JOURNAL OF BACTERIOLOGY, June 2005, p. 4163–4172 Vol. 187, No. 12
0021-9193/05/$08.00+0 doi:10.1128/JB.187.12.4163–4172.2005

Evolution of Transcription Regulatory Genes Is Linked to Niche Specialization in the Bacterial Pathogen *Streptococcus pyogenes*†

Debra E. Bessen,* Anand Manoharan, Feng Luo, John E. Wertz, and D. Ashley Robinson

Department of Microbiology and Immunology, New York Medical College, Valhalla, New York

Received 24 January 2005/Accepted 3 March 2005

Streptococcus pyogenes **is a highly prevalent bacterial pathogen, most often giving rise to superficial infections at the throat or skin of its human host. Three genotype-defined subpopulations of strains exhibiting strong tropisms for either the throat or skin (specialists) or having no obvious tissue site preference (generalists) are recognized. Since the microenvironments at the throat and skin are distinct, the signal transduction pathways leading to the control of gene expression may also differ for throat versus skin strains of *S. pyogenes*. Two loci (*mga* and *rofA*/*nra*) encoding global regulators of virulence gene expression are positioned 300 kb apart on the genome; each contains alleles forming two major sequence clusters of ~25 to 30% divergence that are under balancing selection. Strong linkage disequilibrium is observed between sequence clusters of the transcription regulatory loci and the subpopulations of throat and skin specialists, against a background of high recombination rates among housekeeping genes. A taxonomically distinct commensal species (*Streptococcus dysgalactiae* subspecies *equisimilus*) shares highly homologous *rof* alleles. The findings provide strong support for a mechanism underlying niche specialization that involves orthologous replacement of regulatory genes following interspecies horizontal transfer, although the directionality of gene exchange remains unknown.**

Many microbial pathogens exhibit strong tropisms for a narrow range of hosts or for specific tissues within a host. This type of niche specialization is an early, key step in the formation of new species (12, 46). An understanding of the molecular basis for host or tissue tropisms may provide insights on the steps leading to the emergence of genetically discrete populations of microorganisms.

Streptococcus pyogenes is a human pathogen having high prevalence throughout the world. Although *S. pyogenes* can cause severe invasive disease, it most often causes a mild infection at superficial, epithelial tissue sites. Infection of the throat and skin leads to pharyngitis ("strep" throat) and impetigo, respectively. Importantly, these two tissues constitute the primary habitat for *S. pyogenes*, where it is most successful in reproductive growth and transmission of progeny to new hosts. Decades of field epidemiology based on the M- and *emm*-typing schemes have led to the recognition of distinct throat and skin strains (2, 9, 10, 13, 15, 16, 30, 38, 39, 49, 65). The prevalence of pharyngitis versus impetigo caused by *S. pyogenes* varies widely throughout the world, largely in accordance with climatic conditions. Thus, the spatial and temporal (seasonal) distances between throat strains and skin strains are amplified even further by distinct epidemiological trends.

A genetic marker displaying statistically significant nonrandom associations with tissue site of isolation has been defined for *S. pyogenes*. The genetic marker for tissue site preference is designated *emm* pattern, and it is based on the chromosomal arrangement of *emm* genes, which, in turn, encode a diverse family of surface protein fibrils (27). The *emm* pattern A-C and D strains are regarded as niche specialists, having strong preference for just one tissue site (throat and skin, respectively), whereas pattern E strains are readily recovered from both tissues and are considered generalists. For example, nearly all pharyngitis isolates (>99%) collected from hospitals in Rome (Italy) are of *emm* types typically found in *emm* patterns A-C or E, whereas <1% are of *emm* types associated with pattern D (14). Similarly, in a recent report on >1,900 *S. pyogenes* isolates collected from cases of pharyngitis in the United States over 2 years (56), 53 and 47% are of *emm* types typical of pattern A-C and E strains, respectively, with <1% represented by pattern D *emm* types (41). In a rural aboriginal community in tropical Australia where impetigo is hyperendemic, no cases of pharyngitis were detected during a 25-month surveillance period; of the impetigo isolates recovered, 46 and 41% are either *emm* pattern D or E, respectively (6).

In contrast to the strong linkage observed between *emm* pattern and tissue site of isolation, the distribution of neutral housekeeping alleles among strains of *S. pyogenes* is highly random. In fact, *S. pyogenes* ranks among the most highly recombinogenic of several bacterial species examined (19), whereby genetic recombination is the consequence of a horizontal gene transfer (HGT) event. Numerous statistical tests point to an ample flow of housekeeping genes between the three *emm* pattern-defined subpopulations and between isolates known to be recovered from the throat versus skin (33). Thus, against a background of random associations between housekeeping genes, genotypes exhibiting strong linkage disequilibrium with the *emm* pattern-defined subpopulations are good candidates for having a key role in tissue-specific adaptations.

Since the microenvironments at the throat and skin are distinct in many ways, the signal transduction pathways leading to the control of gene expression may also differ for throat versus skin strains of *S. pyogenes*. In this report, the phylogeny and distribution of alleles at two genetically diverse loci (*mga*

* Corresponding author. Mailing address: Department of Microbiology and Immunology, New York Medical College, Valhalla, NY 10595. Phone: (914) 594-4193. Fax: (914) 594-4176. E-mail: debra_bessen@nymc.edu.

† Supplemental material for this article may be found at http://jb.asm.org/.

and *rofA*/*nra*), encoding global regulators of transcription, are evaluated with respect to *emm* pattern subpopulations. The gene products of both *mga* and *rofA*/*nra* regulate the expression of numerous genes encoding virulence factors of *S. pyogenes* that interface directly with the human host (36).

MATERIALS AND METHODS

Bacterial strains. The 114 *S. pyogenes* isolates under study are listed in Table S1 in the supplemental material. Thirty-three group C and G streptococci isolated from humans were previously described (32). A human isolate of *Streptococcus equi* subspecies *zooepidemicus* (5371) was recovered in association with an outbreak of acute glomerulonephritis in Brazil (45). Additional non-*S. pyogenes* strains were kindly provided by R. Facklam (Centers for Disease Control and Prevention [CDC], Atlanta, GA). Group carbohydrate was determined for groups A, C, and G using a latex agglutination test (Murex Biotech Ltd., United Kingdom). Group L streptococci underwent group carbohydrate analysis at the CDC.

Genotyping. The *emm* type was ascertained according to previously described methods (4) and is based on the 5′ end of the central *emm* gene within the *emm* chromosomal region; a complete listing of *emm* types found in association with *S. pyogenes* is available (www.cdc.gov/ncidod/biotech/strep/strains.html). *emm* pattern was determined by methods previously described (41), using a PCR-based mapping approach that utilizes oligonucleotide primers specific for each of the four major lineages that arise from phylogenetic trees based on the 3′ ends of *emm* genes (26, 27). Chromosomal DNA used as a template for PCR was prepared from freshly grown bacteria according to previously described methods for bacterial cell lysis (7).

Multilocus sequence typing was performed as previously reported (17). In brief, internal fragments of the glucose kinase (*gki*), glutamine transporter protein (*gtr*), glutamate racemase (*murI*), DNA mismatch repair protein (*mutS*), transketolase (*recP*), xanthine phosphoribosyl transferase (*xpt*), and acetyl-coenzyme A acetyltransferase (*yqiL*) genes were amplified by PCR and subjected to nucleotide sequence determination. The relative positions of the seven housekeeping loci on the *S. pyogenes* genome (20) are depicted in Fig. S1 in the supplemental material. For each locus, every different sequence is assigned a distinct allele number, and each isolate is defined by a series of seven integers (the allelic profile) corresponding to the alleles at the seven loci in the following order (alphabetical): *gki-gtr-murI-mutS-recP-xpt-yqiL*. Isolates with the identical allelic profile are assigned to the same sequence type (ST).

PCR-based screening. Using bacterial DNA as the template, PCR amplification was performed with an initial denaturation at 95°C for 4 min followed by 29 cycles at 95°C, 55°C, and 72°C for 1 min each. PCR amplification products corresponding to internal portions of the designated genes were generated with the following oligonucleotide primer pairs: for *rofA*, 5′-CTA RCC TAA AAG AGC AAA AGG CTA GTT TAG-3′ (forward) and 5′-CTT GGA TAG ACA GAA TCG ATT C-3′ (reverse) (amplicon size, 521 bp); for *nra*, 5′-GCA ATT AAA CCA TTC TAA ACA AGA CCT TA-3′ (forward) and 5′-TGA ATT GAA GCA ATA GAG TAG TCA GGS TTA-3′ (reverse) (amplicon size, 482 bp); for *mga-1*, 5′-CAA CGG GCT GTC GAA AAG TGA CCA ACT GGG TTC ATC TYC TTA-3′ (forward) and 5′-GCG ATG AAA GTC CAA GGG GTT CTT GAT GGG-3′ (reverse) (amplicon size, 350 bp); for *mga-2*, 5′-CAT CAG GAG GCA GAC AAG TAA CCA ACT GGA TCC ATC TAT TAG-3′ (forward) and 5′-GTC ACT ATG AGA TTT TGA AGA GGA AGG GGC TTC GAG GTT-3′ (reverse) (amplicon size, 650 bp); and for *mgc*, 5′-CTC TTT TAC CTC AAA TAT TTT TCG GAA GCC TAT A-3′ (forward) and 5′-ACA TCT GTC AAA ATG ACA TCA TAC TCT TTG GCA AGG-3′ (reverse) (amplicon size, 900 bp). Each PCR amplification was independently repeated two or more times and scored as positive or negative for product following agarose gel electrophoresis; initial data yielding ambiguous results were often repeated using the DNA template obtained by boiling ~20 pooled single colony picks.

Nucleotide sequence determination. The nucleotide sequence was determined (on both strands) for the complete open reading frames (ORFs) of several *rofA*/*nra* and *mga* alleles by PCR amplification and primer walking using overlapping amplicons. The extreme 5′ and 3′ end primers used to amplify each gene, plus flanking sequences, are as follows: for *rofA*, 5′-GGA GAA TAC ACT TAT CAA AGA CT-3′ (forward), 5′-ATC TGG TTG GCG ATC AAG GTA CGG CCA AGC GCA A-3′ (reverse, in *S. pyogenes*), and 5′-ATC TGG TTG GCG ATC AAG GTA CG-3′ (reverse, in non-*S. pyogenes*); for *nra*, 5′-TAA TAG CAC TGA ATA GCT ATT CTA ATA GTG-3′ (forward) and 5-'ATC TGG TTG GCG ATC AAG GTA CGG CCA AGC GCA A-3′ (reverse); for *mga-1*, 5′-GGT CGT ACT GAC TTA ACG AAA TAC CTC ACG-3′ (forward) and 5′-CCT GTT TTT AAT TTT CTA AGC GAA TA-3′ (reverse); for *mga-2*, 5′-GGA GTA AAT TGA CTG AAG TAT GAT AGA ATT TTA ATG-3′ (forward) and 5′-CCT GTT TTT AAT TTT CTA AGC GAA TA-3′ (reverse).

Computations and statistics. The number of polymorphic (segregating) sites and mutations, nucleotide diversity (π), McDonald-Kreitman test with Yates' corrected *G*-values, and Tajima's *D* test statistic were calculated using DnaSP (version 4.0) (52). Coalescent simulations (1,000 replicates) were used to calculate confidence intervals for the Tajima's *D* test, assuming either free recombination or no recombination. Sequence alignments for these calculations were performed using the ClustalW algorithm in Megalign of the DNAStar package (Lasergene version 5.0; Madison, WI), and alignment gaps were excluded. Tests for independence, used to establish nonrandom relationships between loci (linkage disequilibrium), were performed with Fisher's exact test (two-tailed) using DnaSP.

Phylogenetic trees were constructed by the neighbor-joining method using PAUP version 4.0b10 (Sinauer Associates, Sunderland, MA) and the maximum likelihood distance measure. The optimal model of DNA substitution and the parameters were derived using hierarchical likelihood ratio tests (28), with the aid of MODELTEST version 3.06 (48).

Nucleotide sequence accession numbers. The following new sequences have been deposited in the GenBank database: 15 new *mga* sequences under accession numbers AY905500 to AY905514, 17 new *rofA*/*nra* sequences under accession numbers AY905515 to AY905531, and 6 new *rofCG* sequences under accession numbers AY905532 to AY905537.

RESULTS

Phylogeny of transcription regulatory genes, *mga and rofA*/*nra*. Of the many genes present among *S. pyogenes* that encode regulators of transcription (20, 57), two loci are known to exhibit extensive genetic diversity within the *S. pyogenes* population. These are *mga* and *rofA*/*nra*, encoding the "stand-alone" response regulators Mga and RofA/Nra, for which the interacting sensory elements remain unknown (36, 50). The *mga* and *rofA*/*nra* loci are positioned ~300 kb apart on the *S. pyogenes* genome (see Fig. S1 in the supplemental material). The complete nucleotide sequence was determined at each locus from ~20 to 25 *S. pyogenes* strains representing each of the three major *emm* pattern-defined subpopulations.

A phylogenetic tree of *mga* alleles displays two major sequence clusters having strong bootstrap support (Fig. 1A), designated *mga-1* and *mga-2*. The maximal nucleotide sequence divergence for any two alleles belonging to the same lineage is 3.3 and 2.5%, respectively, for *mga-1* and *mga-2* (Table 1). In sharp contrast, the maximal nucleotide sequence divergence for any two alleles belonging to different lineages is 24.5%. Similarly, the maximal amino acid sequence divergence within each cluster is 2.9 and 1.5% for *mga-1* and *mga-2*, respectively, whereas the maximal divergence between *mga-1* and *mga-2* alleles is 20.7%. The high number of fixed nucleotide differences (308 out of 449 polymorphic sites), combined with only five shared polymorphisms, is indicative of relatively low levels of intragenic recombination between alleles of the two divergent *mga* lineages. Furthermore, polymorphic sites are distributed across the length of the genes, with no clear evidence for mosaic structures (data not shown).

Like *mga*, a phylogenetic tree of *rofA*/*nra* alleles also displays two major sequence clusters (Fig. 1B), designated *rofA* and *nra*. The maximal nucleotide sequence divergence for any two alleles belonging to the same lineage is 1.6 and 1.0%, respectively, for *rofA* and *nra* (Table 1). However, the maximal nucleotide sequence divergence for any two alleles belonging to different lineages is 33.5%. Similarly, the maximal amino acid sequence divergence within each cluster is 2.4 and 1.2% for

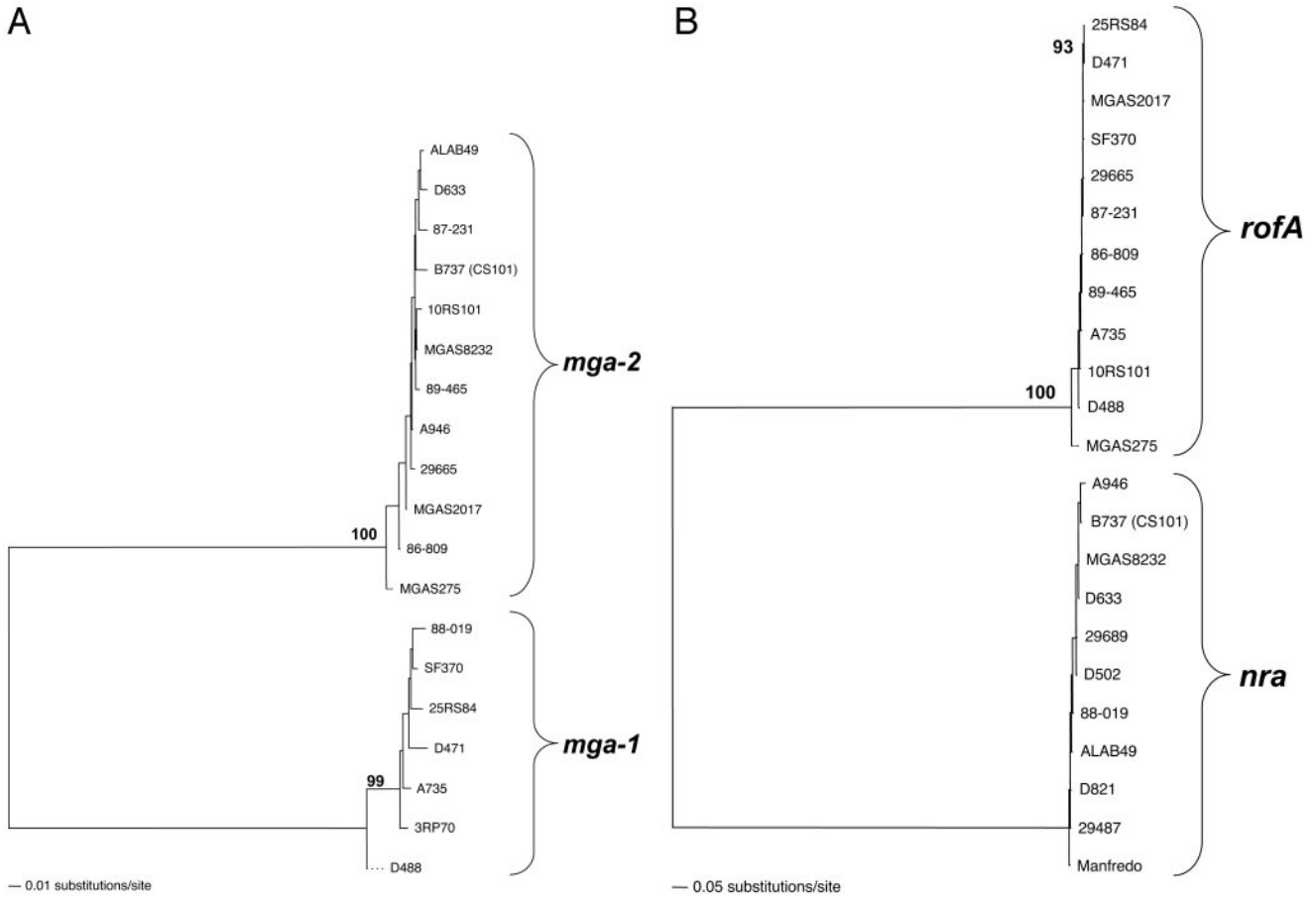

FIG. 1. Phylogenetic trees of complete ORFs of transcriptional regulatory genes of *S. pyogenes*. Neighbor-joining tree (mid-point rooted) with maximum likelihood distance for *mga* derived from 19 isolates (A) and *rofA*/*nra* derived from 23 isolates (B). The rate matrices were optimized to the best-fit models, using the hierarchical likelihood ratio test: TrN+G (for *mga*) and TVM+G (for *rofA*/*nra*). Bootstrap values showing confidence intervals of ≥90% are indicated (1,000 replicates). Excluding alignment gaps, the total number of nucleotide sites is 1,579 for panel A and 1,482 for panel B. Three alleles have a premature termination signal or readthrough, relative to the stop codon of the majority of alleles for that locus (A735 and Manfredo for *mga*; MGAS8232 for *rofA*/*nra*); these sequences were trimmed to the same position as the majority of ORFs for that allele. For *rofA*/*nra*, the genome map position adjacent to the highly conserved *hsp33* locus was confirmed by PCR. Taxon labels indicate *S. pyogenes* strains and are listed and described further in Table S1 in the supplemental material, except for strains D471, MGAS8232, and SF370, for which the GenBank accession numbers are M58461, AE010111, and AE006624 for *mga* and U01312, AE009963, and AE006482, for *rofA*/*nra*, respectively. GenBank accession numbers for strain B737 (CS101) are X68501 (for *mga*) and SPU49397 (for *nra*); for strain A735, the GenBank accession number is AF447492 (for *rofA*/*nra*); for strain Manfredo, sequences were obtained from www.sanger.ac.uk. Note that for *rofA*, the ORFs used for analysis begin with a codon for leucine, not methionine (25). The GenBank accession numbers for new *mga* and *rofA*/*nra* sequences are given in Materials and Methods.

rofA and *nra*, respectively, whereas the maximal divergence between *rofA* and *nra* alleles is 38.2%. Like *mga*, the high number of fixed nucleotide differences (456 out of 527 polymorphic sites) and the few shared polymorphisms ($n = 3$) are indicative of relatively low levels of intragenic recombination between alleles of the *rofA* and *nra* lineages, further supported by a well-spread distribution of polymorphic sites across the entire alignment (data not shown).

TABLE 1. Sequence diversity within and between lineages of transcriptional regulatory genes of *S. pyogenes*

Comparison[a]	No. of sequences	Nucleotide diversity (π)	Maximal % nucleotide divergence	Maximal % amino acid divergence	No. of polymorphic sites	No. of fixed nucleotide differences	No. of shared polymorphisms
Within *mga-1*	7	0.0218	3.29	2.85	94	NA	NA
Within *mga-2*	12	0.0145	2.53	1.52	80	NA	NA
Between *mga-1* and *mga-2*	19	NA	24.45	20.72	449	308	5
Within *rofA*	12	0.0104	1.62	2.44	63	NA	NA
Within *nra*	11	0.0050	1.01	1.15	31	NA	NA
Between *rofA* and *nra*	23	NA	33.54	38.21	527	456	3

[a] Total sites excluding alignment gaps is 1579 and 1482, for *mga* and *rofA*/*nra* alignments, respectively. NA, not applicable.

TABLE 2. Distribution of sequence clusters of transcriptional regulator genes among a diverse set of 114 *S. pyogenes* strains

Genotype	*emm* pattern	*mga-1* or *mga-2*	*rofA* or *nra*	No. of strains represented	*emm* types represented
1	A-C	*mga-1*	*rofA*	17	1, 6, 12, 14, 17, 23, 26, 29, 37, 39, 55, 57, 38/40, st3765, st980584, stCK401, stNS90
2	A-C	*mga-1*	*nra*	2	3, 5
3	A-C	*mga-2*	*rofA*	0	
4	A-C	*mga-2*	*nra*	1	18
5	D	*mga-1*	*rofA*	0	
6	D	*mga-1*	*nra*	0	
7	D	*mga-2*	*rofA*	6	32, 59, 81, 85, 95, st204
8	D	*mga-2*	*nra*	32	34, 36, 41, 42, 52, 53, 56, 67, 70, 71, 74, 83, 91, 93, 98, 99, 105, 111, 116, 119, 65/69, st2037, st2917, st2940, st369, st5282, st809, st854, stCK249, stD432, stD633
9	E	*mga-1*	*rofA*	0	
10	E	*mga-1*	*nra*	0	
11	E	*mga-2*	*rofA*	52	2, 4, 8, 9, 11, 15, 22, 25, 28, 48, 58, 60, 63, 66, 68, 75, 76, 77, 79, 82, 87, 88, 89, 90, 92, 96, 102, 104, 106, 107, 109, 110, 112, 113, 114, 118, 124, 13L, 27G, 44/61, 50/62, st1207, st1389, st213, st2147,st2460, st6735, stMTH81, stNS292, stNS554
12	E	*mga-2*	*nra*	4	49, 73, 94, 117

Other recognized RofA-like proteins are far more divergent, exhibiting amino acid sequence identities to RofA (or Nra) of <30% (data not shown). The low-homology RofA-like proteins include paralogs of *rofA*/*nra* (in *S. pyogenes*) and orthologs (in *Streptococcus pneumoniae*) (8, 25).

Linkage analysis of *mga and rofA*/*nra* lineages. Oligonucleotide primers specific for each of the two major sequence clusters found at the *mga* and *rofA*/*nra* loci were used to screen a diverse set of 114 *S. pyogenes* strains by PCR amplification (see Table S1 in the supplemental material). All strains yielded an amplicon of the expected size with either the *mga-1*- or *mga-2*-specific primer pairs, and there were no strains showing evidence for the presence of alleles belonging to both lineages. Similarly, all 114 strains yielded an amplicon of the expected size with either the *rofA*- or *nra*-specific primer pairs, and no strain yielded an amplicon with both primer pairs. Therefore, among this set of 114 strains, *rofA* and *nra* appear to be mutually exclusive. However, this finding is inconsistent with another study that used a different set of 62 strains, wherein both genes are reported for a small number of strains (35); the observed differences may be due to the selected strains or the methodological approach. That *rofA* and *nra* occupy the same relative position (i.e., locus) on the genome is supported by whole genome maps of several *S. pyogenes* strains (3, 5, 20, 44, 57; http://www.sanger.ac.uk.) and by our finding that all *rofA*/*nra* alleles examined at the 3′ flanking region display a high level of nucleotide sequence identity (>97%), extending through the first 100 bp of the 5′ end of the *hsp33* gene (data not shown). Paralogous genes arise by duplication and occupy different positions on the genome. Taken together, the data strongly suggest that *rofA* and *nra* are not paralogs, nor are *mga-1* and *mga-2* paralogous pairs.

The 114 *S. pyogenes* strains selected for PCR-based screening represent a genetically diverse set and include both highly prevalent clones and rare clones. Nearly all isolates under study have a unique *emm* type (*emm25* and *emm66* strains each occur twice), and all isolates share five or fewer of the seven housekeeping loci (see Table S1 in the supplemental material). A matrix of pairwise distances between strains was constructed based on the proportion of housekeeping loci having shared alleles, by cluster analysis using the unweighted-pair group method using average linkages (see Fig. S2 in the supplemental material). The dendrogram shows that there is a general lack of concordance between *emm* pattern and the genetic relatedness of strains inferred using allelic profiles of neutral housekeeping genes. The finding for this particular subset of strains is consistent with previous phylogenetic and statistical analyses, which consistently show a high degree of recombination between housekeeping genes belonging to strains of different *emm* pattern-defined subpopulations (19, 33). Importantly, there is a general lack of concordance between the major sequence clusters at either the *mga* or *rofA*/*nra* locus and genetic relatedness based on multilocus sequence typing using housekeeping genes (see Fig. S2 in the supplemental material).

Despite the random associations between housekeeping alleles, strong nonrandom associations were observed between the major sequence clusters of the *mga* locus and the *emm* pattern-defined subpopulations for tissue site preference (Table 2). Of the *emm* pattern A-C subpopulation, 19 of 20 (95%) strains had an *mga-1*-lineage allele, whereas 100% of *emm* pattern D ($n = 38$) and E ($n = 56$) strains had an *mga-2*-lineage allele. This difference in *mga* allelic lineage content between the pattern A-C subpopulation and either the pattern D or E subpopulation was highly significant ($P < 0.00001$, Fisher's exact test, two-tailed). Given that *mga* maps immediately upstream of *emm* (see Fig. S1 in the supplemental material), the finding for coevolution of *mga* and *emm* pattern may be a consequence of their tight physical linkage.

The *rofA*/*nra* locus maps ~300 kb from the *emm* region (see Fig. S1 in the supplemental material). This distance represents ~15% of the single circular ~1.9 Mb chromosome of *S. pyogenes*, and therefore physical linkage between the *emm* and *rofA*/*nra* loci is expected to be far weaker, compared to *mga* and *emm*. Yet strong nonrandom associations were observed between the major sequence clusters of the *rofA*/*nra* locus and the *emm* pattern-defined subpopulations (Table 2). Furthermore, the nature of the distribution of the two gene lineages differed from that observed for the *mga* locus. Of the *emm* pattern D subpopulation, 32 of 38 (84%) strains had an *nra*-lineage allele, whereas 17 of 20 (85%) and 52 of 56 (93%) *emm* pattern A-C and E strains, respectively, had a *rofA*-lineage allele instead. The difference in the distribution of *rofA* and *nra* between the *emm* pattern D subpopulation and either the

pattern A-C or E subpopulation, was highly significant ($P < 0.00001$, Fisher's exact test, two-tailed).

There are 12 possible combinations of *emm* pattern (A-C, D, and E), *mga* lineage (*mga-1* and *mga-2*) and *rofA*/*nra* lineage (*rofA* and *nra*) that can exist under conditions of unconstrained gene flow. Among the 114 genetically distinct strains of *S. pyogenes*, 101 (89%) are accounted for by only three combinations of *emm* pattern, *mga* lineage, and *rofA*/*nra* lineage (Table 2). However, 8 of the 12 possible genotypes are observed among this set of 114 strains. Rare genotypes can be associated with highly prevalent clones, such as the pattern A-C/*mga-1*/*nra* and pattern A-C/*mga-2*/*nra* combinations found in the *emm3*-ST15/ST16 and *emm18*-ST62 clones, respectively (see Table S1 in the supplemental material) (41, 56, 64). Highly prevalent clones having a rare genotype may owe their success to compensatory mutations. For the *emm18*-ST62 clone, compensatory mutations may include the genetic variation that leads to copious capsule production, an important phenotype in respiratory tract infection, as well as a frameshift mutation within the *nra* gene, leading to premature termination of translation and a truncated Nra protein (Fig. 1B) (1, 30, 57, 66).

In summary, the data show strong nonrandom associations between the *emm* pattern genotype and lineage-specific alleles of loci encoding global regulators of transcription, against a background of random associations among housekeeping genes. Taken together, the population findings and linkage analysis provide evidence for a role of *mga*, *emm*, and/or *rofA*/*nra* in conferring tissue-specific adaptations in the majority of strains.

Evidence for interspecies HGT of transcription regulatory genes. Nucleotide sequence data for both the *mga* and *rofA*/*nra* loci indicate that the level of divergence within a lineage is relatively low, compared to divergence between lineages (Table 1). This finding is consistent with the idea that a gene corresponding to one of the two lineages was acquired by HGT from another species and replaced the ancestral gene.

Attempts to identify possible donor species of *mga* and/or *rofA*/*nra* genes were made. A phylogenetic tree based on 16S rRNA sequences of the *Streptococcus* genus places the beta-hemolytic streptococcal species in a cluster that closely corresponds to the "pyogenic" group (18). Isolates of some of these streptococcal species have been recovered from humans. The close taxonomic relatives of *S. pyogenes* were examined for the presence of genes having high sequence homology to the *mga* and *rofA*/*nra* alleles recovered from *S. pyogenes*.

Fifty-four isolates of streptococci, which included 12 species or subspecies, and 33 isolates of *Streptococcus dysgalactiae* subspecies *equisimilis* known to be recovered from humans were screened by PCR using lineage-specific primers for the *mga-1*/*mga-2* and *rofA*/*nra* alleles identified in *S. pyogenes* (see Table S2 in the supplemental material). None of the 54 isolates yielded an amplicon with primers specific for *mga-1* or *mga-2* alleles, indicating that a possible donor source for one of the two *mga* lineages remains to be established. As a positive control, primers specific for *mgc* (22), an ortholog of *mga* displaying ~43% nucleotide sequence divergence with both *mga-1* and *mga-2* alleles, yielded an amplicon for the majority of isolates designated *S. dysgalactiae* subsp. *equisimilis*.

Among the set of 54 non-*S. pyogenes* isolates of streptococci, none yielded an amplicon with the *nra*-specific primers (see Table S2 in the supplemental material). However, 33 isolates produced a PCR-generated amplicon with the *rofA*-specific primers. Included among the positive isolates were all 33 isolates of *S. dysgalactiae* subsp. *equisimilis* having the group C or G carbohydrate and known to be recovered from humans. The *rofA*-like genes recovered from group C and G streptococci are herein designated *rofCG*. It should be noted that isolates of the human pathogen *Streptococcus agalactiae*, another beta-hemolytic organism of the pyogenic group, were not analyzed by PCR in this study; however, in silico analysis of whole genome sequences (23, 62) shows a lack of genes with high sequence homology to either the *mga* or *rofA*/*nra* alleles present in *S. pyogenes* strains.

The nucleotide sequence was determined for the complete ORF of *rofCG* genes derived from six isolates of *S. dysgalactiae* subsp. *equisimilis*. A phylogenetic tree, which includes the *rofA* and *nra* alleles derived from *S. pyogenes*, shows that the *rofCG* alleles lie within the same cluster as *rofA* alleles (Fig. 2). Furthermore, all *rofA* and *rofCG* alleles examined display a high level of nucleotide sequence identity (>97%) over the 5′ end of the *hsp33* gene that lies immediately downstream from the *rofA*/*nra* locus (8), suggesting that *rofA* and *rofCG* alleles occupy the same relative position on their respective genomes.

The maximal nucleotide sequence divergence between *rofA* and *rofCG* alleles is only 2.0% (Table 3). Similarly, the maximal amino acid sequence divergence between *rofA* versus *rofCG* alleles is only 2.9%. There are no fixed nucleotide differences between *rofA* and *rofCG* alleles over 74 polymorphic sites, and 29 polymorphisms are shared between the two populations. The extent of sequence differences within the *rofA* and *rofCG* sets of alleles is roughly of the same order as the differences between the two populations. The data suggest that *rofA* and *rofCG* alleles share a recent common ancestor, since they are highly homologous in nucleotide sequence.

Evidence for selection within transcription regulatory genes. In order to ascertain whether there was evidence for selection within the *mga* and *rofA*/*nra* genes, a test for neutrality of polymorphic (segregating) sites was employed. The Tajima's *D* statistic tests the hypothesis that all mutations are selectively neutral, whereby $D = 0$ under neutrality (61). The data in Table 4 indicate that when *mga-1* lineage alleles are analyzed by themselves, the *D* statistic is unable to reject neutrality; a similar result is found for the set of *mga-2* lineage alleles. However, when the combined set of *mga-1* and *mga-2* alleles is considered, *D* values are significant in a positive direction. Parallel findings are obtained for *rofA* and *nra* alleles, both for within lineages and combined lineage calculations (Table 4). The significant positive value for *D*, for the combined lineage alleles, is consistent with a role for balancing selection in maintaining a relatively large number of polymorphic sites at intermediate frequencies. Balancing selection is also supported by the phylogenetic tree topologies observed for both *mga* and *rofA*/*nra* (Fig. 1A and B).

A negative value of the Tajima's *D* statistic arises when there are more polymorphic sites with rare alleles than expected under neutral genetic drift. Negative *D* values were observed for each lineage considered separately (*mga-1*, *mga-2*, *rofA*, and *nra*) when calculated under the assumption that there is no recombination between alleles of the same lineage, but these values are statistically nonsignificant (Table 4). However, *S.*

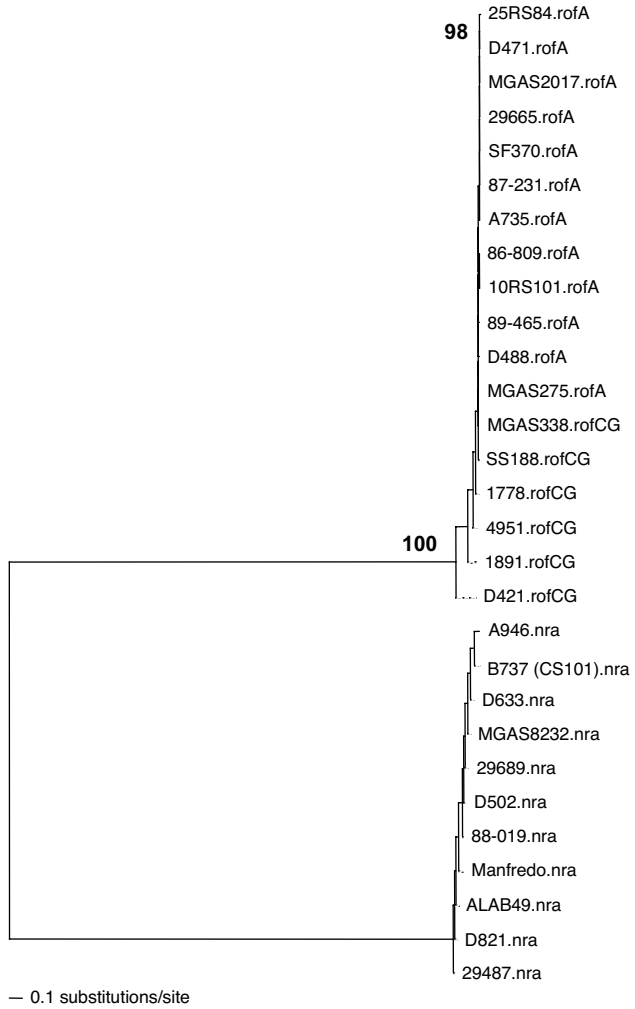

FIG. 2. Phylogenetic tree of *rofA* and *rofCG*. The phylogenetic tree is as described in the legend of Fig. 1B, with an additional six *rofCG* alleles. The rate matrix was optimized to the best-fit model (GTR+G), using the hierarchical likelihood ratio test. Taxon labels indicate streptococcal strains, as listed and described further in Table S1 (for *S. pyogenes*) or Table S2 in the supplemental material (32); included are three strains each of group C and G streptococci. For *rofCG*, the genome map position adjacent to the highly conserved *hsp33* locus was confirmed by PCR. The GenBank accession numbers for six new *rofCG* sequences are given in Materials and Methods.

pyogenes displays evidence for high levels of recombination among housekeeping genes (19, 33), and recombination decreases the variance of Tajima's *D* (54), making it more difficult to reject the hypothesis of neutrality. Confidence intervals for Tajima's *D* were derived by coalescent simulation allowing for free recombination (Table 4), and they show statistically significant departures from neutrality in the negative direction. A negative value of the *D* statistic can indicate purifying (negative) selection or a population bottleneck that purges genetic diversity genome-wide. However, a recent population bottleneck within *S. pyogenes* is not likely, based on the observed diversity in housekeeping alleles at multiple loci (see Table S1 in the supplemental material). Thus, the data are most consistent with a role for purifying selection acting on alleles within each lineage. In biological terms, purifying selection is often the signature of the deleterious effects of mutations on high fitness alleles.

TABLE 4. Selection within and between lineages of transcription regulatory genes of *S. pyogenes*[a]

Locus or lineage	No. of sequences	Tajima's *D* value	95% CI of Tajima's *D* value by coalescent simulations, assuming:	
			No recombination	Free recombination
mga-1	7	−0.60	(−1.49, 1.72)	(−0.33, 0.36)
mga-2	12	−0.63	(−1.76, 1.65)	(−0.44, 0.45)
mga-1 and *mga-2*	19	2.25*	(−1.75, 1.58)	(−0.22, 0.23)
rofA	12	−0.97	(−1.75, 1.64)	(−0.53, 0.54)
nra	11	−1.29	(−1.75, 1.67)	(−0.67, 0.68)
rofA and *nra*	23	3.43*	(−1.76, 1.71)	(−0.21, 0.21)

[a] Statistical significance of *D* was assessed using the beta distribution. For values marked with an asterisk, *P* was <0.05 by a two-tailed test.

Since *rofA* and *nra* appear to be orthologous in origin (Fig. 2) and *mga-1* and *mga-2* alleles display similar key features despite our failure to establish its presence in a second species, the McDonald-Kreitman neutrality test for interspecies variation was used to examine the genetic differences within and between lineages at each locus. Under neutrality, the ratio of synonymous (i.e., silent) to nonsynonymous (i.e., leading to amino acid substitution) fixed nucleotide differences between orthologous genes should be the same as the ratio of synonymous to nonsynonymous nucleotide polymorphisms within the lineages. The neutrality index indicates the extent to which the observed levels of amino acid polymorphism depart from expected levels under neutral evolution. The neutrality index is <1 for both regulatory genes (Table 5) and is consistent with the biological findings based on Tajima's *D* test. The *G* test of independence shows that deviations on the ratio of synonymous to nonsynonymous, for fixed substitutions between lineages versus polymorphisms within lineages, are statistically significant at the *rofA*/*nra* locus but not significant for the *mga* locus (Table 5). When the six *rofCG* alleles are included in the

TABLE 3. Sequence diversity within and between *rofA* alleles derived from *S. pyogenes* and *rofCG* alleles derived from other streptococcal species[a]

Comparison	No. of sequences	Nucleotide diversity (π)	Maximal % nucleotide divergence	Maximal % amino acid divergence	No. of polymorphic sites	No. of fixed differences	No. of shared polymorphisms
Within *rofA* (from *S. pyogenes*)	12	0.0102	1.62	2.44	58	NA	NA
Within *rofCG* (from non-*S. pyogenes*)	6	0.0124	1.75	2.24	43	NA	NA
Between *rofA* and *rofCG*	18	NA	2.02	2.85	74	0	29
Between *rofA*/*rofCG* and *nra*	29	NA	33.54	38.21	530	447	3

[a] NA, not applicable.

TABLE 5. McDonald-Kreitman test for neutrality[a]

Lineages	Synonymous substitutions		Nonsynonymous substitutions		McDonald-Kreitman test	
	Fixed differences between lineages	Polymorphic sites within lineages	Fixed differences between lineages	Polymorphic sites within lineages	Neutrality index	*P*
mga-1 and *mga-2*	NA	122	NA	46	NA	NA
	194	NA	108	NA	0.677	0.078
rofA and *nra*	NA	59	NA	34	NA	NA
	226	NA	230	NA	0.566	0.019

[a] *P* values were determined by a *G* test with Yates' correction. NA, not applicable.

calculation, for a total of 29 *nra* and *rofA*/*CG* alleles, the departure from neutrality remains statistically significant ($P = 0.04$, *G* test with Yates' correction; data not shown).

Several of the fixed nucleotide differences between alleles of different lineages (Table 5) were examined in greater depth for synonymous versus nonsynonymous changes. A key functional region of transcription regulatory proteins is the helix-turn-helix (HTH) DNA-binding motif. Positive selection in the HTH region may reflect an adaptation that results in the regulation of different genes. The putative HTH motif identified in RofA (21, 25) is highly divergent when compared to the aligned region of Nra, displaying 12 nonsynonymous fixed nucleotide changes, at sites 115, 118, 121, 123, 130, 133, 148, 149, 151, 162, 163, and 172. These nucleotide differences result in nine amino acid changes over the 21-residue region. In contrast, the 20-residue HTH motifs of Mga that were previously shown by experiment to be critical for autoregulated *mga* expression (42) are highly conserved among the *mga-1* and *mga-2* lineages, with only two fixed nucleotide differences (at sites 196 and 209) leading to amino acid replacements within one domain (HTH-3) and no fixed replacements within the other domain (HTH-4). The high number of nonsynonymous fixed nucleotide differences in *mga* lying outside of the HTH-coding regions suggests that any positive selection on *mga* appears to act elsewhere in the gene.

In summary, there is evidence that balancing selection acting on the *mga* and *rofA*/*nra* genes played an important role in shaping the nucleotide sequence differences observed between *mga-1* and *mga-2* alleles and between *rofA* and *nra* alleles, whereas purifying selection appears to have contributed to the low level of genetic diversity that is found within each lineage.

DISCUSSION

That the throat and skin represent distinct ecological niches for *S. pyogenes* is supported by the finding that many strains exhibit a strong preference for one tissue site over the other. There are several scales of spatial-temporal distance that keep the throat and skin strain specialists apart: distinct tissue habitats within a single host, geographic partitioning on a global scale, and seasonal peaks in disease incidence. Such physical barriers to HGT can act to reduce the flow of genetic information. Yet for housekeeping genes, discrete sequence clusters corresponding to the *emm* pattern-defined subpopulations of *S. pyogenes* are not evident (19, 33). The reason may be that the generalist subpopulation serves as a genetic shuttle between the specialists (7) or that insufficient time has elapsed for a strong phylogenetic signal to emerge.

Nonetheless, in instances where housekeeping alleles are randomly distributed with respect to ecologically distinct populations, genetic variation that is strongly associated with the different ecological populations may be directly responsible for adaptation to the ecological niche. Thus, products of the *rofA*/*nra* and/or *mga* genes (and/or tightly linked genes, such as *emm*) are strong candidates for having a direct role in conferring tissue-specific tropisms.

The site-frequency distributions among *rofA* and *nra* alleles and among *mga-1* and *mga-2* alleles show that polymorphic sites are maintained at intermediate frequencies, which suggests that both loci are under balancing selection. The finding by coalescent simulation of purifying selection within each lineage is consistent with the notion that each sequence cluster of the transcription regulatory genes corresponds to a distinct fitness peak.

Each of the regulatory gene loci under study may have a long history of evolutionary divergence and adaptation within separate bacterial species, generating alleles of discrete lineages, followed by a more recent replacement of the *S. pyogenes* ancestral gene with an ortholog, via interspecific HGT (24, 29, 34). Newly acquired orthologous genes may potentially provide a rich and ready source for new bacterial phenotypes. During the process of speciation, sites within an ancestral gene that are critical for adaptation to a new niche will undergo positive (diversifying) selection, while constrained functions can be preserved via negative (purifying) selection. Following replacement of the ancestral *S. pyogenes* allele with an ortholog, the recipient strain appears to have undergone genetic diversification at many loci and/or the orthologous allele to have spread via interstrain HGT and localized recombination at sites of highly homologous flanking DNA (37, 40).

At the *rofA*/*nra* locus, *rofCG* may have been acquired by *S. pyogenes* from a human commensal species of streptococci (63). This direction of transfer is consistent with the finding that 100% of the human isolates of *S. dysgalactiae* subsp. *equisimilus* examined harbor *rofCG*. It is unlikely that this commensal species underwent a recent population bottleneck because it exhibits extensive mosaicism in its content of highly divergent housekeeping alleles at multiple loci (32). Alternatively, *rof* may be ancestral to both streptococcal species, and, instead, *nra* was acquired by *S. pyogenes* via an orthologous replacement involving an unknown donor. In either scenario, it seems probable that subsequent recombination between *rofA* and *rofCG* has masked any phylogenetic signal.

Another plausible explanation for balancing selection is that within *S. pyogenes* there was a gradual, long-term diversification of an ancestral gene and incremental increases in fitness, whereby each allele of increased fitness subsequently spread

between some *S. pyogenes* strains via localized recombination, resulting in a partial selective sweep. However, it is difficult to evolve high levels of genetic divergence among strains that are engaged in continuous gene exchange. Furthermore, neither *mga* nor *rofA*/*nra* is characteristic of highly mutable genes (43), and many independent steps are probably required to generate high levels of sequence diversity through genetic changes that occur wholly within a species. Thus, the orthologous gene replacement model, invoking a critical role for additional bacterial species, appears to be more parsimonious.

Divergent forms of global regulators of gene transcription may play a pivotal role in niche adaptation by interacting with distinct sets of genes or by differentially affecting the expression of the same set of genes in a strain- or species-dependent manner. Laboratory replacement of the *Escherichia coli* transcription regulatory gene *pmrD*, with its ortholog derived from *Salmonella enterica*, leads to the differential regulation of conserved genes and to the acquisition of a new phenotype (67). RofA and Nra are distinguished by their differential effects—activation, repression, or no effect—on transcription of several virulence genes, as well as other regulatory genes that include *mga* (35). RofA/Nra displays evidence for positive selection at the putative DNA-binding site, whereas Mga-1 and Mga-2 exhibit few or no amino acid replacements at their two proven DNA-binding sites. Thus, it seems plausible that Mga-1 and Mga-2 regulate the same set of genes by binding to identical *cis*-acting sites but respond to signals by different pathways.

Experimental findings, combined with population analysis of a worldwide collection of *S. pyogenes* isolates, provide strong evidence that several virulence factors contribute to host tissue tropisms (31, 53, 58–60). Mutant bacteria inactivated in genes encoding a secreted cysteine proteinase (SpeB), a plasminogen activator (Ska), or a plasminogen-binding M protein (PAM) display a reduction in net growth in an experimental model for superficial skin infection. In addition, each of the phenotypes ascribed to SpeB, Ska, and PAM exhibits a strong association with the *emm* pattern D subpopulation of skin specialists. The SpeB, Ska, and PAM virulence factors can also influence the phenotypic activities of one another. For example, SpeB can degrade Ska, and Ska can act in cooperation with PAM to generate bacterial-bound plasmin (51). Importantly, Mga-1/Mga-2 and RofA/Nra modulate the expression of the genes encoding SpeB, Ska, and PAM, either directly or through other transcription regulatory networks that include their own cross-regulatory circuit (35, 36). Thus, there appears to be an intricate network of interacting proteins and genes that helps orchestrate the adaptation of *S. pyogenes* to narrowly defined, tissue-specific niches.

Like *rofA*/*nra* and *mga*, the *ska* alleles of *S. pyogenes* form discrete sequence clusters. One *ska* lineage (*ska-2a*) is largely restricted to pattern A-C strains and also shares a recent common ancestor with the *skcg* genes present in all *S. dysgalactiae* subsp. *equisimilus* strains examined (31). It is of potential significance that *S. dysgalactiae* subsp. *equisimilus* often colonizes the throat (47). HGT between streptococcal organisms presumably occurs with greatest efficiency during coinfection at the throat or within an impetigo lesion (6, 11), whereby the donor DNA is present in a relatively high concentration due to the viability of its bacterial host cell. Conceivably, this commensal species may be a donor source for *rofA* and *ska-2a* acquisition by *S. pyogenes*, thereby facilitating the adaptation of certain clones of *S. pyogenes* to the throat.

Mga-1 and RofA, along with certain Ska forms (31), may be essential for high fitness at the throat, whereas Mga-2 and Nra plus another Ska form may be optimal for bacterial growth and transmission at the skin. However, the risk factors that promote throat and skin infection differ, and both niches are not universally available to *S. pyogenes*. Mga-2 and RofA together, as observed in the pattern E subpopulation of generalists, may allow for exploitation of both the throat and skin, thereby allowing the organism to survive shifting periods of niche availability but with a tradeoff in the form of suboptimal reproductive success.

In general terms, high levels of recombination allow for a quick exploration of numerous genotype combinations and, thereby, may promote the emergence of complex phenotypes that require several independent genetic changes. The observed linkage disequilibrium between *mga*, *rofA*/*nra*, and *emm* pattern and between *emm* pattern and other genotypes and phenotypes (31, 58) is supportive of a critical role for epistasis in niche adaptation. Orthologous gene replacements leading to the successful exploitation of a new niche may have a particularly high impact on the accelerated evolution of this bacterial species, which is otherwise lacking in pathogenicity islands (55).

ACKNOWLEDGMENTS

The authors thank R. Facklam (CDC, Atlanta, GA) for providing non-*S. pyogenes* isolates and S. Remold and A. Kalia for helpful discussions.

This work was supported by grants from the National Institutes of Health (R01-AI053826 and GM060793) and the American Heart Association (Grant-in-Aid) to D.E.B.

REFERENCES

1. **Alberti, S., C. D. Ashbaugh, and M. R. Wessels.** 1998. Structure of the *has* operon promoter and regulation of hyaluronic acid capsule expression in group A *Streptococcus*. Mol. Microbiol. **28:**343–353.
2. **Anthony, B. F., E. L. Kaplan, L. W. Wannamaker, and S. S. Chapman.** 1976. The dynamics of streptococcal infections in a defined population of children: serotypes associated with skin and respiratory infections. Am. J. Epidemiol. **104:**652–666.
3. **Banks, D. J.** 2004. Progress toward characterization of the group A *Streptococcus* metagenome: complete genome sequence of a macrolide-resistant serotype M6 strain. J. Infect. Dis. **190:**727–738.
4. **Beall, B.** 2004. http://www.cdc.gov/ncidod/biotech/strep/emmtypes.htm.
5. **Beres, S. B., G. L. Sylva, K. D. Barbian, B. Lei, J. S. Hoff, N. D. Mammarella, M. Y. Liu, J. C. Smoot, S. F. Porcella, L. D. Parkins, D. S. Campbell, T. M. Smith, J. K. McCormick, D. Y. Leung, P. M. Schlievert, and J. M. Musser.** 2002. Genome sequence of a serotype M3 strain of group A *Streptococcus*: phage-encoded toxins, the high-virulence phenotype, and clone emergence. Proc. Natl. Acad. Sci. USA **99:**10078–10083.
6. **Bessen, D. E., J. R. Carapetis, B. Beall, R. Katz, M. Hibble, B. J. Currie, T. Collingridge, M. W. Izzo, D. A. Scaramuzzino, and K. S. Sriprakash.** 2000. Contrasting molecular epidemiology of group A streptococci causing tropical and non-tropical infections of the skin and throat. J. Infect. Dis. **182:**1109–1116.
7. **Bessen, D. E., M. W. Izzo, T. R. Fiorentino, R. M. Caringal, S. K. Hollingshead, and B. Beall.** 1999. Genetic linkage of exotoxin alleles and *emm* gene markers for tissue tropism in group A streptococci. J. Infect. Dis. **179:**627–636.
8. **Bessen, D. E., and A. Kalia.** 2002. Genomic localization of a T-serotype locus to a recombinatorial zone encoding for extracellular matrix-binding proteins in *Streptococcus pyogenes*. Infect. Immun. **70:**1159–1167.
9. **Bisno, A. L., and D. Stevens.** 2000. *Streptococcus pyogenes* (including streptococcal toxic shock syndrome and necrotizing fasciitis), p. 2101–2117. *In* G. L. Mandell, R. G. Douglas, and R. Dolin (ed.), Principles and practice of infectious diseases, 5th ed., vol. 2. Churchill Livingstone, Philadelphia, Pa.
10. **Carapetis, J., B. Currie, and E. Kaplan.** 1999. Epidemiology and prevention of group A streptococcal infections: acute respiratory tract infections, skin infections, and their sequelae at the close of the twentieth century. Clin. Infect. Dis. **28:**205–210.

11. **Carapetis, J., D. Gardiner, B. Currie, and J. D. Mathews.** 1995. Multiple strains of *Streptococcus pyogenes* in skin sores of aboriginal Australians. J. Clin. Microbiol. **33:**1471–1472.
12. **Cohan, F. M.** 2001. Bacterial species and speciation. Syst. Biol. **50:**513–524.
13. **Colman, G., A. Tanna, A. Efstratiou, and E. Gaworzewska.** 1993. The serotypes of *Streptococcus pyogenes* present in Britain during 1980–1990 and their associations with disease. J. Med. Microbiol. **39:**165–178.
14. **Dicuonzo, G., G. Gherardi, G. Lorino, S. Angeletti, M. DeCesaris, E. Fiscarelli, D. E. Bessen, and B. Beall.** 2001. Group A streptococcal genotypes from pediatric throat isolates in Rome, Italy. J. Clin. Microbiol. **39:**1687–1690.
15. **Dillon, H., C. Derrick, and M. Dillon.** 1974. M-antigens common to pyoderma and acute glomerulonephritis. J. Infect. Dis. **130:**257–267.
16. **Dillon, H. C.** 1972. Streptococcal infections of the skin and their complications: impetigo and nephritis, p. 571–587. *In* L. W. Wannamaker and J. M. Matsen (ed.), Streptococci and streptococcal diseases. Academic Press, New York, N.Y.
17. **Enright, M. C., B. G. Spratt, A. Kalia, J. H. Cross, and D. E. Bessen.** 2001. Multilocus sequence typing of *Streptococcus pyogenes* and the relationship between *emm* type and clone. Infect. Immun. **69:**2416–2427.
18. **Facklam, R.** 2002. What happened to the streptococci: overview of taxonomic and nomenclature changes. Clin. Microbiol. Rev. **15:**613–630.
19. **Feil, E. J., E. C. Holmes, D. E. Bessen, M.-S. Chan, N. P. J. Day, M. C. Enright, R. Goldstein, D. Hood, A. Kalia, C. E. Moore, J. Zhou, and B. G. Spratt.** 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc. Natl. Acad. Sci. USA **98:**182–187.
20. **Ferretti, J. J., W. M. McShan, D. Ajdic, D. J. Savic, G. Savic, K. Lyon, C. Primeaux, S. Sezate, A. N. Suvorov, S. Kenton, H. S. Lai, S. P. Lin, Y. Qian, H. G. Jia, F. Z. Najar, Q. Ren, H. Zhu, L. Song, J. White, X. Yuan, S. W. Clifton, B. A. Roe, and R. McLaughlin.** 2001. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc. Natl. Acad. Sci. USA **98:**4658–4663.
21. **Fogg, G., and M. Caparon.** 1997. Constitutive expression of fibronectin binding in *Streptococcus pyogenes* as a result of anaerobic activation of *rofA*. J. Bacteriol. **179:**6172–6180.
22. **Geyer, A., and K. H. Schmidt.** 2000. Genetic organisation of the M protein region in human isolates of group C and G streptococci: two types of multigene regulator-like (mgrC) regions. Mol. Gen. Genet. **262:**965–976.
23. **Glaser, P., C. Rusniok, C. Buchrieser, F. Chevalier, L. Frangeul, T. Msadek, M. Zouine, E. Couve, L. Lalioui, C. Poyart, P. Trieu-Cuot, and F. Kunst.** 2002. Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. Mol. Microbiol. **45:**1499–1513.
24. **Gogarten, J. P., W. F. Doolittle, and J. G. Lawrence.** 2002. Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. **19:**2226–2238.
25. **Granok, A., D. Parsonage, R. Ross, and M. Caparon.** 2000. The RofA binding site in *Streptococcus pyogenes* is utilized in multiple transcriptional pathways. J. Bacteriol. **182:**1529–1540.
26. **Hollingshead, S. K., T. Readdy, J. Arnold, and D. E. Bessen.** 1994. Molecular evolution of a multi-gene family in group A streptococci. Mol. Biol. Evol. **11:**208–219.
27. **Hollingshead, S. K., T. L. Readdy, D. L. Yung, and D. E. Bessen.** 1993. Structural heterogeneity of the *emm* gene cluster in group A streptococci. Mol. Microbiol. **8:**707–717.
28. **Huelsenbeck, J., and B. Rannala.** 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science **276:**227–232.
29. **Jain, R., M. C. Rivera, J. E. Moore, and J. A. Lake.** 2002. Horizontal gene transfer in microbial genome evolution. Theoretical Population Biology **61:**489–495.
30. **Johnson, D. R., D. L. Stevens, and E. L. Kaplan.** 1992. Epidemiological analysis of group A streptococcal serotypes associated with severe systemic infections, rheumatic fever, or uncomplicated pharyngitis. J. Infect. Dis. **166:**374–382.
31. **Kalia, A., and D. E. Bessen.** 2004. Natural selection and evolution of streptococcal virulence genes involved in tissue-specific adaptations. J. Bacteriol. **186:**110–121.
32. **Kalia, A., M. C. Enright, B. G. Spratt, and D. E. Bessen.** 2001. Directional gene movement from human-pathogenic to commensal-like streptococci. Infect. Immun. **69:**4858–4869.
33. **Kalia, A., B. G. Spratt, M. C. Enright, and D. E. Bessen.** 2002. Influence of recombination and niche separation on the population genetic structure of the pathogen *Streptococcus pyogenes*. Infect. Immun. **70:**1971–1983.
34. **Koonin, E. V., K. S. Makarova, and L. Aravind.** 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annu. Rev. Microbiol. **55:**709–742.
35. **Kreikemeyer, B., S. Beckert, A. Braun-Kiewnick, and A. Podbielski.** 2002. Group A streptococcal RofA-type global regulators exhibit a strain-specific genomic presence and regulation pattern. Microbiology **148:**1501–1511.
36. **Kreikemeyer, B., K. S. McIver, and A. Podbielski.** 2003. Virulence factor regulation and regulatory networks in *Streptococcus pyogenes* and their impact on pathogen-host interactions. Trends Microbiol. **11:**224–232.
37. **Majewski, J., P. Zawadski, P. Pickerill, F. M. Cohan, and C. G. Dowson.** 2000. Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. J. Bacteriol. **182:**1016–1023.
38. **Martin, D. R., and K. S. Sriprakash.** 1996. Epidemiology of group A streptococcal disease in Australia and New Zealand. Recent Adv. Microbiol. **4:**1–40.
39. **Maxted, W. R.** 1980. Disease association and geographical distribution of the M types of group A streptococci, p. 763–777. *In* S. E. Read and J. B. Zabriskie (ed.), Streptococcal diseases and the immune response. Academic Press, New York, N.Y.
40. **Maynard Smith, J., C. G. Dowson, and B. G. Spratt.** 1991. Localized sex in bacteria. Nature **349:**29–31.
41. **McGregor, K. F., B. G. Spratt, A. Kalia, A. Bennett, N. Bilek, B. Beall, and D. E. Bessen.** 2004. Multilocus sequence typing of *Streptococcus pyogenes* representing most known *emm* types and distinctions among subpopulation genetic structures. J. Bacteriol. **186:**4285–4294.
42. **McIver, K. S., and R. L. Myles.** 2002. Two DNA-binding domains of Mga are required for virulence gene activation in the group A streptococcus. Mol. Microbiol. **43:**1591–1601.
43. **Moxon, E. R., P. B. Rainey, M. A. Nowak, and R. E. Lenski.** 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr. Biol. **4:**24–33.
44. **Nakagawa, I., K. Kurokawa, A. Yamashita, M. Nakata, Y. Tomiyasu, N. Okahashi, S. Kawabata, K. Yamazaki, T. Shiba, T. Yasunaga, H. Hayashi, M. Hattori, and S. Hamada.** 2003. Genome sequence of an M3 strain of *Streptococcus pyogenes* reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. Genome Res. **13:**1042–1055.
45. **Nicholson, M. L., L. Ferdinand, J. S. Sampson, A. Benin, S. Balter, S. W. L. Pinto, S. F. Dowell, R. R. Facklam, G. M. Carlone, and B. Beall.** 2000. Analysis of immunoreactivity to a *Streptococcus equi* subsp. *zooepidemicus* M-like protein to confirm an outbreak of poststreptococcal glomerulonephritis, and sequences of M-like proteins from isolates obtained from different host species. J. Clin. Microbiol. **38:**4126–4130.
46. **Orr, M. R., and T. B. Smith.** 1998. Ecology and speciation. Trends Ecol. Evol. **13:**502–506.
47. **Oster, H., and A. Bisno.** 2000. Group C and G streptococcal infections: epidemiologic and clinical aspects, p. 184–190. *In* V. Fischetti, R. Novick, J. Ferretti, D. Portnoy, and J. Rood (ed.), Gram-positive pathogens. ASM Press, Washington, D.C.
48. **Posada, D., and K. Crandall.** 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14:**817–818.
49. **Potter, E. V., M. Svartman, I. Mohammed, R. Cox, T. Poon-King, and D. P. Earle.** 1978. Tropical acute rheumatic fever and associated streptococcal infections compared with concurrent acute glomerulonephritis. J. Pediatr. **92:**325–333.
50. **Ribardo, D. A., T. J. Lambert, and K. S. McIver.** 2004. Role of *Streptococcus pyogenes* two-component response regulators in the temporal control of Mga and the Mga-regulated virulence gene *emm*. Infect. Immun. **72:**3668–3673.
51. **Ringdahl, U., M. Svensson, A. Wistedt, T. Renné, R. Kellner, W. Müller-Esterl, and U. Sjöbring.** 1998. Molecular co-operation between protein PAM and streptokinase for plasmin acquisition by *Streptococcus pyogenes*. J. Biol. Chem. **273:**6424–6430.
52. **Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas.** 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:**2496–2497.
53. **Scaramuzzino, D. A., J. M. McNiff, and D. E. Bessen.** 2000. Humanized in vivo model for streptococcal impetigo. Infect. Immun. **68:**2880–2887.
54. **Schierup, M. H., and J. Hein.** 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics **156:**879–891.
55. **Schmidt, H., and M. Hensel.** 2004. Pathogenicity islands in bacterial pathogenesis. Clin. Microbiol. Rev. **17**:14–56.
56. **Shulman, S.** 2004. Group A streptococcal pharyngitis serotype surveillance in North America, 2000–2002. Clin. Infect. Dis. **39:**325–332.
57. **Smoot, J. C., K. D. Barbian, J. J. Van Gompel, L. M. Smoot, M. S. Chaussee, G. L. Sylva, D. E. Sturdevant, S. M. Ricklefs, S. F. Porcella, L. D. Parkins, S. B. Beres, D. S. Campbell, T. M. Smith, Q. Zhang, V. Kapur, J. A. Daly, L. G. Veasy, and J. M. Musser.** 2002. Genome sequence and comparative microarray analysis of serotype M18 group A streptococcus strains associated with acute rheumatic fever outbreaks. Proc. Natl. Acad. Sci. USA **99:**4668–4673.
58. **Svensson, M. D., D. A. Scaramuzzino, U. Sjobring, A. Olsen, C. Frank, and D. E. Bessen.** 2000. Role for a secreted cysteine proteinase in the establishment of host tissue tropism by group A streptococci. Mol. Microbiol. **38:**242–253.
59. **Svensson, M. D., U. Sjöbring, and D. E. Bessen.** 1999. Selective distribution of a high-affinity plasminogen binding site among group A streptococci associated with impetigo. Infect. Immun. **67:**3915–3920.
60. **Svensson, M. D., U. Sjöbring, F. Luo, and D. E. Bessen.** 2002. Roles of the plasminogen activator streptokinase and plasminogen-associated M protein in an experimental model for streptococcal impetigo. Microbiology **148:**3933–3945.
61. **Tajima, F.** 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:**585–595.

62. **Tettelin, H., V. Masignani, M. J. Cieslewicz, J. A. Eisen, S. Peterson, M. R. Wessels, I. T. Paulsen, K. E. Nelson, I. Margarit, T. D. Read, L. C. Madoff, A. M. Wolf, M. J. Beanan, L. M. Brinkac, S. C. Daugherty, R. T. DeBoy, A. S. Durkin, J. F. Kolonay, R. Madupu, M. R. Lewis, D. Radune, N. B. Fedorova, D. Scanlan, H. Khouri, S. Mulligan, H. A. Carty, R. T. Cline, S. E. Van Aken, J. Gill, M. Scarselli, M. Mora, E. T. Iacobini, C. Brettoni, G. Galli, M. Mariani, F. Vegni, D. Maione, D. Rinaudo, R. Rappuoli, J. L. Telford, D. L. Kasper, G. Grandi, and C. M. Fraser.** 2002. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc. Natl. Acad. Sci. USA **99:**12391–12396.
63. **Towers, R. J., D. Gal, D. McMillan, K. S. Sriprakash, B. J. Currie, M. J. Walker, G. S. Chhatwal, and P. K. Fagan.** 2004. Fibronectin-binding protein gene recombination and horizontal transfer between group A and G streptococci. J. Clin. Microbiol. **42**:5357–5361.
64. **Veasy, L. G., L. Y. Tani, J. A. Daly, K. Korgenski, L. Miner, J. Bale, E. L. Kaplan, J. M. Musser, and H. R. Hill.** 2004. Temporal association of the appearance of mucoid strains of *Streptococcus pyogenes* with a continuing high incidence of rheumatic fever in Utah. Pediatrics **113:**E168–E172.
65. **Wannamaker, L. W.** 1970. Differences between streptococcal infections of the throat and of the skin. N. Engl. J. Med. **282:**23–31.
66. **Wessels, M. R., and M. S. Bronze.** 1994. Critical role of the group A streptococcal capsule in pharyngeal colonization and infection in mice. Proc. Natl. Acad. Sci. USA **91:**12238–12242.
67. **Winfield, M. D., and E. A. Groisman.** 2004. Phenotypic differences between *Salmonella* and *Escherichia coli* resulting from the disparate regulation of homologous genes. Proc. Natl. Acad. Sci. USA **101:**17162–17167.