www.sciencemag.org/content/348/6238/1019/suppl/DC1

Supplementary Materials for

Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche

Michael J. Rosen, Michelle Davison, Devaki Bhaya,* Daniel S. Fisher*

*Corresponding author. E-mail: dbhaya@stanford.edu (D.B.); dsfisher@stanford.edu (D.S.F.)

Published 29 May 2015, *Science* **348**, 1019 (2015)
DOI: 10.1126/science.aaa4456

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S6
Tables S1 to S4
References

Materials and Methods

Sample collection and DNA processing: Mat core samples collected from Mushroom Spring (Lower Geyser Basin of Yellowstone National Park, September 2008) (*34*, *10*) were processed to isolate total DNA using a protocol modified from (*35*) to optimize DNA yield. Glass beads were added to frozen cores (150-212mm diameter), and subjected to three freeze-thaw cycles (30s in liquid nitrogen, 30-60s thaw at 50°C, vortex pulse). Cells were resuspended in 900µL cold TE buffer and homogenized with a bead-beater (Bio Spec) for 30s on high in a 4°C room. Lysozyme was added at 200µg/mL and incubated at 37°C for 35 minutes, followed by a 200µg/mL proteinase K incubation with 1% sodium dodecyl sulfate at 50°C for 60 minutes. DNA was recovered with a standard phenol:chloroform:isoamyl alcohol extraction at 40°C. The upper aqueous layer was divided equally between two new tubes, with 0.1M sodium acetate and 2.5V 100% ethanol added, and allowed to precipitate in a -20°C freezer overnight. DNA was pelleted by centrifugation at 14,000rpm for 30 minutes, washed with 70% ethanol, and allowed to air dry. Recovered DNA was resuspended in 125µL ddH_2O, quantified by nanodrop (Thermo Scientific), and visualized by gel electrophoresis to confirm results.

Selection of genes and genomic regions of interest: Genomic regions were selected for amplification based on several criteria or characteristics such as genes encoding proteins required for photosynthetic function; stress or temperature adaptation; defined metabolic pathways; specific KEGG annotations; and known phylogenetic markers (Table S4).

Generation of targeted degenerate primers and testing primer specificity: To generate PCR primers specific to regions of interest, but general enough to amplify *Syn* OS-A and *Syn* OS-B′, as well as diversity observed in an earlier metagenome data set (*13*), reads were first aligned to the *Synechococcus* genomes via BLASTn based on percent identity. The program Primer3 (*36*) was

slightly modified to design species-specific PCR primers capable of capturing within-population diversity. This was done with a combination of primer placement and the use of degenerate bases. Primer pairs were individually tested on DNA extracted from cultures of *Syn* OS-A and *Syn* OS-B′ as well as on total DNA extracted from mat samples. Approximately 65 of the primer sets produced a single band of the expected size. Primers that failed to amplify or generated multiple bands were subjected to further PCR protocol optimization, re-designed, or eliminated from further study.

Generation of a metagenomic amplicon library for deep sequencing: The 90 pairs of primers, as well as general bacterial 16S primers (*37*, Table 1) were used to perform individual PCR reactions in technical replicates with DNA extracted from mat samples. The amplified regions fell into two categories: Low Molecular Weight (LMW) regions with amplicon sizes suitable for direct input into 454 Titanium library preparation (27 samples, 460-748bp) and High Molecular Weight (HMW, 771-3007bp) (64 samples) regions with amplicons that required shearing to be within the acceptable input size range. To ensure random breakage, Covaris (Covaris S2, Massachusetts, USA) acoustic shearing was used to fragment HMW amplicons into an acceptable size range. Five replicates of sheared samples were separated on a 0.8% low-melt agarose gel at 100mV for 2 hours. The gel region corresponding to 500-700bp was excised and gel-purified using a Qiaquick Gel Extraction Kit standard protocol (Qiagen, Valencia, CA). Recovered DNA was quantified by Nanodrop. Detailed Covaris Shearing protocol as follows:

1. All HMW PCRs were assumed to have a concentration of 37ng/μL (the average concentration from LMW PCRs).
2. All 0.75-3kb amplicons from 50°C temperature were mixed to yield ~ 10μg/100μL for shearing.
3. Covaris specs: Target Base Pair 500, Duty Cycle 5%, Intensity 3, Cycles per Burst 200, Time 90s, Water Bath 22°C, in 100μL microtubes.

4. Five technical replicates were performed.

454 Library production with MID tagging and 454 Titanium sequencing: Oligonucleotides and recommended adaptors for *454* library construction were ordered from Integrated DNA Technologies (IDT, http://www.idtdna.com). *454* libraries were created for both HMW and LMW samples and were tagged with MIDs. Libraries were constructed based on the *454* manual. Sequencing was performed at the Stanford Genome Center. HMW and LMW libraries were pooled and sequenced on a full ungasketed plate with *454* Titanium technology. 441.625Mb of data was generated, consisting of 1,242,000 reads. Run statistics met or exceeded all quality checks.

Data preprocessing: 1,071,339 reads exactly matching the 50°C MID tag, `ACGAGTGCGT`, were identified and MID tags were removed. Reads with an exact match to one of the PCR primers at the 5′ or 3′ ends were identified, the primer sequences were removed, and reads reverse complemented if the primer match was only to their 3′ end. Reads shorter than 400nt and those containing Ns were discarded, and the remaining 160,033 reads were trimmed to 400nt. At least one read was obtained for 90 of the 91 amplicons. The 16S rRNA amplicon was short enough (438nt in the *Syn* OS-B′ genome) that, rather than trimming to 400nt, we retained only the 3532 reads that matched both primers. Sequences were then reverse complemented as necessary to put all in the same orientation.

Error correction and inference of alleles: We applied the *DADA* algorithm (*19*) to the collection of sets of 400nt reads originating from each primer. The *DADA* algorithm initially assumes that all reads are errors away from the most abundant sequence. It uses this assumption to train an error model of substitutions that are dominated by PCR. Using this error model, it then goes back and tests whether each sequence has an abundance that is statistically consistent with being an

error. If not, these sequences are inferred to be additional genuine alleles in the sample, and the error model is updated accordingly. The inference of more alleles and the updating of the error model parameters continue until convergence is reached. *DADA* relies on two significance threshold parameters when making decisions about whether to call sequences as genuine alleles rather than errors: Ω_a, which affects multiple-read sequences, and Ω_r, which affects primarily singletons. In this paper, we applied $\Omega_a = \Omega_r = 0.01$, with and without context-dependent error probabilities. The results were nearly identical, so the context-independent error model results were considered our inferred alleles. This resulted in 6738 inferred alleles across all the non-16S loci. The inferred error model is very close to the expected symmetry under the complementation of errors (e.g. Prob(A->C) ≈ Prob(T->G)) and the error probabilities for different types of substitutions ranged over two orders of magnitude from $2.8*10^{-5}$ for C->G to $2.2*10^{-3}$ for A->G. *DADA* was then applied separately to the 3532 16S rRNA reads under the same significance parameters ($\Omega_a = \Omega_r = 0.01$) and the error parameters inferred from the rest of the data set, i.e. without further updating of the error model, resulting in 289 16S rRNA alleles.

Given that this data set was sequenced with the *454* platform, substitution errors are expected to be predominantly generated during the PCR amplification step rather than during sequencing. This creates the potential for oversampling, i.e. generating more reads than there were initial DNA molecules that seeded the PCR process, which would result in correlations due to early round PCR errors, rather than the independence assumed by *DADA*. However, a simple estimate suggests that oversampling is not an issue. 35ng of template DNA was used in each PCR reaction, so the typical sample coverage of the genome was $35*10^{-9}$g * ($6.02*10^{23}$nt / 330g) / ($3*10^{6}$nt / genome) ~ $5*10^{6}$x, at least a factor of 1000 above the typical 1000x coverage of the data even if only ~ 25% of the sample DNA is from *Synechococcus*, as was the case for a PCR-free metagenome from the same population (*13*).

454 sequencing is prone to indel errors, especially associated with homopolymer strings (*38*). However, *DADA* does not model the process of indel generation. Instead, sequences that differ only by the presence of indels are lumped together. Examining the collection of reads mapped to the same alleles revealed at least a few dozen situations where a very long indel (almost surely not due to an error) was ignored, demonstrating that some number of alleles with genuine indels have been incorrectly clustered together with other reads. As our analysis does not make use of indels, this should not cause problems.

***d_n/d_s consistency check on inferred alleles*:** *DADA* has been shown to have low false positive and false negatives rates on several mock community training data sets (*19*, *39*). However, performance on rich, natural data is relatively uncharacterized. Thus, we paid particular attention to the possibility of false positives and performed a consistency check on the inferred alleles and errors, utilizing the fact that errors occur independently of whether they alter amino acids, while genuine diversity will depart from independence and is generally skewed towards being synonymous. Therefore, we studied the statistics of d_n and d_s, the rates of non-synonymous (protein-altering) and synonymous (silent) differences between the coding sequences of pairs of inferred alleles.

For convenience, we defined modified versions of d_n and d_s so that averaging over pairs of sequences, $\langle d_n/d_s \rangle$ would be equal to one if differences were due to errors rather than genuine biology. In contrast to standard definitions, our d_n and d_s are defined at the level of 3-letter codons rather than individual nucleotides. Because two codons are always either identical, synonymous (different but coding for the same amino acid), or non-synonymous (coding for different amino acids), focusing on codons rather than nucleotides eliminates the need for *ad hoc* definitions of synonymous and non-synonymous nucleotide mutations and sites (*40*).

Specifically, for a pair of coding sequences, we define:

- d_n, the ratio of the number of amino acid differences to the expected number due to errors, using the substitution error probabilities inferred by *DADA*.
- d_s, the ratio of the number of identical amino acids coded for by different codons to the number of these expected from errors, again using the substitution error probabilities inferred by *DADA*.

With these definitions, we analyzed 1933 *Syn* OS-A/*Syn* OS-B′ homologues and found an approximately linear relationship between d_n and d_s with a best fit of $d_n/d_s = 0.158$. Similarly, 149 pairs of alleles inferred by *DADA* with at least 20 reads each (consequently very likely to be genuine) and at least 10 amino acids of coding sequence were similarly best fit by $d_n/d_s = 0.138$. We therefore took $d_n/d_s \sim 0.15$ to be a reasonable prior for genuine biological diversity.

We then performed pairwise alignments of all alleles inferred by *DADA* to *Syn* OS-B′ and rejected those > 10% diverged. For the remainder, we extracted positions in the alleles that align to coding sequences in *Syn* OS-B′, gapping out entirely any codons containing one or more gaps. The use of the 10% screen is important because more distantly diverged sequences cannot always be reliably aligned, and misalignments can artificially inflate d_n/d_s estimates.

We next considered d_n/d_s between coding sequences of all pairs of *DADA*-inferred alleles as a function of nucleotide divergence. Well-diverged alleles may tend to have d_n/d_s values dominated by genuine biological SNPs, even when one of the alleles contains one or more errors (i.e. it is a false positive). In the presence of a high false positive rate, nearby allele pairs may contain many real/error pairs, and thus have d_n/d_s significantly elevated above our $d_n/d_s \sim 0.15$ biological prior. However, for alleles < 1% diverged, we find that $\langle d_n \rangle / \langle d_s \rangle \sim 0.14$ when the averages are weighted by the frequencies of the allele pairs (each pair of alleles, indexed by *i*, contributes a $d_{n,i}$

and $d_{s,i}$ that are averaged together via $\langle d_n \rangle = (\Sigma_i f_{i,1} * f_{i,2} * d_{n,i})/(\Sigma_i f_{i,1} * f_{i,2})$ and $\langle d_s \rangle = (\Sigma_i f_{i,1} * f_{i,2} * d_{s,i})/(\Sigma_i f_{i,1} * f_{i,2})$ where $f_{i,1}$ and $f_{i,2}$ are the frequencies of the two alleles in the *i*th pair), consistent with most or all of the diversity being genuine. If an average over allele pairs is taken without weighting, $\langle d_n \rangle / \langle d_s \rangle \sim 0.2$: that this ratio is larger reflects that slightly elevated d_n/d_s is correlated with lower allele frequencies.

Similarly, for singletons – often considered suspicious and thrown out during amplicon studies – we find $\langle d_n \rangle / \langle d_s \rangle = 0.15$ (with the averages here and below taken without frequency weighting) relative to their most similar allele. This ratio strongly weights well-isolated singletons that contain many SNPs. But even looking at singletons < 2% from some other allele, we find $\langle d_n \rangle / \langle d_s \rangle = 0.2$. Being singletons, such alleles are expected to often contain one or more SNPs due to errors that cannot be corrected, but they nonetheless must be dominated by genuine diversity to achieve such low d_n/d_s values.

Sequences in the raw data that were inferred by *DADA* to be errors ought to have d_n/d_s values near unity relative to the alleles they were considered to be errors away from. However, we find significant departures from this. For example, apparent singleton errors one coding sequence SNP away from their parent alleles have $\langle d_n \rangle / \langle d_s \rangle \sim 0.86$, suggesting that a significant fraction are likely to be real even though they are individually impossible to discriminate from errors. Given that there are 24482 such one-away singleton reads, this suggests the potential presence of multiple thousands of false negatives. For the 9082 two-away singletons, $\langle d_n \rangle / \langle d_s \rangle \sim 0.71$, suggesting that an even large fraction of these are real.

The ability to estimate accurate false negative rates for the error cloud requires more detailed knowledge of how the biological d_n/d_s for very rare alleles depends on their relationship to abundant ones, and there is little reason to believe that the $d_n/d_s \sim 0.15$ value discussed above has

much bearing here, as rare alleles near to abundant ones are excellent candidates to harbor deleterious mutations, which could drive up d_n/d_s. But note that if d_n/d_s values are in fact larger than 0.15 for such rare alleles, this implies that an even larger fraction of the inferred errors must be real to explain the sub-unity d_n/d_s values for low frequency errors. For the purposes of the analyses we carried out, the effects of missing a large number of singleton or doubleton alleles that differ by only one or a few SNPs from an allele with a much larger number of reads should not be substantial: we therefore erred on the side of caution, by calling alleles as real based on the significance parameters ($\Omega_a = \Omega_r = 0.01$) chosen to be small.

***Multiple alignments*:** The alleles inferred by *DADA* were multiply aligned with *MUSCLE* v3.8.31 (*41*) under default parameters. For all main cloud results presented in the manuscript, prior to *MUSCLE* each allele was pairwise aligned to the most abundant allele and screened out if > 10% diverged. *Syn* OS-B′ was included in all multiple alignments.

***Annotation of synonymous SNPs and SNP pairs*:** We discarded all columns of the *MUSCLE* alignments in which *Syn* OS-B′ contained intergenic sequence, so that only coding sequence remained. In codons where an allele contained either one or two gaps relative to *Syn* OS-B′, the entire codon was replaced by gaps. Then, the most common amino acid was identified at each codon. If this amino acid was two, three, or four-fold degenerate, then nucleotide frequencies at the third site were recorded. If six-fold degenerate (leucine, serine, arginine), nucleotide frequencies at both the first and third site were recorded. In cases where a site had more than two nucleotides segregating in the population, only the top two frequencies were kept and renormalized to add up to unity. This defined a set of synonymous SNP frequencies (with f = 0 for fixed sites) that are shown in Fig. 2B. For Figs. 2A, 2C, and 2E, we defined synonymous SNP pairs as follows: for each pair of synonymous SNPs defined in the way just described, we recorded the number of reads of the *AB*, *Ab*, *aB*, and *ab* haplotypes and divided by the total to get

the haplotype frequencies f_{AB}, f_{Ab}, f_{aB}, and f_{ab}. Note that reads matching the most common amino acid at one site but not at the other do not contribute to these frequencies (even though they did contribute to the single site frequency spectrum).

***Definition of linkage correlations*:** Figs. 2A and S3 show the inverse of $\langle r^2 \rangle = \langle (f_{ab} - f_a f_b)^2 / [f_a(1-f_a) f_b(1-f_b)] \rangle$ as a function of separation x. The averaging, "<>", is done over all pairs of synonymous sites, whether or not they are observed to be polymorphic. Pairs in which one or the other site was not polymorphic, so that f_A or $f_B = 0$, contribute $r^2 = 0$ to the average: this is appropriate because when one of the frequencies, say $f_b \to 0$, the numerator of r^2 decreases as f_b^2 whereas the denominator is proportional to f_b so that $r^2 \to 0$. Normalizing in this way (as opposed to averaging only over polymorphic sites) should produce results similar to the infinite population size average derived by *ER2* (*26*) as long as the sample size is large enough that $N \gg \rho x$: with $N \sim 10^3$ and $\rho \sim 10^{-2} - 10^{-1}$, this is valid well beyond the 300nt maximum separations that we probe.

***Unlinked and asexual null models for OS-B′ rank frequency spectrum*:** Figs. 3 and S6 contain null curves for unlinked and asexual models of the *Syn* OS-B′ and *Syn* OS-A genomes. These were constructed as follows.

The unlinked null curve:

1. At each locus, a random allele was drawn with probability proportional to its frequency in the population and the frequency recorded. These frequencies were ordered from largest to smallest, producing a rank ordered frequency spectrum.
2. Step 1 was repeated 10^5 times, and we recorded the simulated rank ordered frequency spectrum each time.
3. At each rank, we showed the central 95% of frequencies across the 10^5 simulated frequency spectra. Although there are some correlations between the frequencies across ranks for particular simulations introduced by the rank ordering itself – for example, if

we condition on having an anomalously large number of loci with very high frequency, then the higher rank loci will tend to have relatively high frequency *given their rank* – these correlations are weak and the central range characterizes the set of simulations well.

The asexual null curve:

1. A neutral, asexual phylogeny was generated using *ms* (*42*) with 15000 leaves.
2. At each locus, we simulated a set of 15000 sequences based on this phylogeny using the *seq-gen* software package with $\theta = 0.02$ (matching the overall π in the data) and a transition/transversion ratio of 10 under the Kimura 2-parameter substitution model (*43*).
3. At each locus, we randomly sampled a number of leaves equal to the read depth in the real data and pulled out the sequences associated with those leaves. We then collapsed these sequences down into a set of unique alleles and allele frequencies.
4. We chose a random leaf on the phylogeny to be the *genome leaf* and pulled out its simulated sequence at each locus. We then recorded the frequency of the closest simulated allele to this sequence at each locus and rank ordered these frequencies.
5. We repeated steps 1-4 multiple times. Because of correlations between loci and large fluctuations associated with the common phylogeny in each simulation, we presented 20 example simulations rather than showing a frequency range at each rank.

***Identification of putative chimeric sequences*:** The following *ad hoc* procedure was used to determine, for each allele, the two other alleles of which it appears most like a chimeric recombination:

1. Forward and reverse alignments were performed between all pair of alleles at each locus, and the dynamic programming (DP) matrices of all alignments were stored.
2. For each ordered triple of alleles, labeled the (putative) chimera, left parent, and right parent, we found the position in the chimera that maximized the sum of the DP matrix of

the chimera/left parent forward alignment up to this point and the DP matrix of the chimera/right parent reverse alignment beyond this point, conditioning on it being at least 50nt from either end. This determines an optimal breakpoint – i.e. the point that makes the putative chimera most similar to one of the parents on each side – and alignment score – a measure of how close it is to a chimera – for this triple.

3. For each putative chimera, find the left/right parent alleles with the best chimeric alignment score. The concatenation of these parental sequences from their respective sides of the breakpoint is referred to as the optimal chimera.

The relationship of each allele to its optimal chimera and the relationship of the left and right parents to each other are explored in Table S1.

Negligible effects of chimeras on pair statistics: The presence of some PCR chimeras with parents too similar to be detected cannot be ruled out, but even if they are present, they cannot account for the important signals summarized in Fig. 2. The frequency correlations in Fig. 2E should not be affected at all by the presence of PCR chimeras, as recombination alone does not on average alter individual SNP frequencies, f_a or f_b.

Linkage could be decreased by recombination events that formed PCR chimeras, but for our data in which $\rho(x)$ decreases substantially over tens of nts, typical reads would have to be the result of multiple PCR recombination events for these to substantially affect the observed linkage decay. This scenario is incompatible with the high frequencies of many alleles as these would get broken up by such extensive chimera formation.

Computing SNP pair statistics from metagenomic data set: The metagenomic data set used to design the PCR primers (see above) contains 98,816 reads (generated by paired end-Sanger sequencing) collected from a low temperature (53.5-63.4°C) region of the Octopus Spring biofilm

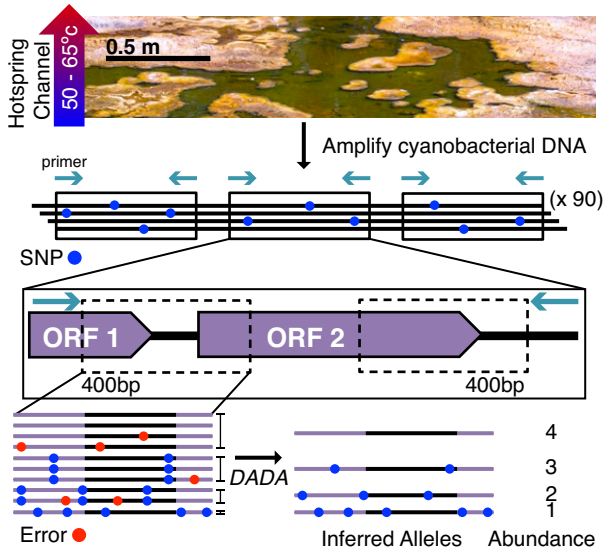
in 2004 (*13*). We BLASTed each open reading frame in *Syn* OS-B′ (with the exception of those with "ISSoc" in their annotation, which are associated with transposable elements that tend to have many similar copies around the genome) against this collection of reads, keeping all hits with e-values <1e-50. Next, for each gene we aligned all hits against the *Syn* OS-B′ sequence, removing insertions in the metagenome reads and gapping out completely any codons with one or more gaps relative to *Syn* OS-B′ (ensuring no frame shifts). From this collection of pairwise alignments for each gene, all pairs of synonymous sites were found with at least 10 reads spanning the pair of sites. The numbers of reads with each possible haplotype was recorded and used to generate Fig. S3B below.

Supplementary Text 1

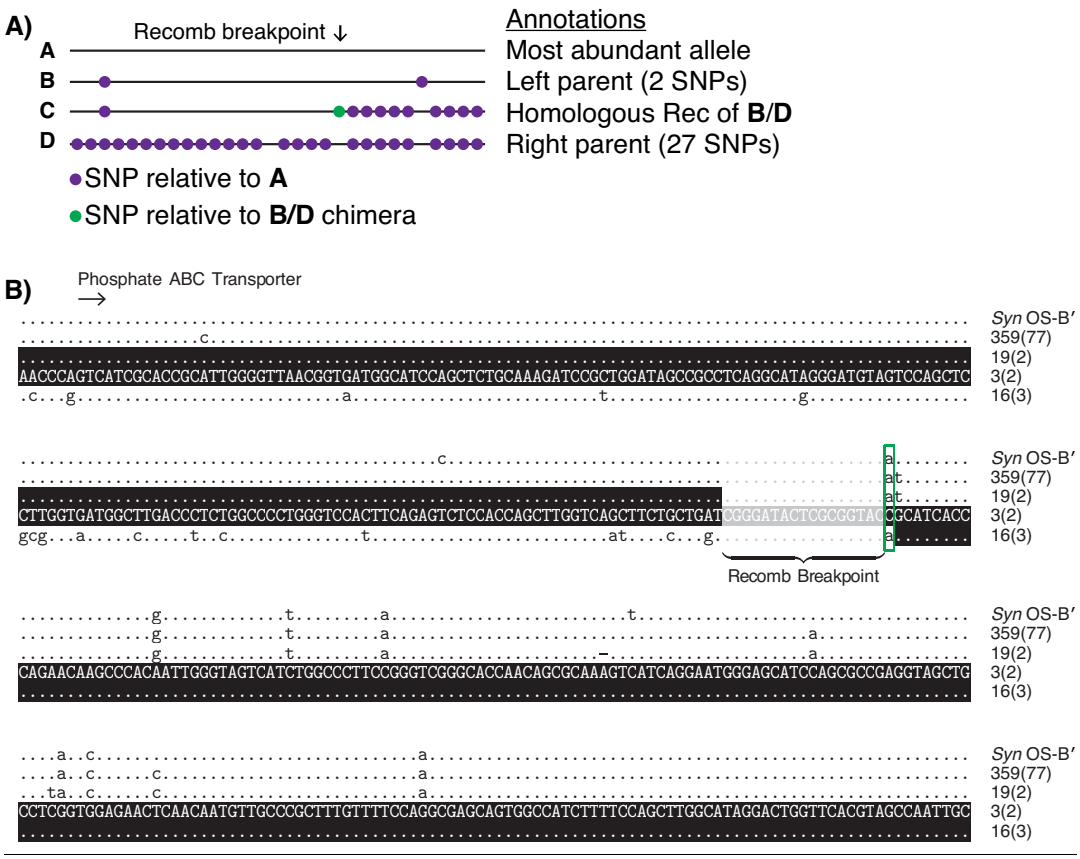
To check if the amplicon reads were primarily amplified from *Synechococcus* sp., we ran BLASTN (ver. 2.2.26+) for all 6738 inferred alleles from the 160,033 amplicon reads against the complete collection of bacterial and archaeal genomes available from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz, obtained on 9 April, 2015), which includes the 20 isolate genomes previously taken as a reference set for these biofilms (Supplementary Table 1 in *44*). We kept the best BLAST hit for each allele, discarding hits with e-values > 1e-40, considering such alleles as absent from the database. We found that 89.9% of alleles (6,058) and 94.6% of reads (151,366) had best BLAST hits to *Syn* OS-B′; 8.1% of alleles (548) and 5.3% of reads (8,472) had best BLAST hits to *Syn* OS-A; 0.6% of alleles (41) and 0.03% of reads (57) had best BLAST hits to Roseiflexus sp. RS-1 (NC_009523.1); 0.3% of alleles (24) and 0.03% of reads (53) had best BLAST hits to 11 other genomes, including a number of known thermophiles; and 1.0% of alleles (67) and 0.5% of reads (85) had no high quality BLAST hit. Within the main cloud, all alleles had best BLAST hits to *Syn* OS-B′ or *Syn* OS-A, with 96.3% of alleles (5157/5356) and 95.8% of reads (141441/147608) nearer to *Syn* OS-B′ than *Syn* OS-A. Of these 5256 main cloud alleles, only 484 had any hit to genomes besides *Syn*

OS-A or *Syn* OS-B′ with an e-value < 1e-40, and these hits all of far lower quality (every e-value was larger by at least a factor of 1e+70 relative to the hit to *Syn* OS-A or *Syn* OS-B′). These results strongly support the assumption that the PCR primers that we designed were specific and essentially amplified DNA from *Synechococcus* sp.

Supplementary Figures

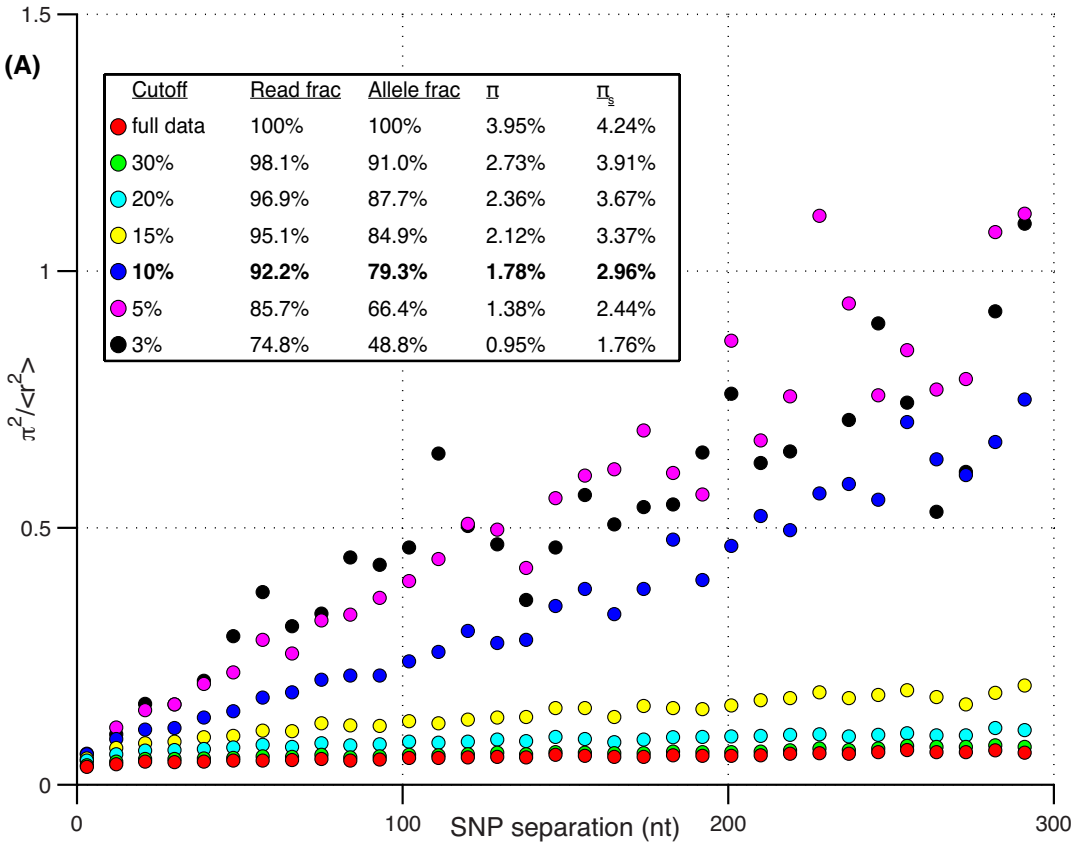


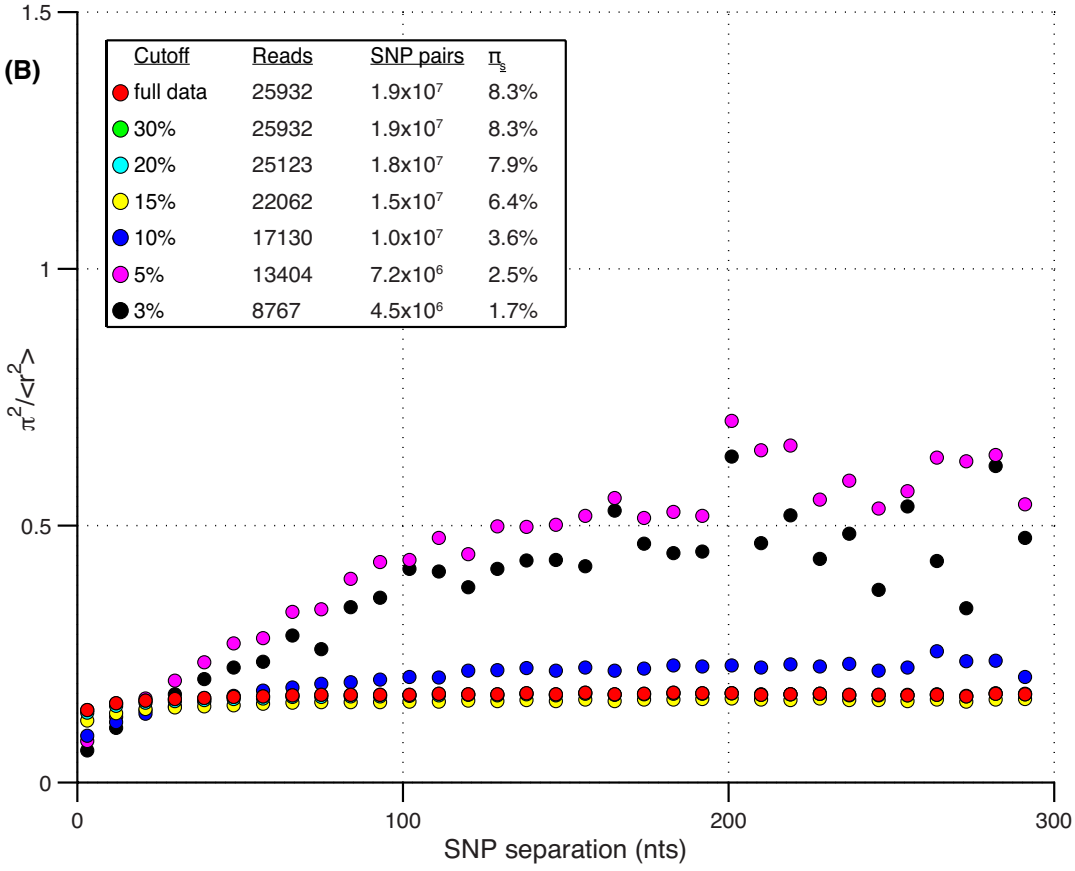
Supplemental Figure 1: **Overview of sample generation and processing; from a natural biofilm to fine-scale *Synechococcus* alleles.** Samples were collected from a microbial biofilm in Mushroom Spring along a hot spring channel. Primers specific to previously observed cyanobacterial diversity were designed and used to amplify DNA from 90 regions. Individual regions contained a mixture of genic (purple) and intergenic (black) regions. Amplicons were sequenced, and 400nt trimmed reads originating from each primer were run through the *DADA* error correction algorithm to produce inferred alleles and allele frequencies.



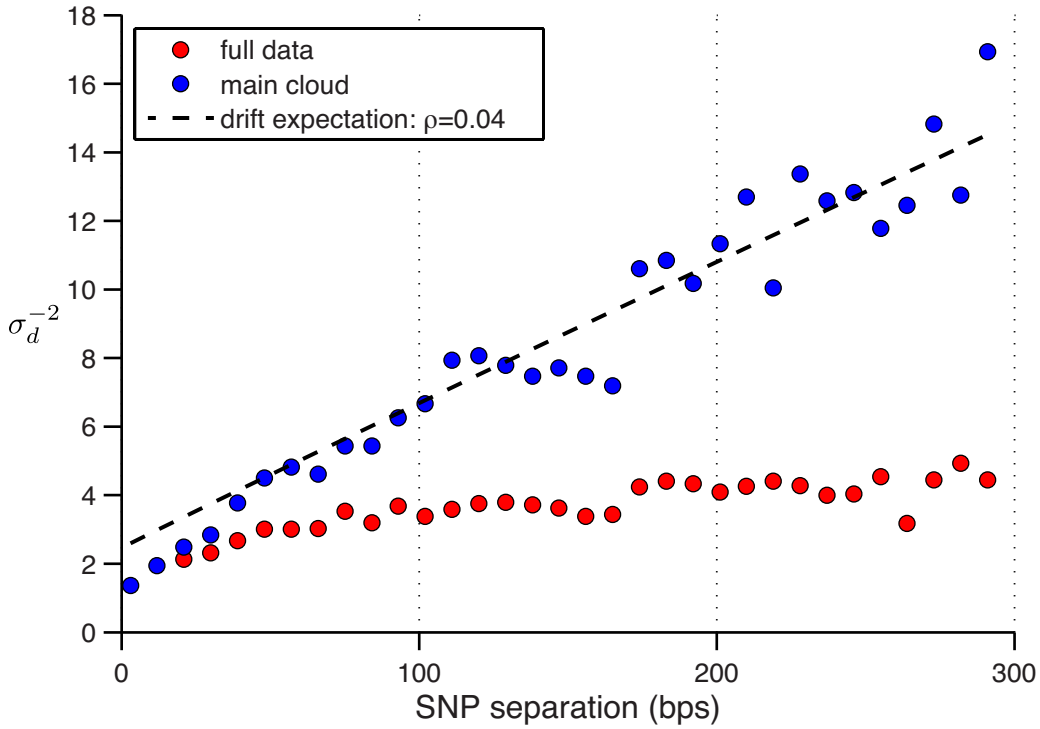
Supplementary Figure 2: Example homologous recombinant allele. (A) Schematic of an alignment of four of the 54 inferred alleles at a locus containing 393nt of the 3′ end of a phosphate ABC transporter (GI:86609371) plus seven intergenic sites and 1353 total reads. The third line is an example of a homologous recombinant, being identical (up to one SNP) of a combination of the two "parent" alleles above (allele B) and below it (allele D). Left of the recombination breakpoint (downward arrow) the left and right parents differ at 17 SNPs and on the right they differ at 10 SNPs, but the recombinant allele contains only one SNP (green circle) relative to a concatenation of the left half of B and the right half of D. The relative fraction of synonymous versus non-synonymous SNPs of this kind is the basis of a check for PCR chimeras (see Table S2). (B) A detailed alignment of these four alleles as well as the homologous region in the *Syn* OS-B′ genome shown on the top line. The number of reads after error correction is given followed by the number of raw reads (i.e. before error correction) exactly matching the sequence in parenthesis. The start location and orientation of the coding sequence is annotated (it does not

terminate within the alignment). The green SNP from (A), an A->C, is boxed in green. The recombinant allele (line 4) is again placed between the two parent alleles, with the half of each matching it highlighted in black.

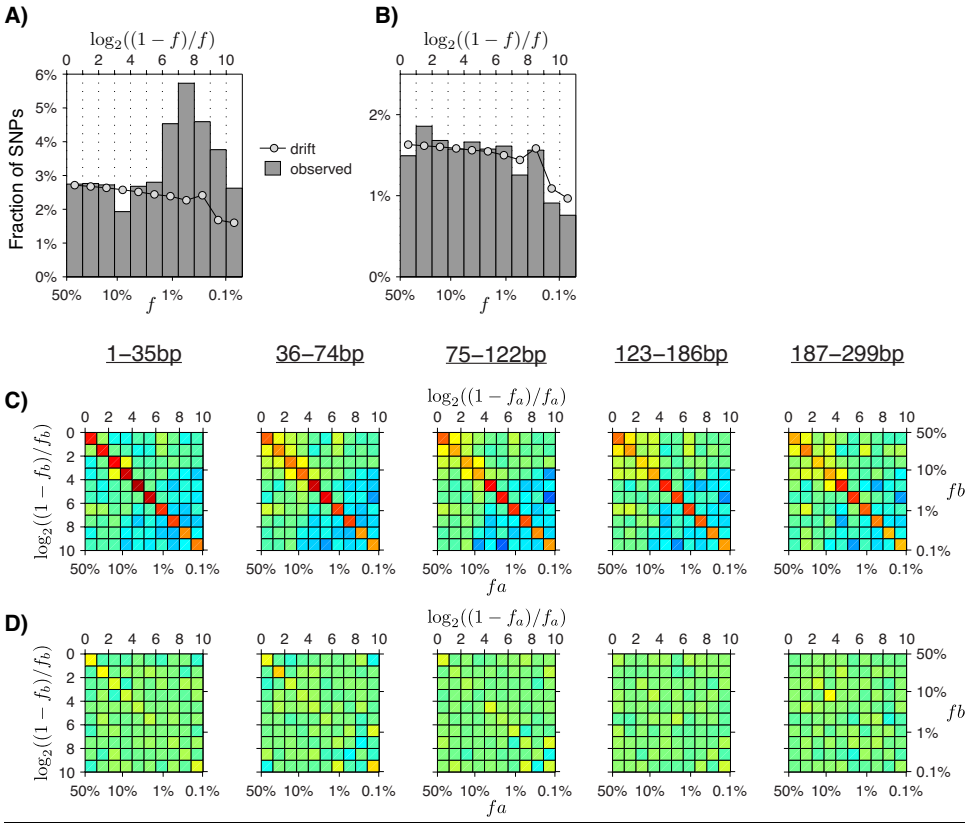

(A)
Cutoff | Read frac | Allele frac | π | πs
full data | 100% | 100% | 3.95% | 4.24%
30% | 98.1% | 91.0% | 2.73% | 3.91%
20% | 96.9% | 87.7% | 2.36% | 3.67%
15% | 95.1% | 84.9% | 2.12% | 3.37%
10% | 92.2% | 79.3% | 1.78% | 2.96%
5% | 85.7% | 66.4% | 1.38% | 2.44%
3% | 74.8% | 48.8% | 0.95% | 1.76%
π2/<r2>
SNP separation (nt)
0
0.5
1
1.5
0
100
200
300


(B)
Cutoff | Reads | SNP pairs | πs
full data | 25932 | 1.9x10^7 | 8.3%
30% | 25932 | 1.9x10^7 | 8.3%
20% | 25123 | 1.8x10^7 | 7.9%
15% | 22062 | 1.5x10^7 | 6.4%
10% | 17130 | 1.0x10^7 | 3.6%
5% | 13404 | 7.2x10^6 | 2.5%
3% | 8767 | 4.5x10^6 | 1.7%
π2/<r2>
SNP separation (nts)
0
0.5
1
1.5
0
100
200
300

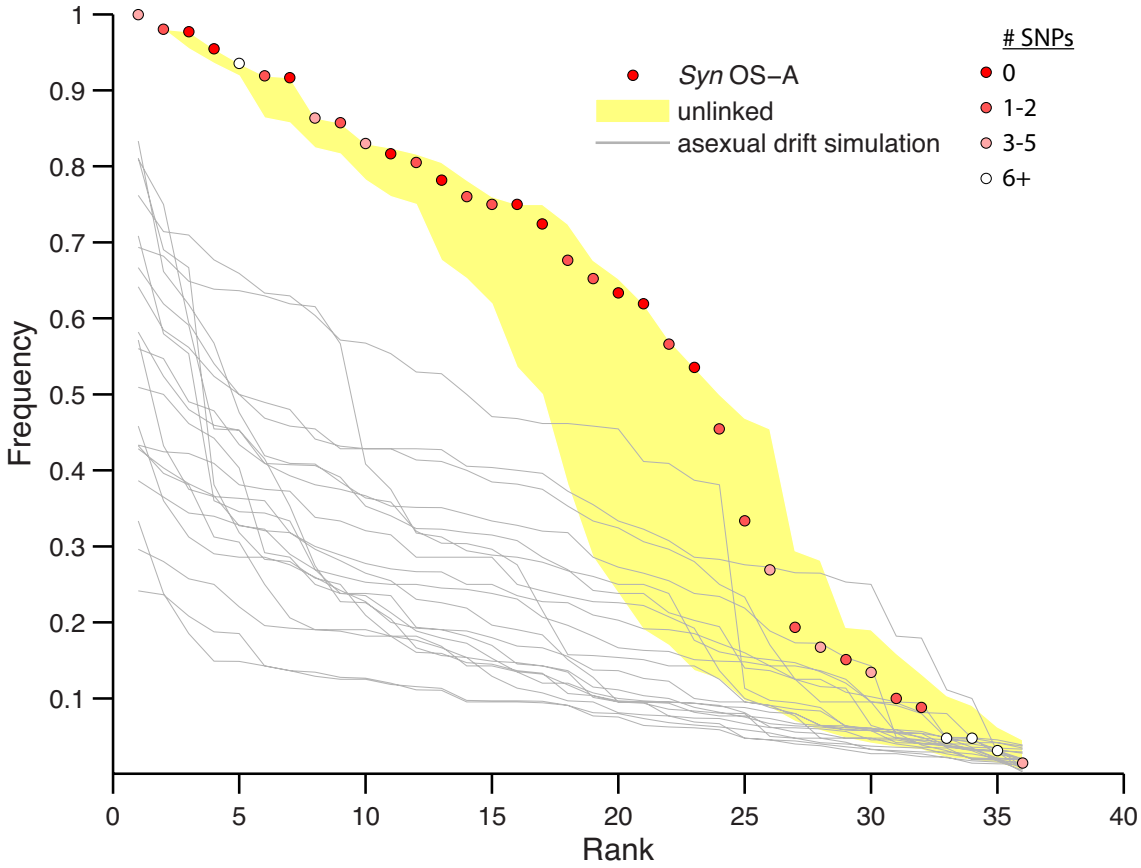
Supplementary Figure 3: π^2 / $\langle r^2 \rangle$ vs cutoff distance to most abundant allele. (A) As in Fig. 2A, we show the inverse of r^2 as a function of distance (here multiplied by π^2 for a convenient normalization, as the number of pairs of polymorphic sites is proportional to this) for a number of screens that exclude alleles outside a cutoff distance from the most abundant allele. As in Fig. 2A, $\langle r^2 \rangle$ was averaged over all pairs of synonymous sites; pairs in which one or both sites were not polymorphic contributed $r^2 = 0$. It is with this averaging that $\langle r^2 \rangle$ has been shown to approach $1/\rho$ in the limit of large ρ (*26*). In the upper left inset, for each cutoff distance we give the fraction of reads and fraction of alleles inside the cutoff, as well as the overall and synonymous average per site heterozygosities, π and π_s. The 10% cutoff is the largest that shows the clear distinction in linkage between the main cloud of the diversity and outliers: we thus chose this to define the main cloud (statistics in bold). The main cloud comprised 79.3% of all alleles (5346/6738), but 92.2% of the reads (147576/160033), as it was enriched for higher frequency alleles. For comparison, 92.6% of the *Synechococcus*-like 16S reads fell within a 3% diameter OTU centered on the *Syn* OS-B′ sequence. However, the main cloud contained only 48.7% of the synonymous SNPs of the full data (3664/7530) and only 35.4% of the number of SNP pairs that are present in the full data (41373/116854). Just over half of those SNPs (51.9%) present in the full data but not the main cloud were also found in the *Syn* OS-A genome. Note that the 10% cutoff is ½ of the typical 20% genetic divergence between the *Syn* OS-A and *Syn* OS-B′ genomes. **(B)** The same statistic for the low temperature Octopus Spring collection of metagenome reads (which were generated by Sanger sequencing of genomic libraries made without any prior PCR amplification step) (*13, 26*). The screen in this case refers to maximum distance to the *Syn* OS-B′ genome. Pairs of sites within the same gene of *Syn* OS-B′ were kept only if at least 10 metagenome reads spanned both positions. The number of reads contributing and the number of SNP pairs, as well as the synonymous heterozygosity, are given in the legend. The saturation of $\langle r^2 \rangle$ at ~150bp is consistent with that expected for the very limited read depth of this data set.



Supplementary Figure 4: alternate linkage measure, σ_d^2. The inverse of the square standard linkage deviation, $\sigma_d^2 = \langle(f_{ab}-f_af_b)^2\rangle / \langle f_a(1-f_a)f_b(1-f_b)\rangle$ is plotted for pairs of SNPs *A/a* and *B/b* as a function of their separation. The drift expectation with infinite sampling depth (*45*) is shown with the observed π and $\rho = 0.04$ providing a reasonable fit. This measure of linkage is similar to r^2, shown in Fig. 2A, except when SNP frequencies are small [e.g. $f_a, f_b < 5\%$, (*46*)]. However, σ_d^2 is less sensitive to low frequency *AB/ab* haplotype doubletons: these each contribute the maximum value of unity to r^2 but have a more limited effect on σ_d^2, to which they contribute proportionally to the square of $f_{ab} = f_a = f_b$ to the numerator and do not dominate the average denominator. This partially explains the better matching achieved between the main cloud results and drift expectation than with r^2, despite the frequency-frequency correlations deviating strongly from drift (Figs. 2C and E).



Supplementary Figure 5: Site frequency spectra (SFS) and frequency-frequency correlations for full data (A,C) and drift simulations (B,D). The full data (A,C) includes all inferred alleles (without screening outliers, as was done with the main cloud), and drift simulations (B,D) were performed with θ = ρ = 0.0305 (θ chosen to match the observed π, analogously to Fig. 2B). (A,B) SFS for the full data and drift simulation, with the drift expectation overlaid on each. (C,D) Frequency-frequency histograms analogous to Fig. 2E, for the full data and drift simulations. The full data averages ~ 20,000 SNP pairs per window and the simulations ~ 4,500 SNP pairs per window. Coalescent simulations were performed in *ms* (*42*) and sequences were generated from the resulting trees with *seq-gen* (*43*).



Supplementary Figure 6: *Syn* OS-A Rank Spectrum. For 36 loci, *Syn* OS-A and *Syn* OS-B′ are > 10% diverged and there are > 20 reads of alleles more similar to *Syn* OS-A than *Syn* OS-B′. The frequency spectrum of the alleles most similar to the *Syn* OS-A isolate genome amongst this OS-A-like subpopulation is shown (red circles). Unlinked and asexual null curves analogous to those in Fig. 3 are also presented (*20*). For the asexual simulations, we applied $\theta = 0.01$ rather than $\theta = 0.02$ as for *Syn* OS-B′ in Fig. 3. This was based on the prior observation (*13*) that there is less diversity around *Syn* OS-A than *Syn* OS-B′ rather than a calculation of π, which is unreliable for this collection due to the sensitivity to outliers and *DADA*'s need for sufficient depth to call alleles as real.

Supplementary Tables

Max # of SNPs between allele and optimal chimera	Min # SNPs between parent alleles on both sides of the breakpoint			
	3	5	8	10
0	583	256	74	37
1	1140	540	175	99
2	1599	**789**	289	159
3	1997	1018	395	218

Supplementary Table 1: Chimeric alleles. We give the number of alleles within X SNPs of their optimal chimera (*20*) and whose parents have at least Y SNPs on either side of the optimal breakpoint. This number depends strongly on the choice of X and Y, so we conservatively took the 789 alleles (shown in bold) that are ≤ 2 SNPs of a perfect chimera with parents ≥ 5 SNPs apart on both sides of the breakpoint; such alleles seem extremely unlikely to have been formed without a recombination event.

	One-away chimeras		Two-away chimeras		
SNP type	*N*	*S*	*NN*	*NS*	*SS*
Observed	126	180	48	85	56
PCR Expectation	217	89	95	78	16

Supplementary Table 2: Chimera d_n/d_s consistency checks. For chimeric alleles (Table S1), defined operationally as having parents ≥ 5 SNPs apart on either side of a putative homologous recombination breakpoint and ≤ 2 SNPs from an optimal chimera, we characterized each SNP between these alleles and the nearest optimal chimera as non-synonymous (*N*), synonymous (*S*), or intergenic. The number of observed *N*s and *S*s for alleles one or two SNPs from being perfectly chimeric are given. Under the *DADA*-inferred mutation matrix, a fraction $f_s = 0.29$ of these coding sequence SNPs are expected to be synonymous if they resulted from PCR recombination followed by errors. Under this assumption we also present the expected number of *N*s and *S*s for one-aways and *NN*s, *NS*s and *SS*s for two-aways. Two-tailed binomial p-values for the null hypothesis that the *N*s and *S*s are all due to errors with synonymous probability $f_s = 0.29$ for the one and two-aways are respectively $p_1 = 2.2*10^{-28}$ and $p_2 = 1.0*10^{-26}$; these alleles cannot consist of PCR chimeras alone. We computed a lower bound on the *real fraction*, or proportion of these alleles not due to PCR, by assuming that the biological $d_n/d_s = 0$, i.e. that real SNPs never alter proteins, yielding minima of 42% for one-aways and 33% for two-aways. Under $d_n/d_s = 0.15$, a reasonable assumption for higher-abundance alleles, the expected synonymous fraction amongst the biological recombinants becomes $f_s / (f_s + 0.15*(1-f_s) \approx 73\%$. This adjusts the estimated real

fraction up to 68% for one-aways and 52% for two-aways. But $d_n/d_s = 0.15$ may be a substantial underestimate for rare diversity (*20*), so given that 33% of one-aways and 46% of two-aways are singletons or doubletons, the real fraction may be higher still.

Parental divergence from most abundant allele	0-5%	5-10%	10-20%	20+%
0-5%	280 (355)	276 (200)	100 (130)	32 (54)
5-10%	-	26 (12)	21 (12)	6 (5)
10-20%	-	-	25 (8)	4 (5)
20+%	-	-	-	19 (8)

Supplementary Table 3: Distribution of distances of putative chimera parents from the most abundant allele. For the 789 putative chimeras identified (in Table S1) as being ≤ 2 SNPs from their optimal chimera with parents ≥ 5 SNPs apart on either side of a putative homologous recombination breakpoint we show the breakdown of the number of each with parent alleles in different ranges of nucleotide divergences from the most abundant allele. In parenthesis, we give the expected number for a uniform recombination null model which assumes that each pair of parents produces chimeras at the same rate without regard to their divergence either from each other or from the most abundant allele, but only in proportion to the product of their frequencies in the overall dataset. Because parents too close together cannot produce chimeras identifiable by our method, for this 5 SNP cutoff we consider as possible pairs of parents only those that are at least 10 SNPs diverged. The observed numbers for which one parent is inside the main cloud (<10% from the most abundant allele) and the other an outlier (>10% from the most abundant allele) is only slightly lower than expected under uniform recombination (159 observed vs. 201 expected). This suggests that chimeras involving outlier parents are not substantially less likely than expected under uniform recombination.

Supplementary Table 4: Amplicons, target genes, and primers.

Syn OS-B' target gene	Gene Annotation	Gene Length (nts)	PDB accession	Left Primer	Right Primer	Amplicon Length (nts)
V1-V3	16s rRNA	465	NA	AGAGTTTGATCMTGGCTCAG	ATTACCGCGGCTGCTGG	465
CYB_0489	circadian clock protein KaiB	476	2QKE	GATGTCyTCAAArCCCTCGAT	CACsTGAGCGAAATGTACCG	532
CYB_0136	photosystem II 12 kDa extrinsic protein PsbU	495	1S5L	srrrATCAGCCGTAGTAbCCyT	CTrTyyCCyGATCATAGTTGTTAAA	538
CYB_1928	transcriptional regulator, ArsR family	433	1R22	AGrrsCmTyyyCGsCTCAGC	yGTCGAACAGTTGGyGAAATAG	543
CYB_0628	ferredoxin-thioredoxin reductase, variable subunit	419	1DJ7	TGAAAGTyGGAGATCGrGTrC	CbyTGGATCCCTACTTGCTCAT	546
CYB_2020	hypothetical protein	556	NA	yGTGAyGsTyAAAACTTCGCTT	AACTGAGGArAATrGGTAGTCATrG	556
CYB_2253	photosystem I reaction center subunit II	570	NA	GATCTGCTACGAAArATTACAGTTC	rrhTyArTCCCAGGGyTTsCC	570
CYB_2479	pentapeptide repeat family protein	626	2F3L	AAGCTGGGCTTGCTCATrGT	ATsCTGCTrGsGGCmAykCT	577
CYB_1245	photosystem I reaction center subunit IV	320	1GXI 1QP2 1PSE 1PSF 1QP2 1QP3	AAAGGsGyGGATCCCTGmAC	AkCsTTCTCACGGACTAGGGTAAT	590
CYB_1598	ferredoxin, 2Fe-2S	351	1RFK	GCyAGrTTsACCGGGATCAG	AGCATrrGmTGCTCTCCTGG	601
CYB_1638	iron-sulfur cluster assembly accessory protein	532	1NWB	AGGTrAymCGAATGGCrTTG	TGGTGAAGCAdTGGGTGrAG	623
CYB_2841	CpeS-like protein	590	3BDR	CGCAAGyTArrATCGGCAAA	AGCCTAGArAGAGCkGGATTC	624
CYB_1938	response regulator	433	NONE	AAGAGkrrCyTCACCAAACAATAG	rTCTTCCACmGTCATCACCTC	632
CYB_2772	phycobilisome degradation family protein	446	1OJH	CACCAGGTGGGAAryGAAAG	GyTCTTCTGTTGAGGCTTTATAGTT	633
CYB_2577	ribulose bisphosphate carboxylase, small subunit [EC:4.1.1.39]	434	1RSC	CyCArmTGAGCGArACrGAT	CTTrTArACCArGAAGsTGAGCAC	649
CYB_1440	phycobilisome 7.8 kDa linker polypeptide, allophycocyanin-associated, core	289	1B33	ArCATyTTrAGGGTvGTGGTCTT	GTACGCGAGCTTTCTGrGGT	651
CYB_1231	ParA homolog, authentic frameshift; this gene contains a frame shift which is not the result of sequencing error	1066	3CWQ	GTAAAACCAGrkTGGGAGGGT	GTCGAGmAGGCGCTyAAACT	675
CYB_1803	plastocyanin	525	1BAW	CCrGAAGCCAACTCGAArAT	GCyGTTrCGTTTTGAGAGGAAT	678
CYB_2426	cytochrome b6-f complex, subunit V	201	1VF5	GTGTGGTCACACTCTTGsCG	ACACyTGCAsyGGATssGAGT	680
CYB_0192	thioredoxin	531	1DBY 2PVO	yAGGAGTTmyCCArGCATGTC	AAAvCCCTGCTCyAGATCCTC	686
CYB_1519	cytochrome c6	527	1KIB	rACrAArTTCTTbCCCTGCTTG	CTkysCTyCACAkCCCTGCTC	687
CYB_0407	hypothetical protein	563	3CSX	CTGATyGAAAAGGArACGGG	ACTGAGCrAATTTGCGTAAGATA	699
CYB_0276	GUN4-like family protein	540	1Z3X	rGCsAGGATCTCrATsAGCA	GTACCCGCCyGAAAsCTGAG	709
CYB_2831	ferredoxin, 2Fe-2S	351	1RFK	GyyAGsrGCCAAACTGsArAC	CATyGyCCTCTACGArATCCAC	713
CYB_2737	photosystem I reaction center subunit XII	268	1JB0	ATsCCyTTGAGAAAyTGCAGC	CrCArGTTTTCACCATCCTG	738
CYB_2740	phycocyanin, beta subunit	632	1I7Y	CATGAyTCAAAACGCTCCTTTAC	TTCATAACCTCTTTTrCTTTCGCTA	748
CYB_2345	copper metallochaperone	294	1SB6	CGGGGCTAyCArGTrGACTA	CTTsAGCTTCmryTCCCGTTCT	771
CYB_2526	phosphate transport system regulatory protein PhoU	642	1SUM	ATTGmTCGGsAGGyTGAGAG	CCCACATyCATGGGAGAAAG	783
CYB_1448	pyridoxamine 5'-phosphate oxidase family protein	530	1VL7	GTCGGyGCwGGAAmAGTTrA	CAGAGGTsTTyTTGGCyTTTG	798

CYB_1630	cytochrome b6-f complex, subunit IV	601	2E74	GyATyTCyGGrCCTTTGTGA	AGGGCwGTrATCCGATAGGT	799
CYB_0594	CRISPR-associated RAMP protein	833	NONE	ArGGGAGACTsrGCmGTkAG	CGGGATCTGGTGGATyTGTA	834
CYB_1499	photosystem I reaction center subunit XI	636	1JB0	TAGAATGAATCsCAAAATTCCAGsC	GGATCCTGCTCCAmTCTAAAGTC	859
CYB_2211	allophycocyanin, beta subunit	574	2VJT	yTrrCTsCsGACAGTTGGCT	CTrGAGyGTTGCAGAAArGTTACA	869
CYB_1912	phosphate ABC transporter, ATP-binding protein	798	2OLK	AAGGGTTTGCmGAGAACAGTAT	GAACTCTAGCTCTTTyTGTGAGTGC	877
CYB_0840	CRISPR-associated RAMP protein, Cmr4 family	845	NONE	CTGGCrmAGGGATTCGTAAAC	GAryTsCThTGGCTsCsTTGG	916
CYB_1498	photosystem II reaction center protein PsbJ	211	1S5L	GmkCChArrCCsryATAGGAG	GCATsCAAmTCCTTATCATTGTGTTA	924
CYB_2396	methyltransferase, UbiE/COQ5 family [EC:2.1.1.-]	799	1XXL	AGGCAGCTCAAACTGAGTCAT	CArrAArCAGCTCAGCCCAG	926
CYB_0947	protein phosphatase 2C family protein	2199	2j82	CTACGGCAAmGACCTCrmyC	yGGCyGGrCATGCTACyTTA	928
CYB_2458	photosystem II reaction center protein PsbI	204	1S5L	CTTTrCCrTGCAAmAGCAAC	yACyTTyTGTAGCTTCAryACCGA	941
CYB_1913	phosphate ABC transporter, permease protein PtsA	894	3FH6	AAGAGTTCGATTCAAGACTTTGTTG	GTAACTTGGGAAGCAGAmAGAAGAT	973
CYB_0884	acid phosphatase SurE [EC:3.1.3.5]	798	1J9J	GGyTwrCTTTACCTTAGCTTCATAA	TymwrAGCCAAArGyyGAGAG	987
CYB_1427	acid phosphatase SurE [EC:3.1.3.5]	783	2V4O	GATCCCACCGmyATCCTGAC	GTGTGGTCATvGCAATCCTC	999
CYB_1793	hypothetical protein	734	3CJ8	ArATCATCCAAACCCCTTGAG	rATbCCCTCyTCrATGGCAA	1004
CYB_2738	phycocyanin, alpha subuni	556	1PHN	GTrATTsCACTCATGyTAATTCCyC	ATyGCCGAAATCGArTCCTACT	1009
CYB_1914	phosphate ABC transporter, permease protein PstC	984	3FH6	CTCCTTyATCTTCTkTCTGCTTCC	GGAAAGATTTTGAACCGAAGATG	1052
CYB_2478	glycerol-3-phosphate ABC transporter, permease	1053	3FH6	GGGyTGGGATCCCTGATyTT	TTATCCsCCrTAsGTyACATTCC	1074
CYB_2856	DNA-binding response regulator	744	1YS7	GAAGTTTGATATyCTAkCGGGsG	rrCGCTACGAAGAyCCyCAG	1094
CYB_0944	phycobilisome rod-core linker polypeptide cpcG1	1689	1W9A	AGAAGCATCACCmrGATTATAACGA	AGAGCGAGCCTTGCAmCTrT	1109
CYB_1551	photosystem II manganese-stabilizing protein	869	2AXT	CAyCAGCTACCGbGTCAACC	AAGCCCrAAATTkGmTCTCTTC	1111
CYB_2268	photosystem II protein, PsbB/PsbC family	1108	1S5L	CCTACCTCGAAACTGCAAGTG	AAAGCTTGTTTAGmAACGCAArAG	1119
CYB_1915	phosphate ABC transporter, phosphate-binding protein	1062	1IXH	mCTCTAGTrGAGTTCAGGCTGC	GAArrTGTTGATTyGCCACAGATA	1121
CYB_2442	hypothetical protein	795	2OC5	CCrATCAGGCCGAACATArA	yGTTCCTCCAGmTCCACCAG	1139
CYB_1439	allophycocyanin, beta subunit	573	1B33	yrGhTAGCACACCAGATCCCT	GGGTATGCTCGyCATACTCTTTA	1140
CYB_1629	cytochrome b6	776	1Q90	CAGCAGTTGCTGCTCTTCyC	TAATsGCCTCGGTyGCTATCAC	1144
CYB_1346	pentapeptide repeat family protein	1061	2J8K 2O6W	CACCAArTTyGCyGCTTTCT	CkrAGyATsGCyTCCTTCAAGT	1163
CYB_1434	dps protein	729	1MOJ	GTGTryTGrCGGATCACrCT	GryGGGTTAGGysArTCAGTTT	1192
CYB_1704	glyceraldehyde-3-phosphate dehydrogenase, type I [EC:1.2.1.-]	1087	2d2i	GwTCCCAAATTTTrTCTTCAGAAT	GACsCTyCTCCTCCAyGACG	1200
CYB_0144	putative zinc ABC transporter, zinc-binding protein	2185	1PQ4	AAyGACACATTyTGyACnGCC	TGTCTCTGTyGTyTTGTGAGGTAyT	1203
CYB_2882	ferredoxin--NADP reductase [EC:1.18.1.2	963	1EWY	GTGTAATTrCGCACATGCCC	AGCAGATCGCmGkTATGGTT	1206
CYB_0490	circadian clock protein KaiA	1259	1R8J	GCCACGTAGAGCTTGAGrACATA	GATGATCAGGGGCACCATAG	1215
CYB_2047	ribulose-phosphate 3-epimerase [EC:5.1.3.1]	787	1TQJ	nGAAGAGCCrGGCTTCAAyT	CAAACyCCGTACCATAGAGGAG	1256

CYB_1894	carboxylesterase, beta-lactamase family [EC:3.5.2.6	1314	2j7v 2JBF 2J8Y 2J9O	sGCCrATyTGTGGGTTTAAT	CddsrGymGACCArAACTCrTAG	1283
CYB_0698	phycocyanobilin:ferredoxin oxidoreductase	1212	2G18	CTCTCGAAAAGCGTTGTGrC	GTTCAyGAAACCyTGGCrCT	1342
CYB_2200	putative adenylate cyclase [EC:4.6.1.1]	1080	NONE	GAGGTGGATCCsCAGCTTTT	rCCTTAAyGTATCTACACAAAATGA	1347
CYB_0843	CRISPR-associated protein, TM1812 family	1516	NONE	rAAAAsCsrGAACCTGArGsyG	CTATGGCyAAwACCCTTTTGAC	1360
CYB_0092	pyridoxamine 5'-phosphate oxidase family protein	701	2I51	GCGCAwCAAGATCACyArsC	GrrATTrCmGGTGAArCAGC	1375
CYB_1895	adaptive-response sensory histidine kinase SasA	1207	1t4y	ATGGTrCGyAArGGCmArAC	CTAGCsmyyGCAGAGyCGATAC	1376
CYB_0593	CRISPR-associated RAMP protein, Csx10 family	1539	NONE	rATkTGCCGyTGGATCTyrA	CrGATCCCTGCAGCTTTTrC	1478
CYB_1230	trypsin domain lipoprotein	1312	1OKX	GTTTTCGCCAATCGTCACTT	CGCyGGTAsGATATCCTTGG	1479
CYB_0940	phycocyanin, alpha subunit	583	2VJR	GsTGGTGCTTTGACCrAACAT	rrAAArACCArAAGGAGATCCA	1488
CYB_1808	putative soluble hydrogenase, tritium exchange subunit	1241	2DR1	GAGGTCTGCAAAACCTCTATGC	GTrTGyAGGTGAAAGTTGTACTGC	1499
CYB_2579	ribulose bisphosphate carboxylase, large subunit [EC:4.1.1.39]	1512	1RSC	TTTTCTGGTTyTGTrATTCGTTyAGAG	rrsCAGyrGGGCAAGAGArC	1511
CYB_2826	photosystem I reaction center subunit XI	572	1JB0	AAGAAkCTGGAGAGCAAAATCC	CTAsCTyAAGGACACATCsTGCyAT	1514
CYB_1680	glutamyl-tRNA synthetase [EC:6.1.1.17]	1520	2CFO	rAAGGCwTrrCAGCAAGmrAT	GATCyCvmyGCyGACTAGGAT	1538
CYB_0263	putative lignostilbene-alpha,beta-dioxygenase	1522	2BIW	sTsTAGCCrATrsTAGCAACTGyTC	rAAAGTTTAAGCAACAATAGArGGC	1615
CYB_0858	sensor histidine kinase [EC:2.7.13.3]	1260	$I5S	GArrTCTTGCTbsTGCCCAC	yyGCrATGTCCTGATTGGTC	1640
CYB_0726	GAF domain protein	1301	1MC0	CTTGGTGTkGATCGAkAGGGT	CTCGCyTCCwCAGsTAAAGTA	1682
CYB_2477	glycerol-3-phosphate ABC transporter,	1353	4AQ4	ATAACCTGCTCTTAATGTyGsTGC	ATrAArATCAGGGATCCCArCC	1684
CYB_1493	putative exopolyphosphatase	1791	2FLO	GGCTArACkGCTGCCTCAAA	rAmAAGGGGsrAAAGyrCACAGTAT	1784
CYB_1198	alkaline phosphatase [EC:3.1.3.1]	1795	1K7H	ACTAssGCyTbAGCTCGCTGG	ArGCyTGCCTrGAmCACTTG	1795
CYB_2266	photosystem II protein, PsbB/PsbC family	1746	1S5L	yTCAAAAGmrrGCGCAGAAG	CTGACGGAAAAACGAATTAAAGAGT	1896
CYB_2736	phycocyanin alpha subunit phycocyanobilin lyase, CpcE subunit [EC:4.-.-.-]	908	1B5O	AGAATyAvAGCAGCTCGTCGAT	GrAAGGTTGTyGAGGGCAAC	1935
CYB_0684	alkaline phosphatase	1728	2YEQ	yCmmyrCCCACACTCTAGCA	GyTGCGGATCCCTTTACATT	1968
CYB_1081	replicative DNA helicase	2049	2R6A	ArrCAyTTGCCAAAACGCTG	CGArGTTACyGCTACGCTGC	2121
CYB_0837	CRISPR-associated RAMP protein, Cmr6 family	2281	NONE	AvrATGGCCAGyGTrATyGAG	TTGCGyrAATTTCGGsAAAC	2139
CYB_0281	photosystem II P680 chlorophyll A apoprotein	1603	1S5L	GAGGAyCGAGCTTTATGGGACTAC	ACCAGyAGCACykTTGCCAT	2176
CYB_0943	phycocyanin alpha subunit phycocyanobilin lyase, CpcF subunit	764	1B3U	AAkCArCGGAAGrGwCATAG	sAGCTGGTACATyGTrGCTCTG	2299
CYB_2082	polyphosphate kinase [EC:2.7.4.1]	2281	1XDO	GAGAGCTTGCAGTCTTCGyTG	sCykyTGTTGACGTGTTTGAG	2387
CYB_0589	CRISPR-associated RAMP protein	981	NONE	GGAAAsGGGATCAGCTCrTAG	TArCCTGATTTGAGGGAACrATGT	2444
CYB_0842	CRISPR-associated protein, Crm2 family	3059	NONE	TCbCGAAACATCCACACATC	CTyTGGGGyCTGyTrCATGAT	3007