MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 3371–3383 doi: 10.1111/mec.12812

Divergence with gene flow in a population of thermophilic bacteria: a potential role for spatially varying selection

CHRISTOPHER A. WALL, GREGORY J. KONIGES and SCOTT R. MILLER

Division of Biological Sciences, 32 Campus Dr. #4824, The University of Montana Missoula, MT 59812-4824, USA

Abstract

A fundamental goal of evolutionary biology is to understand how ecological diversity arises and is maintained in natural populations. We have investigated the contributions of gene flow and divergent selection to the distribution of genetic variation in an ecologically differentiated population of a thermophilic cyanobacterium (*Mastigocladus laminosus*) found along the temperature gradient of a nitrogen-limited stream in Yellowstone National Park. For most loci sampled, gene flow appears to be sufficient to prevent substantial genetic divergence. However, one locus (*rfbC*) exhibited a comparatively low migration rate as well as other signatures expected for a gene experiencing spatially varying selection, including an excess of common variants, an elevated level of polymorphism and extreme genetic differentiation along the gradient. *rfbC* is part of an expression island involved in the production of the polysaccharide component of the protective envelope of the heterocyst, the specialized nitrogen-fixing cell of these bacteria. SNP genotyping in the vicinity of *rfbC* revealed a ~5-kbp region including a gene content polymorphism that is tightly associated with environmental temperature and therefore likely contains the target of selection. Two genes have been deleted both in the predominant haplotype found in the downstream region of White Creek and in strains from other Yellowstone populations of *M. laminosus*, which may result in the production of heterocysts with different envelope properties. This study implicates spatially varying selection in the maintenance of variation related to thermal performance at White Creek despite on-going or recent gene flow.

Keywords: population genetics – empirical, ecological genetics, adaptation, bacteria

Received 18 February 2014; revision received 2 May 2014; accepted 6 May 2014

Introduction

Understanding how selection and gene flow jointly shape population variation is a central concern of evolutionary biology (Haldane 1948; Mayr 1963; Antonovics 1968; Ehrlich & Raven 1969; Endler 1973; Slatkin 1987; Morjan & Rieseberg 2004). Because even a modest amount of gene flow can homogenize variation in the absence of selection (Wright 1931), the process of genetic or phenotypic divergence has generally been considered to occur principally between geographically isolated (i.e., allopatric) populations (Mayr 1942, 1963; Endler 1977). Although it has been long recognized that the exchange of genes does not necessarily prevent divergence (see Harrison 2012 for a historical perspective), the relative importance of divergence in the presence of gene flow (either during initial divergence itself or following secondary contact) continues to be debated (Coyne & Orr 2004; Bolnick & Fitzpatrick 2007).

Divergence may occur or be maintained despite gene flow if the strength of divergent selection at a locus is sufficient to counter the opposing force of migration, indicated by a reduction in effective migration rate compared with unlinked neutral loci (e.g., Wu 2001). At issue is how frequently this condition is met. A recent survey (Pinho & Hey 2010) of 49 studies of recently diverged plant and animal taxa using the isolation with

Correspondence: Scott R. Miller, Fax: 406-243-4184;
E-mail: scott.miller@umontana.edu

migration model (Nielsen & Wakeley 2001; Hey & Nielsen 2004) to jointly estimate migration rates and population divergence time suggests that the maintenance of gene flow may not be uncommon. In addition, tracts of differentiation flanked by regions of comparatively low divergence are frequently observed in genome-wide analyses of sister taxa. The presence of such regions, termed 'islands' (Turner *et al.* 2005) or 'continents' (Michel *et al.* 2010) of speciation, respectively, depending on their size, is typically interpreted as evidence for reduced gene flow at targets of diversifying selection and linked loci (reviewed by Nosil *et al.* 2009). However, alternative mechanisms (such as lineage sorting of ancestral polymorphism) may cause the amount of divergence to vary in different parts of the genome even in the absence of gene flow (e.g., Noor & Bennett 2009).

It is likewise not clear how often natural populations of microorganisms diverge despite ongoing gene flow. Both homologous and illegitimate mechanisms of recombination can be important sources of evolutionary innovation for many bacteria and archaea and, in some taxa, may contribute more to novel genetic variation than do new mutations (Ochman *et al.* 2000; Spratt *et al.* 2001; Vos & Didelot 2009; Didelot & Maiden 2010). Theories on the origins of microbial diversity vary, however, regarding the role of recombination during divergence. In the absence of selection, simulations indicate that bacterial populations diverge only if the rate of recombination is low relative to mutation or in the presence of a recombination barrier (Fraser *et al.* 2007), be it geographic, ecological or molecular in nature. The ecotype model of ecological specialization (Cohan & Perry 2007) argues that recombination is weak relative to periodic selection, resulting in genome-wide high levels of linkage disequilibrium and the purging of variation within populations during the recurrent selective sweeps that drive divergence between populations. Consequently, islands of divergence are not expected. By contrast, the 'fragmented speciation' divergence with gene flow model (Retchless & Lawrence 2007) predicts that diverging genomes gradually become more isolated over time as sequential selection events restrict recombination at niche-defining loci. This process is expected to produce islands of divergence of different ages.

The ecotype model predicts the pattern of clustering of closely related sequences separated by long branches that is frequently observed for gene trees reconstructed for individual bacterial sequences, although alternative explanations for this genealogical pattern exist (e.g., Shapiro *et al.* 2012). However, several studies suggest that divergence with gene flow may be common among bacteria and archaea. For example, the fragmented speciation model may explain patterns of genome divergence data between *Escherichia coli* and *Salmonella enterica* (Retchless & Lawrence 2007) and among enteric bacteria more generally (Retchless & Lawrence 2010), although Luo *et al.* (2011) reported no evidence for islands of divergence for a comparison of the divergent genomes of enteric and environmental *Escherichia* strains. Among more recently diverged taxa, the divergence of early vs. late colonizing genotypes of the bacterium *Leptospirillum* in an acid mine drainage system has involved a history of recombination (Denef *et al.* 2010), and islands or continents of divergence have been observed between ecologically differentiated, sympatric populations of marine bacteria (Shapiro *et al.* 2012) and thermophilic archaea (Cadillo-Quiroz *et al.* 2012), respectively.

At White Creek, a geothermally influenced stream in the Lower Geyser Basin of Yellowstone National Park, the thermophilic, multicellular cyanobacterium *Mastigocladus* (also called *Fischerella*) *laminosus* is distributed along a ~1 km long thermal gradient (mean annual temperature spanning 39–54 °C). This population provides an ideal system for investigating the contributions of gene flow and divergent (e.g., spatially varying) selection to the ecological diversity of microorganisms. Previous work has found extensive variation in temperature performance among laboratory strains isolated from along the gradient (Miller *et al.* 2009). Specifically, whereas upstream strains (defined as strains isolated from sites with a mean annual temperature >46 °C) are ecological generalists that perform similarly at low (37 °C) and high (55 °C) temperatures, downstream strains are ecological specialists that exhibit a much faster average growth rate at 37 °C (~50% greater than upstream strains) than at 55 °C (20% of the upstream strain growth rate). Despite these differences in the degree of ecological specialization, a limited sequence data set for four loci suggested that gene flow may still be ongoing along the gradient (Miller *et al.* 2009). Here, we analysed a larger data set to more fully address whether *M. laminosus* has diverged in the face of gene flow at White Creek and to identify loci that have potentially contributed to the ecological divergence observed in the population.

Materials and methods

DNA isolation

A sample of 24 randomly selected strains of *Mastigocladus laminosus* that had been previously isolated from five sites spanning the population range at White Creek (Miller *et al.* 2009) was grown at 50 °C in 25 mL of D medium under 75 μmol of photons/m^2/s provided by cool-white fluorescent lights under a 12-h/12-h light/dark cycle. Genomic DNA was isolated as previously described (Miller & Castenholz 2000).

Multilocus sequence analysis data set

Because no strain genome data were initially available for primer design, we utilized a White Creek metagenome database (Klatt *et al.* 2013) to develop primer sets for 30 random regions of the genome with no a priori expectation of being under selection at White Creek. These regions were randomly selected from a list of metagenome contigs with closest sequence identity to a heterocyst-forming cyanobacterium by BLAST analysis (Altschul *et al.* 1990). In addition, primer sets were designed for eleven candidate loci previously shown to be upregulated during heat and/or nutrient stress in transcriptome studies of model cyanobacteria (Ehira *et al.* 2003; Suzuki *et al.* 2005). Primer sets for the amplification of each locus are provided in Table S1 (Supporting information). All fragments were amplified as 50-μL reactions with an MJ Research PTC-100 thermal cycler for 40 cycles of 94 °C for 1 min, 50 °C for 1.5 min and 72 °C for 1 min. Amplified fragments were cleaned by QIAquick PCR purification (QIAGEN) and then Sanger-sequenced either with an ABI 3130 genetic analyzer at the University of Montana Murdock Lab DNA Sequencing Facility or at the DNA Sequencing and Gene Analysis Center at the University of Washington.

Population genetic analyses

Nucleotide diversity (π) and Tajima's D (Tajima 1989) for sequences in the multilocus sample were estimated with DNASP version 5.10 (Librado & Rozas 2009). The degree of genetic differentiation between upstream (sites WC3, WC4 and WC5) and downstream (sites WC1 and WC2) strains of *M. laminosus* was estimated by F_{ST} using ARLEQUIN version 3.5.1.3 (Excoffier & Lischer 2010).

Migration rate estimation

Migration rate (*Nm*) between upstream and downstream sites was estimated by $(1/F_{ST} - 1)/2$ and, for a subset of loci (6882b, *htpG*, *cpcG3*, *recQ*, *rfbC*), with the 'isolation with migration' (IMa) model of population divergence (Hey & Nielsen 2004), as implemented by the IMa program (http://lifesci.rutgers.edu/heylab/HeylabSoftware.htm). IMa uses a Markov chain Monte Carlo approach to fit a coalescent-based model to the aligned sequence data and estimate the likelihoods of mutation-scaled model parameters (effective population sizes, population splitting time and population migration rates). A migration prior of 200 was used. The Markov chain was initiated with a burn-in length of 500 000 steps to achieve independence of starting conditions, followed by a chain length of 250 000 000 steps. To attain good chain mixing, the authors recommend an effective sample size (the number of independent parameter values) of 500 for each estimated parameter and multiple independent runs of the Markov chain. Under the conditions used, the effective sample sizes of the model parameters ranged between 800 and 29 000, and the results for three independent runs were nearly identical.

SNP genotyping

We used PCR-RFLP to genotype strains at 22 biallelic SNP loci. The data set consisted of: 11 loci from the multilocus sequence analysis data set for which suitable SNP-distinguishing restriction sites were available for genotyping by gel visualization; additional candidate loci (*nifB*, *nifH*, *nifN* and *hik34*), sequenced and screened for SNPs in a restriction site that could be scored on a gel; additional random loci that were designed from the White Creek metagenome and screened as above; and three previously available SNP loci (*devH*, *narB* and *nirA*; Miller *et al.* 2009). Cycling conditions for amplification were as above, and the primer set and restriction enzyme used for SNP genotyping at each locus are provided in Table S3 (Supporting information). Restriction digests were performed according to manufacturer specifications and visualized by agarose gel electrophoresis. F_{ST} between upstream and downstream strains was estimated for each SNP with ARLEQUIN version 3.5.1.3 (Excoffier & Lischer 2010).

Genome data, assembly and annotation

High molecular weight genomic DNAs were extracted from White Creek *M. laminosus* strains WC111, WC344 and WC542 by the method of Inskeep *et al.* (2013). A sample of the pooled DNAs was delivered to Cofactor Genomics (St. Louis, MO) for library construction and 60-bp paired-end sequencing on a lane of an Illumina GAII. Average insert size of the library was 360 bp, and sequencing resulted in 18 857 562 reads that passed the Illumina quality filter, for a total of >2.2 Gbp of sequence. *De novo* assembly by Velvet (Zerbino & Birney 2008; kmer 30 inslength 200) resulted in an assembly with an N50 of 84 kbp. Contigs >1 kbp were automatically annotated on the RAST server (http://rast.nmpdr.org/).

Evolutionary history of the rfbC region

Neighbor-net splits networks for *rfbC* and *galE* sequence data from White Creek *M. laminosus* and available genomes for other members of the *Mastigocladus*/

Fischerella group (Dagan *et al*. 2013; Shih *et al*. 2013) were inferred with Splittstree (Huson & Bryant 2006) using default settings. To map genetic architecture of this region of the heterocyst envelope polysaccharide expression island onto a *Mastigocladus* phylogeny, 1406 nucleotides of 16S rRNA gene sequence data obtained from genome data for the *Mastigocladus*/*Fischerella* group and out-group taxa were aligned as described previously (Miller & Castenholz 2000). A maximum-likelihood tree was reconstructed using PAUP* version 4.0b (Swofford 1998) according to a GTR + I model of sequence evolution selected by AIC using Modeltest (Posada & Crandall 1998). Heuristic searches for the likelihood analysis were performed using the tree-bisection–reconnection branch-swapping algorithm for five independent starting trees obtained by random stepwise sequence addition. A single ML tree was obtained in all five analyses and was bootstrap replicated 1000 times. MRBAYES version 3.1.2 (Huelsenbeck & Ronquist 2001) was used to generate phylogenies by Bayesian inference using a GTR + I model. Two replicate analyses of Metropolis-coupled MCMC were run for 1 000 000 generations, at which point the average standard deviation of split frequencies was ~0.0015, indicating that the chains had converged. The chains were sampled every 100 generations and the first 10% of sampled trees were discarded as burn-in.

RT-PCR

Cells grown under standard conditions were washed with ND medium (which lacks a source of combined nitrogen) and subsequently transferred to triplicate flasks of ND medium. After 12 h of nitrogen limitation, ~500 μL of cells were snap-frozen in liquid nitrogen. RNA was isolated (Campbell *et al.* 2007) and tested for genomic DNA contamination prior to cDNA synthesis (Invitrogen). An ~275-bp *rfbC* fragment and an ~800-bp fragment of the *alr2828* ortholog were, respectively, amplified in 50-μL reactions containing 2 μL cDNA for 35 cycles with an MJ Research PTC-100 thermal cycler under the following conditions: 94 °C for 1 min, 52 °C for 1 min, 72 °C for 30 s.

Transmission electron microscopy

Strains were grown at 37 and 55 °C, respectively, pelleted by centrifugation and then fixed in 2.0% EM-grade glutaraldehyde in cacodylate buffer at pH 7.2 overnight at 4 °C. The cells were gently centrifuged to a pellet, rinsed in dH2O, resuspended in 1% osmium tetroxide for 1 h at room temperature, and rinsed twice in dH2O. The cells were pelleted by centrifugation and dehydrated in a standard ethanol series. The ethanol was replaced with propylene oxide (PO), and the pellet was incubated in a 1:1 solution of Embed 812 epoxy and PO for 2 h followed by a 2:1 solution of Embed 812:PO overnight at room temperature. The samples were placed in 100% Embed 812 for 4 h, placed in BEEM capsules, and cured in a 60 °C oven for 24 h. Samples were sectioned on a Boeckeler MTXL ultramicrotome and 50-nm sections were placed on 400-mesh nickel grids. The sections were stained with 2% solutions of uranyl acetate followed by lead citrate and were imaged at the University of Montana EMtrix electron microscopy facility on a Hitachi H7100 TEM at 75 kV.

Results

Mastigocladus laminosus genetic variation at White Creek

Sequence data were obtained for 41 loci (Table 1) from a sample of 24 laboratory strains of *Mastigocladus laminosus* randomly selected from the White Creek culture collection of Miller *et al.* (2009), which includes strains from five sites (WC1 – WC5) along the stream channel. Most loci were randomly selected from *M. laminosus* contigs in a White Creek metagenome data set (Klatt *et al*. 2013; Table S1, Supporting information) and therefore had no a priori expectation of an association with environmental temperature, but the data set also included 11 candidate loci regulated in response to environmental stress (temperature, nutrient, light) in model cyanobacteria (Suzuki *et al.* 2005; Murata & Los 2006; Nakamoto & Honma 2006; Singh *et al.* 2006).

Most loci exhibited little or no genetic variation in the sample (Table 1): 24 loci were monomorphic, and mean nucleotide diversity was approximately 0.001. However, two loci (*rfbC* and serine/threonine protein kinase 6882b) exhibited comparatively elevated levels of polymorphism ($\pi > 0.01$). In addition, few loci exhibited a high level of genetic differentiation between phenotypically divergent upstream (strains from sites WC3-5) and downstream (sites WC1-2) strains, with the *rfbC* locus being a notable exception (Table 1).

Estimating migration rates for the White Creek Mastigocladus laminosus population

F_{ST}-derived estimates of the population-scaled migration rate *Nm* (Table 2) ranged from 0.05 copies per generation (*rfbC*) to free migration (i.e. the absence of population structure). Although F_{ST}-based estimates of gene flow remain useful (Neigel 2002), the assumptions of the infinite island model (Wright 1931) on which they are based are rarely met in natural biological systems (e.g., Whitlock & McCauley 1999). To relax these

Table 1 Summary of White Creek *Mastigocladus laminosus* polymorphism data

Locus	*N*	nt	π	F_{ST}	*D*
Candidates					
rfbC	24	220	0.0125	0.91	2.45**
htpG	21	421	0.0039	0.02	2.07*
dnaJ	20	345	0.0062	0	1.46
nifX	23	435	0.0011	0	1.28
cpcG2	21	498	0.0004	0.03	−0.49
psaK	24	666	0.0036	0.59	1.50
nifK	21	775	0		
iscS	22	701	0		
apcA	20	697	0		
apcB	23	494	0		
psaL	23	546	0		
Random					
dapB	24	520	0.0006	0	0.14
hoxU	24	676	0.0018	0	−0.29
recQ	23	514	0.0010	0.33	1.58
6882b	22	423	0.0123	0	3.05**
ubiH	21	691	0.0007	0.07	1.20
ksgA	24	637	0.0014	0.16	−0.42
infB	22	667	0.0003	0.02	−0.64
recF	24	640	0.0008	0	−2.08*
tkt	22	679	0.0010	0	−1.04
19284 g	22	669	0.0014	0	1.76
degT	24	646	0.0027	0.20	1.79
923 g	20	675	0		
minD/E	21	658	0		
8013b	24	649	0		
thiD	21	394	0		
9992 g	22	576	0		
11178 g	24	512	0		
14455 g	23	626	0		
18232 g	24	567	0		
clpB	24	668	0		
21238 g	24	601	0		
21253b	23	603	0		
dprA	22	565	0		
rnc	23	593	0		
rnb	24	614	0		
28386b	23	580	0		
28762b	24	560	0		
29132b	24	612	0		
hoxY	23	651	0		
33552b	23	659	0		

N, number of sequences; π, average pairwise nucleotide diversity; *D*, Tajima's D. *$P < 0.05$; **$P < 0.01$.
See Supporting Information for annotation and primer details.

assumptions, we therefore also used the isolation with migration (IMa) model (Hey & Nielsen 2004; Hey 2006) for a subset of loci spanning the range of π and F_{ST} estimates in the sample to estimate migration and distinguish ongoing gene flow from recent isolation. The results were qualitatively similar to those of the F_{ST}-based approach (Table 2). Although the 95% credibility intervals were generally broad, reflecting the expectation that migration rates between recently diverged populations are difficult to estimate (Won *et al.* 2005), the models nonetheless do indicate different migration rates among loci. However, we obtained no conclusive evidence for asymmetric gene flow between upstream and downstream samples.

Table 2 Estimates of population migration rates

		Nm (IMa)	
Locus	*Nm* (F_{ST})	D → U (95% CI)	U → D (95% CI)
6882b	—	41.0 (15, 423)	2.8 (1.2, 42)
htpG	24.5	28.0 (20, 519)	2.0 (0.9, 30)
cpcG3	16.2	17.0 (5, 2200)	0.9 (0.5, 110)
recQ	1.0	4.0 (1.7, 22)	1.1 (0.6, 12)
rfbC	0.05	1.8 (0.6, 2.7)	0.4 (0.01, 1.8)

–, Free migration estimated for this unstructured locus ($F_{ST} = 0$); D, downstream sites WC1-2; U, upstream sites WC3-5.

Locus-specific selection along White Creek

Genomic regions contributing to the observed variation in temperature performance among *M. laminosus* strains are expected to exhibit particular molecular population genetic signatures indicative of spatially varying selection. Specifically, a selective target (and linked variation) is predicted to have both an excess of common alleles compared with the expectation under selective neutrality and an elevated level of polymorphism compared with divergence from a sister taxon. Three of the sampled loci (*rfbC*, serine/threonine kinase 6882b and *htpG*) show an excess of common variants, indicated by significantly positively skewed values of Tajima's *D* (Table 1). These same three loci also exhibited an excess of polymorphism compared with divergence from a sister lineage strain (*M. laminosus* CCMEE 5318 from El Salvador; Miller *et al.* 2007) in a multilocus HKA analysis (http://lifesci.rutgers.edu/heylab/HeylabSoftware.htm; Hudson *et al.* 1987) for twelve loci ($P = 0.0004$ by coalescent simulation; Table S2, Supporting information), accounting for 49% of the χ^2 test statistic. These results suggest that these regions of the genome may be subject to some form of balancing selection. However, of the three loci, only *rfbC* exhibits the high degree of genetic differentiation and reduced migration rate expected for a candidate under spatially varying selection. Together, these results suggest that the *rfbC* marker or a linked locus may have been maintained over a long period of time by spatially varying selection related to temperature performance.

SNP genotyping

To further investigate the possibility that the *rfbC* marker is under spatially varying selection at White Creek, we genotyped the entire collection of 142 White Creek *M. laminosus* strains by PCR-RFLP for 22 SNP markers (including *rfbC*) segregating in the population (Table S3, Supporting information). The sample included 11 loci from the multilocus sequence analysis data set for which suitable SNP-distinguishing restriction sites were available for genotyping. The remainder of the sample consisted of additional candidate loci (nitrogenase genes *nifB*, *nifH* and *nifN* and histidine kinase *hik34*, which has been shown to be involved in thermotolerance in the cyanobacterium *Synechocystis* PCC 6803; Suzuki *et al*. 2005) and a combination of additional random markers that were either designed from the White Creek metagenome or previously reported (*devH*, *narB* and *nirA*; Miller *et al*. 2009).

Whereas the pattern of variation for genetic loci that are not under selection will reflect a population's demographic history, selection for alternative alleles between environments is expected to increase the amount of genetic differentiation between locations relative to this neutral background (Barton 1999). To avoid making assumptions about population demographic history that may result in the misspecification of a theoretical null distribution (Gavrilets 2003; Bolnick & Fitzpatrick 2007), we treated the observed frequency distribution of F_{ST} as an empirical null distribution with the prediction that *rfbC* would exhibit an extreme level of genetic differentiation compared with other loci (Luikart *et al.* 2003). Similar to the random sample (Table 1), genetic differentiation between upstream and downstream strains in the collection was generally low, and the *rfbC* marker emerged as the strongest outlier for the White Creek population (Fig. 1; F_{ST} = 0.88). One *rfbC* allele is fixed in the upstream sample, whereas the alternative allele is at high frequency (75%) downstream.

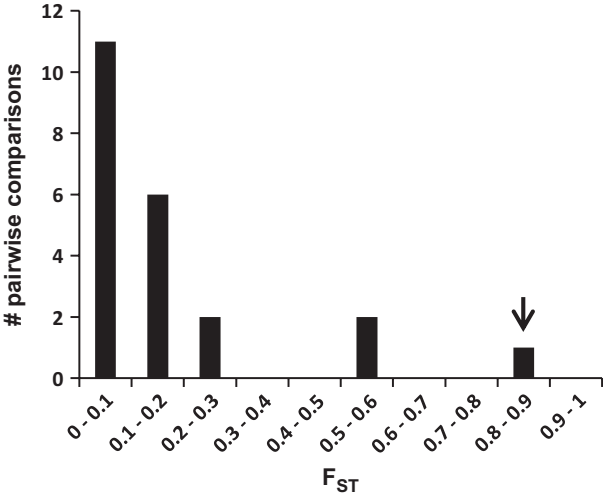

Fig. 1 F_{ST} frequency distribution for 22 biallelic SNP loci genotyped by PCR-RFLP for 142 laboratory strains of *Mastigocladus laminosus* isolated from along White Creek. The *rfbC* marker exhibits the most extreme value (arrow).

Genomic context of the rfbC marker and linked indel polymorphism

rfbC encodes dTDP-4-deoxyrhamnose-3,5-epimerase, an enzyme in the biosynthetic pathway of the sugar rhamnose, a common component of cell wall lipopolysaccharide in gram-negative bacteria. It was selected as a candidate in our study because its homolog in the cyanobacterium *Synechocystis* PCC 6803 is strongly up-regulated during heat shock (Singh *et al.* 2006). A BLAST analysis indicated that the ortholog of *M. laminosus rfbC* (*alr2830* in the model cyanobacterium *Anabaena* PCC 7120) maps to the Heterocyst Envelope Polysaccharide (HEP) expression island, a 28-kbp region of 21 CDS that is strongly up-regulated during heterocyst development (Ehira *et al.* 2003; Flaherty *et al.* 2011). Heterocysts are terminally differentiated cells specialized for the oxygen-sensitive process of nitrogen fixation that are produced during nitrogen limitation by members of a clade of multicellular cyanobacteria that includes *M. laminosus*. White Creek is an N-limited stream, and *M. laminosus* filaments contain abundant heterocysts and fix nitrogen *in situ* (Miller *et al.* 2006). A key structural feature of the heterocyst is the presence of an envelope of polysaccharide (HEP) and glycolipid that acts as a nonselective barrier to gas flow, thereby protecting nitrogenase from oxygen (Walsby 1985). Transposon-mutagenesis screens have demonstrated that several of the genes in the HEP island are required for both the deposition of the HEP layer and the ability to fix nitrogen in the presence of oxygen (Huang *et al.* 2005).

Genome data provided additional insight into the nature of the genetic variation that is segregating in the White Creek population in the vicinity of *rfbC*. Approximately 2.2 Gbp of Illumina sequence data were obtained from genomic DNAs pooled from two upstream strains (WC344 and WC542) and one downstream strain (WC111). The pooled data assembled into a ~5.4-Mbp draft genome of 100 contigs of length >1 kbp and an N50 of 84 kbp. In the draft assembly, HEP island genes were found on two contigs. Most (including *rfbC*) were found on contig 24813 (~144 kbp, ~450× coverage). A total of two genes upstream of *rfbC* in *Anabaena* PCC 7120 (*alr2828* and *alr2829*), however, were located on a short contig with lower coverage (contig 4308; ~2.6 kbp, ~300× coverage) that was nested

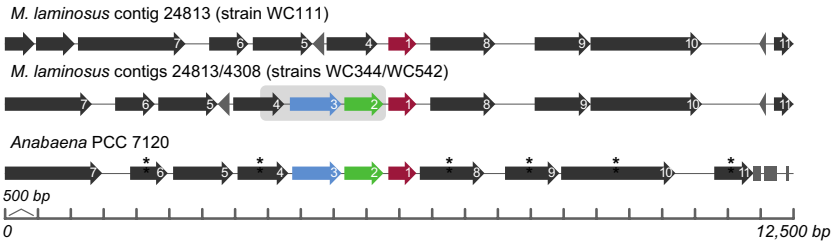

Fig. 2 Genetic architecture of a portion of the Heterocyst Envelope Polysaccharide (HEP) gene expression island for the contig 24813 (strain WC111) and 24813/4308 (strains WC344 and WC542) haplotypes, respectively, and for the model cyanobacterium *Anabaena* PCC 7120. Contig 4308 is indicated by the grey box and is flanked by contig 24813 in the *Mastigocladus laminosus* assembly. Key to CDS: 1, *rfbC* (*alr2830*); 2, hypothetical protein (*alr2829*); 3, glycosyltransferase (*alr2828*); 4, *galE* (*alr2827*); 5, *ygaF* (*alr2826*); 6, *rfbF* (*alr2825*); 7, hypothetical protein (*alr2824*); 8, NAD(P)-dependent oxidoreductase (*alr2831*); 9, glycosyltransferase (*alr2832*); 10, *wzc* (*alr2833*); 11, *hepC* (*alr2834*). The unlabelled putative genes between *galE* and *ygaF* and between *wzc* and *hepC* in *M. laminosus* are hypothetical protein ORFans that may not be genic. Fox genes in *Anabaena* PCC 7120 identified by Huang *et al.* (2005) are indicated by asterisks.

within contig 24813 (i.e. the paired reads of sequences near the respective contig 4308 boundaries map to contig 24813) and adjacent to *rfbC*. This suggested that contig 4308 sequence was not present in the genome of one of the three strains, and a subsequent PCR-based presence-absence assay showed that these genes were missing in the genome of strain WC111 (not shown). The genomic architectures in the *rfbC* region of the HEP island for WC111 and WC344/WC542, respectively, are shown in comparison with *Anabaena* PCC 7120 in Fig. 2 and indicate that the two CDS on contig 4308 have been deleted in WC111. Additional presence-absence genotyping of contig 4308 for the entire strain collection demonstrated that this indel polymorphism is in complete linkage disequilibrium with the *rfbC* marker (not shown), that is, the genes are present (4308+) in all upstream strains but missing (4308−) in most individuals from lower temperatures.

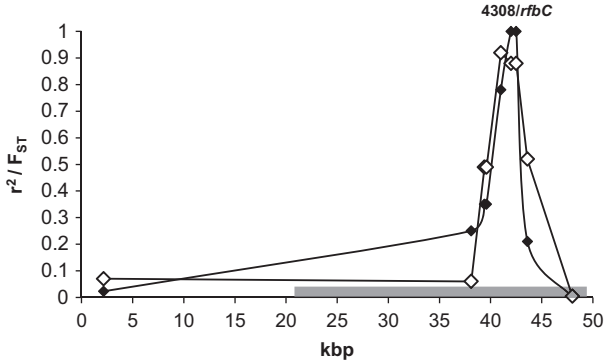

Fig. 3 Estimates of linkage disequilibrium (r^2; closed diamonds) and F_{ST} (open diamonds) between 4308/*rfbC* and flanking biallelic SNP alleles genotyped for the White Creek *Mastigocladus laminosus* strain collection. The grey bar spans the location of *M. laminosus* genes annotated as belonging to the Heterocyst Envelope Polysaccharide island.

Fine mapping the rfbC region

To attempt to better resolve the putative target of spatially varying selection, we used the genome data to develop biallelic SNP markers (Table S4, Supporting information) flanking *rfbC* for genotyping of the White Creek strain collection by PCR-RFLP. Both LD between a SNP and *rfbC* and genetic differentiation between upstream and downstream strains decreased rapidly within approximately 5 kbp of the *rfbC* marker (Fig. 3). The restriction of the temperature association to this region of the HEP island suggests that it is the target of selection.

The 4308 deletion is not restricted to White Creek

High-quality draft genomes for two strains of *Mastigocladus* from different geothermal areas in Yellowstone, strain JSC-11 (Chocolate Pots, ~20 km from White Creek; GenBank accession no. AGIZ00000000.1) and strain PCC 7521 (Mammoth Hot Springs, ~50 km from White Creek; Dagan *et al.* 2013), revealed that these strains also have a deletion of the two genes on contig 4308. These strains share the same genetic architecture in the HEP region as the White Creek deletion allele (Fig. 4), and sequences for flanking genes *rfbC* and *galE* are either identical to the predominant downstream allele at White Creek or cluster with it in a haplotype network (Fig. 5). We conclude that this pattern is the result of a single-deletion event, that the 4308 deletion is widespread in Yellowstone and that its origin either pre-dates the White Creek population or has since spread from the population.

This appears to have generally been a dynamic region of the genome during the diversification of heterocyst-forming cyanobacteria, based on apparent additional

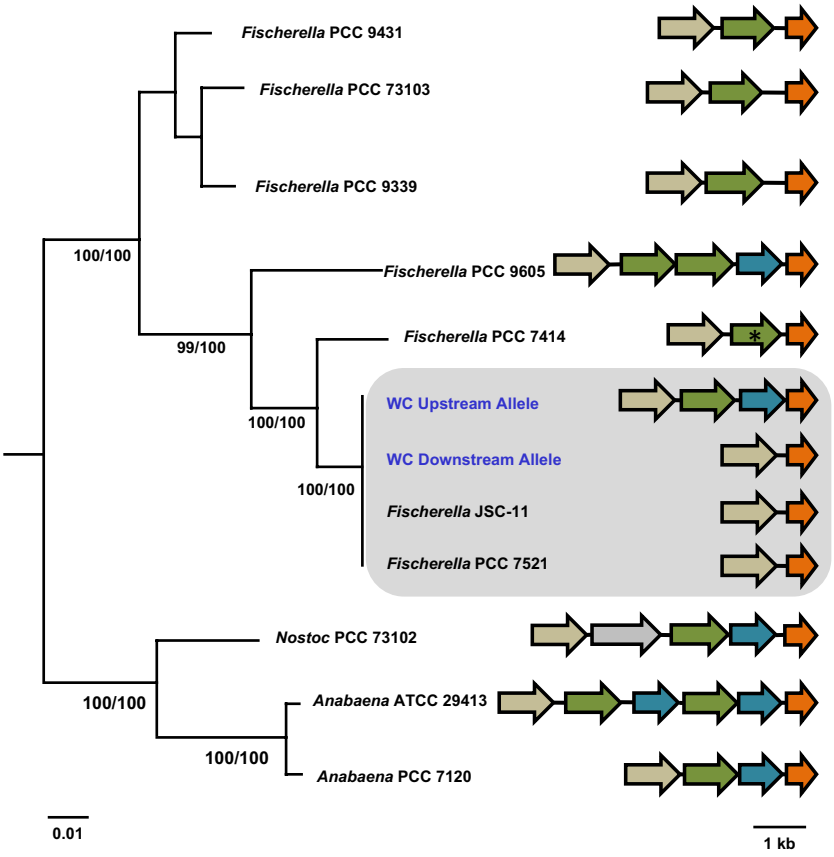

Fig. 4 Genetic architecture of the *rfbC* region in heterocyst-forming cyanobacteria. Maximium-likelihood phylogeny of the 16S rRNA gene for *Mastigocladus laminosus* and related *Fischerella* strains, out-group rooted with representative Nostocales strains. Labelled nodes indicate bootstrap percentage followed by posterior probability in a Bayesian analysis. Orthologous genes are colour-coded as follows (*Anabaena* PCC 7120 designations): *alr2827* (tan), *alr2828* (green), *alr2829* (blue) and *alr2830* (orange). The grey-labelled locus in *Nostoc* PCC 73102 is annotated as a methyltransferase. An inferred pseudogene is denoted by an asterisk.

deletion, insertion and tandem duplication events (Fig. 4). A strain of *M. laminosus* isolated from a New Zealand hot spring (PCC 7414; Dagan *et al.* 2013) has lost the *alr2829* ortholog, while the *alr2828* ortholog appears to have been pseudogenized. By contrast, the *alr2828* ortholog has been duplicated in the mesophilic sister taxon *Fischerella* PCC 9605. Strains within the clade comprising *Fischerella* strains PCC 9339, PCC 9431 and PCC 73103 (isolated from a rice field in India) have retained a copy of the *alr2828* ortholog but have deleted the *alr2829* ortholog. BLAST analyses indicated that these genes are not found elsewhere in these genomes (not shown). Representative members of the Nostocales out-group are similarly variable, including the inferred insertion of an annotated methyltransferase in *Nostoc punctiforme* PCC 73102 and a duplication of the *alr2828* and *alr2829* orthologs in *Anabaena variabilis* ATCC 29413.

4308 deletion strains can produce a HEP layer

We confirmed by RT-PCR that the *alr2828* ortholog is expressed during heterocyst development by 4308+ *M. laminosus* strains (Fig. S1, Supporting information). *alr2828* is functionally annotated as a glycosyltransferase (*alr2829* encodes a hypothetical protein of unknown function). Glycosyltransferases catalyse the transfer of a monosaccharide from a nucleotide sugar donor to an acceptor during, for example, polysaccharide synthesis. In the three heterocyst-forming cyanobacteria that have been investigated, HEP has been shown to be a complex, branching polysaccharide with a backbone of glucosyl and mannosyl residues and strain-specific terminal-side branch residues (Cardemil & Wolk 1976, 1979, 1981). The 4308 deletion could therefore potentially alter HEP composition, structure and/or physical properties, for example, by the loss of a specific monosaccharide or side branch.

The principal function of the HEP is believed to be the stabilization of the underlying glycolipid layer of the envelope (Walsby 1985; Wolk *et al.* 1994). Consistent with this model, HEP mutants that are unable to fix nitrogen in the presence of oxygen (the Fox- phenotype) exhibit not only the absence of envelope polysaccharide but also the loss of an intact, laminated glycolipid layer (Huang *et al.* 2005). Transposon-mutagenesis screens for Fox genes in *Anabaena* PCC 7120 have not recovered the contig 4308 loci (e.g., Huang *et al.* 2005), which suggests that inactivation of these genes does not result in a Fox- phenotype. This is to be expected, because nitrogen

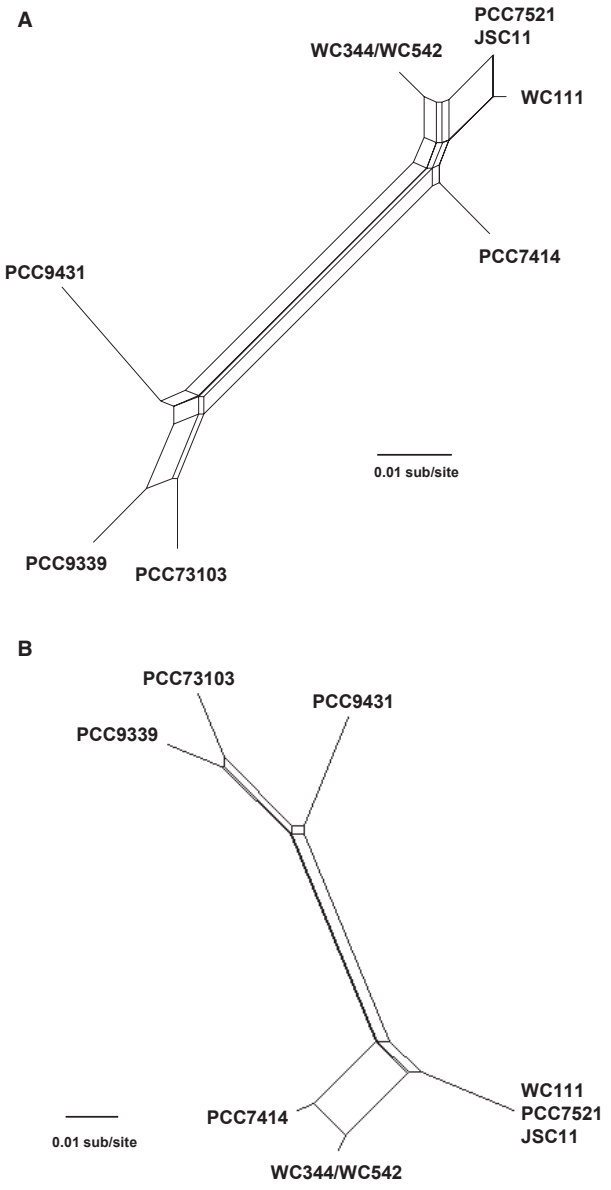

Fig. 5 Neighbor-net splits genealogical networks for (A) *rfbC* (*alr2830*) and (B) *galE* (*alr2827*) for representative White Creek *Mastigocladus laminosus* strains (WC111, WC344 and WC542), other Yellowstone strains (PCC7521 and JSC11) and other members of the *Mastigocladus*/*Fischerella* group. Cycles in the networks represent reticulate events such as recombination in the evolutionary histories of these loci.

fixation in the presence of oxygen is known to be an important part of metabolism at White Creek (Miller *et al.* 2006). However, previous transposon-mutagenesis studies have not been saturating (e.g., Huang *et al.* 2005). Therefore, to test whether the 4308 deletion results in an observable aberration of the HEP and/or glycolipid layers of the heterocyst envelope, we used transmission electron microscopy to image heterocyst ultrastructure of representative strains with and without the 4308 deletion, respectively. In all cases, strains produced HEP and an intact glycolipid layer at both 37 and 55 °C (Fig. 6).

Discussion

Estimated levels of gene flow among ecologically divergent *M. laminosus* at White Creek appear to be sufficient to prevent substantial genetic divergence for many loci, yet migration is clearly restricted in certain regions of the genome (e.g., *rfbC*; Table 2). These differences in estimated migration rates among loci are expected if selection has played a role in the divergence process or in its reinforcement (Hey 2006). The rapid decay of LD in the vicinity of *rfbC* further indicates a generally strong impact of historic recombination events on the evolutionary dynamics of the White Creek population. This is not the pattern expected for the ecotype model (Cohan & Perry 2007) but, rather, is consistent with the specific suppression of gene flow at niche-defining loci predicted by the fragmented speciation (Retchless & Lawrence 2007) and other divergence with gene flow models.

In addition to the reduced migration rate at the *rfbC* marker, other lines of evidence suggest that *rfbC* or a linked locus is a target of selection. These include its high amount of genetic variation (Table 1), an excess of common variants (Table 1) and elevated polymorphism compared with divergence (Table S2, Supporting information). The high amount of genetic differentiation between upstream and downstream strains (Table 1; Fig. 1) specifically implicates spatially varying selection as a mechanism for actively maintaining putatively locally adaptive variation in the HEP region along the White Creek thermal gradient. The presence in other Yellowstone populations of *Mastigocladus laminosus* of both the 4308 deletion (Fig. 4) and the predominant downstream allele at the adjacent *galE* locus (Fig. 5B) likewise supports the interpretation that this haplotype is selectively favoured in particular environments.

The small size (~5 kbp) of the island of divergence between upstream and downstream sites in the HEP region reduces the potential targets of spatially varying selection to only a few genes (Figs 2 and 3). We speculate that the 4308 indel polymorphism is the likely target based on the expectation that deletion of these two genes impacts heterocyst development, structure and/or function, given their strong induction during heterocyst differentiation (Ehira *et al.* 2003; Flaherty *et al.* 2011). Because strains that lack these genes can still produce an intact heterocyst envelope (Fig. 6) and can grow in a nitrogen-limited environment in the presence of oxygen (Miller *et al.* 2009), the phenotypic effect (s) of the deletion appear to be more subtle than the

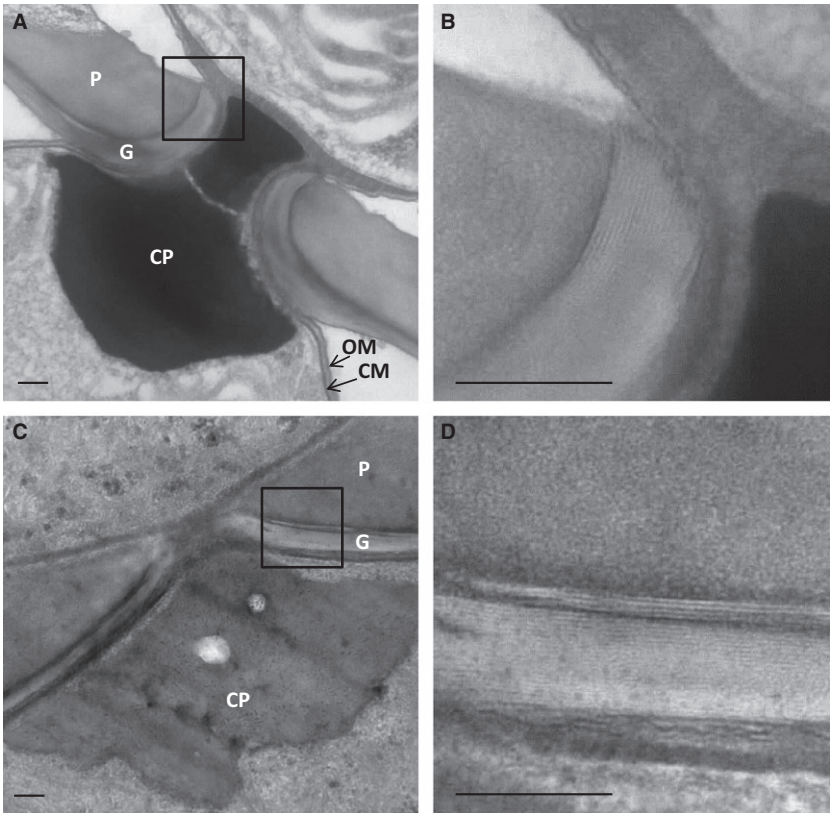

Fig. 6 Transmission electron micrographs of heterocyst ultrastructure for representative 4308+ (A, B) and 4308− (C, D) strains of *Mastigocladus laminosus* showing intact polysaccharide and glycolipid layers of the heterocyst envelope. Panels B and D are magnifications of the boxed areas in panels A and C, respectively. The image for 4308+ cells was for cells grown at 37 °C, and the image for 4308− cells for cells grown at 55 °C. Note the laminated glycolipid layers in panels B and D. Scale bar is 100 nm in all panels. P, heterocyst envelope polysaccharide layer; G, heterocyst glycolipid layer; OM, outer membrane; CM, cytoplasmic membrane; CP, nitrogen storage granule cyanophycin. In panel A, the white space between the glycolipid layer and the outer membrane is the result of dehydration during sample preparation.

strongly deleterious Fox- phenotype observed for laboratory mutants of *Anabaena* sp. PCC 7120 with inactivated copies of several of the surrounding HEP island genes (Huang *et al*. 2005; Fig. 2). These may include an alteration of the chemical composition and structure of the HEP layer itself. However, several of the genes in the HEP island, including the *alr2828* ortholog, *galE* and *rfbC*, are homologous to genes involved in the synthesis of the lipopolysaccharide (LPS) component of the outer membrane of the gram-negative bacterial cell wall (Huang *et al*. 2005). Therefore, an alternative possibility is that the 4308 glycosyltransferase is involved in the remodelling of cell wall LPS for export of HEP outside the cell. Regardless of the mechanism, if it proves to be the case that the indel polymorphism is the selective target, such an observation would be in accord with the growing appreciation of the importance of null mutations as a mechanism of bacterial adaptation to environmental change (Hottes *et al*. 2013). We cannot, however, rule out the possibility that variation within *rfbC* itself is the selective target: although most of the variation between alleles at this locus is synonymous (12/13 SNPs), there is a single amino acid polymorphism (Ala/Thr at codon 128).

Other loci in our data set exhibited some, but not all, of the signatures of selection observed for *rfbC*. Notably, both *htpG* and an annotated signal transduction gene (serine/threonine kinase 6882b) were also characterized by an excess of polymorphism and a positively skewed value of Tajima's *D* (Table 1; Table S2, Supporting information). The former encodes a molecular chaperone that is a homolog of the eukaryotic heat shock protein Hsp90. We had selected it as an a priori candidate in our study, because in cyanobacteria it is both up-regulated and essential under heat stress (Tanaka & Nakamoto 1999) and has been demonstrated to stabilize proteins and prevent their aggregation (Sato *et al*. 2010; Minagawa *et al*. 2011). However, unlike *rfbC*, variation at both of these loci was not spatially structured along the White Creek gradient (Table 1; F_{ST} for both was 0.02 in the Fig. 1 data set). This lack of an association with temperature suggests that these loci do not contribute to the differences among upstream and downstream strains in temperature performance but, rather, that they may be maintained by some other unidentified environmental or biological factor(s). Taken together, our identification of several loci in the sample with a similar signature of selection may indicate contributions of spatial heterogeneity and, perhaps, other forms of balancing selection to the maintenance of polymorphism during *M. laminosus* diversification, potentially over long evolutionary time scales. The acquisition of more extensive genome sequence data for strains of *M. laminosus* from the White Creek population

will enable us to identify additional islands of divergence. Estimation of the time to coalescence and the rate of decay of LD in these regions will further our understanding of both the timing of niche-defining events and the relative contributions of older vs. new variation to the population divergence process along the White Creek environmental gradient.

Acknowledgements

We thank the three anonymous reviewers for their comments and suggestions. We thank Darla Carvey and Michelle Ganoza for their technical assistance, Kayli Anderson and Patrick Hutchins for their assistance with the RT-PCR experiment, and Jim Driver of the University of Montana EMtrix electron microscopy facility. This research was supported by NSF IOS-1110819 and an REU supplement for NSF EF-0801999 to SRM and by NSF DDIG DEB-1110819 to SRM and CAW.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Antonovics J (1968) Evolution in closely adjacent plant populations. VI. Manifold effects of gene flow. *Heredity*, **23**, 507–524.

Barton N (1999) Clines in polygenic traits. *Genetical Research*, **74**, 223–236.

Bolnick D, Fitzpatrick B (2007) Sympatric speciation: models and empirical evidence. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 459–487.

Cadillo-Quiroz H, Didelot X, Held N *et al.* (2012) Patterns of gene flow define species of thermophilic Archaea. *PLoS Biology*, **10**, e1001265.

Campbell EL, Summers ML, Christman H, Martin ME, Meeks JC (2007) Global gene expression patterns of *Nostoc punctiforme* in steady-state dinitrogen-grown heterocyst-containing cultures and at single time points during the differentiation of akinetes and hormogonia. *Journal of Bacteriology*, **189**, 5247–5256.

Cardemil L, Wolk C (1976) The polysaccharides from heterocyst and spore envelopes of a blue-green alga. Methylation analysis and structure of the backbones. *Journal of Biological Chemistry*, **251**, 2967–2975.

Cardemil L, Wolk C (1979) The polysaccharides from heterocyst and spore envelopes of a blue-green alga. Structure of the basic repeating unit. *Journal of Biological Chemistry*, **254**, 736–741.

Cardemil L, Wolk C (1981) Polysaccharides from the envelopes of heterocysts and spores of the blue-green algae *Anabaena variabilis* and *Cylindrospermum licheniforme*. *Journal of Phycology*, **17**, 234–240.

Cohan F, Perry E (2007) A systematics for discovering the fundamental units of bacterial diversity. *Current Biology*, **17**, R373–R386.

Coyne J, Orr H (2004) *Speciation*. Sinauer Associates, Inc., Sunderland, MA.

Dagan T, Roettger M, Stucken K *et al.* (2013) Genomes of Stigonematalean cyanobacteria (Subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome Biology and Evolution*, **5**, 31–44.

Denef V, Kalnejais L, Mueller R *et al.* (2010) Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 2383–2390.

Didelot X, Maiden M (2010) Impact of recombination on bacterial evolution. *Trends in Microbiology*, **18**, 315–322.

Ehira S, Ohmori M, Sato N (2003) Genome-wide expression analysis of the responses to nitrogen deprivation in the heterocyst-forming cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Research*, **10**, 97–113.

Ehrlich P, Raven P (1969) Differentiation of populations. *Science*, **165**, 1228–1232.

Endler J (1973) Gene flow and population differentiation. *Science*, **179**, 243–250.

Endler J (1977) *Geographic Variation, Speciation, and Clines*. Princeton University Press, Princeton, NJ.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Flaherty B, van Nieuwerburgh F, Head S, Golden J (2011) Directional RNA deep sequencing sheds new light on the transcriptional response of *Anabaena* sp. strain PCC 7120 to combined-nitrogen deprivation. *BMC Genomics*, **12**, 332–341.

Fraser C, Hanage W, Spratt B (2007) Recombination and the nature of bacterial speciation. *Science*, **315**, 476–480.

Gavrilets S (2003) Perspective: models of speciation: what have we learned in 40 years? *Evolution*, **57**, 2197–2215.

Haldane J (1948) Theory of a cline. *Journal of Genetics*, **28**, 277–284.

Harrison R (2012) The language of speciation. *Evolution*, **66**, 3643–3657.

Hey J (2006) Recent advances in assessing gene flow between diverging populations and species. *Current Opinion in Genetics & Development*, **16**, 592–596.

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **167**, 747–760.

Hottes A, Freddolino P, Khare A *et al.* (2013) Bacterial adaptation through loss of function. *PLoS Genetics*, **9**, e1003617.

Huang X, Fan Q, Lechno-Yossef S *et al.* (2005) Clustered genes required for the synthesis of heterocyst envelope polysaccharide in *Anabaena* sp. strain PCC 7120. *Journal of Bacteriology*, **187**, 1114–1123.

Hudson R, Kreitman M, Aquadé M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.

Huelsenbeck J, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

Huson D, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**, 254–267.

Inskeep W, Jay Z, Tringe S *et al.* (2013) The YNP metagenome project: environmental parameters responsible for microbial distribution in the Yellowstone geothermal ecosystem. *Frontiers in Microbiology*, **4**, 67.

Klatt C, Inskeep W, Herrgard M *et al.* (2013) Community structure and function of high-temperature chlorophototrophic

microbial mats inhabiting diverse geothermal environments. *Frontiers in Microbiology*, **4**, 106.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

Luikart G, England P, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Luo C, Walk S, Gordon D *et al.* (2011) Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 7200–7205.

Mayr E (1942) *Systematics and the Origin of Species*. Columbia University Press, New York, NY.

Mayr E (1963) *Animal Species and Evolution*. Harvard University Press, Cambridge, MA.

Michel A, Sim S, Powell T *et al.* (2010) Widespread genomic divergence during sympatric speciation. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 9724–9729.

Miller S, Castenholz R (2000) Evolution of thermotolerance in hot spring cyanobacteria of the genus *Synechococcus*. *Applied and Environmental Microbiology*, **66**, 4222–4229.

Miller SR, Purugganan MD, Curtis SE (2006) Molecular population genetics and phenotypic diversification of two populations of the thermophilic cyanobacterium *Mastigocladus laminosus*. *Applied and Environment Microbiology*, **72**, 2793–2800.

Miller S, Castenholz R, Pedersen D (2007) Phylogeography of the thermophilic cyanobacterium *Mastigocladus laminosus*. *Applied and Environmental Microbiology*, **73**, 4751–4759.

Miller S, Williams C, Strong AL, Carvey D (2009) Ecological specialization in a spatially structured population of the thermophilic cyanobacterium *Mastigocladus laminosus*. *Applied and Environmental Microbiology*, **75**, 729–734.

Minagawa S, Kondoh Y, Sueoka K, Osada H, Nakamoto H (2011) Cyclic lipopeptide antibiotics bind to the N-terminal domain of the prokaryotic Hsp90 to inhibit the chaperone activity. *Biochemical Journal*, **435**, 237–246.

Morjan C, Rieseberg L (2004) How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Molecular Ecology*, **13**, 1341–1356.

Murata N, Los D (2006) Histidine kinase Hik33 is an important participant in cold-signal transduction in cyanobacteria. *Physiologia Plantarum*, **126**, 17–27.

Nakamoto H, Honma D (2006) Interaction of a small heat shock protein with light-harvesting cyanobacterial phycocyanins under stress conditions. *FEBS letters*, **580**, 3029–3034.

Neigel J (2002) Is F_{ST} obsolete? *Conservation Genetics*, **3**, 167–173.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Noor M, Bennett S (2009) Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, **103**, 439–444.

Nosil P, Funk D, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.

Pinho C, Hey J (2010) Divergence with gene flow: models and data. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 215–230.

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Retchless A, Lawrence J (2007) Temporal fragmentation of speciation in bacteria. *Science*, **317**, 1093–1096.

Retchless A, Lawrence J (2010) Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 11453–11458.

Sato T, Minagawa S, Kojima E, Okamoto N, Nakamoto H (2010) HtpG, the prokaryotic homologue of Hsp90, stabilizes a phycobilisome protein in the cyanobacterium *Synechococcus elongatus* PCC 7942. *Molecular Microbiology*, **76**, 576–589.

Shapiro B, Friedman J, Cordero O *et al.* (2012) Population genomics of early events in the ecological differentiation of bacteria. *Science*, **336**, 48–51.

Shih P, Wu D, Latifi A *et al.* (2013) Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 1053–1058.

Singh A, Summerfield T, Li H, Sherman L (2006) The heat shock response in the cyanobacterium *Synechocystis* sp. strain PCC 6803 and regulation of gene expression by HrcA and SigB. *Archives of Microbiology*, **186**, 273–286.

Slatkin M (1987) Gene flow and geographic structure of natural populations. *Science*, **236**, 787–792.

Spratt B, Hanage W, Feil E (2001) The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Current Opinion in Microbiology*, **4**, 602–606.

Suzuki I, Kanesaki Y, Hayashi H *et al.* (2005) The histidine kinase Hik34 is involved in thermotolerance by regulating the expression of heat shock genes in *Synechocystis*. *Plant Physiology*, **138**, 1409–1421.

Swofford DL (1998) *PAUP*. Phylogenetic Analysis Using Parsimony (* and other methods). Version 4*. Sinauer Associates, Sunderland, Massachusetts, USA.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tanaka N, Nakamoto H (1999) HtpG is essential for the thermal stress management in cyanobacteria. *FEBS letters*, **458**, 117–123.

Turner T, Hahn M, Nuzhdin S (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, **3**, e285.

Vos M, Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. *ISME Journal*, **3**, 199–208.

Walsby A (1985) The permeability of heterocysts to the gases nitrogen and oxygen. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **226**, 345–366.

Whitlock M, McCauley DE (1999) Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm+1)$. *Heredity*, **82**, 117–125.

Wolk CP, Ernst A, Elhai J (1994) Heterocyst metabolism and development. In: *The Molecular Biology of Cyanobacteria* (ed. Bryant D.), pp. 769–823. Kluwer Academic Publishers, Dordrecht.

Won Y, Sivasundar A, Wang Y, Hey J (2005) On the origin of Lake Malawi cichlid species: a population genetic analysis of divergence. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 6581–6586.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

S.R.M. and C.A.W. designed the research, C.A.W., G.J.C. and S.R.M. performed the research, C.A.W. and S.R.M. analysed the data, and C.A.W. and S.R.M. wrote the paper.

Data accessibility

DNA sequences: GenBank accession nos KJ710710-KJ710774 and KJ737440-KJ738304; NCBI SRA: SRX517500; Aligned sequence data for the multilocus data set: Dryad doi:10.5061/dryad.1g3v3; genome sequence assembly draft for the pooled sample of *M. laminosus* strains WC111, WC344 and WC542: Dryad doi:10.5061/dryad.1g3v3; SNP genotype data for the White Creek *M. laminosus* strain collection (Fig. 1 and Fig. 3): Dryad doi:10.5061/dryad.1g3v3.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Annotation and primers for the *Mastigocladus laminosus* polymorphism data set.

Table S2 Results of HKA analysis.

Table S3 Primers and restriction enzymes for distinguishing biallelic SNPs by PCR-RFLP for the *Mastigocladus laminosus* SNP genotyping data set.

Table S4 Primers and restriction enzymes for distinguishing biallelic SNPs by PCR-RFLP for fine mapping the HEP candidate region.

Fig. S1 RT-PCR of cDNAs from 4308+ (WC344, WC538) and 4308− (WC119, WC249) strains during heterocyst development 12 h after the onset of nitrogen limitation.