nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-017-0435-9

Ancient balancing selection on heterocyst function in a cosmopolitan cyanobacterium

Emiko B. Sano, Christopher A. Wall, Patrick R. Hutchins and Scott R. Miller*

The conventional view of bacterial adaptation emphasizes the importance of rapidly evolved changes that are highly repeatable in response to similar environments and subject to loss in the absence of selection. Consequently, genetic variation is not expected to persist over long time scales for these organisms. Here, we show that a geographically widespread gene content polymorphism has surprisingly been maintained for tens of millions of years of diversification of the multicellular cyanobacterium *Fischerella thermalis*. The polymorphism affects gas permeability of the heterocyst—the oxygen-sensitive, nitrogen-fixing cell produced by these bacteria—and spatial variation in temperature favours alternative alleles due to thermodynamic effects on both heterocyst function and organism fitness at physiological temperature extremes. Whether or not ancient balancing selection plays a generally important role in the maintenance of microbial diversity remains to be investigated.

The maintenance of a genetic polymorphism for millions of years by balancing selection is thought to be extraordinarily rare[1]. Despite the longstanding theoretical and empirical interest in balancing selection, there are still few clear examples of the persistence of distinct alleles within or across species over such long time scales. These are typically involved in self-recognition and/or host–pathogen interactions and include: self-incompatibility loci, which prevent self-fertilization in hermaphroditic plants[2,3]; heterokaryon incompatibility loci in filamentous fungi[4]; major histocompatibility loci[5], the ABO system[6] and recently identified regulatory variation in primates that is implicated in host–pathogen interactions[7]; and resistance and susceptibility alleles at the *rpm1* locus in *Arabidopsis thaliana*, which are involved in the recognition of pathogenic bacteria[8].

The persistence of ancient variation is particularly unexpected for bacteria. Bacterial adaptation is generally rapid[9,10], yet adaptive changes may be quickly purged from a population when they are no longer selectively favoured[9,11,12], especially if they bear a cost in alternative environments. This implies that locally adaptive variation typically has a recent origin and that convergent evolution plays an important role in the response of bacterial populations to similar selective pressures. In agreement with this expectation, adaptation in response to antibiotics, carbon limitation, nutrient quality and high temperature is often highly repeatable in laboratory evolved populations[9,11–14]. This may generally be the case for natural populations as well: convergent evolution of antibiotic resistance, immune system evasion, bacteriocin production and heavy metal tolerance has also been observed for clinical isolates of bacterial pathogens[13,15–18].

Consistent with the perceived importance of recent, repeatable and transient innovation, there is only limited evidence for the action of balancing selection in bacteria. As for eukaryotes, the best examples are involved either in self-recognition (for example, the quorum sensing locus *agr* in *Staphylococcus aureus*[19]) or host–pathogen interactions (for example, the surface-antigen locus *ospC* in the *Borrelia burgdorferi* species complex[20]). In these cases, diversity appears to be favoured, but allele or gene turnover may be high. At *ospC*, for example, old mutational variation is principally maintained in alleles of comparatively recent origin that have been generated by recombination both within and between sympatric *Borrelia* species[21,22]. Similarly, closely related individuals within populations of non-pathogenic bacteria and archaea often exhibit vastly different gene contents (that is, the flexible genome)[23,24]. Many of these flexible genes encode proteins related to cell surface structures and it has been proposed that much of this diversity may be involved in phage resistance that is maintained over the short term by negative-frequency-dependent selection but subject to high turnover[25,26]. Whether evolutionarily stable allelic polymorphisms can persist over long time scales during bacterial evolution therefore remains an open question with important implications for our understanding of the origins and maintenance of microbial diversity.

Here, we report that a gene content polymorphism that originated by a unique deletion event has been maintained for tens of millions of years of diversification of the cosmopolitan, thermophilic cyanobacterium *Fischerella thermalis* (also known as *Mastigocladus laminosus*). The polymorphism is located within a gene expression island that contributes to the structure and function of the heterocyst—the specialized nitrogen-fixing cell of these bacteria—and different alleles at this locus are both tightly associated with local ecological variation and distributed globally. With a genetic approach, we show that the polymorphism impacts the physiology of nitrogen fixation through its effect on the gas permeability of the heterocyst and subsequently propose a mechanism for its long-term maintenance in alternative environments.

Results

Population genomics of local adaptation in *F. thermalis*. At White Creek—a nitrogen-limited, geothermally heated stream in Yellowstone National Park—*F. thermalis* is a prominent member of microbial streamer mats that form along an approximately 1.1 km stretch of the channel spanning a mean annual temperature of ~39–54 °C[27]. *F. thermalis* laboratory strains isolated from upstream (47–54 °C) and downstream (39–43 °C) regions have diverged in growth rate at temperatures characteristic of these sites[27] despite high gene flow between regions[28]. To identify genetic loci that potentially contribute to these ecological differences, we assembled draft genomes from Illumina sequencing data obtained for ten randomly selected *F. thermalis* strains each from upstream and downstream sites

Division of Biological Sciences, University of Montana, Missoula, MT, USA. *e-mail: scott.miller@umontana.edu

(Supplementary Table 1). The genomes were very closely related, with ~54% of genes invariant and a mean nucleotide diversity (π) of 0.1%. We inferred a neighbour net genealogical network for a concatenated alignment of 2,297 protein-coding genes (approximately 1.5 megabase pairs and 4,080 single nucleotide polymorphisms (SNPs)), which clustered the strains into three distinct groups (Fig. 1a). Strains with a group 1 genome were restricted to the downstream sample, whereas other strains were principally observed upstream. Based on their close relationships with specific upstream strains, downstream strains from the latter groups (strains 213 and 1110) have probably recently descended from migrants of upstream origin. This conclusion is corroborated by the shared clustered regularly interspaced short palindromic repeats spacer (CRISPR) content among these strains (Supplementary Fig. 1), which reflects the histories of past infections by foreign DNA inherited from a common ancestor.

Although ~90% of genes in the sample pan-genome were conserved among all strains (Supplementary Fig. 2a), strains also clustered into the same three genome groups based on gene content differences (Fig. 1b, non-metric multidimensional scaling stress value = 5.6%, coefficient of determination $R^2 = 0.92$; Supplementary Fig. 2b). These differences in the flexible genome were more subtle than the hundreds of unique genes per genome that have been reported for other bacteria[24]: the greatest number of unique genes within a genome was 10, while 16 genomes exhibited no unique genes. Differences were largely concentrated in a few discrete genomic regions, and most of this flexible genome codes for mobile elements or hypothetical proteins.

The genome groups were highly interconnected by loops in the network (Fig. 1a), which indicates a history of extensive recombination. The method of ref. [29] estimated a minimum of 458 recombination events in the dataset. As expected for a recombining population, genetic differentiation (F_{ST}) between upstream and downstream samples was low to moderate for most loci (mean = 0.15; Fig. 1c). In genomic regions underlying local adaptation, however, divergent selection should increase F_{ST} between sub-populations relative to the demographically determined neutral background[30]. We identified 37 genes with F_{ST} values greater than 0.55 as outliers in the right tail of this frequency distribution with a false discovery rate below 10% (false discovery rates (q-values): 5.9–9.0%; Fig. 1c and Supplementary Table 2). Outliers were distributed throughout the genome (Fig. 2). They also exhibited extreme values for a related metric of population differentiation (Φ_{ST}) and, in many cases, absolute genetic divergence (D_{XY}, the average number of nucleotide differences per site between alleles from upstream and downstream samples; Supplementary Table 2).

Group 1 strains were fixed for different alleles than other strains at 36 loci, most of which ($n = 33$) were detected as F_{ST} outliers (Fig. 2 and Supplementary Table 2). Several of these also co-localized with regions of gene content variation among groups (Fig. 2). Using previously acquired growth data[27], we found that these differences come with apparent fitness consequences: strains with a group 1 genome grew faster on average than other strains under nitrogen-fixing conditions at 37 °C, but much slower at 55 °C (Supplementary Fig. 3; $P < 0.0001$ for the genome group × temperature interaction). Genome group identity explained 23% of the variation in growth rate at 37 °C ($R^2 = 0.76$, $P < 0.0001$) and 51% of the variation at 55 °C ($R^2 = 0.81$, $P < 0.0001$). We conclude that group 1 *F. thermalis* are downstream specialists with functionally important genomic differences that are retained despite extensive genetic exchange with members of the upstream population.

What variation contributes to these fitness differences? Six F_{ST} outliers that are also fixed in group 1 genomes are implicated in the development and function of heterocysts (Fig. 2 and Supplementary Table 2)—the nitrogen-fixing cells produced by these bacteria in nitrogen-limited environments like White Creek[31]. Heterocysts must simultaneously manage oxygen exposure and provide sufficient energy for the oxygen-sensitive and metabolically demanding nitrogen-fixation process. This is accomplished through a suite of morphological and physiological changes during differentiation from a vegetative cell[32], thereby enabling the simultaneous activity of nitrogen fixation and oxygenic photosynthesis through the

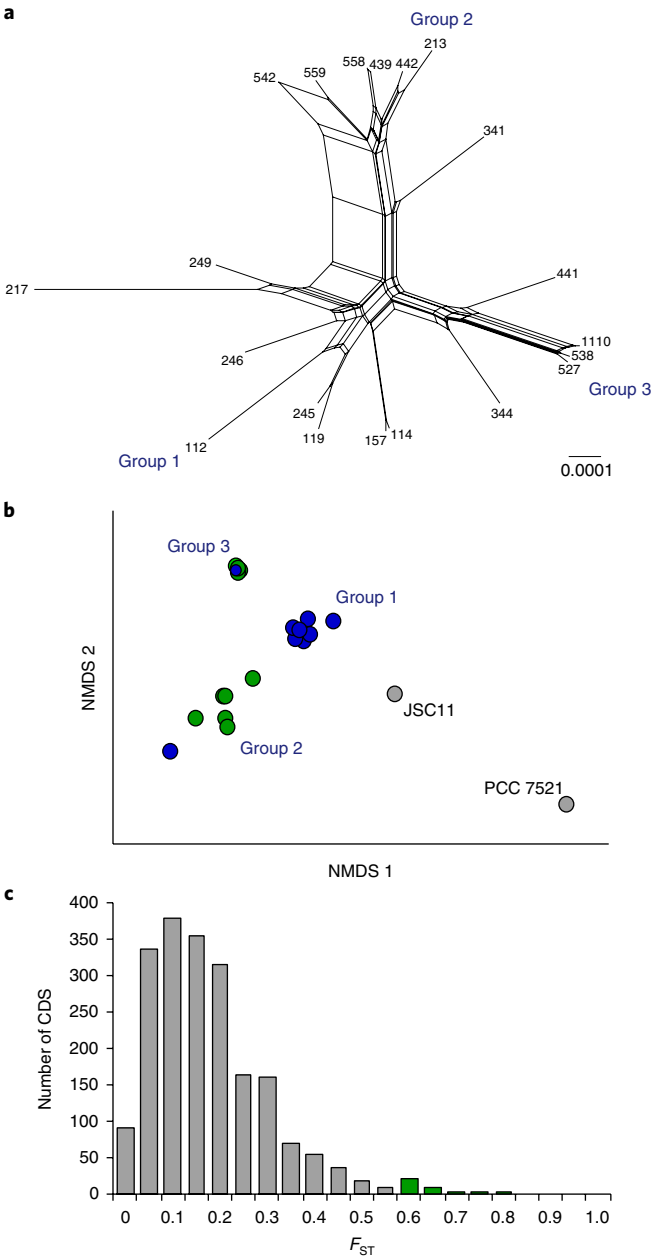

Fig. 1 | Genomic diversity and genetic differentiation of White Creek *F. thermalis*. **a**, A neighbour net analysis of the relationships among strains constructed from a concatenated alignment of 1,503,580 nucleotides resolves three genome groups within the population. The scale bar is in units of nucleotide substitutions per site. **b**, Two-dimensional non-metric multidimensional scaling (NMDS) also distinguishes these three genome groups based on Jaccard distances estimated from gene presence–absence data for downstream (blue) and upstream (green) strains. These are distinct from strains from populations located ~50 km (PCC 7521) and ~65 km (JSC11) from White Creek, respectively. **c**, Frequency distribution of F_{ST} between upstream and downstream strains ($n = 10$ each) for 2,043 polymorphic protein-coding genes. Inferred outlier loci with a q-value of less than 10% are shaded in green.

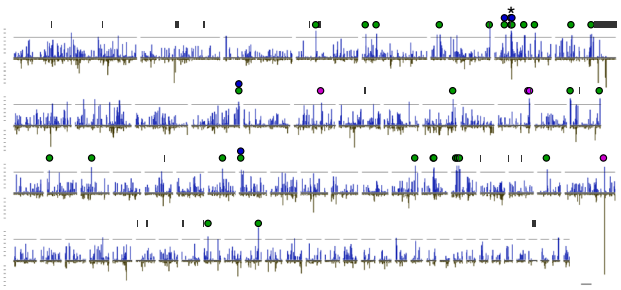

Fig. 2 | Genes differentiated between upstream and downstream White Creek *F. thermalis* are distributed throughout the genome. Genome-wide plot of per-locus nucleotide diversity, π (brown traces, tick mark increments are 5×10^{-3}) and the π-based measure of relative genetic differentiation between upstream and downstream samples, Φ_{ST} (blue traces, tick mark increments are 0.1), mapped to contigs (>15 kilobase pairs; scale bar is 25 kilobase pairs) of the White Creek *F. thermalis* genome draft assembly[28]. Absolute divergence, D_{XY}, is the product of Φ_{ST} and π. The indicated Φ_{ST} threshold value of 0.5 represents the cutoff for the ~1% most extreme values in the genome. Grey boxes indicate indel polymorphisms. F_{ST} outliers in Fig. 1c are distinguished by whether different alleles are fixed between group 1 and other genomes (green circles) or not (purple circles). Heterocyst-associated outliers are indicated by blue circles and the asterisk indicates the HEP island polymorphism.

spatial separation of these biochemically incompatible processes. These include the deposition of an outer envelope that protects nitrogenase from oxygen by restricting gas diffusion into the heterocyst. This protection comes at the potential costs of both N_2 limitation and adenosine triphosphate (ATP) production by aerobic respiration[33,34]. Aerobic respiration makes a substantial contribution ($21\pm1.7\%$ (s.e.); $n=22$ laboratory strains) to the energy provisioned for nitrogen fixation by *F. thermalis*. Consequently, maximizing nitrogen fixation entails allowing as much gas into the heterocyst as possible without irreversibly inactivating nitrogenase[33,35]. Two outliers (108_4197 (RefSeq: WP_009459096.1) and 4_47543 (RefSeq: WP_016871352.1)) encode signal transduction proteins that are differentially regulated during heterocyst development in *Anabaena* PCC 7120 (ref. [36]). The other four (25_24813 (RefSeq: WP_009459734.1), 36_24813 (RefSeq: WP_019495315.1), 37_24813 (RefSeq: WP_009459712.1) and 39_24813 (RefSeq: WP_016867130.1)) are located within a cluster of co-expressed genes involved in the deposition of the heterocyst envelope polysaccharide (HEP), which is required for heterocyst function in the presence of oxygen[37]. In group 1 genomes, there has been a deletion of two genes (hypothetical protein 1_4308 (RefSeq: WP_015112171.1) and annotated glycosyltransferase 2_4308 (RefSeq: WP_012407823.1)) between 37_24813 and 39_24813 that are generally conserved among heterocystous cyanobacteria (Fig. 2). Previously, we showed that a region of around five kilobase pairs surrounding this gene content polymorphism was the most differentiated region of the HEP expression island along White Creek[28].

The HEP island deletion increases heterocyst gas permeability. Based on the characterized function of the HEP expression island, we predicted that the two-gene deletion increases gas flux across the heterocyst envelope. If this is the case, we would expect the deletion to lower the resistance of nitrogen-fixation activity to high oxygen concentrations, but to also enhance activity under oxygen limitation due to greater ATP provision by aerobic respiration. Because genetic tools are not available for *F. thermalis*, we tested the consequences of the deletion in *Anabaena* species PCC 7120, which has the same genetic architecture in the island as *F. thermalis*[28]. A mutant strain (UMT12) with an in-frame deletion of both genes (*alr2828/alr2829* in *Anabaena*; Supplementary Fig. 4a and Supplementary Table 3) exhibited no discernible difference in heterocyst ultrastructure compared with the wild type (Supplementary Fig. 4b), as reported for *F. thermalis* with different HEP genotypes[28]. The two strains also performed similarly at atmospheric O_2 concentration when assayed for nitrogen fixation (Fig. 3a,b) and growth (Supplementary Fig. 4c). However, the nitrogen-fixation activity of the mutant was more sensitive to elevated oxygen concentration than the wild type (Fig. 3a,b). Activity was also greater under micro-oxic conditions, both in the light (Fig. 3b) and the dark (mean ± s.e. = 0.40 ± 0.049 versus 0.11 ± 0.035 parts per million ethylene produced per hour for UMT12 and the wild type, respectively), where the principal source of ATP for nitrogen fixation was aerobic respiration. We conclude that UMT12 heterocysts are more gas permeable. This phenotype requires the deletion of both genes due to epistasis between *alr2828* and *alr2829* (Fig. 3c). The individual in-frame deletions resulted in either a deleterious nitrogen-fixation phenotype (Δ*alr2828*) or nitrogen-fixation activity that is indistinguishable from the wild type (Δ*alr2829*).

Temperature effects on heterocyst function maintain the polymorphism. For haploid organisms, spatially varying selection (different alleles favoured in different environments) and negative-frequency dependence (rare alleles favoured) are the most likely potential mechanisms of balancing selection for maintaining diversity. The divergent HEP allele frequencies between upstream and downstream samples (Figs. 1 and 2) favour the spatial variation hypothesis.

We first considered whether the downstream habitat at White Creek favours a more permeable heterocyst due to generally lower oxygen availability. Gas flux into the heterocyst is determined by the product of the diffusion coefficient, which increases with temperature, and the concentration difference between the environment and the cell. At White Creek, *F. thermalis* can experience extreme fluctuations in oxygen concentration, ranging from midday supersaturation to near-anoxia in the dark (Supplementary Fig. 5a), as a consequence of the shifting balance between the rates of oxygenic photosynthesis and aerobic respiration within microbial mats over the diel cycle. However, we observed no systematic differences in oxygen availability between upstream and downstream sites (Supplementary Fig. 5b).

Instead, we propose that differences in prevailing temperature favour different HEP alleles due to a performance trade-off that arises from the flux of oxygen into the heterocyst and the cell's capacity to consume it. Over much of an organism's physiological range, enzyme kinetic rates increase much faster with increasing temperature than rates of gas diffusion. Consequently, the gap between potential and realized ATP production by aerobic respiration within the heterocyst is expected to widen and favour greater permeability[33]. However, physiological activity becomes negatively temperature dependent as the tolerance limit is approached. Beyond this point, it is expected to become increasingly difficult to maintain a micro-oxic environment within the heterocyst, thereby favouring a more effective diffusion barrier.

According to this model, the ancestral haplotype with a less permeable heterocyst should be favoured where the breakpoint temperature for nitrogen fixation and related processes by *F. thermalis* is exceeded. This is indeed the case at White Creek upstream sites. Nitrogen-fixation rates in the laboratory were lower at 55 °C (near the White Creek *F. thermalis* growth limit[38]) than at 37 °C (Fig. 4a), both in the light and under ATP limitation in the dark, which is indicative of heat stress on aerobic respiration. Furthermore, strains harbouring the HEP deletion ($n=7$) tended to be more sensitive at 55 °C (37 versus 20% lower activity; $P<0.02$; Fig. 4b), as expected if a more permeable heterocyst is at a greater disadvantage at higher temperature. This result was not due to a difference in heterocyst

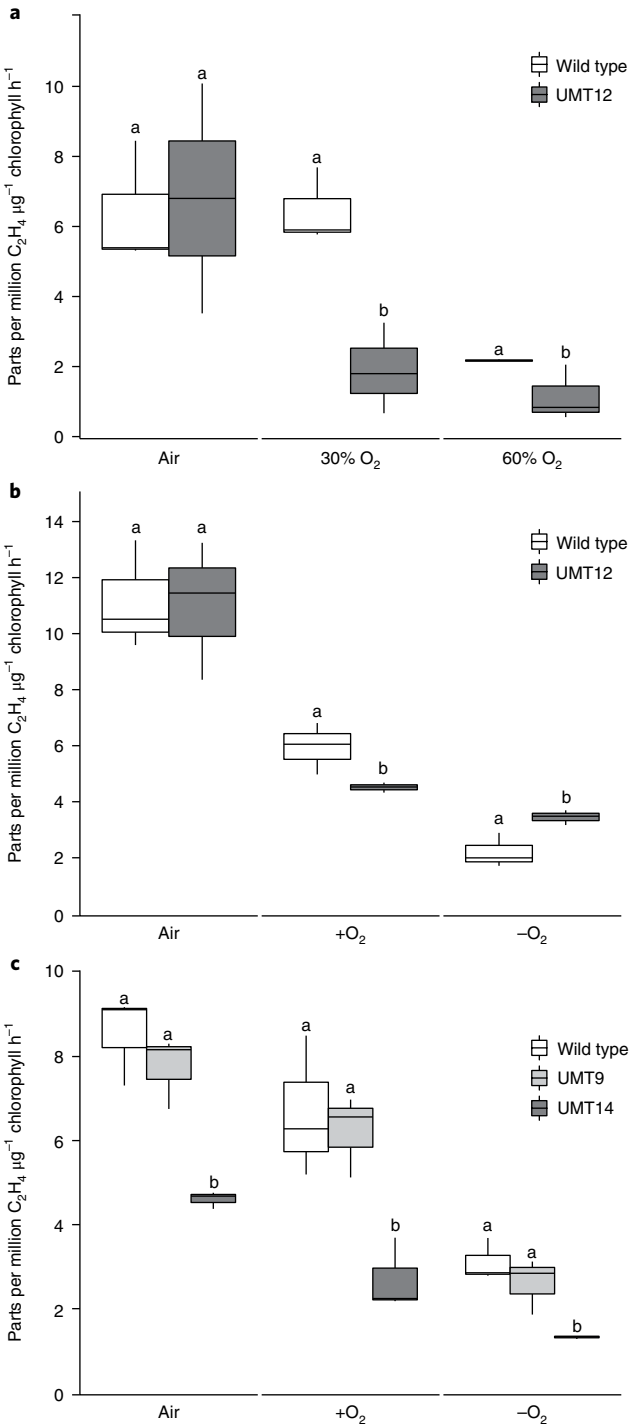

Fig. 3 | Nitrogen-fixation activities of wild-type *Anabaena* PCC 7120 and HEP deletion mutant strains. Activity was measured as chlorophyll-normalized ethylene production in acetylene reduction assays in the light for triplicate cultures for each strain. Box plots indicate the median (dark line) ± 1.5 × the interquartile range along with 25 and 75% quartiles. **a**,**b**, Activities of the wild-type and double-deletion mutant strain UMT12 (Δ*alr2828*/Δ*alr2829*) at increasing levels of atmospheric oxygen (**a**) and for air, 60% oxygen and micro-oxic conditions (**b**). Different letters within an experimental treatment indicate a significant *t*-test difference at the $P < 0.05$ level ($n = 6$). **c**, Activities of wild-type *Anabaena* PCC 7120 and single-deletion mutant strains UMT14 (Δ*alr2828*) and UMT9 (Δ*alr2829*) as in **b**. Different letters within anexperimental treatment indicate a significant difference by Tukey's honest significant difference test at the $P < 0.05$ level ($n = 9$).

content, which did not differ between the two genotypes ($n = 20$; $F_{1,18} = 0.015$; $P = 0.90$). Because carbon fixation was also lower at 55 °C (Fig. 4a), a reduction in the supply of reducing equivalents for nitrogen fixation provisioned to the heterocysts by vegetative cells may also contribute to this effect. Additionally, nitrogen-fixation rates at White Creek were lower at upstream sites (Supplementary Fig. 6), corroborating earlier findings that these mats are stressed at upstream temperatures[39].

The HEP island deletion was a unique and ancient event. Remarkably, a phylogenomic analysis indicated that the two HEP genes were also absent in diverse *F. thermalis* strains from around the world (Fig. 5a). Based on the conventional view of bacterial adaptation, we would expect this region to have been independently deleted multiple times during *F. thermalis* diversification in response to similar local selective pressures. Surprisingly, however, the data instead support a single, ancient origin and subsequent spread of the polymorphism. All of these strains harbour a homologous deletion allele between flanking genes 37_24813 and 39_24813, consisting of ~170 base pairs (bp) of sequence that bears no homology to the rest of the genome and includes a novel predicted promoter with conserved −10 and −35 elements that is absent in the ancestral haplotype (Fig. 5b). The amount of sequence variation within the deletion haplotype is comparable to that of flanking genes ($\pi = 0.035$ versus 0.043 and 0.026 for 37_24813 and 39_24813, respectively). We can therefore rule out a recent origin of the deletion that has spread by recombination into divergent genetic backgrounds, since variation in the region would be expected to stand out as younger than the surrounding sequence under this scenario. We conclude that the polymorphism arose at the very origin of *F. thermalis* diversification. In addition to White Creek, we also observed both alleles in a New Zealand population from which multiple strains have been isolated (Culture Collection of Microorganisms from Extreme Environments (CCMEE) 5196 and CCMEE 5198; Fig. 5a). This suggests that the co-occurrence of alleles may be common.

To estimate the age of the HEP polymorphism, we implemented uncorrelated gamma, white noise and log-normal autocorrelated Bayesian relaxed clock models with three fossil calibration points (see Methods) for a maximum likelihood topology of 15 taxa spanning the diversity of heterocyst-forming cyanobacteria (Supplementary Fig. 7). Models with both hard[40] and soft[41] calibration constraints were employed. The uncorrelated gamma model with soft constraints estimated the origin of the polymorphism (that is, the ancestor of *F. thermalis* strains) to be 74 million years ago (Ma) with a 95% confidence interval of 29–194 Ma (Fig. 6a). We obtained similar results under hard constraints as well as when only root calibration was used. In contrast, log-normal and white noise model estimates were considerably older (Supplementary Table 4). It is not clear whether our estimates are biased by recombination within *F. thermalis*. Although recombination is expected to distort the terminal branches of a phylogeny, the time to the most recent common ancestor does not appear to be impacted by recombination under maximum likelihood[42]. Our results are also in line with an estimate of 50–100 Ma for the origin of *F. thermalis* based on 16S ribosomal RNA divergence rate calibration[38], and our estimated age of 166 Ma for the divergence of *F. thermalis* and *Fischerella muscicola* (Fig. 6a) agrees with a previous analysis[43]. We can conclude that the HEP polymorphism has been maintained for tens of millions of years.

The sequence variation surrounding the HEP polymorphism also meets three predictions from population genetic theory for genomic regions either directly experiencing long-term balancing selection or genetically linked to them. First, we expect such regions to exhibit an excess of intermediate-frequency alleles compared with the expectation under selective neutrality[44]. This is indicated by positively skewed values of Tajima's *D*, a summary of the mutation

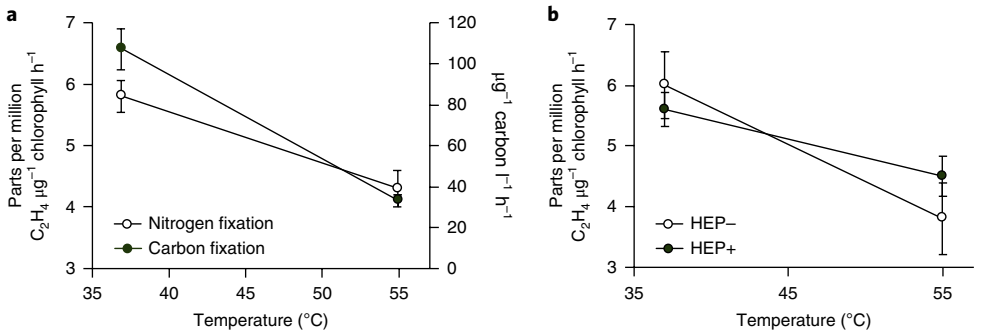

Fig. 4 | Temperature dependence of *F. thermalis* physiological rates. a, The nitrogen-fixation activity of culture homogenates was measured as ethylene production in acetylene reduction assays and carbon fixation was measured as the incorporation of radiolabelled bicarbonate for triplicate cultures of 22 randomly selected strains of White Creek *F. thermalis* at 37 and 55 °C, respectively. Cell homogenates were adjusted to an OD_{750} of 0.05 before the assay. Error bars are s.e. **b**, Nitrogen-fixation activity for strains with the HEP deletion (HEP–) or the ancestral haplotype (HEP+). Error bars are s.e.

site-frequency spectrum, and 37_24813 and 39_24813 are among the top 15 most positively skewed genes in the White Creek population (Fig. 6b). Furthermore, near the target of balancing selection, we also expect to observe elevated levels of polymorphism compared with divergence from a sister taxon[45]. Both 37_24813 and 39_24813 rank in the top 0.5% of genes for the ratio of polymorphism to divergence from *F. muscicola* PCC 7414 (Fig. 6b). Finally, for ancient balancing selection, we expect the region of excess polymorphism to only be very close to the target of selection, as recombination events break up linkage at greater distance over time[46].

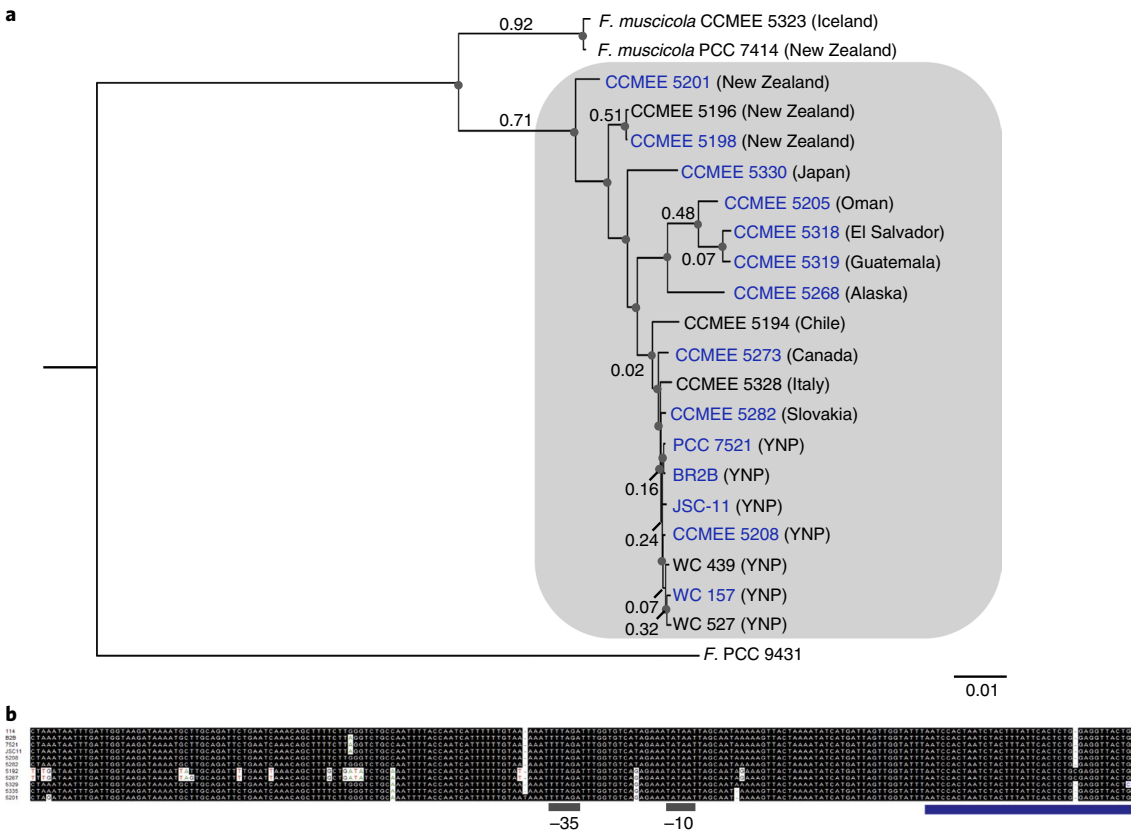

Fig. 5 | The HEP deletion was a unique event that occurred early during *F. thermalis* diversification. a, Maximum likelihood phylogeny of *F. thermalis* (grey box), *F. muscicola* and outgroup strain *Fischerella* PCC 9431 reconstructed from a concatenated alignment of 1,764 protein-coding genes according to a mixture model of nucleotide substitution with gamma rate heterogeneity. Strains are distinguished by whether HEP genes 1_4308 and 2_4308 are present (black) or absent (blue). Grey circles at the nodes indicate 100% ultrafast bootstrap support, and internode certainty values greater than 0.05 are indicated. The scale bar unit is the expected number of nucleotide substitutions per site. YNP, Yellowstone National Park. **b**, Alignment of intergenic sequences of the deleted region between 37_24813 and 39_24813 for divergent, geographically widespread *F. thermalis* strains indicates that the deletion allele is a distinct structural variant with a single origin. Completely conserved nucleotide positions are shown in black. The region exhibits no detectable homology to the rest of the genome, with the exception of the 3′ 37 nucleotides (blue bar), which is homologous to DNA 5′ of the start codon of 37_24813 in the ancestral haplotype. The −10 and −35 elements of a predicted promoter that is absent in the ancestral haplotype are indicated.

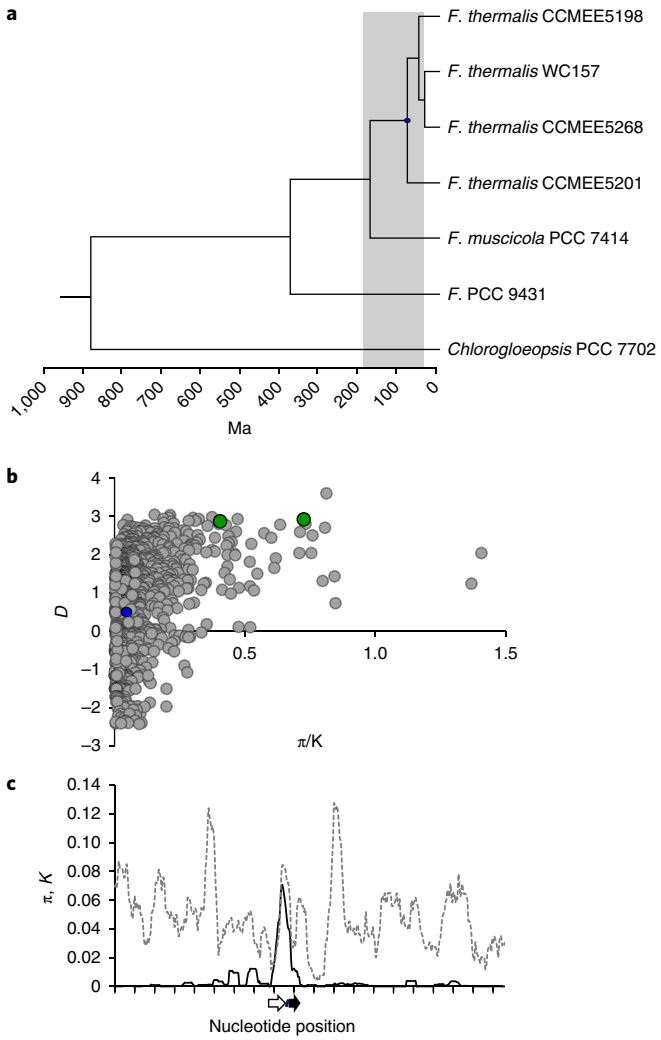

Fig. 6 | Long-term balancing selection on the HEP indel polymorphism. **a**, Chronogram of the origin of the HEP polymorphism at the ancestral node for *F. thermalis* (blue circle) with the 95% confidence interval shaded. The results shown are for an uncorrelated gamma Bayesian relaxed clock model. **b**, Tajima's *D* values for polymorphic genes in the White Creek sample ($n = 20$ genomes) are plotted against the ratio of polymorphism (π) to divergence (K) from orthologues in the genome of *F. muscicola* strain PCC 7414 ($n = 1{,}623$ loci). Genes flanking the HEP deletion polymorphism are green. Genome-wide averages are denoted by the blue circle. **c**, Sliding window of π (solid line) among White Creek *F. thermalis* and divergence from *F. muscicola* strain PCC 7414 (K, dashed line) for a roughly 20-kilobase region (axis ticks are 1-kilobase increments) surrounding the HEP deletion. Positions of flanking genes 39_24813 (open arrow) and 37_24813 (closed arrow) are indicated. Non-homologous DNA between indel polymorphism variants was removed from the analysis (hash mark). The window length was 100 nucleotides with a step size of 25 nucleotides.

A sliding window analysis confirmed that the variation maintained by selection is restricted to the region immediately flanking the HEP polymorphism (Fig. 6c). This is in accordance with a previous fine-mapping study that showed that linkage disequilibrium with flanking biallelic SNPs decayed within 5 kilobase pairs of the HEP polymorphism[28].

Discussion

Bacterial evolution is a dynamic process, with most variation expected to be fixed in or lost from populations on short evolutionary time scales. To our knowledge, the remarkably long persistence of distinct alleles observed here is unprecedented for a bacterium. This single, ancient origin contrasts with the prevailing view of the process of adaptation distilled from laboratory evolution experiments with microbial populations, which can be marked by rapid and repeatable change; for example, for deletions mediated by homologous recombination between insertion sequence elements[12,14]. It is likewise opposite to the pattern of recurrent mutations typically observed during the adaptation of human pathogens to individual hosts[13,15–18] and of *Escherichia coli* to mouse gut[47]. Our study raises the question of whether balancing selection acting on functional variation at niche-defining loci over extraordinarily long time periods and across broad spatial scales may play a more important role in the maintenance and distribution of microbial diversity than has been previously recognized. We note that this insight was only possible because of our sampling of both population and global diversity, together with our use of functional validation to explicitly connect genotype and phenotype. Future integrative studies with a similar cross-section of divergence will help to address how pervasive ancient variation is for the evolutionary dynamics of bacteria.

Methods

Genomics and bioinformatics. High-molecular-weight genomic DNA was extracted from independent, directly isolated White Creek *F. thermalis* strains[27] and from *Fischerella* strains obtained from the University of Oregon's CCMEE, as described in ref. [48]. Indexed libraries with an insert size of ~400 bp were prepared from fragmented genomic DNA using the NEXTflex DNA-Seq Kit and barcodes (Bioo Scientific). Libraries were sequenced on an Illumina HiSeq 2000 (50-bp paired-end) by the Genomic Services Lab at the HudsonAlpha Institute for Biotechnology (Huntsville, AL), which resulted in ~1 gigabase pairs of sequence per library. CCMEE strains of *F. thermalis* and *F. muscicola* were sequenced on an Illumina HiSeq 2000 (150-bp paired-end) by the University of Pittsburgh Center for Evolutionary Genomics Research. Sequence reads that failed the Illumina chastity filter (CASAVA version 1.8) were removed and remaining FASTQ files for each library were trimmed of leading and trailing low-quality bases and Ns (ambiguous nucleotides), as well as filtered based on read length and sequence quality with Trimmomatic version 0.22 (ref. [49]). Draft genome sequences for each strain were subsequently assembled with Velvet[50] version 1.2.03 using parameter settings (hash length, coverage cutoff, and so on) that were manually optimized to maximize the N50 (Supplementary Table 1). Assembled contigs for each strain were annotated with the National Center for Biotechnology Information (NCBI) Prokaryotic Genome Annotation Pipeline and deposited in GenBank.

Protein coding sequences (CDS) were extracted from GenBank files for the individual genomes, and those with the same best hit as the pooled White Creek *F. thermalis* reference genome[28] in local blastn queries were assigned as orthologues. For White Creek strains, population parameters were estimated for aligned full-length sequence data for each orthologue with at least 10 sequences in the sample ($n = 4{,}766$) using custom Perl scripts, including: the number and frequency of alleles segregating in the population; F_{ST}, the allele-based measure of relative genetic differentiation between sub-populations ($\frac{H_T - H_S}{H_T}$, where H is the haplotype diversity at a haploid locus, $1 - \sum_{i=1}^{j} p_i^2$, p_i is the frequency of allele i, H_S is the mean within sub-populations and H_T is the diversity of the total sample); the number of SNPs; mean nucleotide diversity, π; the π-based measure of relative genetic differentiation between sub-populations, Φ_{ST}[51]; the absolute genetic differentiation between sub-populations, D_{XY} (that is, the numerator of Φ_{ST}); and Tajima's D[44]. In addition, average Jukes–Cantor[52] corrected sequence divergence from the closely related sister taxon *F. muscicola* PCC 7414 (ref. [53]) was estimated for CDS that shared > 70% sequence identity and 75% sequence length overlap (the mean divergence was 5% for 3,250 loci). F_{ST} outliers were identified by the likelihood approach of ref. [54] implemented in the OutFLANK R package available at https://github.com/whitlock/OutFLANK. OutFLANK infers a neutral distribution for F_{ST} from a trimmed dataset and uses this distribution to assign q-values for each locus. The approach has an improved false discovery rate compared with other methods[54].

To estimate the core and pan-genomes of the White Creek *F. thermalis* population, a non-redundant reference database of CDS from the 20 genomes with a length greater than 150 nucleotides ($n = 5{,}473$) was queried with a blastn search (E-value < 0.01, percent identity >90% and >50% length overlap) against genome contig databases for individual strains to create a CDS presence–absence array for each genome. Core and pan-genome collector's curves (that is, the number of shared and total CDS as a function of the number of genomes analysed) were estimated with a custom Perl script for 1,000 permutations of array addition in random input order. To analyse clustering of genomes based on gene content, a Jaccard distance matrix was first produced from a gene content presence–absence matrix of the

White Creek genomes using the 'vegdist' function in the vegan package of R (http://CRAN.R-project.org/package=vegan). Jaccard similarity values were in the range of 95–99.9%. Two-dimensional non-metric multidimensional scaling analysis of this distance matrix was performed using the 'isoMDS' function in the R 'MASS' package (convergence was attained after 15 iterations).

To analyse the clustered regularly interspaced short palindromic repeats spacer content of genomes, CRISPRFinder[55] was first used to identify a total of 5,203 spacers in the 20 genomes. From these data, custom Perl scripts were then used to create a database of 3,427 non-redundant spacers (100% identical sequence in region of overlap) observed in the population sample. This was queried with blastn searches against individual genome databases to obtain presence–absence and location information for the individual spacers for each genome.

The genealogical relationships of White Creek strains were inferred with a neighbour net analysis implemented in SplitsTree version 4.14.4 (www.splitstree.org) for a concatenated alignment of 1,503,580 nucleotides (and 4,080 SNPs) from 2,297 genes for which complete sequence data at a locus was available for at least 18 of the 20 White Creek strains. Highly similar results were obtained with different models of estimating evolutionary distance between sequences (uncorrected, Jukes–Cantor, and Hasegawa–Kishino–Yano), as well as for other network-building approaches such as a median network. The minimum number of recombination events, R_M[29], was estimated from the 4,075 two-variant SNPs in the dataset using DnaSP version 5.

Media and growth conditions. *Anabaena* species PCC 7120 and derivative strains were grown in BG-11 or BG-11$_0$ (BG-11 lacking $NaNO_3$) medium buffered with 10 mM HEPES (final) at pH 8.0. Cultures were grown at 30 °C (unless otherwise stated) with gentle agitation and continuous illumination from cool white fluorescent bulbs at ~40 μmol photons m^{-2} s^{-1} (low light) unless otherwise noted. When appropriate, streptomycin and spectinomycin were added at 2.5 μg ml^{-1} each, or neomycin at 45 μg ml^{-1}. *E. coli* strains were grown in liquid lysogeny broth at 37 °C. Antibiotics were added when appropriate at the following final concentrations: streptomycin (25 μg ml^{-1}), chloramphenicol (30 μg ml^{-1}), kanamycin (50 μg ml^{-1}) and ampicillin (100 μg ml^{-1}).

Construction of HEP deletions in *Anabaena*. Double and single chromosomal in-frame deletion mutants were constructed for *alr2828* and *alr2829* in *Anabaena* PCC 7120. All constructs were built using two rounds of polymerase chain reaction and cloned into a suicide vector (pRL277 or pRL278) using standard cloning protocols. Plasmid pUMT3 is a suicide vector that carries a deletion that spans most of *alr2828* and *alr2829*. This construct results in a deletion with the first few bases of *alr2828* and the last few bases of *alr2829* (resulting in a fusion of the two genes). A 1,563-bp region, including 1,496 bp upstream of *alr2828*, was amplified by polymerase chain reaction using the primers Pes1 and Pes9 (Supplementary Table 3). A 1,770-bp region containing 1,705 bp downstream of *alr2829* was amplified using the primers Pes8 and Pes10. These two fragments were used to finish the deletion construct using overlapping polymerase chain reaction extension with Pes1 and Pes8. This deletion construct was cloned as a fragment (with SpeI restriction enzyme sites introduced with primers) into pRL277 (ref. [56]). Single gene deletions were constructed similarly. The suicide vector pUMT1 carries the single *alr2829* deletion. Both the upstream (1,585 bp, using primers Pes5 and Pes6) and downstream regions (2,191 bp, primers Pes7 and Pes8) were amplified. The overlapping polymerase chain reaction was finished using primers Pes5 and Pes8 and the product was cloned into pRL278 (ref. [57]) using a primer-introduced SpeI site. The suicide vector pUMT2 (carrying the in-frame deletion of *alr2828*) was constructed using primers Pes1 and Pes2 for the upstream region (1,554 bp) and primers Pes3 and Pes4 for the downstream region (1,699 bp). The products were used to finish the deletion using primers Pes1 and Pes4 and cloned into pRL278 using the SpeI site. The integrity of all constructs was verified with sequencing. Each plasmid was introduced to the *Anabaena* wild-type strain via conjugation as described previously[58], with the plasmid first introduced to UC585 (ref. [59]), which carries the helper plasmids pRK24 and pRL528 required for biparental conjugation.

Growth experiments. Wild-type and HEP deletion mutant *Anabaena* cultures were grown in BG-11 (+N) or BG-11$_0$ (−N) at 30 °C with agitation. The light intensity used was either 40 μmol photons m^{-2} s^{-1} (low light) or ~100 μmol photons m^{-2} s^{-1} (high light). Each experiment was performed in biological triplicates. Generation times were estimated from the exponential growth phase of each culture.

Transmission electron microscopy. *Anabaena* strains were grown in BG-11 medium to a chlorophyll *a* concentration of 2–4 μg ml^{-1}, determined by methanol extractions[60]. Cells were washed twice in BG-11$_0$ medium and incubated for 2 days at 30 °C with agitation with light in BG-11$_0$. Cells were prepared for transmission electron microscopy as described previously[28] and imaged using a Hitachi H7100 TEM at 75 kV at the University of Montana EMtrix electron microscopy facility.

***Anabaena* acetylene reduction assays.** Nitrogen-fixation rates were estimated by the acetylene reduction method[61]. All *Anabaena* cultures were grown in triplicate with a starting OD_{750} of 0.005 in nitrogen-free BG-11$_0$. Exponentially growing cells were harvested after four days. After adjusting the chlorophyll *a* concentration to 5 μg ml^{-1}, 5 ml samples from each culture were sealed in 20 ml crimp-sealed vials. Different oxygen environments were provided by the replacement of a corresponding air volume from the headspace with oxygen. To obtain microoxic conditions, we added 10 μM of 3-(3,4-dichlorophenyl)-1,1-dimethylurea to inhibit oxygenic photosynthesis and flushed with argon for 3 min before the addition of acetylene. Samples were incubated with agitation at 30 °C for 2 h following the addition of 5 ml of acetylene gas (generated by the addition of 5 g of calcium carbide to 100 ml of deionized water). For experiments to determine the resistance of nitrogen fixation to increasing oxygen concentrations (air, 30% O_2 or 60% O_2), strains were grown and assayed at a light intensity of 40 μmol photons m^{-2} s^{-1}. For experiments investigating nitrogen fixation at both low and high oxygen concentrations (microxia, air or 60% O_2), strains were grown and assayed at a light intensity of 200 μmol photons m^{-2} s^{-1} or in the dark. Assays were terminated by injecting ~15 ml of sample headspace into a pre-evacuated 5 ml crimp vial. Ethylene production was measured using flame-ionization detection gas chromatography with a Shimadzu GC-2014 and estimated using a standard curve, blank corrected against parallel incubation vials that contained only BG-11$_0$ medium.

***F. thermalis* strain acetylene reduction assays and carbon fixation assays.** Nitrogen fixation was assayed at the temperature of growth (either 37 or 55 °C) for growing cultures of White Creek *F. thermalis* strains incubated in the absence of combined nitrogen (ND Medium) with a 12/12 h light/dark cycle and at a light intensity of 105 ± 5 μmol photons m^{-2} s^{-1} provided by cool white fluorescent bulbs. In an experiment, three replicate cultures derived from the same inoculum were assayed from each temperature for each strain, and two experiments were performed for each strain using independent starting cultures as inoculum. Cultures were homogenized using a tissue grinder, which broke up large clumps of trichomes but kept long chains containing both vegetative cells and heterocysts intact. Homogenates were adjusted to an OD_{750} of 0.05 ± 0.003 with fresh ND Medium; optical density has a linear relationship with *F. thermalis* cell dry mass (Pearson correlation coefficient (r) = 0.97, $P < 0.001$). Acetylene reduction assays were carried out in 10 ml of ND Medium in 20 ml crimp-sealed vials, with a light and a dark replicate for each sub-line. Samples were incubated for 4 h following the addition of 5 ml of acetylene gas. Assays were terminated and ethylene production was measured as described above. Carbon fixation rates were concurrently estimated by the incorporation of ^{14}C-bicarbonate as previously described[62].

In situ acetylene reduction assays. In July and August 2013, triplicate streamer tufts collected from White Creek sites WC1 (44.537° N 110.804° W), WC3 (44.533° N 110.799° W) and WC5 (44.531° N 110.794° W) were assayed for acetylene reduction under ambient mid-afternoon conditions. Assays were carried out in 10 ml of White Creek water in 20 ml crimp-sealed vials following the addition of 5 ml of acetylene gas. Incubations were terminated after 1 h and ethylene production was measured as described above. Estimates were blank corrected against parallel incubation vials containing only White Creek water and were normalized by the amount of chlorophyll *a* in the assay.

Oxygen measurements. We measured in situ oxygen concentrations at White Creek through a 24 h period on 5–6 September 2014 with a Unisense PA2000 picoammeter and an OX-500 Clark-type oxygen microsensor calibrated according to the manufacturer's specifications. Nine streamer mats each at sites WC1 and WC5 were sampled at distal, mid-point and base regions of the mat (see picture inset of Supplementary Fig. 5). The oxygen solubility at saturation reflects both environmental temperature and oxygen partial pressure (78.5 kPa) at the elevation of the field site (2,225 m). Diffusion coefficient values for oxygen at different temperatures were estimated using gas tables available at http://www.unisense.com/technical_information/. For a given heterocyst diffusion barrier efficiency and assuming an anoxic heterocyst, the estimated oxygen flux is proportional to the product of the environmental concentration and the diffusion coefficient of oxygen[33].

Phylogenomic analyses. Separate FASTA-formatted files for CDS were created as above from genome assemblies for 22 strains (including 18 obtained as a part of this study), representing the diversity of thermophilic *Fischerella*[38] and the outgroup strain *Fischerella* PCC 9431 (NCBI GenBank assembly accession GCA_000447295.1; the average nucleotide divergence from thermophilic *Fischerella* is ~0.15). These were subsequently aligned with CLUSTAL W using a customized BioPerl script. A concatenated alignment was produced with Sequence Matrix version 1.8 for the 1,764 CDS (1,156,218 bp) for which sequence data were available for at least 20 strains. A maximum likelihood tree with 1,000 ultrafast bootstrap replicates was reconstructed with IQ-TREE version 1.5.5 (ref. [63]) according to a mixture model of substitution (the Hasegawa–Kishino–Yano and General Time Reversible models)—a discrete approximation of a gamma distribution with four rate categories and empirical base frequencies. That is, the probability of belonging to each of the eight possible mixture classes (the product of the two substitution models and four gamma rate categories) was computed

for each nucleotide position in the alignment. Support for bipartitions in the concatenated tree was additionally provided with internode certainty values[64]—an entropy measure of the degree of conflict for each internal branch due to discordance between gene trees in the concatenated alignment. Discordance can arise by several mechanisms, both technical and biological (including incomplete lineage sorting and horizontal transfer). An internode certainty value of 1 indicates that all of the individual gene trees support the concatenated topology, whereas a value of 0 indicates equal support for a bipartition that conflicts with the concatenated tree. Following ref. [65], individual maximum likelihood trees for each gene were inferred with IQ-TREE version 1.5.5 with 1,000 ultrafast bootstrap replicates according to the nucleotide substitution model selected for each gene with ModelFinder. Bootstrap consensus trees for the 1,000 ultrafast bootstrap replicate trees for each gene were produced with the CONSENSE programme of the PHYLIP package (http://evolution.genetics.washington.edu/), and these were used to calculate internode certainty values with RAxML version 8.2.4 (ref. [66]) that were corrected for partial gene trees (that is, those with different subsets of taxa with available sequence data)[67].

Dating the origin of the HEP polymorphism. A maximum likelihood amino acid phylogeny was first reconstructed with IQ-Tree version 1.5.5 from genome data for 15 taxa spanning the diversity of heterocyst-forming cyanobacteria (including four *F. thermalis* strains and *F. muscicola* PCC 7414) and the outgroup strain *Chroococcidiopsis* PCC 7203. A total of 1,316 individual loci that were at least 50 amino acids in length and for which data were available for at least 14 strains were aligned and concatenated as above (317,324 amino acids in the final alignment). To infer a concatenated tree, a guide tree was first reconstructed using the LG general matrix[68] with empirical amino acid frequencies and discrete gamma rate heterogeneity with four rate categories (that is, the LG + F + G model). Given the guide tree, we next estimated the parameters of a posterior mean site-frequency model[69]. The model (LG + C60 + F + G) accounts for site-specific patterns of amino acid substitutions in the concatenated alignment by applying an LG matrix for 60 classes of amino acid profiles and has been shown to not be prone to long branch attraction artefacts. The final maximum likelihood analysis was conducted with the inferred posterior mean site-frequency model and with 1,000 ultrafast bootstrap replicates, and internode certainty values were estimated as above.

Next, we used Bayesian relaxed clock models implemented in PhyloBayes version 4.1 (ref. [70]) to estimate the divergence time of the *F. thermalis* clade. The models require a fixed topology, for which we used the topology of the maximum likelihood amino acid phylogeny above. Site-specific variation in evolutionary rates was incorporated with a CAT-Poisson substitution model with discrete (four categories) gamma rate heterogeneity for uncorrelated gamma[71], white noise[72] and log-normal autocorrelated[73] relaxed clock models, respectively. The uncorrelated gamma model has previously been shown to be the best-fitting molecular clock model for cyanobacteria[74]. We used three fossil calibrations: (1) the divergence at the root of heterocyst-forming cyanobacteria and *Chroococcidiopsis* PCC 7203, with a minimum age of 1.96 billion years for the presence of heterocyst-forming cyanobacteria in the fossil record[75] and a maximum age of 2.45 billion years based on oxygenation of the planet; (2) a minimum age of 400 Ma for the origin of the true-branching multicellular cyanobacteria, *Fischerella* and *Chlorogloeopsis*[76]; and (3) a maximum age of 110 Ma for the bifurcation between *Richelia intracellularis* HH01 (a symbiont of the diatom *Hemiaulus hauckii*) and *R. intracellularis* HM01 (a symbiont of *Hemiaulusmembranaceus*) from fossil evidence for *Hemiaulus*[77]. Models with both hard[40] and soft[41] constraints on fossil calibrations were used. Models with soft constraints allowed 2.5% of the posterior probability mass to fall outside the calibration points. The prior distributions for divergence times were either a uniform distribution (for hard constraint models) or a birth-death process (as required for soft constraint models). Two independent Markov chain Monte Carlo chains were run for each model until chain convergence was attained, as assessed by a maximum discrepancy less than 0.3 and an effective sample size greater than 50 for each parameter.

Statistical models. To estimate the effect of fixed differences between group 1 and other genomes on growth rate, strains assayed at 37 and 55 °C under nitrogen-fixing conditions in ref. [27] were assigned to a genome group (group 1 or other) based on their HEP deletion genotypes. A within-temperature model with genome group and strain nested within genome group as random effects was analysed with JMP version 10 (SAS Institute), and variance components were estimated by restricted maximum likelihood. A full model including data for both temperatures was used to estimate the genome group × temperature interaction. For the light and dark acetylene reduction assays with *F. thermalis* strains, we used a mixed-effects model with strain nested within genome group and genome group as random effects, and temperature as a fixed effect.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code availability. Custom Perl scripts used in genome analyses are available at http://www.nature.com/ismej/journal/v11/n1/abs/ismej2016105a.html.

Data availability. Annotated genome assemblies produced for this study have been deposited in GenBank under the following accessions: NGZS00000000; NGZT00000000; NGZU00000000; NGZV00000000; NGZW00000000; NGZX00000000; NGZY00000000; NGZZ00000000; NHAA00000000; NHAB00000000; NHAC00000000; NHAD00000000; NHAE00000000; NHAF00000000; NHAG00000000; NHAH00000000; NHAI00000000; NHAJ00000000; NHAK00000000; NQMD00000000; NQME00000000; NQMF00000000; NMQA00000000; NMQB00000000; NMQC00000000; NMQD00000000; NMQE00000000; NMQF00000000; NMQG00000000; NMQH00000000; NMQI00000000; NMQJ00000000; NMQK00000000; and NRQW0000000. Raw genome sequence data are available at the NCBI Sequence Read Archive (SRP075623).

Received: 13 March 2017; Accepted: 29 November 2017;
Published online: 15 January 2018

References

1. Hedrick, P. W. Genetic polymorphism in heterogeneous environments: the age of genomics. *Ann. Rev. Ecol. Evol. Syst.* **37**, 67–93 (2006).
2. Castric, V. & Vekemans, X. Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances. *Mol. Ecol.* **13**, 2873–2889 (2004).
3. Delph, L. & Kelly, J. On the importance of balancing selection in plants. *New Phytol.* **201**, 45–56 (2014).
4. Wu, J., Saupe, S. J. & Glass, N. L. Evidence for balancing selection operating at the *het*-c heterokaryon incompatibility locus in a group of filamentous fungi. *Proc. Natl. Acad. Sci. USA* **95**, 12398–12403 (1998).
5. Klein, J., Satta, Y., O'hUigin, C. & Takahata, N. The molecular descent of the major histocompatibility complex. *Annu. Rev. Immunol.* **11**, 269–295 (1993).
6. Ségurel, L. et al. The ABO blood group is a trans-species polymorphism in primates. *Proc. Natl Acad. Sci. USA* **109**, 18493–18498 (2012).
7. Leffler, E. et al. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**, 1578–1582 (2013).
8. Stahl, E., Dwyer, G., Mauricio, R. & Kreitman, M. Dynamics of disease resistance polymorphism at the Rpm1 locus of *Arabidopsis*. *Nature* **400**, 667–671 (1999).
9. Elena, S. F. & Lenski, R. E. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet.* **4**, 457–469 (2003).
10. Koskella, B. & Vos, M. Adaptation in natural microbial populations. *Ann. Rev. Ecol. Evol. Syst.* **46**, 503–522 (2015).
11. Andersson, D. I. & Hughes, D. Gene amplification and adaptive evolution in bacteria. *Annu. Rev. Genet.* **43**, 167–195 (2009).
12. Lee, M.-C. & Marx, C. J. Repeated, selection-driven genome reduction of accessory genes in experimental populations. *PLoS Genet.* **8**, e1002651 (2012).
13. Laehnemann, D. et al. Genomics of rapid adaptation to antibiotics: convergent evolution and scalable sequence amplification. *Genome Biol. Evol.* **6**, 1287–1301 (2014).
14. Tenaillon, O. et al. The molecular diversity of adaptive convergence. *Science* **335**, 457–461 (2012).
15. Lieberman, T. D. et al. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat. Genet.* **43**, 1275–1280 (2011).
16. Roumagnac, P., Weill, F., Dolecek, C., Baker, S. & Brisse, S. Evolutionary history of *Salmonella typhi*. *Science* **314**, 1301–1304 (2006).
17. Croucher, N., Harris, S., Fraser, C. & Quail, M. Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434 (2011).
18. Kos, N. V. et al. Comparative genomics of vancomycin-resistant *Staphylococcus aureus* strains and their positions within the clade most commonly associated with methicillin-resistant *S. aureus* hospital-acquired infection in the United States. *mBio* **3**, e00112-12 (2012).
19. Thomas, J. C., Godfrey, P. A., Feldgarden, M. & Robinson, D. A. Candidate targets of balancing selection in the genome of *Staphylococcus aureus*. *Mol. Biol. Evol.* **29**, 1175–1186 (2012).
20. Seifert, S. N., Khatchikian, C. E., Zhou, W. & Brisson, D. Evolution and population genomics of the Lyme borreliosis pathogen, *Borrelia burgdorferi*. *Trends Genet.* **31**, 201–207 (2015).
21. Jacquot, M. et al. Comparative population genomics of the *Borrelia burgdorferi* species complex reveals high degree of genetic isolation among species and underscores benefits and constraints to studying intra-specific epidemiological processes. *PLoS ONE* **9**, e94384 (2014).
22. Haven, J. et al. Pervasive recombination and sympatric genome diversification driven by frequency-dependent selection in *Borrelia burgdorferi*, the Lyme disease bacterium. *Genetics* **189**, 951–966 (2011).
23. Kashtan, N. et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* **344**, 416–420 (2014).
24. Shapiro, B. J. et al. Population genomics of early events in the ecological differentiation of bacteria. *Science* **336**, 48–51 (2012).

25. Cordero, O. X. & Polz, M. F. Explaining microbial genomic diversity in light of evolutionary ecology. *Nat. Rev. Microbiol.* **12**, 263–273 (2014).
26. Rodriguez-Valera, F. et al. Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.* **7**, 828–836 (2009).
27. Miller, S. R., Williams, C., Strong, A. L. & Carvey, D. Ecological specialization in a spatially structured population of the thermophilic cyanobacterium *Mastigocladus laminosus*. *Appl. Environ. Microbiol.* **75**, 729–734 (2009).
28. Wall, C. A., Koniges, G. J. & Miller, S. R. Divergence with gene flow in a population of thermophilic bacteria: a potential role for spatially varying selection. *Mol. Ecol.* **23**, 3371–3383 (2014).
29. Hudson, R. R. & Kaplan, N. L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164 (1985).
30. Barton, N. Clines in polygenic traits. *Genet. Res.* **74**, 223–236 (1999).
31. Miller, S. R., Purugganan, M. D. & Curtis, S. E. Molecular population genetics and phenotypic diversification of two populations of the thermophilic cyanobacterium *Mastigocladus laminosus*. *Appl. Environ. Microbiol.* **72**, 2793–2800 (2006).
32. Kumar, K., Mella-Herrera, R. A. & Golden, J. W. Cyanobacterial heterocysts. *Cold Spring Harb. Perspect. Biol.* **2**, a000315 (2010).
33. Staal, M., Metsman, F. J. R. & Stal, L. J. Temperature excludes N_2-fixing heterocystous cyanobacteria in the tropical oceans. *Nature* **425**, 504–507 (2003).
34. Walsby, A. The permeability of heterocysts to the gases nitrogen and oxygen. *Proc. R. Soc. Lond. B* **226**, 345–366 (1985).
35. Stal, L. J. Is the distribution of nitrogen-fixing cyanobacteria in the oceans related to temperature? *Environ. Microbiol.* **11**, 1632–1645 (2009).
36. Flaherty, B. L., van Nieuwerburgh, F. V., Head, S. R. & Golden, J. W. Directional RNA deep sequencing sheds new light on the transcriptional response of *Anabaena* sp. strain PCC 7120 to combined-nitrogen deprivation. *BMC Genom.* **12**, 332 (2011).
37. Huang, G. et al. Clustered genes required for the synthesis of heterocyst envelope polysaccharide in *Anabaena* sp. strain PCC 7120. *J. Bacteriol.* **187**, 1114–1123 (2005).
38. Miller, S. R., Castenholz, R. W. & Pedersen, D. Phylogeography of the thermophilic cyanobacterium *Mastigocladus laminosus*. *Appl. Environ. Microbiol.* **73**, 4751–4759 (2007).
39. Stewart, W. D. P. Nitrogen fixation by blue-green algae in Yellowstone thermal areas. *Phycologia* **9**, 261–268 (1970).
40. Kishino, H., Thorne, J. L. & Bruno, W. J. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* **18**, 352–361 (2001).
41. Yang, Z. & Rannala, B. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.* **23**, 212–226 (2005).
42. Schierup, M. H. & Hein, J. Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879–891 (2000).
43. Schirrmeister, B. E., Gugger, M. & Donoghue, P. C. J. Cyanobacteria and the Great Oxidation Event: evidence from genes and fossils. *Palaeontology* **58**, 769–785 (2015).
44. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
45. Hudson, R. R. & Kaplan, N. L. The coalescent process in models with selection and recombination. *Genetics* **120**, 831–840 (1988).
46. Charlesworth, D. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* **2**, e64 (2006).
47. Barroso-Batista, J. et al. The first steps of adaptation of *Escherichia coli* to the gut are dominated by soft sweeps. *PLoS Genet.* **10**, e1004182 (2014).
48. Inskeep, W. P. et al. The YNP metagenome project: environmental parameters responsible for microbial distribution in the Yellowstone geothermal ecosystem. *Front. Microbiol.* **4**, 67 (2013).
49. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
50. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
51. Nei, M. Evolution of human races at the gene level. *Prog. Clin. Biol. Res.* **103**, 167–181 (1982).
52. Jukes, T. H. & Cantor, C. R. in *Mammalian Protein Metabolism* (ed. Munro, H. N.) 21–132 (Academic Press, New York, 1969).
53. Dagan, T. et al. Genomes of stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome Biol. Evol.* **5**, 31–44 (2013).
54. Whitlock, M. C. & Lotterhos, K. E. Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F_{ST}. *Am. Nat.* **186**, S24–S36 (2015).
55. Grissa, I., Vergnaud, G. & Pourcel, C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* **35**, W52–W57 (2007).
56. Black, T. A., Cai, Y. & Wolk, C. P. Spatial expression and autoregulation of *hetR*, a gene involved in the control of heterocyst development in *Anabaena*. *Mol. Microbiol.* **9**, 77–84 (1993).
57. Cai, Y. P. & Wolk, C. P. Use of a conditionally lethal gene in *Anabaena* sp. strain PCC 7120 to select for double recombinants and to entrap insertion sequences. *J. Bacteriol.* **172**, 3138–3145 (1990).
58. Elhai, J. & Wolk, C. P. Conjugal transfer of DNA to cyanobacteria. *Methods Enzymol.* **167**, 747–754 (1988).
59. Liang, J., Scappino, L. & Haselkorn, R. The *patB* gene product, required for growth of the cyanobacterium *Anabaena* sp. strain PCC 7120 under nitrogen-limiting conditions, contains ferredoxin and helix-turn-helix domains. *J. Bacteriol.* **175**, 1697–1704 (1993).
60. Meeks, J. C., Wycoff, K. L., Chapman, J. S. & Enderlin, C. S. Regulation of expression of nitrate and dinitrogen assimilation by *Anabaena* species. *Appl. Environ. Microbiol.* **45**, 1351–1359 (1983).
61. Stewart, W. D. P., Fitzgerald, G. P. & Burris, R. H. In situ studies on N_2 fixation using the acetylene reduction technique. *Proc. Natl Acad. Sci. USA* **58**, 2071–2078 (1967).
62. Miller, S. R., Wingard, C. E. & Castenholz, R. W. Effects of visible light and UV radiation on photosynthesis in a population of a hot spring cyanobacterium, a *Synechococcus* sp., subjected to high-temperature stress. *Appl. Environ. Microbiol.* **64**, 3893–3899 (1998).
63. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2014).
64. Salichos, L., Stamatakis, A. & Rokas, A. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol. Biol. Evol.* **31**, 1261–1271 (2014).
65. Salichos, L. & Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327–331 (2013).
66. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
67. Kobert, K., Salichos, L., Rokas, A. & Stamatakis, A. Computing the internode certainty and related measures from partial gene trees. *Mol. Biol. Evol.* **33**, 1606–1617 (2016).
68. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
69. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst. Biol.* https://doi.org/10.1093/sysbio/syx068 (2017).
70. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).
71. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
72. Lepage, T., Bryant, D., Philippe, H. & Lartillot, N. A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.* **24**, 2669–2680 (2007).
73. Thorne, J. L., Kishino, H. & Painter, I. S. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**, 1647–1657 (1998).
74. Sánchez-Baracaldo, P., Ridgwell, A. & Raven, J. A. A neoproterozoic transition in the marine nitrogen cycle. *Curr. Biol.* **24**, 652–657 (2014).
75. Tomitani, A., Knoll, A. H., Cavanaugh, C. M. & Ohno, T. The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. *Proc. Natl Acad. Sci. USA* **103**, 5442–5447 (2006).
76. Croft, W. N. & George, E. A. Blue-green algae from the Middle Devonian of Rhynie, Aberdeenshire. *Bull. Br. Mus. Nat. Hist. Geol.* **3**, 339–353 (1959).
77. Sims, P. A., Mann, D. G. & Medlin, L. K. Evolution of the diatoms: insights from fossil, biological and molecular data. *Phycologia* **45**, 361–402 (2006).

Acknowledgements

We thank the Cory Cleveland laboratory group for use of their gas chromatograph, J. Meeks and his laboratory group for strains and technical advice, and J. Driver at the University of Montana EMtrix electron microscopy facility. We also thank D. Vanderpool for advice and sharing custom Python scripts used in the phylogenomics analyses. We are grateful to L. Fishman, J. McCutcheon, M. Polz, F. Rosenzweig and A. Woods for reading and commenting on earlier versions of the manuscript. Field work was conducted under National Park Service research permit YELL-5482. This work was supported by US National Science Foundation award IOS-1110819 and by NASA Astrobiology Institute award NNA15BB04A to S.R.M.

Author contributions

S.R.M. conceived the study. C.A.W., E.B.S. and S.R.M. performed the genome sequencing, assembly and annotation. P.R.H. assayed nitrogen fixation by *F. thermalis* in

the laboratory and field. E.B.S. constructed the *Anabaena* mutant strains and conducted physiological assays of *Anabaena* strains. E.B.S. measured the oxygen concentration at White Creek. S.R.M. performed the population genomic, phylogenomic and molecular clock dating analyses. S.R.M. and E.B.S. wrote the manuscript. All authors read and commented on the manuscript.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-017-0435-9.

Correspondence and requests for materials should be addressed to S.R.M.

nature research

Corresponding author(s): Scott R. Miller

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

The sample size of of 10 F. thermalis strains each from upstream and downstream White Creek sub-populations was not based on prior power analysis but was sufficient to distinguish significant outliers in genetic differentiation tests for subsequent functional validation.

A sample size of 9 microbial mats for diel oxygen concentration measurements in the field was determined based on personnel and equipment constraints on the ability to monitor samples efficiently over the duration of the monitoring.

In physiological experiments we used biological triplicates (i.e., derived from independent cultures) for each strain, which both provided sufficient statistical power in our analyses and could be accommodated by growth chamber space.

2. Data exclusions

Describe any data exclusions.

We performed three phylogenomics analyses involving 20, 22, and 16 genomes, respectively. In advance of the analyses, we decided to exclude sequence data for individual genes for which we did not have data for at least 18/20, 20/22 and 14/16 genomes. This resulted in good representation from all genomes while still retaining a large amount of data for analysis.

3. Replication

Describe whether the experimental findings were reliably reproduced.

All attempts at replication were successful.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

To obtain random samples of 10 F. thermalis strains each from upstream and downstream White Creek sub-populations from our culture collection, strains were selected with the use of random number generation.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Blinding was not relevant to this study.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

General statistical testing: JMP version 10

Genome data quality filtering and assembly: Trimmomatic version 0.22; Velvet version 1.2.03

Population genetics analyses: OutFLANK (to find Fst outliers based on an inferred distribution of neutral Fst); 'vegan' and 'R MASS' R packages (for NMDS analysis); CRISPRFinder (to identify clustered regularly interspaced short palindromic repeats); custom Perl scripts (available at http://www.nature.com/ismej/journal/v11/n1/abs/ismej2016105a.html.) and BioPerl code to prepare CDS data files from GenBank files and to estimate basic population parameters (e.g., Fst).

Phylogenomics analyses: SplitsTree version 4.14.4 (haplotype network analysis); DnaSP version 5 (estimation of the minimum number of recombination events); Sequence Matrix version 1.8 (concatenation of sequence data for individual genes); IQ-TREE v. 1.5.5 (phylogeny reconstruction); CONSENSE program of the PHYLIP package (to produce bootstrap consensus trees for calculation of internode certainty values); RAxML v. 8.2.4 (calculation of internode certainty values); Phylobayes v. 4.1 (Bayesian relaxed clock models)

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

All unique materials (strains) are deposited in the University of Montana Culture Collection for Cyanobacteria and are available from the authors.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

Not applicable.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used. — No eukaryotic cell lines were used.

b. Describe the method of cell line authentication used. — No eukaryotic cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination. — No eukaryotic cell lines were used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use. — No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study. — No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants. — The study did not involve human research participants.