Published in final edited form as:
Infect Genet Evol. 2011 March ; 11(2): 343–348. doi:10.1016/j.meegid.2010.11.005.

Ongoing purifying selection on intergenic spacers in 1 group A streptococcus

Haiwei Luo[1], **Jijun Tang**[2], **Robert Friedman**[1], and **Austin L. Hughes**[1]

[1]Department of Biological Sciences, University of South Carolina, Columbia 29208, USA

[2]Department of Computer Science and Engineering, University of South Carolina, Columbia 29208, USA

Abstract

Bacterial intergenic spacers are non-coding genomic regions enriched with *cis*-regulatory elements for gene expression. A population genetics approach was used to investigate the evolutionary force shaping the genetic diversity of intergenic spacers among 13 genomes of group A streptococcus (GAS). Analysis of 590 genes and their linked 5' intergenic spacers showed reduced nucleotide diversity in spacers compared to synonymous nucleotide diversity in protein-coding regions, suggestive of past purifying selection on spacers. Certain spacers showed elevated nucleotide diversity indicative of past homologous recombination with divergent genotypes. In addition, analysis of the difference between mean nucleotide difference and number of segregating sites showed evidence of an excess of rare variants both at nonsynonymous sites in genes and at sites in spacers, which is evidence that there are numerous slightly deleterious variants in GAS populations with potential effects on both protein sequences and gene expression.

Keywords

bacterial spacers; Group A streptococcus; purifying selection; slightly deleterious mutations; *Streptococcus pyogenes*

Introduction

In bacterial genomes, intergenic spacers host *cis*-regulatory sequences regulating gene expression, as well as encoding small RNAs (sRNAs) with regulatory function, including riboswitches, small regulatory RNAs and others (Perez *et al.*, 2009). Because intergenic regions in prokaryotes are much shorter on average than in eukaryotes, the density of regulatory elements per intergenic region is expected to be much higher in prokaryotes than in eukaryotes (Hughes and Friedman, 2004; Rogozin *et al.*, 2002). A number of studies have focused on the evolutionary dynamics of *cis*-regulatory elements and sRNAs, since changes in transcriptional and translational regulation are likely to play an important role in phenotype changes (Horler and Vanderpool, 2009; Madan Babu and Teichmann, 2003; Moses *et al.*, 2003; Rodriguez-Trelles *et al.*, 2003; Sridhar and Rafi, 2008; Wray *et al.*,

Corresponding author: Austin L. Hughes, Department of Biological Sciences, University of South Carolina, Columbia SC 29208 USA; Tel. +1-803-777-9186; austin@biol.sc.edu.

2003). In group A streptococcus (GAS, *Streptococcus pyogenes*), transcriptional regulation is known to play an important role in virulence; for example, the Mga regulator activates 98 and represses 103 virulence-related genes (Ribardo and McIver, 2006), illustrating the potential for mutations in *cis*-regulatory elements to have a major impact on phenotypes of clinical relevance in this important human pathogen.

Evolutionary biologists distinguish two types of natural selection: (1) positive selection, which favors advantageous mutations; and (2) purifying selection, which acts to eliminate deleterious mutations (Hughes, 2007). In the case of protein-coding genes, there is evidence that purifying selection is ubiquitous, whereas positive selection is rarer (Kimura, 1983). The predominance of purifying selection is supported by the observation that the number of synonymous nucleotide substitutions per synonymous site (d_S) substantially exceeds the number of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) in the vast majority of protein-coding genes (Kimura, 1977). The extent of purifying selection on intergenic spacers has been less frequently studied, but comparisons of nucleotide substitution in 5' intergenic spacers with linked protein-coding genes in pairs of completely sequenced genomes of 10 bacterial species showed lower levels of substitution in spacers than at synonymous sites in linked genes (Hughes and Friedman, 2004). These results support the hypothesis that intergenic spacers of bacteria are subject to strong purifying selection, as might be expected given the important functional role of *cis*-regulatory elements (Hughes and Friedman, 2004).

Strongly deleterious mutations are eliminated quickly by purifying selection, but slightly deleterious mutations may persist in populations for long times, since the effectiveness of purifying selection depends on effective population size and the frequency of recombination (Ohta, 1973, 1976; Ohta, 2002). In bacterial populations, there is evidence that nonsynonymous single nucleotide polymorphic variants are rare in comparison to synonymous variants in the same genes (Hughes, 2005). Since nonsynonymous mutations are more likely to be deleterious than are synonymous mutations, this observation supports the hypothesis that slightly deleterious variants, subject to ongoing purifying selection, are common in protein-coding genes of bacteria (Hughes, 2005). However, there is little information available regarding the occurrence of slightly deleterious variants in intergenic spacers in bacterial genomes.

We addressed this question in the case of GAS by examining the pattern of nucleotide polymorphism in 5' intergenic spacers and at synonymous and nonsynonymous sites in associated genes. A recent study identified riboswitches and small regulatory RNAs in one GAS genome (MGAS5005) using a combination of bioinformatic and tiling microarray approaches (Perez *et al.*, 2009). We identified the orthologous sequences in the other 12 GAS genomes, and compared the pattern of nucleotide polymorphism in spacers including sRNAs with those in the remainder of the intergenic spacers, in order to test the hypothesis that the former are subject to especially strong purifying selection in GAS. In addition, because unusually high levels of nucleotide diversity in specific regions of bacterial genomes can indicate past events of homologous recombination that have introduced divergent allelic sequences (Hughes and Friedman, 2005; Hughes and French, 2007; Hughes and Langley, 2007; Hughes, 2008), we compared nucleotide diversity in spacers with that in linked genes in order to test for evidence of past homologous recombination events involving spacers.

Materials and Methods

Sequence data

Genomic DNA sequences of 13 GAS isolates were downloaded from NCBI: M1 GAS (AE004092), MGAS10270 (CP000260), MGAS10394 (CP000003), MGAS10750 (CP000262), MGAS2096 (CP000261), MGAS315 (AE014074), MGAS5005 (CP000017), MGAS6180 (CP000056), MGAS8232 (AE009949), MGAS9429 (CP000259), NZ131 (CP000829), SSI-1 (BA000034), and Manfredo (AM295007); and all sequences were submitted to online RAST server for annotation (Aziz *et al.*, 2008).

Intergenic spacers shorter than 50 bp were not included because they are less likely to contain a complete promoter (Lewin, 2003) and detection of orthologous spacers becomes less reliable using BLAST (Altschul *et al.*, 1997). The complete set of spacers ≥ 50 bp from each genome was queried using a reciprocal all-versus-all BLASTN (Altschul *et al.*, 1997) with an E-cutoff value of 0.1. Likewise, each possible pair of proteomes was searched using a reciprocal all-versus-all BLASTP (Altschul *et al.*, 1997) with an E-cutoff value of 0.1. The reciprocal BLAST output files were formatted along with their genomic position and subsequently served as input files for MSOAR software (Fu *et al.*, 2007; Jiang, 2007) for ortholog prediction. MSOAR is a two-step procedure which identifies homologous genes based on sequence similarity followed by ortholog identification using genome context information (Fu *et al.*, 2007; Jiang, 2007). Here, MSOAR was applied to identify orthologous gene and spacer sequences separately. The pairwise orthologous genes and spacers were further assembled to obtain orthologous sets across the 13 GAS genomes (Luo *et al.*, 2009). The orthologous genes and spacers were aligned separately using PRANK software (Loytynoja and Goldman, 2005, 2008), which models insertions and deletions as distinct evolutionary events and thus reduces bias (Loytynoja and Goldman, 2005, 2008). Protein-coding sequences were aligned at the amino acid level and the alignment imposed on the DNA sequences.

The analyses reported below were based on a set of 590 pairs of protein-coding gene and 5' spacer, which met the following criteria: (1) at least one single nucleotide polymorphism (SNP) was present in the spacer and both synonymous and nonsynonymous SNPs (see below) were present in the coding region; and (2) the orthologous gene and spacer pair were found in at least 10 of the 13 genomes. In the case of genes with a head-to-head orientation, causing them to share spacers, only one gene (the larger in coding sequence length) was chosen for analysis. Of the 590 pairs, 140 were found in all 13 genomes, 88 in 12 genomes, 330 in 11 genomes, and 32 in 10 genomes. In a preliminary analysis, median levels of all measures of nucleotide sequence polymorphism (below) showed no significant differences among the four categories of gene and spacer pairs based on the number of genomes in which the pair was found (Kruskal-Wallis tests; not shown). Thus only results from the pooled data (Supplementary Table S1) are reported below. Using the list of sRNAs identified by Perez and colleagues (Perez *et al.*, 2009), we compared polymorphism in the spacers including sRNAs with the remainder of spacers.

Analysis of polymorphism

The number of synonymous nucleotide substitutions per synonymous site (d_S) and the number of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) were estimated for all pairwise comparisons in each orthologous gene set using Nei and Gojobori method (Nei and Gojobori, 1986) in the PAML software package (Yang, 1997). Preliminary analyses showed that the Yang and Nielsen method (Yang and Nielsen, 2000) making more assumptions yielded essentially identical results, as has been observed in many other studies when the number of substitutions per site is small (Hughes and French, 2007). The number

of nucleotide substitutions per site (d) in intergenic regions was estimated using Jukes-Cantor model (Jukes and Cantor, 1969). Within each orthologous gene set, we computed the mean of d_S in all pairwise comparisons (i.e., the synonymous nucleotide diversity, symbolized π_S) and the mean of d_N in all pairwise comparisons (i.e., the nonsynonymous nucleotide diversity, symbolized π_N) (Nei and Kumar, 2000). Likewise, the mean of d in all pairwise comparisons provided an estimate of the nucleotide diversity in intergenic regions (symbolized π). In the estimation of d_S and d_N, we excluded codons having ambiguous sites at which both synonymous and nonsynonymous variants occurred or at which the polymorphism could be considered either synonymous or nonsynonymous depending on the pathway taken by evolution. The standard errors of π_S, π_N, and π were estimated by the bootstrap method (Nei and Kumar, 2000); 1000 bootstrap pseudo-samples were used.

SNPs in coding regions were classified either as synonymous or as nonsynonymous depending on the coding effect of nucleotide change; we excluded ambiguous sites at which both synonymous and nonsynonymous changes occurred or at which either synonymous or nonsynonymous mutations could account for the polymorphism considered depending on the evolutionary pathway. There were 873 such ambiguous polymorphic sites out of 20,105 total polymorphic sites in the 590 protein-coding genes (4.3%).

In order to examine the relative frequency of rare alleles at synonymous and nonsynonymous sites, we compared the average number of nucleotide differences and the number of segregating sites (Tajima, 1989) separately for synonymous and nonsynonymous sites (Hughes, 2005; Hughes and Hughes, 2007b; Rand and Kann, 1996). For each orthologous gene, we computed the difference $K_S - S_S^*$. K_S is the mean number of synonymous nucleotide difference for all pairwise comparisons among the n allelic sequences in the alignment; n varies depending on the number of intergenic sequences that were matched to coding sequences. If S_S is the number of synonymous segregating sites, then

$$S_S^* = S_S / a_1 \tag{1}$$

The divisor a_1 in equation (2) is a factor adjusting the number of sample size (n) and is given by the following (Tajima, 1989):

$$a_1 = \sum_{i=1}^{n-1} 1/i \tag{2}$$

Likewise, we also computed the difference $K_N - S_N^*$, where K_N is the mean number of nonsynonymous nucleotide difference for all pairwise comparisons in an alignment; S_N^* is the adjusted number of nonsynonymous segregating sites. In intergenic regions, we computed the difference $K_{spacer} - S_{spacer}^*$, where K_{spacer} is the mean number of nucleotide difference; S_{spacer}^* is the adjusted number of segregating sites.

The differences $K_S - S_S^*$, $K_N - S_N^*$, and $K_{spacer} - S_{spacer}^*$ constitute the numerator of Tajima's (Tajima, 1989) D statistic computed separately for synonymous, nonsynonymous, and intergenic polymorphisms, respectively. We then computed the ratio of this difference to the absolute value of the minimum possible value of the difference, which would occur if all polymorphisms were singletons (Schaeffer, 2002). We designate this ratio Q_S in the case of synonymous polymorphisms, Q_N in the case of nonsynonymous polymorphisms, and Q_{spacer} in the case of polymorphisms of intergenic sequences. Comparing Q_S, Q_N, and Q_{spacer}

provides an index of the relative abundance of rare alleles at synonymous, nonsynonymous, and intergenic sites, with a strongly negative value suggesting an excess of rare variants (Hughes and Hughes, 2007a; Hughes and Hughes, 2007b; Hughes *et al.*, 2008). We used these statistics instead of Tajima's D (Tajima, 1989) because the latter is dependent of sample size (n) and thus not directly comparable among data sets (Hughes and Hughes, 2007b; Hughes *et al.*, 2008). If ongoing purifying selection is occurring at numerous nonsynonymous SNP sites, we expect that Q_N will be lower (more strongly negative) than Q_S, indicating an abundance of rare nonsynonymous variants (Hughes and Hughes, 2007b; Hughes *et al.*, 2008). Likewise, if ongoing purifying selection is occurring at numerous SNP sites in spacers, we expect that and Q_{spacer} will be lower (more strongly negative) than Q_S, indicating an abundance of rare variants in intergenic regions.

Because nucleotide diversity measures and Q_S, Q_N, and Q_{spacer} were not normally distributed across the 590 pairs of genes and linked 5' spacers, we used nonparametric methods in hypothesis testing. All P values reported are two-tailed and corrected for multiple testing by the conservative Bonferroni procedure.

Phylogenetic trees of genes and spacers were constructed by the neighbor-joining method (Saitou and Nei, 1987) on the basis of the maximum composite likelihood (MCL) distance (Tamura *et al.*, 2007) and by the quartet maximum likelihood (QML) method using the HKY distance, as implemented in TREEPUZZLE 5.2 (Schmidt et al., 2007). The reliability of clustering patterns in NJ trees was assessed by bootstrapping (Felsenstein, 1985); 1000 bootstrap pseudo-samples were used. The Kishino-Hasegawa test in TREEPUZZLE 5.2 was used to test the identity of topologies of trees based on spacers and coding regions.

Results

Characteristics of Spacers

There were a total of 20434 spacers in the 13 GAS genomic sequences, with a mean length of 134.4 bp and a median length of 107 bp (Fig. 1). 67.4% of the spacers were greater than 50 bp in length, while 95.4% were less than 400 bp in length. Thus very short and very long spacers were rare. When spacers less than 50 bp in length were excluded, the mean length of the remaining spacers was 190.6 bp, and the median length was 151 bp. In the 590 spacers used in our analyses, the mean length was 180.4 bp, and the median length was 150 bp. Thus our sample of spacers was very representative of the overall length distribution of spacers greater than 50 bp in length.

Nucleotide Diversity

Using pairwise comparisons, we compared synonymous nucleotide diversity (π_S) in coding regions, nonsynonymous nucleotide diversity (π_N) in coding regions, and nucleotide diversity (π) in linked 5- spacers (Fig. 2). In coding regions, median π_N 0.0025) was significantly lower than median π_S (0.0195; $P < 0.001$; Sign test; Fig. 2A). Similarly, median π in 5' spacers (0.0084) was significantly lower than median π_S ($P < 0.001$; Sign test; Fig. 2). These results support the hypothesis that purifying selection has acted to eliminate deleterious mutations at nonsynonymous sites in genes and at certain sites within spacers. However, there was evidence that purifying selection acting on nonsynonymous sites in gene has been stronger overall than that acting on spacers, since median π_N in genes was significantly lower than median π in linked 5' spacers ($P < 0.001$; Sign test; Fig. 2). Median π was slightly lower in spacers including sRNAs (0.0066) than in other spacers (0.0085); but the difference was not significant (Mann-Whitney test, N.S.).

Rare Polymorphisms

In addition to examining past purifying selection that has eliminated deleterious mutations, comparison of allelic sequences can provide evidence of ongoing purifying selection acting to reduce the frequency of slightly deleterious variants in a population (Hughes *et al.*, 2003). We tested for ongoing purifying selection by comparing median Q_S, Q_N, and , Q_{spacer} for the 590 pairs of ortholgous genes and 5' spacers. Median Q_S (0.0631) was slightly positive, but not significantly difference from zero (Sign test; N.S.; Fig. 3). Median Q_N was negative (−0.5491) and significantly different from zero (P < 0.001; Sign test; Fig. 3). Median Q_{spacer} was also negative (−0.2610) and significantly different from zero (< 0.001; Sign test; Fig. 3).

In pairwise comparisons, median Q_N was significantly lower than median Q_S (P < 0.001; Sign test; Fig. 3). Likewise, median Q_{spacer} was significantly lower than median Q_S (P < 0.001; Sign test; Fig. 3). However, median Q_{spacer} was significantly greater than median Q_N (P < 0.001; Sign test; Fig. 3). These results indicate ongoing purifying selection at both synonymous sites in coding regions and at sites in 5' spacers, leading to an excess of rare variants in both cases. However, because median Q_N was significantly more negative than median Q_{spacer} (Fig. 3), the results supported the hypothesis that this selection is on average stronger on nonsynonymous sites than on sites in spacers. Median Q_{spacer} was slightly lower in spacers including sRNAs (−0.3132) than in other spacers (−0.2369); but the difference was not significant (Mann-Whitney test, N.S.).

Differences in Nucleotide Diversity

There were only 4 genes in our sample for which π_N was greater than π_S; and in none of these cases was there a significant difference between π_N and π_S by the Z-test. These four genes were in unusually short genes, with a median length of 183 bp (range 126–363 bp). The median length of these genes was significantly lower than that of the other 586 genes (871.5 bp, range 117–4383 bp). Thus, the evidence suggested that the pattern of $\pi_N > \pi_S$ in these genes was probably due to stochastic error because of the small number of sites, rather than to some form of positive selection.

There were 89 cases in our sample in which π in the spacers was greater than π_S in the linked gene. In three of these cases the difference was significant by the Z-test (in each case P < 0.001 level): (1) In the case of the *RuvB* gene, encoding a Holiday junction DNA helicase (NP_268453) that promotes strand exchange during homologous recombination (Sharples *et al.*, 1999), π in the 5' spacer was 0.222 ± 0.027, while π_S was 0.011 ±. 0.004. (2) In the case of the gene encoding the D-alanyl-D-alanine-carboxypeptidase or penicillin-binding protein (NP_269251), which functions in peptidoglycan biosynthesis (Yocum *et al.*, 1980), π in the 5' spacer was 0.244 ± 0.037, while π_S was 0.089 ±. 0.011. In the case of a gene encoding a putative cytoplasmic protein (NP_665543), π in the 5' spacer was 0.156 ± 0.020, while π_S was 0.020 ±. 0.011.

Phylogenetic trees of spacers and coding regions in these three cases revealed a similar pattern (Fig. 4). In each case, the spacer sequences formed two very distinct groups, separated by a strongly supported internal branch (100% bootstrap support in each case; Fig. 4A–C). In each case, branches within the two clusters were very short compared to the long branch between the two clusters (Fig. 4A–C). In each case, a very different pattern was seen in phylogenetic trees of the coding regions (Fig. 4D–F). There were no strongly separately clusters, and the topologies of the trees based on coding regions (Fig. 4D–F) did not match that of the trees based on the spacers (Fig. 4A–C). The QML trees (not shown) had topologies similar to the NJ trees. The topologies of the trees for spacers and coding regions were significantly different by the Kishino-Hasgawa test for the three cases illustrated in Figure 4 (P < 0.001 in each case). Note also that the widely separated clusters formed by the

spacer sequences did not include the same genomes in each case. These results thus provide evidence that numerous past events of homologous recombination have served to introduce alternative spacer sequences into different GAS genomes.

Discussion

Analysis of 590 paired data sets consisting of orthologous 5' spacers and linked orthologous protein-coding genes of GAS showed that nucleotide diversity both at nonsynonymous sites in genes and at sites within spacers is reduced in comparison to synonymous sites in genes. These results provide evidence that purifying selection has acted to eliminate deleterious mutations both at nonsynonymous sites and at sites within spacers. However, median nucleotide diversity was significantly higher in spacers than at nonsynonymous sites in the linked genes, indicating that purifying selection has been on average stronger at the latter. Spacers including known sRNAs showed slightly reduced median nucleotide diversity than other spacers, but the difference was not statistically significant, indicating that the presence of sRNAs is not the only factor responsible for purifying selection on spacers. Rather, it is likely that *cis*-regulatory elements are also subject to purifying selection (Hughes *et al.*, 2005). Consistent with this hypothesis, there is evidence that a minimal change in *cis*-regulatory regions can lead to a substantial alteration of patterns of gene expression (Wittkopp, 2006) For example, a single base pair deletion or replacement in *cis*-regulatory elements leads to reduction of expression by 80–90% at the act promoter in *Myxococcus xanthus* (Gronewold and Kaiser, 2007).

Our analyses provided evidence of alternative, highly divergent forms of the 5' spacer in the genes encoding the RuvB DNA helicase, D-alanyl-D-alanine-carboxypeptidase, and a putative cytoplasmic protein. One explanation for these observations would be recombination events that resulted in the presence in the GAS population of two highly divergent forms of the 5' spacer. It is possible that their might also be some form of balancing selection on the spacer, maintaining the two alternative forms. However, the fact that the phylogeny of the spacers did not match that of the genes themselves (Fig. 4) indicates the importance of recombination events in explaining the observed pattern. Whether selectively maintained or not, such divergent forms of a given spacer may be functionally distinct, thus resulting in differences among GAS genomes with respect to expression patterns of these genes. The possibility of differences among GAS strains with respect to the expression of D-alanyl-D-alanine-carboxypeptidase may be of particular interest, since in other bacterial species that enzyme is known to play a role in susceptibility to certain antimicrobials (Yocum *et al.*, 1980; Yocum *et al.*, 1982).

In addition to evidence that purifying selection has acted in the past, our analyses provided evidence of an excess of rare variants both at nonsynonymous sites in genes and at sites in spacers. This implies that in GAS populations there are numerous deleterious variants at nonsynonymous sites and at sites within spacers which are subject to ongoing purifying selection that has decreased their population frequency but has not yet succeeded in eliminating them. Since homologous recombination is the major way to purge deleterious alleles in a population, the presence of numerous slightly deleterious variants in GAS, as in other bacteria, is evidence that homologous recombination has been limited (Hughes, 2008). Moreover, our results suggest that a substantial fraction of the heritable phenotypic variation in GAS – both at the level of protein sequence and at the level of gene expression – is likely to be slightly deleterious to the bacterium. Understanding the nature and phenotypic effects of such deleterious variation in important human pathogens such as GAS may suggest novel strategies for prophylaxis and treatment, because it may open previously unexplored windows on the pathogen's vulnerability; for example, by targeting pathways in which slightly deleterious variants are particularly abundant.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Richard Vogt for helpful discussion. This research was supported by grant GM43940 from the National Institute of Health to A.L.H and grant GM078991 from the National Institute of Health to J.T. Acknowledgment is also made to the University of South Carolina's High Performance Computing Group for the computing time on a 128-core shared memory computer used in this research.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]

Aziz R, Bartels D, Best A, DeJongh M, Disz T, Edwards R, Formsma K, Gerdes S, Glass E, Kubal M, Meyer F, Olsen G, Olson R, Osterman A, Overbeek R, McNeil L, Paarmann D, Paczian T, Parrello B, Pusch G, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. The RAST server: rapid annotations using subsystems technology. BMC Genomics. 2008; 9:75. [PubMed: 18261238]

Felsenstein J. Confidence limits on phylogenies: An approach using the bootstrap. Evolution. 1985; 39:783–791.

Fu Z, Chen X, Vacic V, Nan P, Zhong Y, Jiang T. MSOAR: A high-throughput ortholog assignment system based on genome rearrangement. J. Comput. Biol. 2007; 14:1160–1175. [PubMed: 17990975]

Gronewold TMA, Kaiser D. Mutations of the Act Promoter in Myxococcus xanthus. J. Bacteriol. 2007; 189:1836–1844. [PubMed: 17189369]

Horler RSP, Vanderpool CK. Homologs of the small RNA SgrS are broadly distributed in enteric bacteria but have diverged in size and sequence. Nucleic Acids Res. 2009; 37:5465–5476. [PubMed: 19531735]

Hughes A, Packer B, Welch R, Chanock S, Yeager M. High level of functional polymorphism indicates a unique role of natural selection at human immune system loci. Immunogenetics. 2005; 57:821–827. [PubMed: 16261383]

Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M. Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc. Natl. Acad. Sci. U. S. A. 2003; 100:15754–15757. [PubMed: 14660790]

Hughes AL, Friedman R. Patterns of Sequence Divergence in 5' Intergenic Spacers and Linked Coding Regions in 10 Species of Pathogenic Bacteria Reveal Distinct Recombinational Histories. Genetics. 2004; 168:1795–1803. [PubMed: 15611157]

Hughes AL. Evidence for Abundant Slightly Deleterious Polymorphisms in Bacterial Populations. Genetics. 2005; 169:533–538. [PubMed: 15545641]

Hughes AL, Friedman R. Nucleotide substitution and recombination at orthologous loci in Staphylococcus aureus. J. Bacteriol. 2005; 187:2698–2704. [PubMed: 15805516]

Hughes AL. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity. 2007; 99:364–373. [PubMed: 17622265]

Hughes AL, French JO. Homologous recombination and the pattern of nucleotide substitution in Ehrlichia ruminantium. Gene. 2007; 387:31–37. [PubMed: 17005333]

Hughes AL, Hughes MAK. More effective purifying selection on RNA viruses than in DNA viruses. Gene. 2007a; 404:117–125. [PubMed: 17928171]

Hughes AL, Hughes MAK. Coding sequence polymorphism in avian mitochondrial genomes reflects population histories. Mol. Ecol. 2007b; 17:1369–1376.

Hughes AL, Langley KJ. Nucleotide usage, synonymous substitution pattern, and past recombination in genomes of Streptococcus pyogenes. Infect. Genet. Evol. 2007; 7:188–196. [PubMed: 17000138]

Hughes AL. Near Neutrality: Leading Edge of the Neutral Theory of Molecular Evolution. Ann. NY. Acad. Sci. 2008; 1133:162–179. [PubMed: 18559820]

Hughes AL, Friedman R, Rivailler P, French JO. Synonymous and Nonsynonymous Polymorphisms versus Divergences in Bacterial Genomes. Mol. Biol. Evol. 2008; 25:2199–2209. [PubMed: 18667439]

Jiang, T. Proceedings of the 18th annual symposium on Combinatorial Pattern Matching. London, Canada: Springer-Verlag; 2007. A combinatorial approach to genome-wide ortholog assignment: beyond sequence similarity search.

Jukes, T.; Cantor, C. Evolution of protein molecules. In: Munro, H., editor. Mammalian Protein Metabolism. New York: Academic Press; 1969. p. 21-132.

Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature. 1977; 267:275–276. [PubMed: 865622]

Kimura, M. The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press; 1983.

Lewin, B. Genes VIII. Benjamin Cummings; 2003.

Loytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. Proc. Natl. Acad. Sci. U.SA. 2005; 102:10557–10562.

Loytynoja A, Goldman N. Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis. Science. 2008; 320:1632–1635. [PubMed: 18566285]

Luo H, Sun Z, Arndt W, Shi J, Friedman R, Tang J. Gene order phylogeny and the evolution of methanogens. PLoS ONE. 2009; 4:e6069. [PubMed: 19562076]

Madan Babu M, Teichmann SA. Evolution of transcription factors and the gene regulatory network in Escherichia coli. Nucleic Acids Res. 2003; 31:1234–1244. [PubMed: 12582243]

Moses A, Chiang D, Kellis M, Lander E, Eisen M. Position specific variation in the rate of evolution in transcription factor binding sites. BMC Evol. Biol. 2003; 3:19. [PubMed: 12946282]

Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 1986; 3:418–426. [PubMed: 3444411]

Nei, M.; Kumar, S. Molecular Evolution and Phylogenetics. Oxford Univerity Press; 2000.

Ohta T. Slightly deleterious mutant substitutions in evolution. Nature. 1973; 246:96–98. [PubMed: 4585855]

Ohta T. Role of very slightly deleterious mutations in molecular evolution and polymorphism. Theor Popul Biol. 1976; 10:254–275. [PubMed: 1013905]

Ohta T. Near-neutrality in evolution of genes and gene regulation. Proc. Natl. Acad. Sci. U.S.A. 2002; 99:16134–16137. [PubMed: 12461171]

Perez N, Trevino J, Liu Z, Ho SCM, Babitzke P, Sumby P. A genome-wide analysis of small regulatory RNAs in the human pathogen group A streptococcus. PLoS ONE. 2009; 4:e7668. [PubMed: 19888332]

Rand DM, Kann LM. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. Mol. Biol. Evol. 1996; 13:735–748. [PubMed: 8754210]

Ribardo DA, McIver KS. Defining the Mga regulon: comparative transcriptome analysis reveals both direct and indirect regulation by Mga in the group A streptococcus. Mol. Microbiol. 2006; 62:491–508. [PubMed: 16965517]

Rodriguez-Trelles F, Tarrio R, Ayala FJ. Evolution of cis-regulatory elements versus codifying regions. Int. J. Dev. Biol. 2003; 47:665–673. [PubMed: 14756342]

Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, Wolf YI, Yin J, Koonin EV. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. Nucleic Acids Res. 2002; 30:4264–4271. [PubMed: 12364605]

Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 1987; 4:406–425. [PubMed: 3447015]

Schaeffer SW. Molecular population genetics of sequence length diversity in the Adh region of Drosophila pseudoobscura. Genet. Res. 2002; 80:163–175. [PubMed: 12688655]

Schmidt, HA.; von Haeseler, A. Maximum-Likelihood Analysis Using TREE-PUZZLE. In: Baxevanis, AD.; Davison, DB.; Page, RDM.; Stormo, G.; Stein, L., editors. Current Protocols in Bioinformatics (Supplement 17), Unit 6.6. New York: Wiley and Sons; 2007.

Sharples GJ, Ingleston SM, Lloyd RG. Holliday junction processing in bacteria: insights from the evolutionary conservation of RuvABC, RecG, and RusA. J. Bacteriol. 1999; 181:5543–5550. [PubMed: 10482492]

Sridhar J, Rafi ZA. Functional annotations in bacterial genomes based on small RNA signatures. Bioinformation. 2008; 2:284–295. [PubMed: 18478081]

Tajima F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics. 1989; 123:585–595. [PubMed: 2513255]

Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. Mol. Biol. Evol. 2007; 24:1596–1599. [PubMed: 17488738]

Wittkopp PJ. Evolution of cis-regulatory sequence and function in Diptera. Heredity. 2006; 97:139–147. [PubMed: 16850038]

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. 2003; 20:1377–1419. [PubMed: 12777501]

Yang Z. PAML: A program package for phylogenetic analysis by maximum likelihood. Bioinformatics. 1997; 13:555–556.

Yang Z, Nielsen R. Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. Mol. Biol. Evol. 2000; 17:32–43. [PubMed: 10666704]

Yocum RR, Rasmussen JR, Strominger JL. The mechanism of action of penicillin. Penicillin acylates the active site of Bacillus stearothermophilus D-alanine carboxypeptidase. J. Biol. Chem. 1980; 255:3977–3986. [PubMed: 7372662]

Yocum RR, Amanuma H, O'Brien TA, Waxman DJ, Strominger JL. Penicillin is an active-site inhibitor for four genera of bacteria. J. Bacteriol. 1982; 149:1150–1153. [PubMed: 7061385]

Research Highlights

- We examine sequence variation in 590 genes and their linked 5' intergenic spacers in group A Streptococcus (GAS).
- Spacers show evidence of past purifying selection, as do nonsynonymous sites in coding regions.
- There is an excess of rare variants both at nonsynonymous sites in genes and at sites in spacers, which is evidence that there are numerous slightly deleterious variants in GAS populations with potential effects on both protein sequences and gene expression.

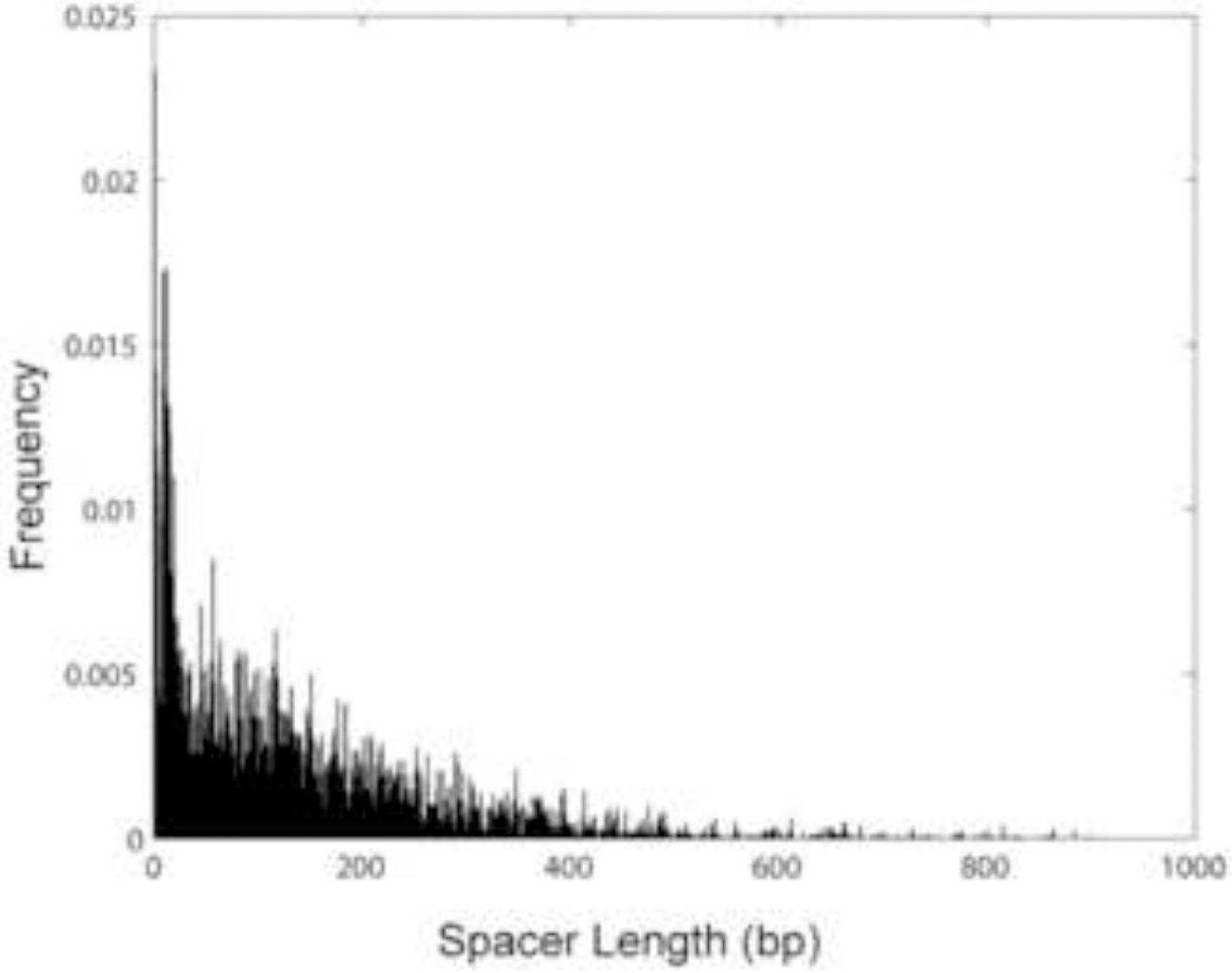

Figure 1.
Distribution of the lengths of intergenic spacers in 13 genomes of the group A streptococcus.

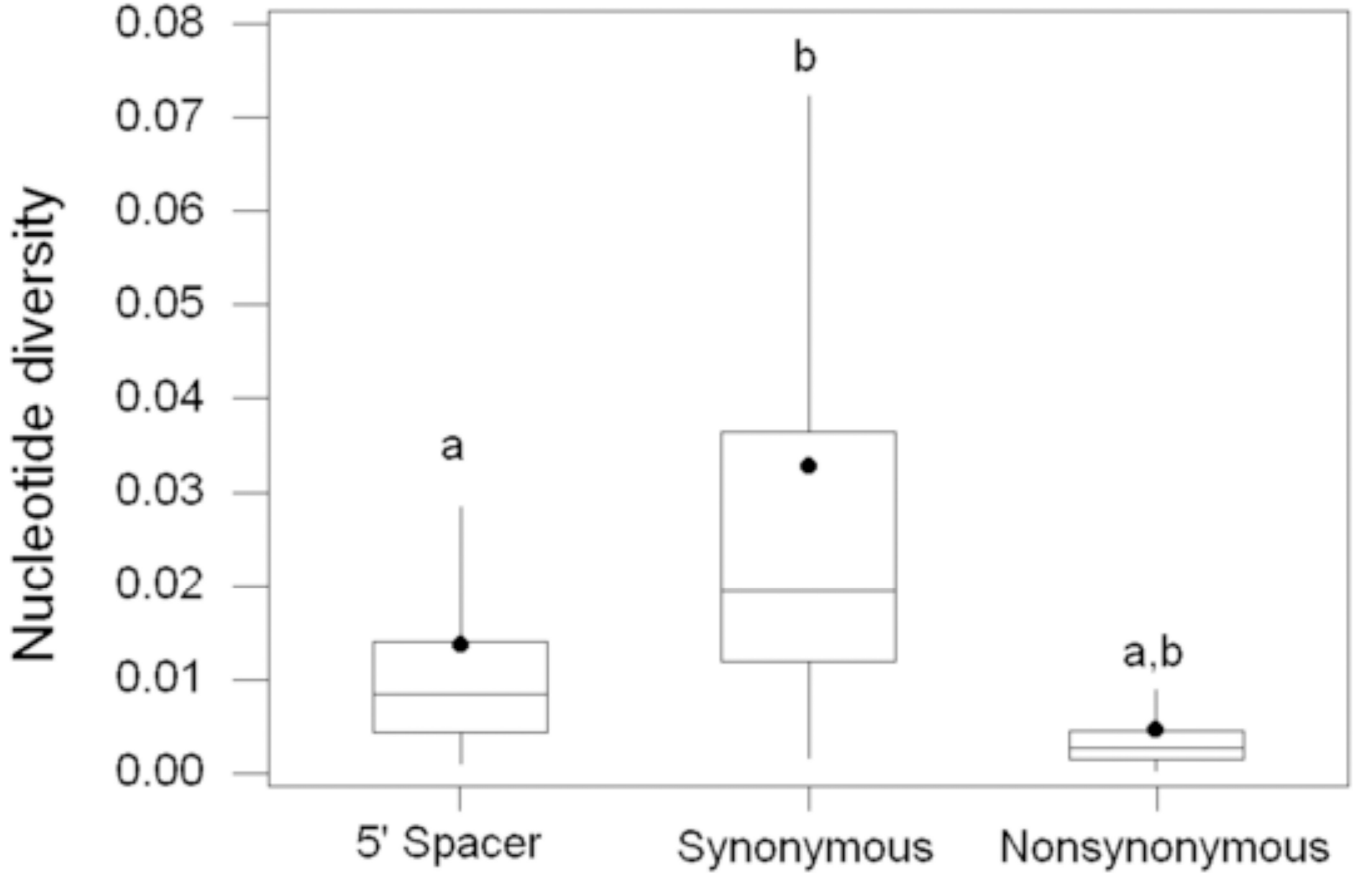

Figure 2.

Box-and-whiskers plot showing mean (*solid circle*) and median (*horizontal line*) nucleotide diversity in 5' spacers (π) and at synonymous (π_S) and and nonsynonymous (π_N) sites in linked genes. Sign tests of the hypothesis that median nucleotide diversity equals that at synonymous sites: [a] $P < 0.001$. Sign tests of the hypothesis that median nucleotide diversity equals that in spacers: [b] $P < 0.001$. The *box* indicates the first quartile (Q1) and third quartile (Q3), and the *whiskers* the range from Q1 − 1.5(Q3−Q1) to Q3 + 1.5(Q3−Q1).

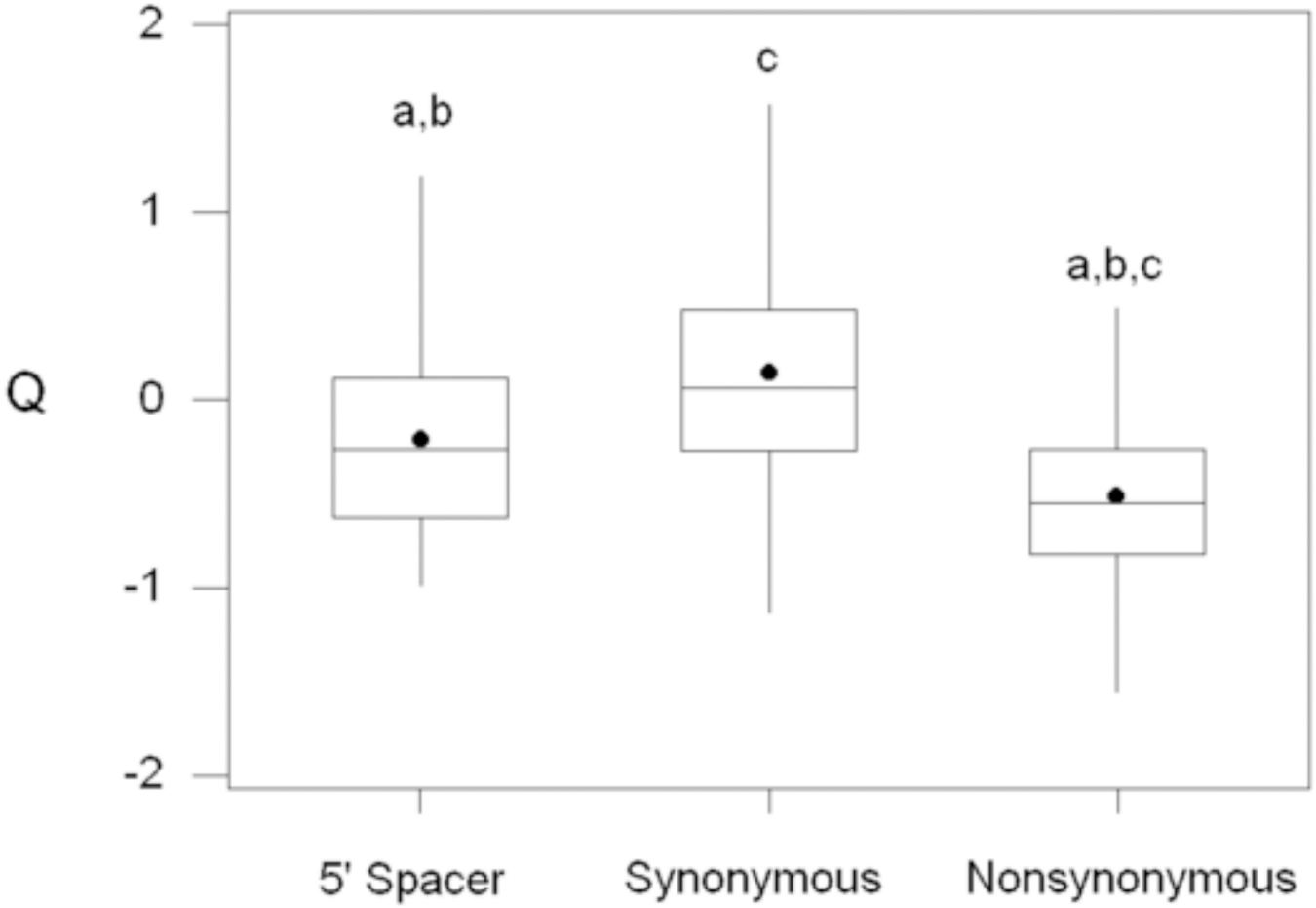

Figure 3.

Box-and-whiskers plot showing mean (*solid circle*) and median (*horizontal line*) values of Q, a measure of the relative abundance of rare polymorphisms, in 5' spacers(Q_{spacer})and at synonymous (Q_S) and and nonsynonymous (Q_N) sites in linked genes. Sign tests of the hypothesis that median Q equals zero: [a] $P < 0.001$. Sign tests of the hypothesis that median Q equals that at synonymous sites: [b] $P < 0.001$. Sign tests of the hypothesis that median Q equals that in spacers: [c] $P < 0.001$. The *box* indicates the first quartile (Q1) and third quartile (Q3), and the *whiskers* the range from Q1 − 1.5(Q3−Q1) to Q3 + 1.5(Q3−Q1).

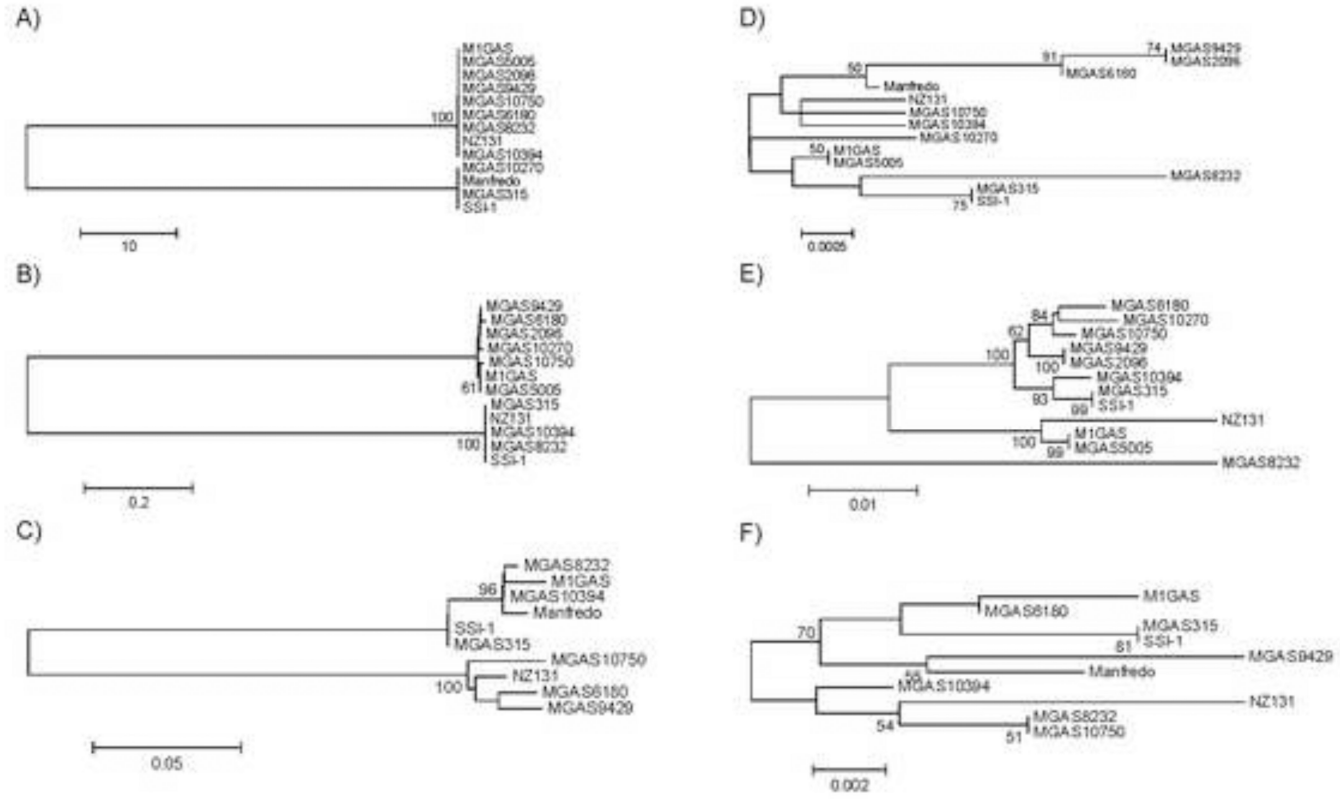

Figure 4.

Phylogenetic trees of 5’ spacers for the genes encoding (A) RuvB DNA helicase, (B) D-alanyl-D-alanine-carboxypeptidase, and (C) a putative cytoplasmic protein; and of the coding regions of the corresponding genes: (D) RuvB DNA helicase, (E) D-alanyl-D-alanine-carboxypeptidase, and (F) a putative cytoplasmic protein. All trees were constructed by the NJ method based on the MCL distance. Numbers on the branches represent the percentage of 1000 bootstrap samples supporting the branch; only values ≥ 50% are shown.