ia} - \overline{P_i})(P_{ib} - \overline{P_i})\rangle$, where $\overline{P_i} = N^{-1}\sum_{a=1}^{L} P_{ia}$. Note the use of the mean abundance for each SAG rather than $p = (NL)^{-1}\sum_{i,a} P_{ia}$ to account for the large variability in coverage between SAGs. **(C)** shows the distribution of OG presences $\overline{P_a} = N^{-1}\sum_{i=1}^{N} P_{ia}$. The solid orange line shows the Gaussian fit under the assumption that each gene is a Bernoulli random variable with success probability p. **(D)** shows the distribution of the entries in the covariance matrix from (B) together with the predictions for the diagonal (orange) and off-diagonal (green) components under the assumption that each gene is independent and present in the final assembly with probability p. Note the value of p was obtained from panel (C) and there are no free parameters for the predicted curves.

Beyond the indirect presence of hybridization given by the presence of mixed-species orthogroups, we also tested for any effect of cross-contamination on the presence of linkage blocks. We applied the same algorithm described in Sec. VI to the subsample of 48 α cells from OS2005 and compared the results with those from random samples of α cells from samples with mixtures of α and β cells. We found that the number of blocks detected in OS2005 was similar to the bootstrapped distribution from mixed $\alpha-\beta$ samples (Fig. S31A). We also found the distribution of block lengths and divergences between block haplotypes to be similar as well (Fig. S31BC).

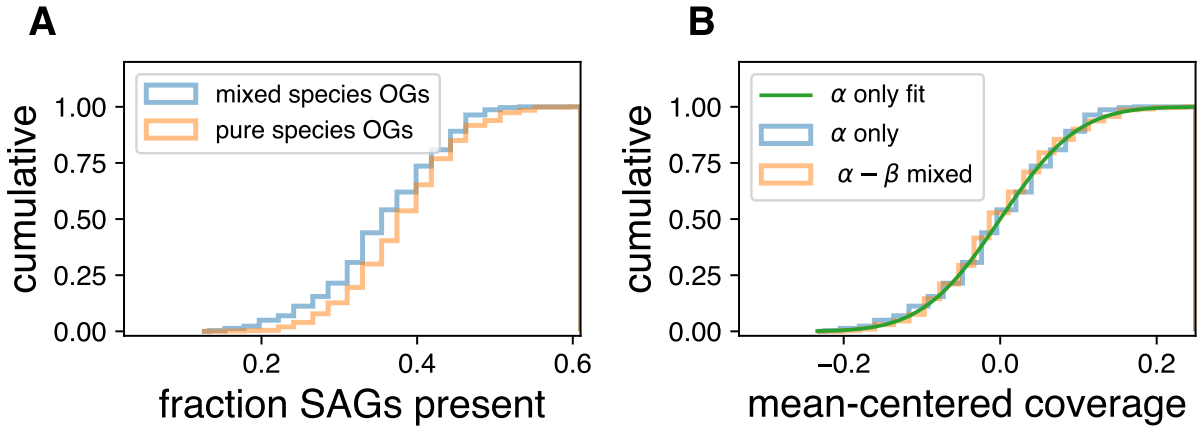

Figure S30: Variability in core gene presence is inconsistent with observed hybridization being dominated by artifacts. (A) compares the presence distributions among the control SAG sample of the test OG group (blue) with the control group (orange), as defined in the main text. **(B)** compares the presence distributions of the test OG group in the control (blue) and test (orange) SAG samples. The green line shows the fit from S29C. All distributions were centered at zero to control for batch effects (see main text).

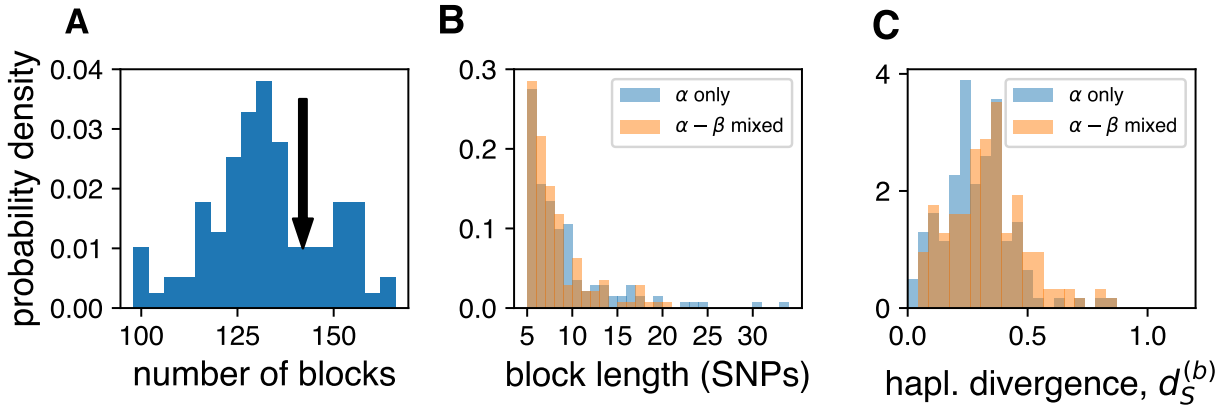

Figure S31: Linkage blocks are robust to large variations in species composition across samples. (A) Bootstrapped distribution of the number of linkage blocks detected in alignments across 4-fold degenerate sites. The distribution was obtained by randomly choosing groups of 48 SAGs from samples containing mixtures of α and β SAGs (MS2004, MS2006, OS2006, OS2007, and OS2009). The arrow shows the number of blocks detected using the 48 SAGs from OS2005, which only contained α SAGs. Comparison of the distributions of block lengths (**B**) and divergence between the two haplotypes from each block (**C**) are shown in the other two panel. OS2005 (shown in blue) is compared to a randomly chosen sample from panel A (shown in orange).

Based on the evidence presented above, we concluded that cross-contamination between reaction wells and genome assembly artifacts do not contribute significantly to the presence of mixed-species orthogroups or linkage blocks. While we cannot definitively exclude that some of the sequences we observe are the result of artifacts, the evidence above shows that they are rare enough to not have a significant effect on our main conclusions.

References

1. C. Rinke *et al.*, *Nature Protocols* **9**, 1038–1048, ISSN: 1750-2799 (May 2014).
2. R. Stepanauskas *et al.*, *Nature Communications* **8**, 84, ISSN: 2041-1723 (July 2017).
3. C. Rinke *et al.*, *Nature* **499**, 431–437, ISSN: 1476-4687 (July 2013).
4. E. V. Kriventseva, N. Rahman, O. Espinosa, E. M. Zdobnov, *Nucleic Acids Research* **36**, D271–D275, ISSN: 0305-1048 (Oct. 2007).
5. D. M. Emms, S. Kelly, *Genome Biology* **16**, 157, ISSN: 1474-760X (Aug. 2015).
6. W. Ding, F. Baumdicker, R. A. Neher, *Nucleic Acids Research* **46**, e5–e5, ISSN: 0305-1048 (Jan. 2018).
7. A. J. Page *et al.*, *Bioinformatics* **31**, 3691–3693, ISSN: 1367-4803 (Nov. 2015).
8. A. J. Enright, S. Van Dongen, C. A. Ouzounis, *Nucleic Acids Research* **30**, 1575–1584, ISSN: 0305-1048 (Apr. 2002).
9. T. R. Gibbons, S. M. Mount, E. D. Cooper, C. F. Delwiche, *BMC bioinformatics* **16**, 1–14, ISSN: 1471-2105 (2015).
10. K. Katoh, D. M. Standley, *Molecular Biology and Evolution* **30**, 772–780, ISSN: 0737-4038 (Apr. 2013).
11. M. N. Price, P. S. Dehal, A. P. Arkin, *PLOS ONE* **5**, e9490 (Mar. 2010).
12. M. J. Rosen, M. Davison, D. S. Fisher, D. Bhaya, *PLOS ONE* **13**, e0205396 (Nov. 2018).
13. P. Virtanen *et al.*, *Nature Methods* **17**, 261–272, ISSN: 1548-7105 (Mar. 2020).
14. C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, S. Aluru, *Nature Communications* **9**, 5114, ISSN: 2041-1723 (Nov. 2018).
15. F. M. Cohan, *Annual Review of Microbiology* **56**, 457–487, ISSN: 0066-4227 (Oct. 2002).
16. C. Fraser, E. J. Alm, M. F. Polz, B. G. Spratt, W. P. Hanage, *Science* **323**, 741 (Feb. 2009).
17. P. Yarza *et al.*, *Nature Reviews Microbiology* **12**, 635–645, ISSN: 1740-1534 (Sept. 2014).
18. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *Journal of Molecular Biology* **215**, 403–410, ISSN: 0022-2836 (Oct. 1990).
19. G. W. Tyson *et al.*, *Nature* **428**, 37–43, ISSN: 1476-4687 (Mar. 2004).
20. I. Lo *et al.*, *Nature* **446**, 537–541, ISSN: 1476-4687 (Mar. 2007).
21. M. Wu, J. A. Eisen, *Genome Biology* **9**, R151, ISSN: 1474-760X (Oct. 2008).
22. M.-Y. Chen *et al.*, *The ISME Journal*, ISSN: 1751-7370 (Sept. 2020).
23. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, *Bioinformatics* **25**, 1972–1973, ISSN: 1367-4803 (Aug. 2009).
24. C. Quast *et al.*, *Nucleic Acids Research* **41**, D590–D596, ISSN: 0305-1048 (Jan. 2013).
25. D. R. Meyer-Dombard *et al.*, *Environmental Microbiology* **13**, 2216–2231, ISSN: 1462-2912 (Aug. 2011).
26. B. Langmead, S. L. Salzberg, *Nature Methods* **9**, 357–359, ISSN: 1548-7105 (Apr. 2012).
27. H. Li *et al.*, *Bioinformatics* **25**, 2078–2079, ISSN: 1367-4803 (June 2009).

28. A. R. Quinlan, I. M. Hall, *Bioinformatics* **26**, 841–842, ISSN: 1367-4803 (Mar. 2010).

29. T. Ohta, M. Kimura, *Genetics* **63**, 229–238, ISSN: 0016-6731, PMID: 5365295 (Sept. 1969).

30. R. R. Hudson, *Genetics* **109**, 611 (Mar. 1985).

31. Y. S. Song, J. S. Song, *Theoretical Population Biology* **71**, 49–60, ISSN: 0040-5809 (Feb. 2007).

32. N. R. Garud, B. H. Good, O. Hallatschek, K. S. Pollard, *PLOS Biology* **17**, e3000102 (Jan. 2019).

33. M. J. Rosen, M. Davison, D. Bhaya, D. S. Fisher, *Science* **348**, 1019 (May 2015).

34. T. Sakoparnig, C. Field, E. van Nimwegen, *eLife* **10**, ed. by A. Nourmohammad, A. M. Walczak, e65366, ISSN: 2050-084X (Jan. 2021).

35. P. W. Messer, D. A. Petrov, *Trends in Ecology & Evolution* **28**, 659–669, ISSN: 0169-5347 (Nov. 2013).

36. T. H. Jukes, C. R. Cantor, in *Mammalian Protein Metabolism*, ed. by H. N. Munro (Academic Press, Jan. 1969), pp. 21–132, ISBN: 978-1-4832-3211-9.

37. M. Nei, T. Gojobori, *Molecular Biology and Evolution* **3**, 418–426, ISSN: 0737-4038 (Sept. 1986).

38. Z. Chen *et al.*, *Nature Ecology & Evolution* **6**, 183–194, ISSN: 2397-334X (Feb. 2022).

39. E. P. C. Rocha, *Molecular Biology and Evolution* **35**, 1338–1347, ISSN: 0737-4038 (June 2018).

40. H. Lee, E. Popodi, H. Tang, P. L. Foster, *Proceedings of the National Academy of Sciences* **109**, E2774–E2783 (2012).

41. P. D. Sniegowski, P. J. Gerrish, R. E. Lenski, *Nature* **387**, 703–705, ISSN: 1476-4687 (June 1997).

42. N. A. Moran, H. J. McLaughlin, R. Sorek, *Science* **323**, 379–382 (2009).