JOURNAL OF BACTERIOLOGY, Jan. 2004, p. 110–121 Vol. 186, No. 1
0021-9193/04/$08.00+0 DOI: 10.1128/JB.186.1.110–121.2004

Natural Selection and Evolution of Streptococcal Virulence Genes Involved in Tissue-Specific Adaptations

Awdhesh Kalia† and Debra E. Bessen*

Department of Ecology & Evolutionary Biology, Yale University, New Haven, Connecticut

Received 24 June 2003/Accepted 29 September 2003

The molecular mechanisms underlying niche adaptation in bacteria are not fully understood. Primary infection by the pathogen group A streptococcus (GAS) takes place at either the throat or the skin of its human host, and GAS strains differ in tissue site preference. Many skin-tropic strains bind host plasminogen via the plasminogen-binding group A streptococcal M protein (PAM) present on the cell surface; inactivation of genes encoding either PAM or streptokinase (a plasminogen activator) leads to loss of virulence at the skin. Unlike PAM, which is present in only a subset of GAS strains, the gene encoding streptokinase (*ska*) is present in all GAS isolates. In this study, the evolution of the virulence genes known to be involved in skin infection was examined. Most genetic diversity within *ska* genes was localized to a region encoding the plasminogen-docking domain (β-domain). The gene encoding PAM displayed strong linkage disequilibrium ($P \ll 0.01$) with a distinct phylogenetic cluster of the *ska* β-domain-encoding region. Yet, *ska* alleles of distant taxa showed a history of intragenic recombination, and high intrinsic levels of recombination were found among GAS strains having different tissue tropisms. The data suggest that tissue-specific adaptations arise from epistatic coselection of bacterial virulence genes. Additional analysis of *ska* genes showed that ~4% of the codons underwent strong diversifying selection. Horizontal acquisition of one *ska* lineage from a commensal *Streptococcus* donor species was also evident. Together, the data suggest that new phenotypes can be acquired through interspecies recombination between orthologous genes, while constrained functions can be preserved; in this way, orthologous genes may provide a rich and ready source for new phenotypes and thereby play a facilitating role in the emergence of new niche adaptations in bacteria.

Beta-hemolytic group A streptococci (GAS) (*Streptococcus pyogenes*) are common bacterial pathogens whose host range is restricted to humans. The primary tissue sites for infection are the mucosal epithelial lining of the upper respiratory tract (URT) and the epidermal layer of the skin, where the organism can cause pharyngitis and impetigo, respectively. It is at these two superficial tissue sites that the organism is most successful in reproduction and transmission to new hosts. However, many strains of GAS differ widely in the ability to cause throat and skin infections, giving rise to the concept that there are distinct subpopulations of throat and skin strains (2, 8, 32, 47).

Both population and experimental studies have been used to better understand the molecular basis for tissue-specific adaptations among GAS. Organisms exhibiting high fitness for just one of the tissue sites have an increased frequency of tissue-specific adaptive alleles in their gene pool relative to the frequency in the other subpopulations. The *emm* pattern is a genetic marker that distinguishes many throat- and skin-tropic strains of GAS (5, 7, 15); the *emm* pattern is defined by the chromosomal arrangement of *emm* subfamily genes. *emm* pattern A-C strains are usually recovered from the URT, whereas *emm* pattern D isolates are mostly found in association with impetigo. As a group, the *emm* pattern E strains display no clear-cut preference for tissue site of infection. Despite niche separation, there is an ample flow of neutral housekeeping genes between *emm* pattern groups (27), and there are high rates of genetic recombination within the species as a whole (18). In instances where neutral housekeeping alleles are randomly distributed with respect to ecologically distinct populations (27), genetic variation that is strongly associated with the different populations may be directly responsible for adaptation to an ecological niche, and thus, *emm* gene products (or closely linked genes) may have a direct role in tissue tropism.

The *emm* genes encode M surface proteins, which display extensive heterogeneity in terms of structure and function (14, 20). More than 150 distinct *emm* types are recognized, where an *emm* type is based on nucleotide sequence differences near the 5′ end of the *emm* gene (17). Plasminogen (Plg)-binding group A streptococcal M protein (PAM) is encoded by an *emm* gene that is uniquely associated with *emm* pattern D strains (40). Many, but not all, *emm* pattern D isolates contain PAM, and a high-affinity Plg-binding site is localized to the central portion of the M protein surface fibril. By using an experimental model for impetigo that measures net bacterial reproductive growth at a superficial skin site, a role for PAM in impetigo has been demonstrated (41). When considered together, the experimental, epidemiological, and population genetics findings provide strong evidence that PAM contributes to the establishment of tissue tropism for the skin.

Host Plg presented in a PAM-bound form interacts with streptokinase, a GAS-secreted Plg activator, yielding bacterium-bound plasmin activity; plasmin is a broad-spectrum proteinase involved in blood clot dissolution and cellular migration. Insertional inactivation of the gene encoding strep-

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

† Present address: Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110.

TABLE 1. GAS isolates studied

Strain[a]	Site of isolation	*emm* type	*emm* pattern	β-Domain *ska* allele	β-Domain *ska* cluster or subcluster	PAM site[b]	Designation in dendrogram[c]
86-779	URT	1	A-C	*ska66*	2a	ND[e]	ska66
MGAS2109	URT	1	A-C	*ska66*	2a	Negative	
MGAS2110	Unknown	1	A-C	*ska67*	2a	Negative	ska67
SF370	Invasive	1	A-C	*ska66*	2a	Negative	
MGAS2120	Impetigo	1	A-C	*ska66*	2a	Negative	
MGAS2123	Impetigo	1	A-C	*ska66*	2a	ND	
1GL90	URT	3	A-C	*ska22*	2a	ND	
88-019	URT	3	A-C	*ska22*	2a	Negative	ska22a
1RP144	URT	5	A-C	*ska68*	1	ND	
Manfredo	URT	5	A-C	*ska68*	1	Negative	ska68a
1RP112	URT	6	A-C	*ska25*	2a	Negative	ska25a
D471	Unknown	6	A-C	*ska25*	2a	Negative	
A374	URT	12	A-C	*ska23*	1	ND	ska23a
25RS84	URT	14	A-C	*ska78*	1	ND	ska78
1GL217	URT	17	A-C	*ska25*	2a	Negative	ska25b
1RP268	URT	18	A-C	*ska82*	2a	Negative	ska82
1GL205	URT	19	A-C	*ska65*	1	Negative	
1RP118	URT	19	A-C	*ska65*	1	ND	ska65a
1RP284	URT	24	A-C	*ska65*	1	Negative	ska65b
11RS100	URT	26	A-C	*ska69*	1	Negative	ska69a
3RP70	URT	29	A-C	*ska65*	1	ND	
SS53	URT	37	A-C	*ska68*	1	ND	ska68b
19RS14	URT	39	A-C	*ska76*	1	Negative	ska76
SS642	URT	46	A-C	*ska69*	1	Negative	ska69b
A291	Unknown	51	A-C	*ska58*	1	Negative	ska58
D488	Impetigo	55	A-C	*ska22*	2a	Negative	ska22b
D306	Unknown	57	A-C	*ska37*	2a	ND	ska37
2RSC3	URT	38 & 40	A-C	*ska26*	1	ND	ska26
1RP31	URT	st1RP31	A-C	*ska65*	1	Negative	ska65c
SS1445	URT	st854	A-C	*ska71*	1	ND	ska71
10RS101	URT	32	D	*ska38*	1	Negative	ska38
29487	Impetigo	33	D	*ska39*	2b	Positive	ska39
C142	URT	34	D	*ska52*	1	ND	
A457	Unknown	36	D	*ska40*	1	Negative	ska40
1RS79	URT	42	D	*ska35*	1	Negative	ska35
D407	Invasive	43	D	*ska84*	2b	Positive	ska84
A946	Impetigo	52	D	*ska27*	2b	Positive	ska27b
ALAB49	Impetigo	53	D	*ska41*	2b	Positive	ska41
D795	Unknown	67	D	*ska51*	1	Negative	ska51a
D998	Unknown	70	D	*ska42*	2b	Positive	ska42
SS1098	Impetigo	71	D	*ska72*	1	Negative	ska72
SS1144	Impetigo	72	D	*ska45*	2b	Positive	ska45a
CT95-104	Invasive	80	D	*ska27*	2b	ND	
29689	Impetigo	83	D	*ska43*	2b	Positive	ska43
D964	Impetigo	86	D	*ska44*	2a	Positive	ska44
D821	Impetigo	91	D	*ska45*	2b	Positive	ska45b
D466	Unknown	93	D	*ska64*	2b	Positive	ska64
D502	Impetigo	93	D	*ska46*	2b	Positive	ska46
MGAS2111	Unknown	95	D	*ska30*	2b	Negative	ska30
D626	Impetigo	97.1	D	*ska53*	1	Negative	ska53
SS1434	Impetigo	98	D	*ska80*	2b	Positive	ska80
SS1433	Impetigo	99	D	*ska79*	1	Negative	ska79
D641	Impetigo	101	D	*ska54*	2b	Positive	ska54
MGAS308	Invasive	105	D	*ska23*	1	Negative	ska23b
29486	Impetigo	116	D	*ska47*	2b	Positive	ska47
MGAS341	Invasive	119	D	*ska28*	2b	Positive	ska28
SS1096	Impetigo	65 & 69	D	*ska73*	1	Negative	ska73
SS1497	Impetigo	st3757	D	*ska81*	2b	Positive	ska81
D997	Unknown	st4973	D	*ska29*	2b	Negative	ska29
D432	Unknown	stD432	D	*ska27*	2b	Positive	ska27c
D631	Impetigo	stD631	D	*ska48*	2b	Positive	ska48
D633	Impetigo	stD633	D	*ska49*	2b	Positive	ska49
89-465	URT	2	E	*ska57*	1	Negative	ska57a
CT98-529	Invasive	4	E	*ska62*	1	ND	ska62
29740	Impetigo	8	E	*ska20*	1	Negative	ska20
D733	Unknown	9	E	*ska63*	1	ND	ska63a
CT95-126	Invasive	11	E	*ska60*	1	ND	ska60
MGAS275	Invasive	22	E	*ska55*	1	Negative	ska55
D316	Invasive	25	E	*ska21*	1	ND	ska21
CT95-189	Invasive	28	E	*ska31*	1	ND	ska31

Continued on following page

TABLE 1—*Continued*

Strain[a]	Site of isolation	*emm* type	*emm* pattern	β-Domain *ska* allele	β-Domain *ska* cluster or subcluster	PAM site[b]	Designation in dendrogram[c]
B737	Impetigo	49	E	*ska32*	1	Negative	ska32a
29454	Impetigo	58	E	*ska24*	1	Negative	ska24a
4500-S	Impetigo	60	E	*ska24*	1	ND	ska24b
A956	Unknown	63	E	*ska50*	1	Negative	ska50a
ALAB53	Impetigo	66	E	*ska56*	1	ND	ska56
5552-S	Impetigo	68	E	*ska33*	1	Negative	ska33
86-809	URT	75	E	*ska34*	1	Negative	ska34
CT95-159	Invasive	77	E	*ska61*	1	ND	ska61
D812	Impetigo	78	E	*ska74*	1	ND	ska74
29665	Impetigo	81	E	*ska36*	1	ND	ska36
D424	Invasive	89	E	*ska63*	1	ND	ska63b
89-456	URT	90	E	*ska75*	1	Negative	ska75
4426-S	Impetigo	92	E	*ska50*	1	Negative	ska50b
CT95-122[d]	Invasive	94	E	*ska27*	2b	ND	ska27a
CT95-169	Invasive	102.1	E	*ska83*	1	Negative	ska83
6250-S	Impetigo	110	E	*ska51*	1	Negative	ska51b
1RP18rn	URT	44 & 61	E	*ska77*	1	ND	ska77
5569-S	Impetigo	44 & 61	E	*ska59*	1	Negative	
D938	Impetigo	44 & 61	E	*ska32*	1	Negative	ska32b
MGAS2140	URT	st833	E	*ska57*	1	Negative	ska57b

[a] For each strain, additional information concerning the year and place of isolation, as well as disease association, can be found at http://www.cdc.gov/ncidod/biotech/strep/strepindex.html (for SS strains) or www.mlst.net (for all other strains).
[b] As established by Svensson et al. (40) and/or by predicted amino acid sequence analysis. Although all strains underwent *emm* typing, only the strains whose saved sequences included data up to the C repeat region were analyzed for PAM.
[c] Dendrogram presented in Fig. 4.
[d] *emm* type 94 = *emm* type 13W.
[e] ND, not determined.

tokinase (*ska*) also leads to attenuated infection in the experimental model for GAS impetigo (41). It is postulated that during impetigo lesion formation, the combined action of streptokinase and PAM-bound Plg leads to fibrinolysis, which retards scabbing and prevents the lesion from drying out. This, in turn, expands the window of opportunity for GAS reproduction and transmission to new hosts.

In this report, the evolution of streptococcal virulence genes involved in tissue-specific adaptations is examined in depth. The nucleotide sequences of *ska* genes derived from GAS isolates characterized for the presence of PAM were determined, and phylogenetic analysis was performed.

MATERIALS AND METHODS

Bacterial strains. The 90 GAS isolates which we studied are listed in Table 1. Nearly all isolates known to be recovered from the URT are also known to be disease associated (pharyngitis and/or nonsuppurative sequelae) and not associated with asymptomatic carriage. The *emm* pattern was determined by a PCR-based method, as previously described (5). The 34 group C streptococci (GCS) and group G streptococci (GGS) isolated from humans were described previously (26). The *emm* type was ascertained by previously described methods (5).

Primers. The following oligonucleotide primers were used for PCR amplification and/or nucleotide sequence determination: for *ska*, 5′-AACCTTGCCGACCCAACCTGT-3′ (SKNF2), 5′-TTATTCTAATAATGGGGATTGAAACTTAA-3′ (SKNF3), 5′-TGAAACTTAACTTTTAGGAGGTTT-3′ (SKNF4), 5′-ATCGCAGTCACTTGAAACTGTTCAC-3′ (SKNF5), 5′-GTGAACAGTTTCAAGTGACTGCGAT-3′ (SKNR2), 5′-GCTGTTAAGAGCTGCTCGCTT-3′ (SKNR3), 5′-AATCTCATCGTTTTAGAAGATCG-3′ (SKNR4), 5′-AATCTCATCRTTTTAGAAGATCG-3′ (SKNR5), and 5′-ACAGGTTGGGTCGGCAAGGTT-3′ (SKNR6); for *dppA*, 5′-TCAAATGATGTGCGCGGCTTAT-3′ (DPPAF) and 5′-ATAAGCCGCGCACATCATTTGA-3′ (DPPAR); for *lmb*, 5′-TTCGGCTTGAAACAACTTGGTATCTCGGG-3′ (LMBF) and 5′-CCCGAGATACCAAGTTGTTTCAAGCCGAA-3′ (LMBR); for *nrd*, 5′-TCWGGCAACAAAAACTTTAAYCAYCAGTATT-3′ (NRDF2) and 5′-AATACTGRTGRTTAAAGTTTTTGTTGCCWGA-3′ (NRDR2); for *pabP*, 5′-GACCTCAACTATTGTTGGTGACCTCAA-3′ (PABPF) and 5′-TTGAGGTCACCAACAATAGTTGAGGTC-3′ (PABPR); and for *scpA*, 5′-ATCTTGCTCAATGCACAATCAG-3′ (SCPAF) and 5′-CTGATTGTGCATTGAGCAGAT-3′ (SCPAR). PCR were performed for large and small amplification products as previously described (6).

Statistical and computational analyses. (i) Phylogenetic trees. All trees were constructed by the neighbor-joining method by using MEGA, version 2.1; the Kimura two-parameter distance measure was used for nucleotide sequences, and the Poisson-corrected distance measure was used for amino acid sequences. The maximum-likelihood method was employed for trees analyzed by PAML (phylogenetic analysis by maximum likelihood) (see below) by using PAUP, version 4.0 beta 10. The desired evolutionary model of DNA substitution and the parameters were optimized by using hierarchical likelihood ratio tests (24) with MODELTEST, version 3.0 (36).

(ii) Gene conversion. Geneconv, version 1.81, was used to detect gene conversion events among full-length *ska* alleles; the default settings were used (37). Bonferroni-corrected Karlin-Altschul *P* values that were less than 0.05 are reported below for global fragments. The analysis included full-length *ska* sequences ($n = 13$), as defined in Fig. 1; strain 89-465 was not included because of the presence of an indel, and strain D998 was not included because it was nearly identical to D633.

(iii) Cluster analysis. Cluster analysis, based on seven housekeeping gene alleles, was performed by using the unweighted pair group method with arithmetic averages and the percent disagreement distance measure (Statistica, version 5.5; StatSoft, Tulsa, Okla.). The data are presented below as a dendrogram.

(iv) PAML. A maximum-likelihood approach was used to examine selection pressures acting on *ska*. The ratios of nonsynonymous nucleotide substitutions (d_N) to synonymous nucleotide substitutions (d_S) (ω ratios) were determined codon by codon by using several models of codon substitution that differ in how the ω ratios are allowed to vary along the sequence. Six models of codon substitution were used (see below). All models were implemented with the codeml program of the PAML package (version 3.13) (11, 42, 44, 45, 49, 50). Nested models were compared by using the likelihood ratio test; in this test twice the difference in log likelihood (ln *L*) between two models was compared to the value obtained under a χ^2 distribution, and the degrees of freedom was equal to the difference in the number of parameters used in each model. Positive selection could be inferred when a group of codons having a ω ratio of more than 1 was identified and the likelihood of the codon substitution model in question was

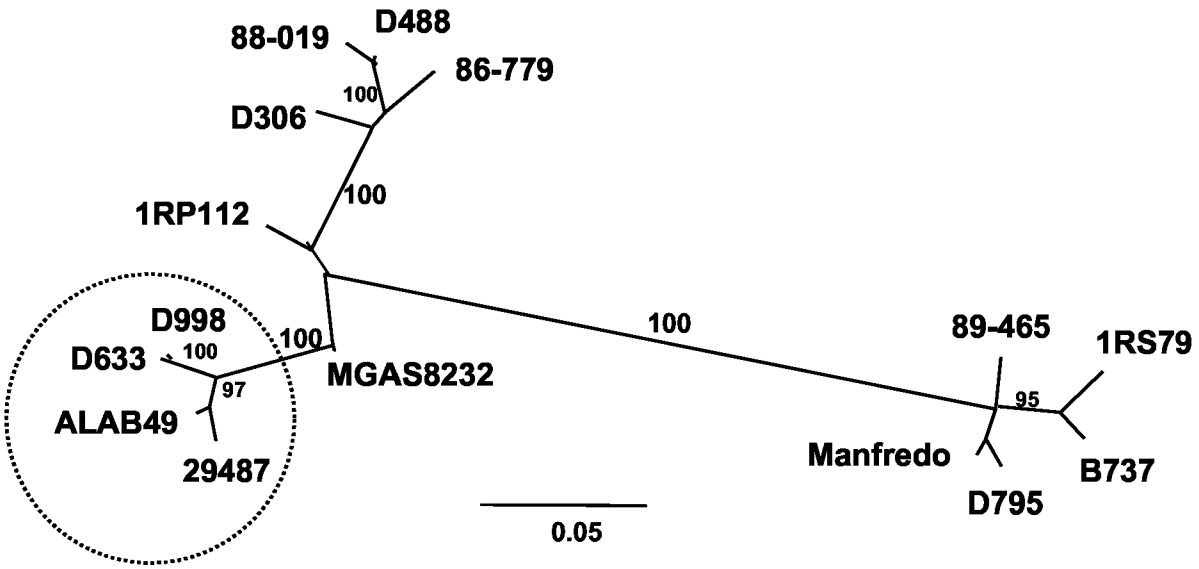


FIG. 1. Phylogenetic tree for full-length *ska* alleles. The relationships of 1,320-bp nucleotide sequences of the *ska* gene derived from 15 strains of GAS are indicated by an unrooted radial tree constructed by the maximum-likelihood method, in which the rate matrix was optimized to a submodel of GTR+G+I (K81uf). Bootstrap values of ≥90% (1,000 replicates) are indicated at the nodes. Taxon designations indicate GAS strains that are listed in Table 1, except for MGAS8232 (SPyM18-2042; GenBank accession no. AE009940). The *ska* genes derived from strains MGAS315 (SPyM3-1698; GenBank accession no. AE014074) and SF370 (SPy1979; GenBank accession no. AE004092) exhibit 100% nucleotide identity with the genes of strains 88-019 (*emm* type 3) and 86-779 (*emm* type 1), respectively. Bar = 0.05 substitution per site. The tree topology was very similar when the neighbor-joining method was used. The *ska* lineage corresponding to PAM-positive strains is indicated by the dotted circle. The GenBank accession numbers for 14 new *ska* sequences are AY234128 to AY234141.

significantly higher ($P < 0.01$) than the likelihood of a nested model which did not take positive selection into account. Bayesian methods implemented (automatically) in PAML identify any codons under positive Darwinian selection.

The M0 model assumes that all codons are subject to the same selection pressure, so that a single ω ratio value is estimated. Model M1 divides codons into two categories; one category represents the codons that are invariant (p_0), with ω_0 fixed at 0, and the other represents codons that are neutral (p_1), with ω_1 set to 1. The M2 model accounts for positive selection by addition of a third category of codons (p_2) with ω_2, which can take on any value (including 1) estimated from the data; however, this model cannot simultaneously account for sites with $0 < \omega$ ratio < 1 and sites with an ω ratio of >1. The M3 model estimates ω ratios for three codon site classes and provides a more sensitive test for positive selection, such that all ω ratios are estimated from the data and all values may be greater than 1. The M7 model uses a discrete β distribution, whose shape varies depending on the parameters p and q, to model ω ratios of codons; in the M7 model, no class of codons can have an ω ratio of >1. Model M8 also uses a β distribution, but an extra class of codons is incorporated, in which the ω ratio can be more than 1. A likelihood ratio test of a comparison of the M7 and M8 models is much less affected by the presence of recombination than tests for the other comparisons (1).

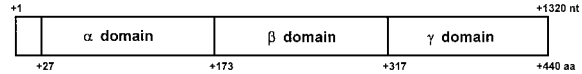
(v) Tests for independence. Tests for independence, used to establish nonrandom relationships (linkage disequilibrium), were performed with Fisher's exact test (DnaSP, version 3.52).

RESULTS

Phylogeny of streptokinase genes. The complete *ska* sequence was determined for 14 strains of GAS. The findings are shown in a maximum-likelihood phylogenetic tree in Fig. 1, which includes previously published data for an additional *ska* allele, so that there were 15 distinct taxa. Three well-supported, major sequence lineages are evident. The relationships of strains 1RP112 (*emm* type 6) and MGAS8232 (*emm* type 18) to the three major *ska* lineages are less certain. The four GAS strains containing the high-affinity PAM have *ska* alleles that form a single cluster.

Based on extensive structural and functional studies, streptokinase is recognized as having three principal domains, α, β, and γ (46). The lengths of the three domains are approximately equal (146, 144, and 123 amino acid residues, respectively) (Fig. 2). The β-domain of streptokinase displayed the highest level of predicted amino acid sequence divergence when the 15 *ska* genes (Fig. 1) were compared. During plasmin formation, the β-domain of at least one form of streptokinase has direct molecular contact with Plg and docks Plg as an initial step in the formation of the streptokinase-plasmin(ogen) activation complex in the fluid phase (10, 31).

Relationships among streptokinase β-domain, PAM, and tissue site preference. Given the tight association observed



Maximal % nt divergence	14.8	38.1	10.1
Maximal % aa divergence	23.1	54.3	17.2

FIG. 2. Domain structure of streptokinase. The sequence positions of the three principal domains of streptokinase (α, β, and γ) are illustrated (46). The maximal nucleotide (nt) sequence divergence and the maximal amino acid (aa) sequence divergence between *ska* alleles shown in Fig. 1 are indicated for each of the three major streptokinase domains. Since strain 89-465 has a deletion within the α-domain, it was not included in the α-domain analysis.

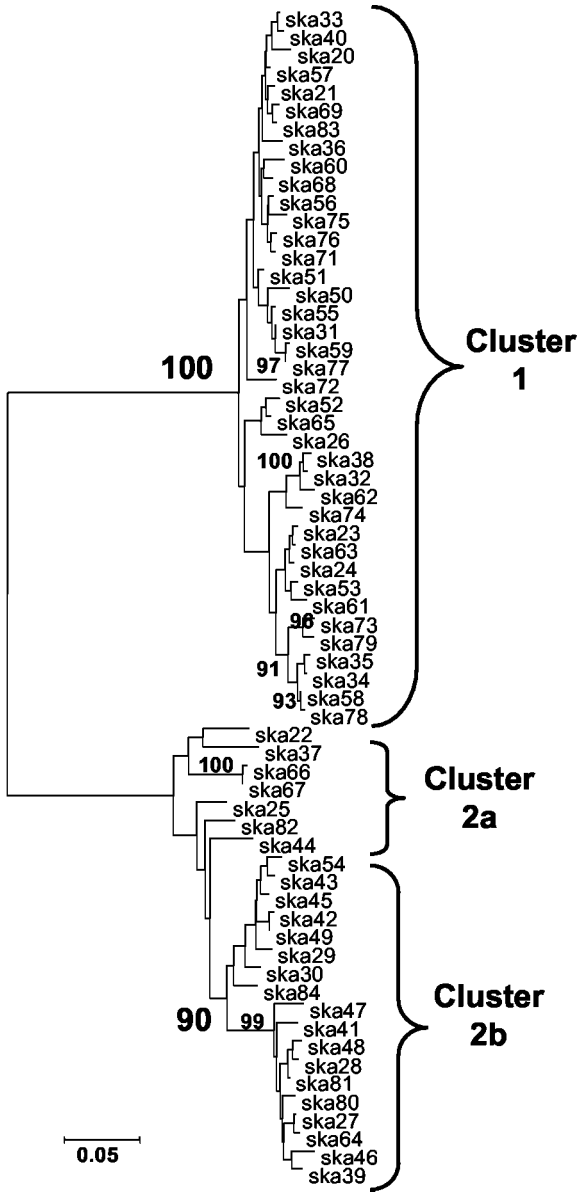


FIG. 3. Phylogenetic tree for the β-domain-encoding region of *ska*. The relationships of the nucleotide sequences of a 423-bp portion of *ska* encoding amino acid residues 173 through 316 of the streptokinase protein (the first residue of the leader peptide is designated residue 1) (Fig. 2), derived from 90 strains of GAS, are indicated by a neighbor-joining tree. For visual clarity, the tree is midpoint rooted. Bootstrap values of ≥90% (500 replicates) are indicated at the nodes. The designations indicate the *ska* alleles, which are listed in Table 1. Bar = 0.05 substitution per site. The GenBank accession numbers for 64 new partial *ska* sequences are AY234261 to AY234324.

between PAM and one of the major *ska* lineages (Fig. 1), it was of interest to ascertain whether a strong association between PAM and the *ska* lineage extended to the broader GAS population. Since the β-domain of streptokinase displays the highest level of sequence heterogeneity (Fig. 2) and thereby makes a large contribution to the phylogenetic signal (Fig. 1), this domain was chosen for further in-depth studies.

The phylogeny of the portion of the *ska* locus encoding the β-domain was examined for GAS strains representing a broad spectrum of genetic diversity (Fig. 3). For 90 GAS isolates, representing 78 *emm* types, 64 distinct (partial) *ska* alleles encoding the β-domain were identified (Table 1). Two major sequence clusters that had strong bootstrap support were clearly evident (clusters 1 and 2) (Fig. 3). Both major clusters, clusters 1 and 2, contained several smaller subclusters of alleles having strong bootstrap support.

TABLE 2. Relationship between β-domain form of *ska*, *emm* pattern group, and PAM site

emm pattern group	PAM site	No. of unique strains in *ska* β-domain clusters[a]				
		Cluster 1	Cluster 2	Total for clusters 1 and 2	Subcluster 2a	Subcluster 2b
A-C	NA[b]	14	8	22	8	0
D	Negative	9	2	11	0	2
D	Positive	0	19	19	1	18
E	NA	26	1	27	0	1

[a] A unique strain was defined as an isolate having a unique *emm* type and less than five of seven housekeeping alleles in common with another isolate of that *emm* type. Only pattern D strains whose PAM status was known were included.
[b] NA, not applicable.

PAM is the product of a subset of *emm* genes and is defined by the ability to bind Plg with high affinity. Precise mapping of the Plg-binding site within PAM led to recognition of consensus sequences at both the amino acid (9) and nucleotide (40) levels. Previous studies of a set of 81 genetically diverse strains of GAS showed that high-affinity binding of Plg was restricted to *emm* pattern D strains, and furthermore, there was a strong correlation between Plg binding and the PAM consensus sequence (40). In this study, many additional GAS strains were included in the analysis; data in Table 1 show that the PAM site was absent from all 33 *emm* pattern A-C and E isolates examined, confirming our previous findings with a different strain set. For *emm* pattern D strains ($n = 30$) (Table 1), a neighbor-joining tree was constructed by using input amino acid sequences corresponding to the amino terminus of the predicted mature M protein, up to the C repeat region (4); PAM-positive and PAM-negative sequence clusters were delineated by a branch point having strong bootstrap support (99% confidence) (data not shown). However, within the PAM-positive cluster, 11 of the 17 branch points had more than 50% bootstrap support, which was indicative of a high degree of sequence diversity among PAM molecules derived from different GAS strains. The partial amino acid sequences of 19 M proteins having a PAM consensus sequence, as indicated in the sequence alignment, revealed the A1 and/or A2 repeat region (9) and confirmed the relationship between the predicted PAM consensus sequence and the percentage of Plg bound (40).

The relationship between the β-domain-encoding region of *ska* and the *emm* pattern marker for tissue site preference was examined. Each of the three *emm* pattern groups (pattern A-C [throat preference], pattern D [skin preference], and pattern E [no preference]) was represented by numerous strains having cluster 1 alleles (Tables 1 and 2). Strikingly, all nine of the *emm* pattern D isolates having a cluster 1 *ska* allele lacked PAM.

Cluster 2 *ska* alleles were found in many *emm* pattern A-C

and D strains, whereas only 1 of the 27 pattern E strains examined had a cluster 2 *ska* allele (Table 2). Furthermore, the vast majority (19 of 21 strains; 90%) of the *emm* pattern D strains harboring a cluster 2 *ska* allele also had PAM. Based on the branches of the phylogenetic tree having strong bootstrap support, 18 of the 19 PAM-positive strains were shown to harbor an *ska* allele falling in the major *ska* subcluster that was designated subcluster 2b (Fig. 3). For all eight *emm* pattern A-C strains harboring a cluster 2 *ska* allele, the *ska* allele belonged to the other subcluster, designated subcluster 2a.

In summary, nearly all PAM-positive *emm* pattern D strains (18 of 19 strains; 95%) had a subcluster 2b *ska* allele (Table 2). The vast majority of *emm* pattern D strains lacking a PAM site had a cluster 1 *ska* allele (9 of 11 strains; 82%). *emm* pattern D strains harboring *ska* cluster 1 genes also tended to be strains belonging to rarely recovered *emm* types (www.cdc.gov/ncidod/biotech/strep/strepindex.html). The association between subcluster 2b *ska* alleles and *emm* pattern D strains with a PAM site was highly significant ($P = 0.00004$, as determined by Fisher's two-tailed exact test), which was indicative of a strong linkage disequilibrium. None of the strains harboring a subcluster 2b *ska* allele was known to be recovered from the URT (Table 1).

Epistasis and linkage of subcluster 2b *ska* and *pam*. The finding that there is a strong linkage disequilibrium between the subcluster 2b form of the streptokinase β-domain and the presence of PAM strongly suggests that the corresponding genotypes are coinherited. Coinheritance could arise by clonal descent within a population exhibiting low rates of recombination and/or through tight physical linkage (i.e., close proximity on the genome) between the *ska* and *emm* (*pam*) loci. Alternatively, coinheritance could be maintained by epistasis, driven by phenotypic interactions between streptokinase and PAM that give rise to an essential adaptive function. Epistasis can occur against a background of high levels of genetic recombination.

Statistical tests were used to estimate the level of recombination within the GAS population by examining neutral loci. The genetic background of each of the 90 GAS isolates (Table 1) was defined for allelic profiles (sequence types [ST]) based on seven housekeeping loci (16), which yielded 87 unique *emm* type-ST combinations (data not shown). Previous studies have shown that the associations between housekeeping loci of GAS are random, based on a maximum-likelihood method for measuring the extent of congruency between housekeeping gene tree topologies (18, 27). As observed in previous studies performed with slightly different sets of GAS strains (18, 27) and a linkage distance cutoff of 0.55, no significant congruence between gene trees was observed for this particular set of GAS isolates, and the differences in the likelihoods of the trees fell within the 99th percentile of the random distribution of random tree topologies for all 42 possible pairwise comparisons of housekeeping genes (data not shown). Therefore, when deep phylogenetic relationships were considered, the rates of recombination among housekeeping loci are relatively high for this particular set of GAS isolates. There is no evidence that throat- and skin-tropic strains of GAS comprise distinct evolutionary lineages (27).

Despite the strong linkage disquilibrium observed between subcluster 2b *ska* forms and PAM, several individual *ska* alleles ($n = 12$), as defined by the β-domain-encoding region, show a history of horizontal movement between GAS strains having distantly related STs (linkage distance, >0.6) (Fig. 4). In addition, for one clone, as defined by seven of seven identical housekeeping alleles, there were isolates that had highly divergent *ska* alleles (*ska44* and *ska54*); this finding is also indicative of horizontal movement of *ska* between different GAS strains. Of the 18 strains having a PAM site, a subcluster 2b *ska* allele, and unique *emm* type-ST combinations, 12 differed from all other isolates by a linkage distance of >0.6 (Fig. 4). Although some of the isolates having both PAM and a subcluster 2b *ska* allele are close genetic relatives, the majority of the strains are genetically distant in terms of their neutral housekeeping genes.

When a statistical test for detecting gene conversion was used (37), numerous examples of intragenic recombination between full-length *ska* genes (Fig. 1) were evident. A total of 53 pairwise global inner-sequence fragments were found with *P* values of <0.05, indicating that there were as many as 53 intragenic recombinational events. Most of the predicted gene conversion events (44 of 53 events; 83%) had (double) crossover sites that were contained within one of the three major structural domains (as shown in Fig. 2). Crossover sites spanning the α-domain–β-domain junction were most likely to involve alleles corresponding to the PAM-positive cluster and strain MGAS8232, whereas crossover sites spanning the β-domain–γ-domain junction involved alleles of numerous distant taxa (data not shown). The gene conversion findings for *ska* are consistent with the findings for housekeeping genes, indicating that GAS display high levels of genetic recombination.

The complete genome sequences of several GAS strains, containing either a cluster 1 or subcluster 2a *ska* allele, show that the distance between *emm* and *ska* ranges from ~33 to 38 kb (3, 19, 33, 39; www.sanger.ac.uk). By using a PCR-based mapping approach, the genomic content and distance between the *emm* and *ska* loci in a PAM-positive, subcluster 2b *ska*-positive, *emm* pattern D strain (Alab49) were found to be very similar to those of the GAS strains whose complete genome sequences are known (data not shown).

Although the genes encoding PAM and streptokinase are not too far apart on the genome, the combined findings for random associations between housekeeping genes, intragenic recombination between *ska* genes of different strains, the horizontal movement of *ska* alleles to distant strain backgrounds, and high sequence diversity among PAM from different strains argue strongly against coinheritance due to physical proximity. In summary, the β-domain-encoding region of subcluster 2b *ska* maintains strong linkage disequilibrium with PAM-positive *emm* pattern group D. Combined with experimental evidence that streptokinase and PAM play key roles in impetigo (41), the findings suggest that the linkage between PAM and the subcluster 2b form of the streptokinase β-domain arises from strong coselective pressures due to epistasis.

Positive selection within the streptokinase β-domain. The relative proportion of d_N and d_S, leading to a change and no change in amino acid residues, respectively, can provide insight into the role of natural selection in the evolution of a gene. It is widely assumed that ω ratios of more than 1 signify diversifying (positive) selection. The average ω ratio for full-length *ska* genes (Fig. 1) is 0.449, suggesting that purifying (negative)

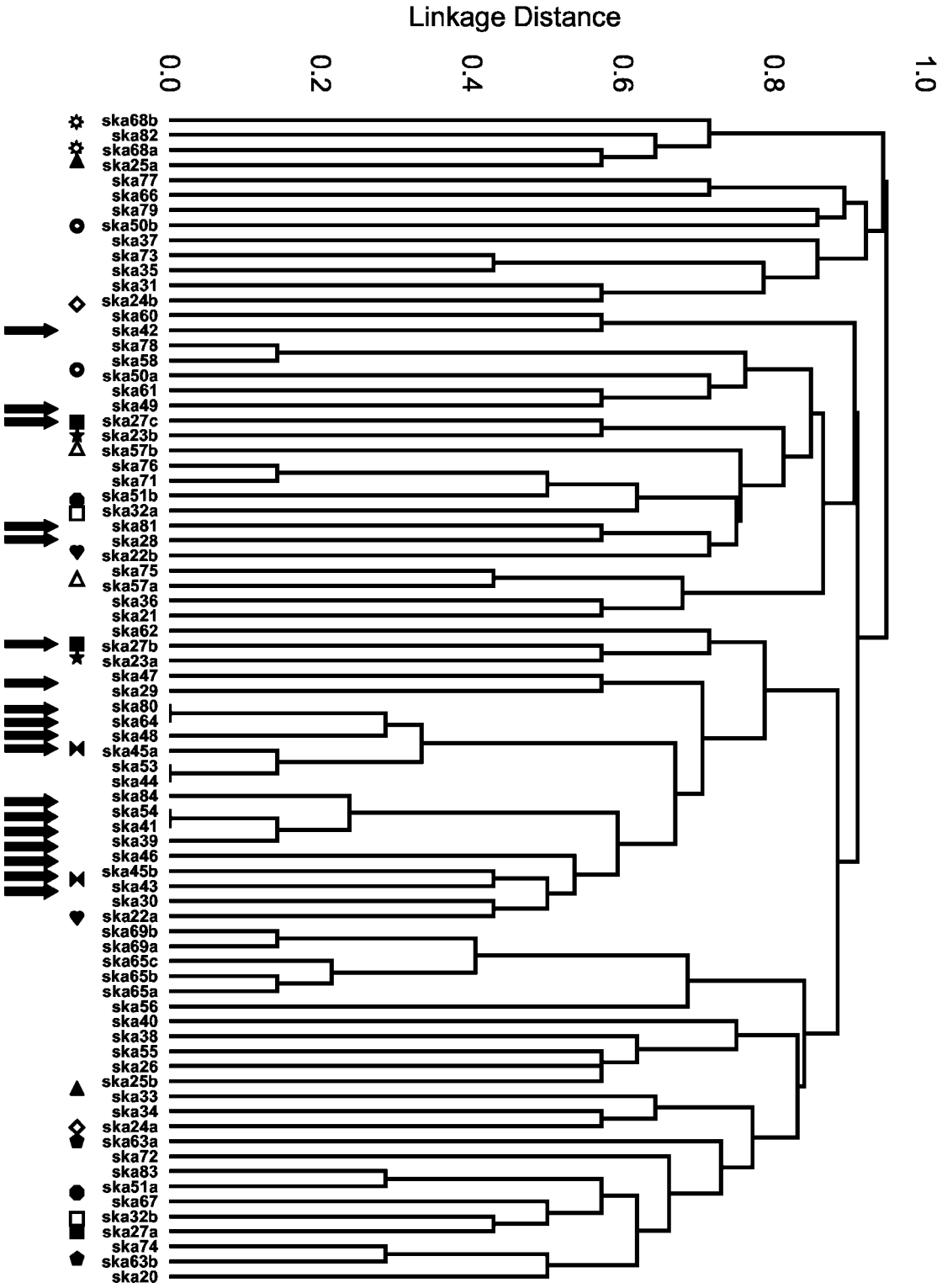


FIG. 4. Unweighted pair group method with arithmetic averages dendrogram based on housekeeping loci. A matrix of pairwise differences in allelic profiles between strains was constructed based on the proportion of housekeeping loci having shared alleles (16). The relationships between housekeeping gene allelic profiles at seven loci are shown for 78 GAS strains having unique *emm* type-ST combinations. For the 90 GAS isolates listed in Table 1 having 87 unique *emm* type-ST combinations, clonal complexes having the same *emm* type are reduced to one representative strain; clonal complexes are defined as groups of clones that share five or more of the seven housekeeping alleles. The branch labels indicate the *ska* allele corresponding to the β-domain-encoding region (Fig. 3; Table 1) for each GAS strain. The various symbols indicate sets of identical *ska* alleles that are distributed among GAS strains that differ at three or more housekeeping loci. The arrows indicate branch tips representing isolates having both a PAM site and subcluster 2b *ska* allele ($n = 18$). Isolates having identical *emm* types and STs also tend to have identical or nearly identical *ska* alleles (Table 1), as follows: *emm1*-ST28, *ska66*; *emm5*-ST99, *ska68*; *emm6*-ST37, *ska25*; and *emm44/61*-ST31, *ska59* and *ska77*. Isolates that have the same *emm* type and differ at only one or two housekeeping alleles (clonal complexes) also tend to have identical *ska* alleles (*emm3-ska22*, *emm19-ska65*, *emm1-ska66*). The multilocus sequence typing raw data were published previously for all isolates except the nine strains whose designations begin with SS (16).

TABLE 3. Parameter estimates for maximum-likelihood analysis of selection pressures acting on streptokinase

Model[a]	ln *L*	d_N/d_S	Estimates of parameters	Likelihood ratio test (*P* value)	Amino acid residues subject to positive selection ($P > 0.99$)[b]
M0 (one-ratio)	−4478.904	0.449	$\omega = 0.449$		
M1 (neutral)	−4383.673	0.404	$p_0 = 0.596$, ($p_1 = 0.404$)		
M2 (selection)	−4354.346	0.625	$p_0 = 0.574$, $p_1 = 0.384$, ($p_2 = 0.041$), $\omega_2 = 5.831$		30, 164, 202, 204, 208, 236, 279, 280, 282, 286, 287, 290, 302, 308, 332, 414
M3 (discrete)	−4353.302	0.581	$p_0 = 0.646$, $p_1 = 0.314$, ($p_2 = 0.040$), $\omega_0 = 0.045$, $\omega_1 = 1.049$, $\omega_2 = 5.538$	M3 vs M0 (<0.0001), M3 vs M1 (<0.0001)	30, 38, 74, 83, 11, 147, 164, 173, 174, 183, 187, 189, 191, 195, 197, 199, 202, 204, 206, 207, 208, 220, 226, 228, 231, 234, 236, 243, 263, 270, 271, 273, 279, 280, 282, 284, 286, 287, 289, 290, 302, 308, 324, 329, 332, 414
M7 (beta)	−4382.473	0.341	$p = 0.057$, $q = 0.114$		
M8 (beta & ω)	−4353.360	0.574	$p = 0.082$, $q = 0.151$, $p_0 = 0.956$, ($p_1 = 0.045$), $\omega = 5.193$	M6 vs M7 (<0.0001)	30, 164, 173, 187, 202, 204, 208, 236, 271, 279, 280, 282, 286, 287, 290, 302, 308, 332, 414

[a] The values for kappa (transition/transversion rate ratio) and tree length (number of nucleotide substitutions along the tree per codon) were fairly homogeneous for each model.
[b] Numbering begins at the amino terminus of the leader sequence (the leader is 26 amino acid residues long).

selection has been a major force in *ska* gene evolution when all codons are considered together. However, this ratio does not consider individual codons, and it was of interest to ascertain whether specific codons of *ska* were under diversifying selection. By using a statistical approach, ω ratios were determined codon by codon. Maximum-likelihood analysis of the selection pressures acting on *ska* by using the tree topology of Fig. 1 and allowing for heterogeneous ω ratios among sites provided evidence that there has been diversifying selection within streptokinase (Table 3).

Parameter estimates from the discrete (M3) model suggest that 64.6% of the sites are under purifying selection ($\omega_0 = 0.045$), 31.4% are under very weak diversifying selection ($\omega_1 = 1.049$), and 4.0% are under strong diversifying selection ($\omega_2 = 5.538$) (Table 3). All models that allow for positively selected sites (M2, M3, and M8) indicated that there are such sites, and ~4% of the codons are under strong positive selection (ω ratio, >5).

Since *ska* genes were found to undergo intragenic recombination and tests for positive selection by the maximum-likelihood method assumed a phylogenetic tree, the ω ratio was also estimated from a star phylogeny (45). For the full-length *ska* genes of the strains shown in Fig. 1 but with a tree in which all sequences diverge from a single node, there was still evidence of significant positive selection. For the M3 model with the star phylogeny, 10.1% of amino acid sites were under strong diversifying selection ($\omega_2 = 5.134$), which was less conservative than the values estimated with the maximum-likelihood tree (Table 3). Therefore, intragenic recombination between distantly related *ska* genes does not appear to weaken the findings for positively selected codons.

The Bayes approach can be used to identify specific amino acid sites likely to be under positive selection. For the M3 model (ω ratio, >1), 46 codons exceed the 99% posterior probability threshold (Table 3). For the M2 and M8 models, 16 and 19 codons, respectively, exceeded this threshold. All of the positively selected codons identified by the M2 model were a subset of the codons identified by both the M3 and M8 models; all codons identified by the M8 model were a subset of the M3 model codons. M3 is more sensitive than the other models and detected more codons under positive selection, because it incorporates more codon site classes (50).

Of the 46 positively selected codons suggested by the M3 model (Table 3), 35 (76%) are in the β-domain-encoding region and comprise 24% of the total β-domain residues. For the M2 and M8 models, 75 and 79% of the positively selected sites, respectively, lie within the β-domain-encoding region. Thus, diversifying selection appears to have played a major role in the evolution of the streptokinase β-domain. The strong purifying selection observed within the α- and γ-domains may be the consequence of functional constraints.

Lineage-specific, fixed amino acid differences in the β-domain. Of the codons identified to be under diversifying selection based on the 15 full-length *ska* alleles (Table 3), the β-domain-encoding regions of 64 partial *ska* alleles (Fig. 3) were assessed for fixed amino acid differences between any two of the three major sequence (sub)clusters. At 11 amino acid sites, all subcluster 2a and 2b predicted products were identical to each other, but they differed from all cluster 1 *ska* products (residues 174, 183, 191, 195, 197, 199, 208, 226, 228, 231, and 234). Cluster 1 and subcluster 2a forms also displayed a fixed amino acid difference at residue 243.

At three codon sites (residues 279, 280, and 282), all subcluster 2b products have a different amino acid sequence than all subcluster 2a products. Site 282 contains a Lys in subcluster 2a streptokinase forms that has been shown by site-specific mutagenesis to be important for Plg activation in the fluid phase (10).

In summary, at least some of the amino acid residues that evolved under diversifying selection (Table 3) also appear to have contributed to the lineage-specific differences observed for the β-domain-encoding region of *ska* (Fig. 3).

Interspecies spread of *ska*-related alleles. Human isolates of GCS and GGS, which are classified as *Streptococcus dysgalactiae* subsp. *equisimilis*, are the closest known genetic relatives of GAS. GCS and GGS are considered to be more commensal-

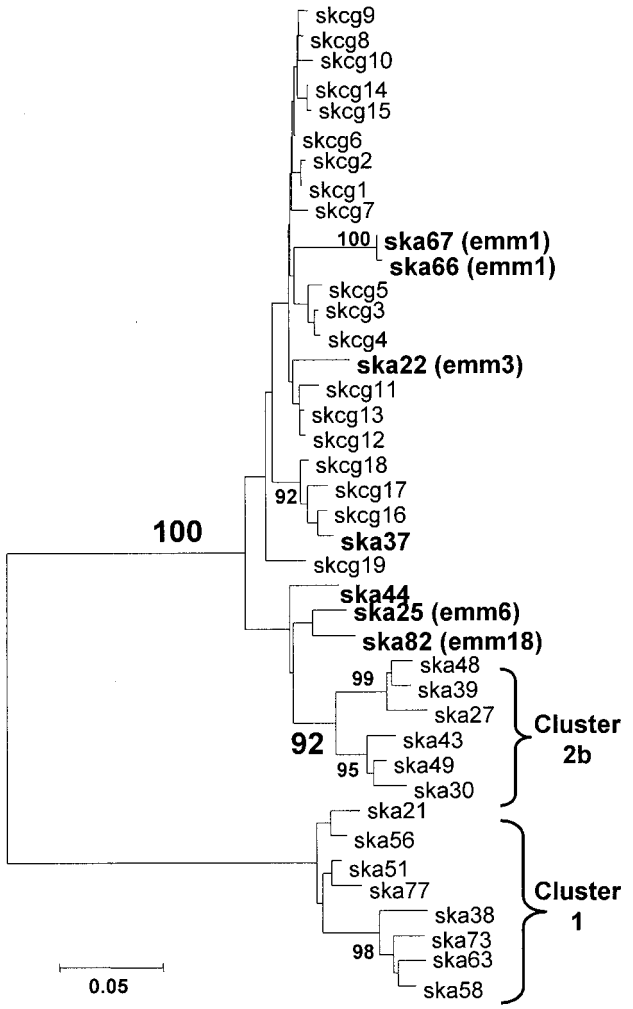


FIG. 5. Phylogenetic tree based on the β-domain-encoding regions of *skcg* and *ska*. The relationships of the nucleotide sequences of a 423-bp portion of *skcg* encoding the β-domain, derived from 34 strains of GCS and GGS, are indicated by a neighbor-joining tree that was obtained by using the Kimura two-parameter distance measure. Bootstrap values of ≥90% (1,000 replicates) are indicated at the nodes. Also included in the analysis were 21 *ska* alleles from cluster 1 and subclusters 2a and 2b (Fig. 3). Subcluster 2a *ska* alleles are indicated by boldface type. The designations indicate *skcg* and *ska* alleles. Bar = 0.05 substitution per site. The GenBank accession numbers for 19 new partial *skcg* sequences are AY234242 to AY234260.

like than GAS, primarily inhabiting the URT, although GCS and GGS can be recovered in association with disease. Since GAS and *S. dysgalactiae* subsp. *equisimilis* show evidence for recent horizontal exchange of housekeeping alleles (26), it was of interest to assess the relationships between the major phylogenetic lineages of GAS *ska* genes and orthologous streptokinase genes derived from GCS and GGS (designated *skcg*).

The nucleotide sequences the β-domain-encoding region of the *skcg* genes of 34 human isolates of GCS and GGS, representing 34 distinct STs as defined by housekeeping alleles (26), were determined. The results obtained are shown in a neighbor-joining tree in Fig. 5 and include results for several cluster 1 and subcluster 2a and 2b *ska* alleles (Fig. 3) for comparison. Among the 34 GCS and GGS isolates, 19 distinct *skcg* alleles corresponding to the β-domain-encoding region were identified. The levels of nucleotide sequence identity among the 19 *skcg* alleles ranged from 94.8 to 99.8%, indicating that there was a relatively high degree of homogeneity. This finding is in marked contrast to the data for the *ska*-encoded β-domains of GAS, in which the maximal nucleotide sequence divergence exceeds 40% (divergence between a cluster 1 allele and a subcluster 2b allele) (data not shown).

Several subcluster 2a alleles of *ska* were more closely related to *skcg* than to known cluster 1 or subcluster 2b *ska* alleles. All subcluster 2a *ska* alleles except one were detected in strains belonging to the throat-tropic, *emm* pattern A-C subpopulation (Tables 1 and 2). The *ska22* allele (subcluster 2a) (Fig. 5), which is derived from *emm3* isolates, is more closely related to *skcg* alleles (96.7% nucleotide identity to *skcg12*) than to any other known *ska* allele. The *ska66* and *ska67* subcluster 2a alleles, which are derived from *emm1* strains and are 99.8% identical to each other, were also more closely related to *skcg* alleles than to known *ska* alleles (95.0% nucleotide identity to *skcg6*). The *ska37* subcluster 2a allele was also more closely related to *skcg16* (98.8% nucleotide identity) than to other known *ska* alleles. Although the majority of *emm* pattern A-C strains (64%) have a cluster 1 *ska* allele (Fig. 3; Tables 1 and 2), several of the more common *emm* types associated with recent cases of pharyngitis (*emm* types 1, 3, 6, and 18) have subcluster 2a *ska* alleles.

The data strongly support the idea that streptokinase alleles underwent interspecies transfer and that most subcluster 2a *ska* alleles and *skcg* alleles have a relatively recent common ancestor.

DISCUSSION

The studies described in this report are part of a multidisciplinary effort to better understand the molecular mechanisms underlying bacterial niche adaptation in general and the molecular basis for tissue site preferences among GAS specifically. Epidemiological surveys indicate that *emm* pattern D strains of GAS have a strong preference for superficial infection of the skin (5, 7, 15). Furthermore, PAM is restricted to a subset of *emm* pattern D strains (40). Experimental work in which an in vivo model for impetigo was used demonstrated that both PAM and the Plg activator, streptokinase, play key roles in virulence and reproductive growth of GAS at the skin (41). The strong linkage disequilibrium observed in skin-tropic GAS strains between PAM and the subcluster 2b form of the streptokinase β-domain in this study cannot be readily explained by physical genetic linkage because of the extensive genetic recombination that occurs among GAS strains. The β-domain of streptokinase makes direct molecular contact with Plg during plasmin formation (10, 31). When the experimental and population findings are taken together, it seems reasonable to conclude that *pam* and the subcluster 2b β-domain form of *ska* underwent coselection, driven by epistatic interactions that a conferred novel phenotype. The novel phenotype, in turn, contributed to high levels of bacterial fitness (i.e., reproduction and transmission) at the skin.

Linkage disequilibrium can be maintained within recombining populations of bacteria through host immune selection (21,

22). Two antigenic epitope regions within the outer membrane protein, PorA, of *Neisseria meningitidis* provide an example of how a strongly cross-protective immune response can lead to the emergence of nonoverlapping combinations of antigenic variants. Like GAS, *N. meningitidis* is highly prevalent and usually found in association with asymptomatic carriage and displays high levels of genetic recombination, as shown by HK loci (18). However, a host protective response to just one of the two PorA epitope regions leads to loss of antigenic variants associated with a strain, and over time the bacterial population can acquire a discrete nonoverlapping structure. However, unlike the outer membrane protein PorA, streptokinase is secreted and diffusible, and thus, host immunity to streptokinase may be far less effective in leading to loss of the entire bacterial cell. On the basis of these findings along with epidemiological and experimental findings (40, 41), we favor the idea that the linkage disequilibrium observed between streptokinase (subcluster 2b) and PAM results from a direct biological interaction.

It is important to emphasize that while *emm* pattern D strains are associated significantly more often with impetigo than with pharyngitis (5, 7, 15), the link between *emm* pattern D strains and the skin is not absolute. This is probably because all (or most) GAS strains can persist in both the throat and the skin to at least some small degree; this is particularly true for the URT, where colonization or secondary infection following impetigo is not uncommon (8). Also, neither PAM nor subcluster 2b *ska* is essential for streptococcal impetigo, because many *emm* pattern E strains are frequently recovered from impetigo lesions (5). Therefore, pattern E strains, which uniformly lack a high-affinity Plg-binding protein (40), appear to use an entirely different molecular strategy for causing this disease. Presumably, nonbullous impetigo caused by *Staphylococcus aureus* involves a different molecular strategy as well. Thus, coselection of PAM and subcluster 2b *ska* is the result of a strong adaptive advantage for GAS reproduction and transmission at the skin, even though bacterial adaptation to the skin can occur by an alternate (although undefined) route.

The evolutionary history of the *ska* lineages within GAS is the result of a series of genetic events, the order of which is not entirely certain. All 34 GCS and GGS isolates have *skcg* alleles that are highly homologous to subcluster 2a *ska* alleles, which is indicative of a recent common ancestor. Furthermore, the 34 GCS and GGS isolates do not appear to have undergone a recent bottleneck, since they are highly variable in terms of the complement of housekeeping alleles (26). Therefore, the most plausible model is that an *skcg* allele from a GCS or GGS donor strain underwent lateral transfer to a GAS recipient strain, yielding a subcluster 2a *ska* allele. Thus, the ancestral form of *ska* within GAS most likely evolved into either the cluster 1 or subcluster 2b *ska* lineage. Given the high level of sequence divergence, it seems likely that either cluster 1 or subcluster 2b *ska*, whichever is not the ancestral form, was also acquired by an interspecies transfer event rather than having been derived from the other form. Since the subcluster 2b *ska* lineage allele is somewhat homologous to *skcg*, it may have been acquired earlier by GAS from a GCS or GGS donor strain as an ancestral *skcg* allele and may have subsequently evolved along a separate path within GAS. Alternatively, subcluster 2b *ska* may have been acquired by GAS from another closely related (but unidentified) streptococcal species.

It is plausible that PAM and subcluster 2b *ska* on occasion may have been packaged together and mobilized between GAS via bacteriophage-mediated generalized transduction. However, since the Plg-binding region of PAM displays extensive sequence diversity, which could have arisen only after an extended period of evolution, it is unlikely that cotransfer of PAM and subcluster 2b *ska* occurred to any significant extent in recent history. The intergenic region between the *emm* and *ska* loci of a PAM- and subcluster 2b *ska*-positive isolate was very similar in terms of both distance and gene content to the intergenic regions of GAS strains containing either cluster 1 *ska* or subcluster 2a *ska* but was markedly different from the *emm-skcg* region of GCS (20a). Thus, importation of both PAM and subcluster 2b *ska* in a single step from another bacterial species donor is also unlikely. Combined with evidence for intragenic recombination within *ska*, the data best support the idea that epistasis had an important role in the observed linkage disequilibrium between PAM and subcluster 2b *ska*.

The data for *skcg* alleles from GCS and GGS strongly support the idea that one or more of the three major *ska* lineages present in contemporary isolates of GAS originated in another bacterial species and recently was laterally transferred to GAS. Orthologous genes can arise by sequence divergence under ecological or sexual isolation conditions. Such isolation can promote speciation following multiple rounds of periodic selection for mutants that are fitter for a particular niche (12, 13). Sites within an ancestral gene that are critical for adaptation to a new niche undergo positive selection. Portions of the ancestral gene that are subject to strong purifying selection tend to have lower levels of nucleotide sequence diversity than regions experiencing strong diversifying selection. Homologous recombination between the donor and recipient (target) genes is favored in stretches where there is low sequence diversity. Through interspecies recombination between orthologous genes, new phenotypes can be acquired, while constrained functions can be preserved. Newly acquired orthologous genes potentially provide a rich and ready source for new bacterial phenotypes, which in turn may provide an adaptive advantage under certain ecological conditions.

The first direct encounter between PAM and a subcluster 2b *ska* product may have occurred following evolution of ancestral *ska* in discrete phylogenetic lineages. Therefore, PAM did not necessarily shape the environment in which the subcluster 2b *ska* lineage evolved. The increased fitness at the skin resulting from the PAM gene and subcluster 2b *ska* being brought into direct contact, by residing within a single genome, may simply have been a chance event. Thus, the strong epistatic coselection observed for PAM and subcluster 2b *ska* is not necessarily a driving force for the positive Darwinian selection that was detected at many of the codon sites for the streptokinase β-domain. Based on our data, the epistatic coselection observed for *pam* and subcluster 2b *ska* and the diversifying selection observed for *ska* could have been either coupled or independent.

Several of the *ska* codons identified as being under diversifying selection also represent fixed amino acid differences among the three major lineages of the β-domain-encoding region of *ska*. Therefore, at least some of the diversifying

selection pressures acting on *ska* likely contributed to the evolution of discrete lineages. Furthermore, one or two of the three major *ska* lineages likely evolved within distinct bacterial species. The unique environment provided by each bacterial species or GAS strain can account for the differential selection pressures encountered during the evolution of each *ska* lineage. During infection, streptokinase has direct interactions with mammalian host Plg, the mammalian host immune response (43), as well as with bacterial proteases (41) and Plg bound via different bacterial proteins (30, 34, 35). Any of these host or bacterial factors has the potential to provide positive selection pressure on the *ska* gene.

In most structural studies of streptokinase the workers have utilized the product of an *skcg* gene (46), which is most closely related to the subcluster 2a form of streptokinase. In the fluid phase, the β-domain of streptokinase is engaged in direct molecular contact with kringle 5 of human-derived Plg (10, 31). One possibility is that subcluster 2b forms of streptokinase are highly adapted to Plg when it is presented in a form that is bound by PAM, which occurs via kringle 2 (48). GAS also express low-affinity Plg-binding proteins on the cell surface (30, 35). The molecular interactions of the β-domain of streptokinase with Plg may be different for a fluid-phase form and a bound form and may be dependent on the type of Plg-binding protein as well. GCS and GGS express Plg-binding proteins that are structurally distinct from PAM and all other known GAS proteins (34). Another possible selective influence is the possibility that one of the GAS *ska* forms had a long history of coevolution with Plg in another mammalian host. Streptokinase-mediated activation of Plg derived from nonhuman sources can be less effective than activation of human Plg (38). There are numerous streptococcal species that infect other animals whose streptokinase genes have yet to be analyzed.

It is potentially significant that the subcluster 2a form of *ska*, present in several throat-tropic strains of GAS (*emm* pattern A-C), probably originated from GCS and GGS, which are commensals of the URT in humans. Several of the GAS strains harboring the subcluster 2a form of *ska*, corresponding to *emm* types 1, 3, 6, and 18, also appear to be responsible for a significant proportion of recent cases of GAS pharyngitis in the United States (23, 25, 28, 29). It remains to be established whether *ska* facilitates colonization in the throat.

Population genetics and phylogenetics are powerful tools that can be used to guide future experimental studies. For example, site-specific mutagenesis at codons under diversifying selection provides a rational approach for studying the effect of each adaptive change on the in vitro functional activity and immunogenicity of streptokinase. Isogenic mutants, generated by directed allelic replacement of the parental *ska* gene with an *ska* allele of another lineage, can be used to measure biological properties of GAS by using in vivo models for infection or colonization. Studies on swapping *ska* alleles are planned.

The molecular basis for niche adaptation by bacteria can be complex. Experimental findings, epidemiological surveys, population genetics, and evolutionary inferences can all contribute to a comprehensive understanding of this complex phenotype. Epistatic coselection arising between bacterial proteins (PAM and subcluster 2b streptokinase) acting on a common host factor (Plg) appears to contribute to tissue-specific adaptation of *emm* pattern D GAS at the skin. Recombination between orthologous genes may also play a facilitating role in the emergence of new adaptive phenotypes in bacteria.

ACKNOWLEDGMENTS

We thank Bernie Beall for providing several strains and Karen McGregor for early release of multilocus sequence typing data.

This work was supported by grants AI-28944, AI-53826, and GM-60793 from NIH and by a grant-in-aid from the American Heart Association to D.E.B. A.K. was a recipient of a Brown-Coxe postdoctoral fellowship.

REFERENCES

1. **Anisimova, M., R. Nielsen, and Z. H. Yang.** 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics **164:**1229–1236.
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. **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. **99:**10078–10083.
4. **Bessen, D., and V. A. Fischetti.** 1990. Differentiation between two biologically distinct classes of group A streptococci by limited substitutions of amino acids within the shared region of M protein-like molecules. J. Exp. Med. **172:**1757–1764.
5. **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.
6. **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.
7. **Bessen, D. E., C. M. Sotir, T. L. Readdy, and S. K. Hollingshead.** 1996. Genetic correlates of throat and skin isolates of group A streptococci. J. Infect. Dis. **173:**896–900.
8. **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.
9. **Carlsson Wistedt, A. C., U. Ringdahl, W. Müller-Esterl, and U. Sjöbring.** 1995. Identification of a plasminogen-binding motif in PAM, a bacterial surface protein. Mol. Microbiol. **18:**569–578.
10. **Chaudhary, A., S. Vasudha, K. Rajagopal, S. S. Komath, N. Garg, M. Yadav, S. C. Mande, and G. Sahni.** 1999. Function of the central domain of streptokinase in substrate plasminogen docking and processing revealed by site-directed mutagenesis. Protein Sci. **8:**2791–2805.
11. **Civetta, A.** 2003. Positive selection within sperm-egg adhesion domains of fertilin: an ADAM gene with a potential role in fertilization. Mol. Biol. Evol. **20:**21–29.
12. **Cohan, F. M.** 2001. Bacterial species and speciation. Syst. Biol. **50:**513–524.
13. **Cohan, F. M.** 2002. Sexual isolation and speciation in bacteria. Genetica **116:**359–370.
14. **Cunningham, M. W.** 2000. Pathogenesis of group A streptococcal infections. Clin. Microbiol. Rev. **13:**470–511.
15. **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.
16. **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.
17. **Facklam, R., B. Beall, A. Efstratiou, V. Fischetti, E. Kaplan, P. Kriz, M. Lovgren, D. Martin, B. Schwartz, A. Totolian, D. Bessen, S. Hollingshead, F. Rubin, J. Scott, and G. Tyrrell.** 1999. Report on an international workshop: demonstration of *emm* typing and validation of provisional M-types of group A streptococci. Emerg. Infect. Dis. **5:**247–253.
18. **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. **98:**182–187.
19. **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. **98:**4658–4663.

20. **Fischetti, V.** 2000. Surface proteins on gram positive bacteria, p. 11–24. *In* V. A. Fischetti, R. P. Novick, J. J. Ferretti, D. A. Portnoy, and J. I. Rood (ed.), Gram positive pathogens. ASM Press, Washington, D.C.

20a. **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.

21. **Gupta, S., N. Ferguson, and R. M. Anderson.** 1998. Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. Science **240:**912–915.

22. **Gupta, S., M. C. J. Maiden, I. M. Feavers, S. Nee, R. M. May, and R. M. Anderson.** 1996. The maintenance of strain structure in populations of recombining infectious agents. Nat. Med. **2:**437–442.

23. **Haukness, H. A., R. R. Tanz, R. B. Thomson, Jr., D. K. Pierry, E. L. Kaplan, B. Beall, D. Johnson, N. P. Hoe, J. M. Musser, and S. T. Shulman.** 2002. The heterogeneity of endemic community pediatric group A streptococcal pharyngeal isolates and their relationship to invasive isolates. J. Infect. Dis. **185:**915–920.

24. **Huelsenbeck, J., and B. Rannala.** 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science **276:**227–232.

25. **Johnson, D. R., J. T. Wotton, A. Shet, and E. L. Kaplan.** 2002. A comparison of group A streptococci from invasive and uncomplicated infections: are virulent clones responsible for serious streptococcal infections? J. Infect. Dis. **185:**1586–1595.

26. **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.

27. **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.

28. **Kaplan, E. L., D. R. Johnson, and P. P. Cleary.** 1989. Group A streptococcal serotypes isolated from patients and sibling contacts during the resurgence of rheumatic fever in the United States in the mid-1980s. J. Infect. Dis. **159:** 101–103.

29. **Kaplan, E. L., J. T. Wotton, and D. R. Johnson.** 2001. Dynamic epidemiology of group A streptococcal serotypes associated with pharyngitis. Lancet **358:** 1334–1337.

30. **Lottenberg, R., C. C. Broder, M. D. P. Boyle, S. J. Kain, B. L. Schroeder, and R. I. Curtiss.** 1992. Cloning, sequence analysis, and expression in *Escherichia coli* of a streptococcal plasmin receptor. J. Bacteriol. **174:**5204–5210.

31. **Loy, J. A., X. Lin, M. Schenone, F. J. Castellino, X. C. Zhang, and J. Tang.** 2001. Domain interactions between streptokinase and human plasminogen. Biochemistry **40:**14686–14695.

32. **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.

33. **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.

34. **Nasr, B., A. Wistedt, U. Ringdahl, and U. Sjobring.** 1994. Streptokinase activates plasminogen bound to human group C and G streptococci through M-like proteins. Eur. J. Biochem. **222:**267–276.

35. **Pancholi, V., and V. A. Fischetti.** 1998. Alpha-enolase, a novel strong plasmin(ogen) binding protein on the surface of pathogenic streptococci. J. Biol. Chem. **273:**14503–14515.

36. **Posada, D., and K. Crandall.** 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14:**817–818.

37. **Sawyer, S.** 1999. GENECONV: a computer package for the statistical detection of gene conversion. Washington University, St. Louis, Mo.

38. **Schroeder, B., M. D. Boyle, B. R. Sheerin, A. C. Asbury, and R. Lottenberg.** 1999. Species specificity of plasminogen activation and acquisition of surface-associated proteolytic activity by group C streptococci grown in plasma. Infect. Immun. **67:**6487–6495.

39. **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. **99:**4668–4673.

40. **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.

41. **Svensson, M. D., U. Sjobring, 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.

42. **Swanson, W. J., Z. Yang, M. F. Wolfner, and C. F. Aquadro.** 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc. Natl. Acad. Sci. **98:**2509–2514.

43. **Tewodros, W., M. Norgren, and G. Kronvall.** 1995. Streptokinase activity among group A streptococci in relation to streptokinase genotype, plasminogen binding, and disease manifestations. Microb. Pathog. **18:**53–65.

44. **Twiddy, S. S., J. J. Farrar, N. Vinh Chau, B. Wills, E. A. Gould, T. Gritsun, G. Lloyd, and E. C. Holmes.** 2002. Phylogenetic relationships and differential selection pressures among genotypes of dengue-2 virus. Virology **298:**63–72.

45. **Urwin, R., E. C. Holmes, A. J. Fox, J. P. Derrick, and M. C. Maiden.** 2002. Phylogenetic evidence for frequent positive selection and recombination in the meningococcal surface antigen PorB. Mol. Biol. Evol. **19:**1686–1694.

46. **Wang, X., X. Lin, J. A. Loy, J. Tang, and X. C. Zhang.** 1998. Crystal structure of the catalytic domain of human plasmin complexed with streptokinase. Science **281:**1662–1665.

47. **Wannamaker, L. W.** 1970. Differences between streptococcal infections of the throat and of the skin. N. Engl. J. Med. **282:**23–31.

48. **Wistedt, A. C., H. Kotarsky, D. Marti, U. Ringdahl, F. J. Castellino, J. Schaller, and U. Sjobring.** 1998. Kringle 2 mediates high affinity binding of plasminogen to an internal sequence in streptococcal surface protein PAM. J. Biol. Chem. **273:**24420–24424.

49. **Yang, Z., and R. Nielsen.** 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. **19:**908–917.

50. **Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen.** 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:**431–449.