APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Apr. 2006, p. 2793–2800 Vol. 72, No. 4
0099-2240/06/$08.00+0 doi:10.1128/AEM.72.4.2793–2800.2006

Molecular Population Genetics and Phenotypic Diversification of Two Populations of the Thermophilic Cyanobacterium *Mastigocladus laminosus*

Scott R. Miller,* Michael D. Purugganan, and Stephanie E. Curtis

Department of Genetics, NC State University, Raleigh, North Carolina 27695

Received 6 October 2005/Accepted 6 February 2006

We investigated the distributions of genetic and phenotypic variation for two Yellowstone National Park populations of the heterocyst-forming cyanobacterium *Mastigocladus* (*Fischerella*) *laminosus* that exhibit dramatic phenotypic differences as a result of environmental differences in nitrogen availability. One population develops heterocysts and fixes nitrogen in situ in response to a deficiency of combined nitrogen in its environment, whereas the other population does neither due to the availability of a preferred nitrogen source. Slowly evolving molecular markers, including the 16S rRNA gene and the downstream internal transcribed spacer, are identical among all laboratory isolates from both populations but belie considerable genetic and phenotypic diversity. The total nucleotide diversity at six nitrogen metabolism loci was roughly three times greater than that observed for the human global population. The two populations are genetically differentiated, although variation in performance on different nitrogen sources among genotypes could not be explained by local adaptation to available nitrogen in the respective environments. Population genetic models suggest that local adaptation is mutation limited but also that the populations are expected to continue to diverge due to low migratory gene flow.

Understanding the origins and maintenance of ecological diversity is a central goal of the study of microbial ecology and evolution that requires linking genetic and phenotypic variation on a geographic scale. Population genetic theory and experimental evolution in the laboratory suggest that spatially structured microbial populations in nature will diverge rapidly provided that migratory gene flow between populations is low (e.g., see references 1, 2, 18, and 33). Although evidence for geographic structuring of microbial genetic diversity is accumulating (e.g., see references 15 and 29), many questions remain, particularly at finer scales. Are populations locally differentiated, genetically and phenotypically? If so, do these differences result from local adaptation to prevailing environmental conditions? Can we predict whether populations will continue to diverge over time, based on their current genetic structures?

There are several challenges to the study of geographic structure and local adaptation in microbial ecosystems, particularly for microorganisms that do not form intimate associations with eukaryotic hosts. One is simply to unambiguously identify distinct populations in nature. Another is to locate habitats with stable environmental differences that consistently affect the expressed phenotypes of the respective populations in a predictable and ecologically interesting way. It is under such conditions that local adaptation might be expected to be most evident. Third, it is often difficult to cultivate an unbiased representation of in situ diversity for investigation under controlled laboratory conditions of the amount of phenotypic variation (and, ultimately, its genetic basis) for traits of ecological interest among individuals both within and between populations.

Here we describe an integrative approach to investigate genetic and phenotypic differentiation within and between two Yellowstone National Park populations of the thermophilic cyanobacterium *Mastigocladus* (*Fischerella*) *laminosus* that exhibit dramatic phenotypic differences in situ as a result of environmental differences in nitrogen availability. Under the nitrogen-limited conditions at White Creek, this large and morphologically distinct multicellular bacterium develops intercalary heterocysts, terminally differentiated cells specialized for enabling the oxygen-sensitive process of nitrogen fixation in an oxic environment. At Boiling River, in contrast, where a preferred source of nitrogen is consistently available, the genetic programs for heterocyst development and nitrogen fixation are not expressed. Because we can directly isolate these bacteria with a high rate of success in laboratory culture without enrichment and selection for particular genotypes, it is possible to genetically and phenotypically characterize clones that are representative of in situ diversity. Given sufficient evolutionary time and genetic variation, we would expect the relative fitness of each population to have increased on its available nitrogen source in response to selection. We tested whether there has been genetic divergence and local adaptation of *Mastigocladus* populations to prevailing environmental conditions and, if not, whether we expect these populations to continue to diverge in the future.

MATERIALS AND METHODS

Field sites, sample collection, and nutrient chemistry. Boiling River is a channel of Mammoth Hot Springs outflow near the north entrance of Yellowstone National Park that emerges from underground approximately 150 m from its confluence with the Gardner River. White Creek is located roughly 50 km from Boiling River in the Lower Geyser Basin of Yellowstone National Park and

* Corresponding author. Present address: Division of Biological Sciences, The University of Montana, 32 Campus Dr., #4824, Missoula, MT 59812-4824. Phone: (406) 243-5149. Fax: (406) 243-4184. E-mail: scott.miller@mso.umt.edu.

is fed by thermal discharge from several geothermal features within its drainage area. At both sites, *Mastigocladus* (*Fischerella*) *laminosus* forms epilithic streamers at temperatures lower than approximately 56°C. In June 2001, multiple collections were sampled from the ends of individual streamers at both sites with sterile 3-ml syringes and stored at ambient temperature in the dark. The temperature of collection was 52°C for all collections from White Creek and ranged between 52 and 56°C for those from Boiling River. Subsamples of each collection were observed by phase-contrast microscopy to verify the presence of *Mastigocladus* and to evaluate the presence or absence of heterocysts in filaments. Macronutrient and micronutrient concentrations in water samples from both sites were analyzed by the Analytical Services Laboratory at NC State University. Total nitrogen (inorganic and organic) was measured as nitrogen mono-oxide chemiluminescence with a total organic carbon-total nitrogen analyzer (Shimadzu) after passing a water subsample through a Pt column at 720°C.

Acetylene reduction assay. In situ nitrogen fixation rates were estimated for both populations by the acetylene reduction method (23). Triplicate streamer tufts collected from each site were incubated under ambient midday light (approximately 1,000 W m^{-2}) for 3 h at 50°C in crimp-sealed serum vials containing 20 ml of the corresponding in situ water and 5 ml of acetylene gas. Triplicate blanks with no *Mastigocladus*, but otherwise identical to the experimental treatments, were included to estimate background acetylene reduction. Assays were performed within 2 hours of sample collection from the respective sites. At the end of the assay, 2.5 ml of headspace gas was withdrawn from each sample and injected into a pre-evacuated crimp-sealed vial until analysis that evening by flame ionization detection with a Shimadzu GC-14A gas chromatograph. The integrated peak area of the ethylene produced by acetylene reduction was converted to a concentration by using an ethylene standard curve and normalized to the chlorophyll *a* concentration of the sample, determined as previously described (13).

Direct isolation of laboratory strains. A small tuft from each collection was streaked onto a petri dish containing D medium (3) solidified with 1.5% agar. With the aid of a dissecting microscope, four filaments of *Mastigocladus* were individually transferred to tubes of liquid D medium with a small plug of agar excised with watchmaker forceps and incubated under 75 mmol photons m^{-2} s^{-1} of cool white fluorescent light at 50°C. Totals of 64 and 60 direct isolation attempts were made from the White Creek and Boiling River collections, respectively. When growth was evident, a small clump of filaments was transferred to the center of a plate of D agar medium and incubated as described above. To obtain axenic clonal cultures, individual filaments (usually a motile dispersal filament called a hormogonium, but occasionally a fragment of outgrowth from the original clump) were transferred within 48 to 72 h to liquid D medium as described above. This process was repeated until the culture was judged to be free of contamination based on microscopic observation and on the absence of contaminating bacterial growth on D agar. Each strain designation indicates the field site of collection (B for Boiling River, W for White Creek), the collection (by number), and the isolation attempt made for a particular collection (A to D).

DNA isolation, gene amplification, and sequencing. Genomic DNAs were isolated from cultures as previously described (12). An approximately 950-bp fragment of the 16S rRNA gene was amplified from genomic DNAs of 25 randomly selected strains from each field site as described by Miller and Castenholz (12). The internal transcribed spacer (ITS) region of the *rrn* operon was amplified with primers complementary to conserved flanking sequences in the 16S and 23S rRNA genes (GCTGCAACTCGCCTRCRTGAAG and AW18 [30], respectively). Amplification conditions for a 50-μl reaction mixture were 94°C for 1 min, 58°C for 1 min, and 72°C for 30 s for 35 cycles. Sequences were also obtained for *trnL*-UAA (Table 1), encoding a leucine tRNA, which was previously shown to harbor a group I intron and evolve rapidly in some cyanobacteria (16). The reaction conditions were 35 cycles with an annealing temperature of 56°C and a 30-s extension step, but otherwise were the same as those described above. In addition, primers were designed for the amplification of several loci involved in nitrogen metabolism or its regulation (Table 1). A roughly 2.3-kb segment of the *nif* operon (*Anabaena* PCC 7120 *nifHD* nucleotide positions 88 to 2346) was amplified in two fragments. The reaction conditions for the first primer set were 40 cycles with an annealing temperature of 52°C and a 1-min extension; the conditions for the second primer set were 32 cycles with an annealing temperature of 58°C and a 1.5-min extension. The conditions for amplifying an approximately 0.5-kb fragment of *ntcA* (*Anabaena* PCC 7120 positions 61 to 573) were 35 cycles with an annealing temperature of 56°C and a 1-min extension, and those for a 1.1-kb fragment of *glnA* (*Anabaena* PCC 7120 positions 241 to 1310) were 32 cycles with an annealing temperature of 58°C and a 1.5-min extension. A ca. 1-kb fragment of the *nir* operon carrying 660 nucleotides (nt) of the 3′ end of the assimilatory nitrite reductase gene *nirA* and approximately 200 nucleotides of the 5′ end of the nitrate transport protein gene *nrtA* was also amplified with 40 cycles with an annealing temperature of 54°C and a 1-min extension. A fragment of the assimilatory nitrate reductase gene *narB* (*Anabaena* PCC 7120 positions 100 to 2151, excluding nucleotides 997 to 1053) was amplified in two fragments. The conditions were 35 cycles with an annealing temperature of 54°C and a 1.5-min extension time for the first primer set and 45 cycles with an annealing temperature of 50°C and a 1.5-min extension time for the second primer set. To amplify most of the open reading frame along with the upstream sequence for the heterocyst development regulatory gene *devH*, the reverse primer in Table 1 was originally paired with the primer GANCARCAN CGNGCNTG, designed from the conserved amino acid signature SCCRAH in the arginyl-tRNA synthetase gene upstream of *devH*. The forward primer in Table 1 was designed from the *Mastigocladus* sequence of this fragment, and an approximately 1.1-kb fragment of *devH* and upstream DNA was then amplified for most strains with 40 cycles with an annealing temperature of 50°C and a 1.5-min extension time.

TABLE 1. Primers designed for use in this study

Locus	Primer sequences (forward/reverse)
nifHD	AATACCCTAGCTGCGATGGC/CGTTGAGGTGTTTTTCGCGC
	GCGCGAAAAACACCTCAACG/GCGAAACCGTCGTAACCG
ntcA	CACAAGATAANGCCCTAGC/CGCARATCCCCTAGTAGCC
devH	AAATAGGTGATTAGGGAGTGGG/AAACGGGTGACGGTAACACG
narB	AACMCTWTGTCCKTAYTGTGG/CCWGCTTCYCTWCCTCCC
	AATCTGCACTTGATGACCGG/GCNCAGGCTTTTARNTCNGG
glnA	GATGGCGTACCTTTNGANGG/AANTCNTANGGATGAGGNCG
nirA	GAAGTCGAACAGCGTTTGGG/TGCTGTGGTTGCACTTGG
trnL-UAA	GCTCTCAAANTCAGGGAAACC/GGACTCTCCCTTTACCCTCG

Amplified products were cleaned either directly with a QIAquick PCR purification kit (QIAGEN) or following gel purification with a QIAquick gel extraction kit (QIAGEN). Cycle sequencing and cleaning were performed as previously described (11), with sequencing done bidirectionally on an ABI 3700 sequencer.

Phylogeny reconstruction. Phylogenies for Yellowstone *Mastigocladus*, *Mastigocladus* (*Fischerella*) CCMEE 5321, *Hapalosiphon* IAM M-264 (AB093485), *Chlorogloeopsis* PCC 6718 (AF132777), *Anabaena* PCC 7120 (X59559), and the outgroup cyanobacteriun *Chroococcidiopsis* PCC 7203 (from the Ribosomal Database Project II website [http://rdp.cme.msu.edu]) were reconstructed from 948 nucleotides of the 16S rRNA gene (*Escherichia coli* positions 360 to 1326) by maximum likelihood, maximum parsimony, and neighbor-joining methods, using PAUP*, version 4.0b (24), following sequence alignment as described previously (12). For the likelihood analysis, the model of DNA sequence evolution was chosen by hierarchical likelihood ratio tests, as implemented in Modeltest (17). The model selected (HKY + G + I) estimates the transition/transversion ratio and incorporates among-nucleotide site rate heterogeneity by estimating both the proportion of invariant sites and the shape of the discrete approximation (n = 4 categories) of a gamma distribution for variable sites. The analysis was bootstrap replicated 1,000 times. Both the neighbor-joining tree and the maximum parsimony tree (obtained by an exhaustive search of all possible trees) were bootstrapped 10,000 times.

Population genetic parameter estimation. Nucleotide sequences aligned by Clustal W (27) were analyzed with respect to the following with DnaSP (20): synonymous and nonsynonymous polymorphic sites; number of haplotypes; the average number of nucleotide differences between a pair of sequences; nucleotide diversity and the population-scaled mutation rate; degree of genetic differentiation between populations, estimated by F_{ST} (9); and the effective number of codons (32), an index of codon usage. Sequence data were also tested for conformity to the Wright-Fisher neutral model with Tajima's (26) D test, as implemented in DnaSP. Hudson et al.'s (8) test of the neutral theory of molecular evolution, as implemented in DnaSP, was used to evaluate whether different genes have had different selective histories.

Migration rates between the two populations and their separation time were estimated for an aligned multilocus data set (for *narB*, *glnA*, *devH*, and *nifHD*) according to the "isolation with migration" (IM) model of population divergence (6, 14), as implemented by the IM program (http://lifesci.rutgers.edu/~heylab/HeylabSoftware.htm). The model considers the current distribution of sequence polymorphisms to be the product of physical separation of an ancestral population into two descendant populations at some time in the past, following which there may or may not have been gene exchange by migration between populations. IM uses the Markov chain Monte Carlo approach to fit the model to the aligned sequence data and estimate the likelihoods of the model parameters (separation time at which the ancestral population split into the two populations, migration rates, population mutation rates, and their four locus-specific mutation scalars) given the sequence data. Because the timescale of sequence divergence in units of generations or absolute time is unknown, the model scales the parameters by the mutation rate. The Markov chain was initiated with a burn-in length of 100,000 steps to achieve independence of starting conditions, followed by a chain length of 250,000,000 steps. To attain good chain mixing, the software 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 690 and 37,000, and the results for three independent runs were nearly identical.

Physiological assays. Ten strains were randomly chosen from each population and assayed for fitness (measured as the exponential growth rate) under standard maintenance conditions in both N-containing and non-N-containing media (D and ND medium, respectively). The compositions of the media used in the two nitrogen treatments differed only with respect to the presence or absence of nitrate as a source of combined N, respectively. Prior to the assay, an exponential-phase stock culture of each strain growing in N-containing medium was split into two subclones and grown in N-containing medium for 48 h. Each exponential-phase subclone was next split into N-containing and non-N-containing cultures. To do so, a portion of each subclone culture was pelleted in a 1.5-ml microcentrifuge tube, rinsed three times in 1 ml non-N-containing medium, and then resuspended in non-N-containing medium to an optical density of 1 at 750 nm. One-hundred-microliter aliquots were then delivered to fresh N-containing and non-N-containing tubes for a total of 2 (populations) × 10 (strains) × 2 (subclones) × 2 (N treatments) = 80 tubes. After several generations of exponential growth, a homogenized sample from each tube was transferred to a fresh tube of corresponding medium to a final optical density of 0.001. These experimental tubes were incubated for 14 days and monitored with a spectrophotometer for changes in optical density, and exponential growth rates were estimated from the linear component of a plot of the logarithms of optical densities as a function of time. Data were analyzed with SPSS, version 8.0, according to a mixed-effects nested analysis of variance (ANOVA) model with growth rate as a dependent variable, nitrogen and population treatments as fixed factors, and the strain as a random variable nested within the population.

Nucleotide sequence accession numbers. The sequences reported in this paper have been deposited in the GenBank database under the following accession numbers: *rrn*-16S rRNA, DQ372835; ITS, DQ372836; *devH*, DQ385930 to DQ385949; *glnA*, DQ372815 to DQ372834; *narB*, DQ385950 to DQ385969; *nifHD*, DQ385910 to DQ385929; *nirA*, DQ385970 to DQ385989; *ntcA*, DQ385890 to DQ385909; and *trnL*-UAA, DQ372837.

RESULTS AND DISCUSSION

Field site characteristics. White Creek and Boiling River are both dominated (in biomass) by *Mastigocladus* (*Fischerella*) *laminosus*, a multicellular cyanobacterium that, under nitrogen-limited conditions, is capable of developing intercalary heterocysts, terminally differentiated cells specialized for enabling the oxygen-sensitive process of nitrogen fixation in an oxic environment. Over the course of several years, it had been observed that *Mastigocladus* filaments in White Creek produced heterocysts, whereas those in Boiling River did not (S. R. Miller, unpublished observations). Because both heterocyst development and nitrogen fixation activity are negatively regulated by the presence of preferred N sources (31), it was suspected that the phenotypic differences observed between the two populations were the results of environmental effects on gene expression. Consistent with this interpretation, an analysis of water samples collected in June 2001 confirmed that dramatic differences in nitrogen content between the two sites can explain the repression or production of heterocysts in the respective populations. Combined nitrogen was abundant in Boiling River, principally as nitrate but also, to a much lesser extent, as organic N (Table 2); in contrast, combined nitrogen was not detectable in White Creek. Similar results were obtained with colorimetric assays in July 1999 (not shown).

TABLE 2. Field data for the *Mastigocladus* populations

Site	Nitrate concn (mg N liter^{-1})	Ammonium concn (mg N liter^{-1})	Total nitrogen concn (mg N liter^{-1})	Heterocysts	Acetylene reduction activity (nmol µg^{-1} Chl h^{-1})[a]
White Creek	<0.1	<0.1	<0.1	+	151.5 ± 31.19
Boiling River	0.13	<0.1	0.15	−	1.5 ± 0.82

[a] Values are means ± standard errors.

In situ nitrogen metabolism differs between populations. To confirm that the heterocyst-producing White Creek population was indeed actively fixing nitrogen in situ, whereas the Boiling River population was not, we estimated the nitrogen fixation activities of both populations by the standard acetylene reduction assay, which measures the reduction of acetylene to ethylene by nitrogenase. As expected, the acetylene reduction rate was high in the White Creek population, but acetylene reduction was not detected in the Boiling River population (Table 2), as the results did not differ from background ethylene production levels (not shown).

Laboratory strain isolation. Multiple live collections were taken in June 2001 for manual isolation of individual *Mastigocladus* filaments in laboratory culture. By avoiding enrichment prior to isolation, we did not bias the representation of the population diversity in the samples. Nearly all isolation attempts from both populations (62/64 from White Creek and 59/60 from Boiling River) were successful, indicating that the isolation process itself did not select against the recovery of particular genotypes in culture.

Phylogeny of *Mastigocladus* strains. Twenty-five strains were randomly chosen from each population for sequencing of approximately 950 bp of the 16S rRNA gene locus. The sequences of all 50 strains were completely identical and were also identical to sequences obtained for two additional *Mastigocladus* strains which had been recently isolated from Yellowstone National Park, i.e., strains CCMEE 5207 (Chocolate Pots) and CCMEE 5208 (Obsidian Pool). The phylogenetic positions of these strains were analyzed along with those of several additional heterocystous cyanobacteria, including a *Mastigocladus* strain isolated from an Icelandic hot spring (CCMEE 5321) and the thermophile *Chlorogloeopsis* PCC 6718. The heterocystous cyanobacteria are a monophyletic group, and all analyses were rooted with the outgroup *Chroococcidiopsis* PCC 7203, which has been suggested to be the closest relative of the heterocystous clade in many previous phylogenies (e.g., see reference 4). The tree topology was consistent across maximum likelihood, parsimony, and neighbor-joining methods (Fig. 1). *Mastigocladus* CCMEE 5321, which is morphologically similar to the Yellowstone strains, was obtained from a similar habitat, and has 98.1% sequence identity at this locus, was

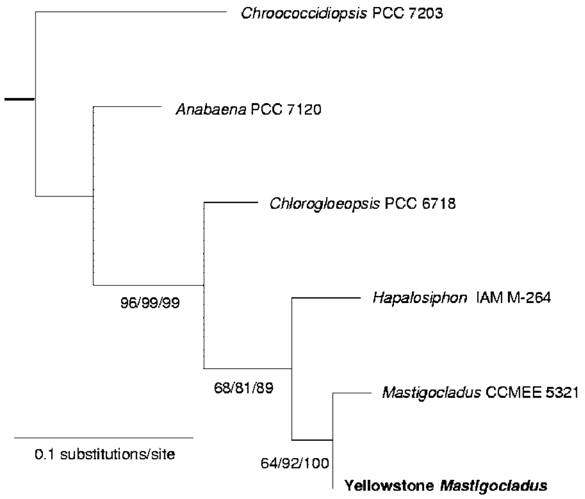

FIG. 1. Maximum likelihood phylogeny of Yellowstone *Mastigocladus* inferred from 948 nucleotides of the 16S rRNA gene. Values at nodes indicate bootstrap frequencies for likelihood, parsimony, and neighbor-joining phylogenies, respectively.

inferred with substantial bootstrap support to be the sister taxon of the Yellowstone strains.

Single nucleotide polymorphism (SNP) identification in Yellowstone populations. To determine whether the White Creek and Boiling River populations are genetically distinct, we examined sequence variation at loci that evolve more rapidly than the extremely conserved 16S rRNA gene. Of the 25 strains originally characterized from each population, 10 were randomly selected for further analysis. As found for the 16S rRNA locus, all strains were completely identical at both the ITS region between the 16S and 23S rRNA genes and the *trnL*-UAA gene, which was previously reported (16) to be a useful marker for the closely related *Nostoc* group (not shown).

We next investigated six protein-coding genes involved in nitrogen metabolism (Table 3). *glnA* encodes glutamine synthetase, a constitutively expressed central nitrogen metabolism enzyme that incorporates ammonium into amino acids (5). *ntcA* encodes a DNA binding protein essential for the transcriptional control of many aspects of nitrogen metabolism, including nitrogen fixation and nitrate assimilation (5). *devH* encodes a DNA binding regulatory protein required for the development of a functional heterocyst (19). *nifHD* encodes two components of the nitrogenase complex, which reduces dinitrogen to ammonium (5). The nitrate reductase (*narB*) and nitrite reductase (*nirA*) genes are key enzymes of the nitrate assimilation pathway (5).

Of the 7,838 nucleotides sequenced for each strain, roughly 0.3% of all sites were variable, for a total of 25 SNPs. The majority of SNPs in protein-coding regions were synonymous, with replacement SNPs only observed for genes involved in nitrate assimilation (Table 3). The average nucleotide diversity (π) at silent sites (the number of nucleotide differences per silent site between two randomly chosen sequences) was 0.0028. For perspective, this value is approximately three times greater than that observed for the human global population at noncoding autosomal loci (~0.0009) (34). More slowly evolving loci typically used in molecular microbial ecology therefore belie substantial genetic diversity in this system.

The nucleotide diversity at silent sites (i.e., the silent site mutation rate estimator π_s) varied between 0 (*ntcA*) and ~0.01 (*narB*) (Table 3). This range of π values is comparable to that observed for primates (34). These differences can arise for several reasons, including stochastic effects, variations in mutation rates among loci due to different selective constraints (e.g., on codon usage), and differences across genomic regions in their recent selective histories. Codon usage bias has been shown to be strongly correlated with levels of gene expression and negatively correlated with the synonymous mutation rate (21). In the present case, differences in the degree of codon usage bias cannot explain the observed differences in the synonymous mutation rate, as the Pearson correlation coefficient between π and the effective number of codons (Table 3), an index of codon usage (32), was not significant ($P = 0.792$).

Tests of the neutral theory, however, provide evidence that different nitrogen metabolism genes have had different selective histories in the recent past for Yellowstone *Mastigocladus* organisms as a whole. The HKA test (8) examines the neutral theory prediction that levels of polymorphism within lineages should be positively correlated with evolutionary divergence between lineages (e.g., fast-evolving genes should have high levels of polymorphism). The test uses chi-square distribution expectations to compare observed ratios of polymorphism to divergence for two or more regions of the genome, and deviations from the null model that these ratios are the same are

TABLE 3. Summary of polymorphism data for nitrogen metabolism genes[a]

Locus	Size (nt)	No. of polymorphic sites		No. of alleles	k	π		θ_w		ENC	D	F_{ST}
		S	N			S	N	S	N			
narB	1,998	10	5	3	6.71	0.0096	0.0014	0.0059	0.0009	53.39	2.17*	0.76
nirA	946	0	1	2	0.44	0	0.0007	0	0.0004	51.29	1.03	0.56
nifHD	2,275	4	0	3	2.06	0.0032	0	0.0018	0	39.76	2.39*	0.94
devH	1,045	2	0	2	0.88	0.0015	0	0.0009	0	48.94	1.33	0.56
glnA	1,061	3	0	2	0.81	0.0032	0	0.0033	0	45.55	−0.13	0.22
ntcA	513	0	0	1	0	0	0	0	0	60.3		

[a] The sample is 10 randomly chosen alleles from each population. Abbreviations: S, silent site substitutions (synonymous codon sites and noncoding regions); N, replacement substitutions at nonsynonymous codon sites; k, average nucleotide difference between a pair of sequences; π, number of nt differences between two randomly chosen alleles per silent and replacement site; θ_w, Watterson's estimator of the scaled mutation rate per site; ENC, effective number of codons; D, Tajima's D value. *, $P < 0.05$.

TABLE 4. HKA test results

Gene	χ^2 value for gene comparison[a]			
	narB	*nifHD*	*devH*	*nirA*
narB		4.4**	6.6**	3.0*
nifHD			0.1	0.0
devH				0.0

[a] **, $P < 0.05$; *, $0.05 < P < 0.10$.

TABLE 5. Multilocus haplotypes and within-population frequencies

Population	Haplotype	Allele					Frequency
		narB	*nirA*	*nifHD*	*devH*	*glnA*	
White Creek	A	2	2	3	2	1	0.5
	B	3	1	2	1	2	0.3
	C	3	2	3	2	1	0.1
	D	3	1	3	1	1	0.1
Boiling River	A	1	1	1	1	1	1.0

evidence that the evolutionary processes shaping the genealogies of the regions are different. We compared polymorphisms within Yellowstone isolates and divergence between Yellowstone *Mastigocladus* and *Mastigocladus* strain CCMEE 5321 for the following four loci: *narB* (1,944 nt), *nirA* (776 nt), *nifHD* (999 nt), and *devH* (1,591 nt). The results presented are for all nucleotide sites (Table 4), but the results based on silent sites alone were qualitatively similar. Significant deviations from the neutral model were observed for the *narB/nifHD* ($P = 0.037$) and *narB/devH* ($P = 0.037$) comparisons. Similarly, the test for *narB/nirA* was of borderline significance ($P = 0.08$). Because these loci are only approximately 5 kb apart on the *Mastigocladus* genome (data not shown), this result suggests that genes in close proximity can evolve with very different evolutionary dynamics.

Although the HKA tests suggested that the *narB* locus has experienced a different selective history from that of the other regions, the results do not indicate the source of deviation from the model. A deficit of polymorphisms in *nifHD* and *devH*, an excess of polymorphisms in *narB*, reduced divergence at *narB*, excessive divergence at *nifHD* and *devH*, or a combination of the above could all potentially produce these results. It appears that polymorphism patterns contribute greatly to the observed deviations from the neutral model because the amount of *narB* polymorphism (0.77% of sites) is approximately four times that of the other loci (0.16 to 0.19%), whereas the levels of between-lineage divergence are more comparable (3.0% for *narB* and 4.1 to 5.9% for other loci). While *narB* in particular appears to be anomalous and is likely the primary source of deviation, the allelic identity of a strain is not obviously associated with its performance on nitrate (see below).

White Creek and Boiling River populations are genetically differentiated. The White Creek and Boiling River populations are genetically differentiated from each other, despite their close proximity. This is most strikingly illustrated by the observation that no haplotype was shared between populations (Table 5). White Creek and Boiling River also have very different population genetic structures. Whereas no molecular variation was observed among the 10 strains analyzed from the Boiling River population, the four multilocus haplotypes detected in the White Creek sample ranged in estimated frequency between 10 and 50% and set a minimum bound on the number of distinct genotypes present at this site (Table 5). Physical differences between the White Creek and Boiling River habitats may contribute to the observed differences in genetic structure. *Mastigocladus* dominates the biomass in White Creek from a mean temperature below 40°C up to approximately 56°C, along a thermal gradient that stretches for more than a kilometer. This gradient creates a highly structured environment with the potential, in theory, to support multiple genotypes with divergent thermal ecologies. The possibility that the different haplotypes isolated from the White Creek site differ in their thermal performance traits requires further investigation. In contrast, Boiling River is a short (ca. 150-m) channel with, consequently, an insignificant thermal gradient and a much smaller population size. Its less heterogeneous environment might be expected to support fewer genotypes than White Creek and to render the population more prone to environmental perturbation (e.g., an increase in temperature). Small populations are also more susceptible to the removal of variation by genetic drift (particularly in the event of a perturbation-induced population bottleneck) or by the selective sweep of a favored genotype. Because mutation introduces new variation into a population, the lack of detectable variation in Boiling River suggests that a recent purging of genetic diversity has indeed occurred. It is not possible, however, to speculate whether this homogenization was due to chance or adaptation without knowledge of the population's diversity levels and demography, respectively, in its recent past.

Although data are limited for other locations in Yellowstone, our results for two additional park strains are consistent with the hypothesis that *Mastigocladus* haplotypes are unique to local populations. The *narB-nifHD-devH-glnA* haplotypes (according to the coding scheme in Table 5, but including alleles not observed for White Creek or Boiling River) of strains CCMEE 5207 and 5208 were 4-1-1-1 and 5-4-1-3, respectively.

The signature of this observed local genetic subdivision between populations should also be evident in the genealogies of the individual loci. A useful null model of the expected pattern of DNA polymorphisms in a sample of alleles can be derived from gene genealogies modeled as a coalescent process according to the neutral Wright-Fisher model (e.g., see reference 7). Briefly, for a large and constant-sized population of N haploid individuals, the model assumes that individuals in one generation are equally likely to leave offspring in the next generation, so the contributions of alleles by individuals to the next generation are approximately Poisson distributed. As a result, for a sample of n alleles, pairs of alleles (which may or may not be genetically identical) coalesce backwards in time (in units of generations) by an approximately exponentially distributed process until the most recent common ancestor (i.e., the root of the gene tree) of the sample is found. For our case of 20 alleles, the probability that the most recent common ancestor of the sample is the root of the entire population N is expected to be approximately 90%. Neutral mutations, which do not affect the likelihood of leaving offspring, are added to the genealogy at a constant rate independent of coalescence events and can occur a maximum of one time at a given nucleotide site. This assumption of an infinite-site model specifies a direct correspondence between the number of mutations and the number of observed polymorphic sites in a sample. Because there is no evidence for repeat mutations at any nucleotide site in our data (e.g., no examples of three-variant sites), this assumption is reasonable.

Different estimators of the population-scaled mutation rate (θ), based on either the number of polymorphic sites (28) or average pairwise nucleotide differences (25) (π is Tajima's estimator expressed on a per-site basis), give identical results when the data conform to the neutral Wright-Fisher model. Geographic subdivision violates the Wright-Fisher model, however, because alleles from the same site are more likely to coalesce than alleles from different sites. In a genealogy, this means that the time to coalescence for alleles from different sites is relatively longer than expected, with the result that there will be more mutations than expected along internal branches of the genealogy and an excess of intermediate-frequency sequence variants in the sample. Because Tajima's estimator of θ weighs mutations on internal branches more heavily than those at the tips of the genealogy, whereas Watterson's θ weighs all mutations equally, significant positive distortion from zero of the normalized difference of the two estimators (D) is evidence for deviation from the neutral model expectation due to population subdivision (26). Ideally, this distortion should be evident at multiple loci. However, because the ability to infer the shape of a genealogy increases with the number of mutations in a sample, our estimate of D will be most reliable for the *narB* data. D was significantly positive for both *narB* and *nifHD* (Table 3), providing genealogical evidence for the subdivision of genetic variation between White Creek and Boiling River.

Finally, the degree of genetic differentiation between populations can be directly estimated by F_{ST}, which takes on values between 0 (when different populations harbor the same alleles in the same proportions) and 1 (when the populations are fixed for different alleles). From a genealogical perspective, F_{ST} measures the magnitude of the difference in expected mean times to coalescence between a pair of alleles from the same location and a pair drawn at random from the entire sample (22). The observed values of F_{ST} for polymorphic loci were considerable and ranged between 0.22 and 0.94 (Table 3). For comparison, the F_{ST} value for two *Yellowstone* populations of the hyperthermophilic archaeon *Sulfolobus* was estimated to be 0.37 (29).

Fitness on different N sources. The genetic differentiation observed between the White Creek and Boiling River sites raises the question of whether these *Mastigocladus* populations also differ with respect to relative fitness on the respective N sources encountered in situ, which would suggest adaptation to local conditions. Given sufficient time and genetic variation, we would expect that performance on assimilated N will improve within a population, whereas the ability to assimilate an unused N source (because it is either not available in situ or not preferred) will decline due to relaxed selective constraints on genes involved in its metabolism. As a first attempt to investigate whether there are locally adaptive differences in nitrogen metabolism between the two populations, we tested whether members of the Boiling River population perform better on average when grown with nitrate as a N source than the White Creek population and, conversely, whether members of the latter population outperform the Boiling River population under nitrogen-fixing conditions. To do this, we assayed growth on both N-containing (nitrate) and non-N-containing (N_2) media for duplicate subclone lines of the 10 randomly selected, genetically characterized strains from each population. The exponential growth rate (h^{-1}) and yield (optical density following 14 days of incubation) were each analyzed with a mixed-effects nested ANOVA model, with N source and population as fixed factors and the random factor (strain) nested within population.

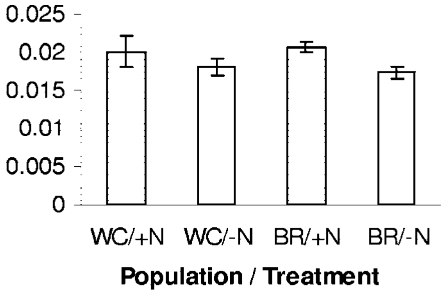

FIG. 2. Exponential-growth-rate constants (means ± standard errors) for White Creek (WC) and Boiling River (BR) populations grown in the presence (+N) or absence (−N) of nitrate. The results for nitrogen treatments were significantly different for both populations ($P < 0.0001$).

Initial inspection of the growth rate data revealed that the two populations performed similarly as a whole on N-containing and non-N-containing media (Fig. 2). In the nested model, there were significant effects of N source ($F = 45.41$; $P < 0.0001$) and strain within population ($F = 6.61$; $P < 0.0001$). Neither the effect of population ($P = 0.90$) nor the population–N-source interaction ($P = 0.13$) was significant, suggesting that there is no evidence for population-wide adaptation to the local N source. The N source effect, with the mean performance on nitrate being superior to that on dinitrogen, was expected given the greater energetic cost of nitrogen fixation. The strain-within-population effect indicated differences in growth rate among strains within at least one of the populations. To test which population contained strains that were significantly different, sets of contrasts were performed for each population. Whereas there was clear evidence for growth rate variation among members of the White Creek population ($F = 12.25$; $P < 0.0001$), strains from the Boiling River population did not differ from each other ($F = 0.96$; $P = 0.48$). The latter was expected, given that we detected no SNP variation among the Boiling River strains.

Each population was further analyzed separately by ANOVA, with N source and strain as factors. Both models explained most of the variation in their respective data ($R^2 = 0.94$ for the White Creek model; $R^2 = 0.81$ for the Boiling River model). Consistent with the above result, only the effect of the N source was significant for the Boiling River data ($F = 60.99$; $P < 0.0001$ [compare with values of 1.71 for F and 0.15 for P for the strain effect and of 1.00 for F and 0.47 for P for the N-source–strain interaction]). That is, all strains performed the same, with substantially higher fitness on N-containing medium. In contrast, N source ($F = 28.91$; $P < 0.0001$), strain ($F = 24.47$; $P < 0.0001$), and their interaction ($F = 6.52$; $P < 0.0001$) were all very highly significant for the White Creek data. Qualitatively similar results were obtained for the yield models (not shown).

The interaction between N source and strain for White Creek is of interest, as it indicates a genotype-environment interaction within this genetically heterogeneous population which can largely be understood in terms of its genetic structure. The interaction principally manifests itself as the different slopes of the norm of reaction between N-containing and non-N-containing treatments for strains with the WC-2 haplotype and for all other strains, respectively. Whereas the mean exponential growth rate (day^{-1}) for other strains was higher with nitrate, as expected (0.023 ± 0.0006 versus 0.019 ± 0.0006 without N), strains belonging to the WC-2 haplotype group actually grew at comparable rates under nitrogen-fixing conditions (0.016 ± 0.0005) and with nitrate (0.014 ± 0.0011). The basis for this difference is not clear; visual inspection of WC-2 strains confirmed the absence of heterocysts when grown with nitrate, indicating that the strains could use the nitrate and were not growing by nitrogen fixation. However, no difference in yield on the respective N sources was observed among White Creek strains (data not shown).

Gene flow between populations is insufficient to prevent their continued divergence. Although there is no clear evidence for local adaptation of the current *Mastigocladus* populations, the question of whether they will continue to diverge and possibly adapt to local conditions in the future remains. The answer depends on whether the rate of gene flow between populations is sufficient to prevent further divergence, but the current distribution of molecular variation within and between the White Creek and Boiling River populations can potentially be explained by competing models which make very different predictions about their evolutionary fates. Specifically, the observed values of F_{ST} (Table 3) can be obtained by assuming either (i) the recent isolation of populations, with no gene flow (migration) between them, or (ii) that the populations are islands that have been separated for a long time and have achieved an equilibrium between gene flow and genetic drift (33). Whereas the isolation model assumes that populations will continue to diverge genetically, the island model assumes that the populations will not diverge more than they already have due to migration. These models represent respective extremes, with neither necessarily being valid for a particular data set.

Because levels of genetic differentiation are typically associated with degrees of phenotypic and ecological diversification, evaluating whether these *Mastigocladus* populations will continue to diverge genetically has important implications for their potential to adapt to local conditions as well as for our general understanding of how the ecological diversity of microorganisms becomes geographically structured. The sharing of alleles (e.g., for *devH* and *glnA*) between populations suggests their recent separation, but the lack of a shared multilocus haplotype also suggests that local differentiation outpaces migratory gene flow. We tested whether these populations have indeed been recently isolated, with negligible gene flow between them, by implementing the IM model (6, 14), a recently developed approach that avoids the restrictive assumptions of the above isolation and island models, including relaxation of the assumption that migration rates between populations are equal. IM uses a Markov chain Monte Carlo approach to estimate the likelihoods of both separation time and migration rate (which are both normalized by the mutation rate) between populations from sequence data for multiple loci.

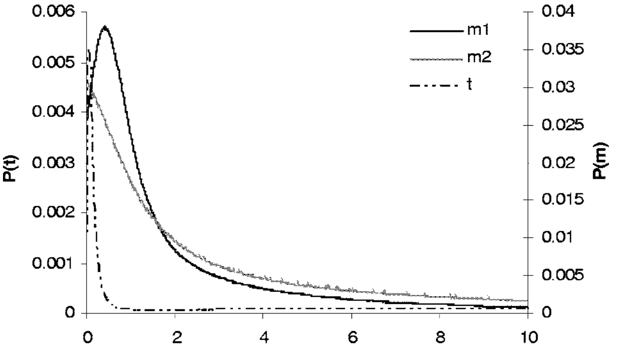

FIG. 3. Likelihood function surfaces for population separation time *t* and migration rate per mutation event from Boiling River to White Creek (m_1) and from White Creek to Boiling River (m_2).

A model of recent divergence time and low (undetectable) gene flow is most compatible with a multilocus data set for *narB*, *glnA*, *devH*, and *nifHD*. The posterior probability distributions (i.e., the likelihood function surfaces) for the separation time and migration parameters are shown in Fig. 3. The time parameter in the model, *t*, is measured in units of mutations since population splitting and takes on a value between 0 (for a single population that has not split) and infinity (i.e., very large values conform to the island model). The maximum likelihood estimate of *t* is close to zero (0.045), indicating a recent separation of the two populations. The marginal posterior probability distributions of the migration parameters m_1 (the rate of migration for each gene copy from Boiling River to White Creek per mutation event) and m_2 (the converse) are maximized at 0.415 and 0, respectively (Fig. 3). The former estimate, however, is not significantly different from zero by a likelihood ratio test [$-2\Delta(\ln L) = 0.78$; $P = 0.38$ for χ^2 distribution with $df = 1$]. Even if this estimate were accurate, this amount of gene flow would be insufficient to prevent continued population divergence. The actual number of migrants per generation, N_m, that is predicted by this estimate can be derived for each locus from the IM model as $m_{\text{WC}}\theta u_i$, where θ is the estimated general population mutation rate parameter and u_i is the estimated relative locus-specific mutation scalar for locus *i* (not shown). Estimated values of N_m for the four loci ranged between 0.02 and 0.14, much lower than the roughly 1 migrant gene copy per generation that is required to prevent substantial divergence (10, 33).

Concluding remarks. We have described two recently isolated bacterial populations that are early in the process of diversifying. The number of mutations that have occurred at the examined nitrogen metabolism loci since population isolation can be estimated from the IM model by the product of *t* and the sum of the mutation scalars for the individual loci (not shown) normalized to sequence length. The maximum likelihood estimate is 0.4 mutations per kb, or approximately 3 mutations over the 6.7-kb sample. This result suggests that much of the observed segregating variation in these populations is ancestral and that geographic sorting of ancestral polymorphisms has probably, to this point, played a larger role in the genetic and phenotypic differentiation of these populations than has the input of new mutations. Altogether, the results suggest that local adaptation to the prevailing nutrient status may be mutation limited but also that locally adaptive mutations may ultimately become fixed in their respective environments given sufficient time.

ACKNOWLEDGMENTS

S.R.M. was supported by a National Science Foundation postdoctoral fellowship in microbial biology (DBI-0102100), NSF Research Starter grant MCB-0354738, and NSF major research instrumentation award DBI-0321329. This work was supported by National Park Service research permits to B. M. Bebout and R. W. Castenholz.

We thank Brad Bebout and Mary Hogan for their help with the acetylene reduction assays and gas chromatography and Deana Pedersen for technical support. We also thank two anonymous reviewers for their helpful suggestions.

REFERENCES

1. **Atwood, K. C., L. K. Schneider, and F. J. Ryan.** 1951. Periodic selection in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **37:**146–155.
2. **Bennett, A. F., and R. E. Lenski.** 1993. Evolutionary adaptation to temperature. II. Thermal niches of experimental lines of *Escherichia coli*. Evolution **47:**1–12.
3. **Castenholz, R. W.** 1988. Culturing methods for cyanobacteria. Methods Enzymol. **167:**68–93.
4. **Fewer, D., T. Friedl, and B. Büdel.** 2002. *Chroococcidiopsis* and heterocyst-differentiating cyanobacteria are each other's closest living relatives. Mol. Phylogenet. Evol. **23:**82–90.
5. **Flores, E., and A. Herrero.** 1994. Assimilatory nitrogen metabolism and its regulation, p. 487–517. *In* D. A. Bryant (ed.), The molecular biology of cyanobacteria. Kluwer Academic Publishers, Dordrecht, The Netherlands.
6. **Hey, J., and R. Nielsen.** 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.
7. **Hudson, R. R.** 1991. Gene genealogies and the coalescent process. Oxford Surv. Evol. Biol. **7:**1–44.
8. **Hudson, R. R., M. Kreitman, and M. Aquadé.** 1987. A test of neutral molecular evolution based on nucleotide data. Genetics **116:**153–159.
9. **Hudson, R. R., M. Slatkin, and W. P. Maddison.** 1992. Estimation of levels of gene flow from DNA sequence data. Genetics **132:**583–589.
10. **Latter, B. D. H.** 1973. The island model of population differentiation: a general solution. Genetics **73:**147–157.
11. **Miller, S. R.** 2003. Evidence for the adaptive evolution of the carbon fixation gene *rbcL* during diversification in temperature tolerance of a clade of hot spring cyanobacteria. Mol. Ecol. **12:**1237–1246.
12. **Miller, S. R., and R. W. Castenholz.** 2000. Evolution of thermotolerance in hot spring cyanobacteria of the genus *Synechococcus*. Appl. Environ. Microbiol. **66:**4222–4229.
13. **Miller, S. R., C. E. Wingard, and R. W. Castenholz.** 1998. 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.
14. **Nielsen, R., and J. Wakeley.** 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics **158:**885–896.
15. **Papke, R. T., N. B. Ramsing, M. M. Bateson, and D. M. Ward.** 2003. Geographic isolation in thermophilic cyanobacteria. Environ. Microbiol. **5:**650–659.
16. **Paulsrud, P., and P. Lindblad.** 1998. Sequence variation of the tRNA(Leu) intron as a marker for genetic diversity and specificity of symbiotic cyanobacteria in some lichens. Appl. Environ. Microbiol. **64:**310–315.
17. **Posada, D., and K. A. Crandall.** 1998. Modeltest: testing the model of DNA substitution. Bioinformatics **14:**817–818.
18. **Rainey, P. B., and M. Travisano.** 1998. Adaptive radiation in a heterogeneous environment. Nature **394:**69–72.
19. **Ramírez, M. E., P. B. Hebbar, R. Zhou, C. P. Wolk, and S. E. Curtis.** 2005. *Anabaena* sp. strain PCC 7120 gene *devH* is required for synthesis of the heterocyst glycolipid layer. J. Bacteriol. **187:**2326–2331.
20. **Rozas, J., J. C. Sánchez-DelBarrio, X. Messeguer, and R. Rozas.** 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:**2496–2497.
21. **Sharp, P. M., and W.-H. Li.** 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24:**28–38.
22. **Slatkin, M.** 1991. Inbreeding coefficients and coalescence times. Genet. Res. **58:**167–175.
23. **Stewart, W. D. P., G. P. Fitzgerald, and R. H. Burns.** 1967. In situ studies on N_2 fixation using the acetylene reduction method. Proc. Natl. Acad. Sci. USA **58:**2071–2078.
24. **Swofford, D. L.** 1998. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods), version 4. Sinauer Associates, Sunderland, Mass.
25. **Tajima, F.** 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105:**437–460.
26. **Tajima, F.** 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:**585–595.
27. **Thompson, J., D. Higgins, and T. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
28. **Watterson, G. A.** 1975. On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. **7:**256–276.
29. **Whitaker, R. J., D. W. Grogan, and J. W. Taylor.** 2003. Geographic barriers isolate endemic populations of hyperthermophilic Archaea. Science **301:**976–978.
30. **Wilmotte, A., G. Van der Auwera, and R. De Wachter.** 1993. Structure of the 16S ribosomal RNA of the thermophilic cyanobacterium *Chlorogloeopsis* HTF ("*Mastigocladus laminosus* HTF") strain PCC7518, and phylogenetic analysis. FEBS Lett. **317:**96–100.
31. **Wolk, C. P.** 2000. Heterocyst formation in *Anabaena*, p. 83–104. *In* Y. V. Brun and L. J. Shimkets (ed.), Prokaryotic development. American Society for Microbiology, Washington, D.C.
32. **Wright, F.** 1990. The "effective number of codons" used in a gene. Gene **87:**23–29.
33. **Wright, S.** 1931. Evolution in Mendelian populations. Genetics **16:**97–159.
34. **Yu, N., M. I. Jensen-Seaman, L. Chemnick, O. Ryder, and W.-H. Li.** 2004. Nucleotide diversity in gorillas. Genetics **166:**1375–1383.