INFECTION AND IMMUNITY, Aug. 2001, p. 4858–4869 Vol. 69, No. 8
0019-9567/01/$04.00+0 DOI: 10.1128/IAI.69.8.4858–4869.2001

Directional Gene Movement from Human-Pathogenic to Commensal-Like Streptococci

AWDHESH KALIA,[1] MARK C. ENRIGHT,[2] BRIAN G. SPRATT,[3] AND DEBRA E. BESSEN[1]*

Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut,[1] and Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY,[2] and Department of Infectious Disease Epidemiology, Imperial College School of Medicine, University of London, St. Mary's Campus, London W2 1PG,[3] United Kingdom

Received 1 March 2001/Returned for modification 19 April 2001/Accepted 1 May 2001

Group A streptococci (GAS) are highly pathogenic for humans, and their closest genetic relatives, group C and G streptococci (GCS and GGS, respectively), are generally regarded as commensals, although they can be found in association with human disease. As part of an effort to better understand the evolution of virulence, the phylogenetic relationships between GAS, GCS, and GGS were examined. The nucleotide sequence was determined for an internal portion of seven housekeeping (neutral) loci among >200 isolates of GAS and 34 isolates of GCS or GGS obtained from human subjects. Genotypic analysis failed to show support for the separation of GCS and GGS into two distinct populations. Unlike GAS, there was poor concordance between *emm* type and genetic relatedness among GCS and GGS. All housekeeping genes within GAS displayed relatively low levels of sequence diversity. In contrast, individual GCS and GGS strains had mosaic genomes, containing alleles at some loci that were similar or identical to GAS alleles, whereas the alleles at other loci were about 10 to 30% diverged. The data provide evidence for a history of recent interspecies transfer of neutral genes that exhibits a strong net directionality from GAS donors to GCS and GGS recipients. A model for the evolution of GAS and of GCS and GGS is described.

The beta-hemolytic, large-colony-forming streptococci which bear group C or G carbohydrate are usually found in association with humans as part of the normal flora, and for this reason, they are widely regarded as being commensal or commensal-like organisms (23, 37). However, there are also numerous reports showing an association of group C streptococci (GCS) or group G streptococci (GGS) with outbreaks of pharyngitis or sporadic cases of severe invasive disease (5, 6, 12, 31, 45). In many ways, the human diseases caused by GCS and GGS—which can also include impetigo, cellulitis, and nephritis—closely resemble infections due to the widely prevalent bacterial pathogen *Streptococcus pyogenes*, also known as group A streptococci (GAS). In some parts of the world, rates for asymptomatic throat carriage of GCS and GGS far exceed the rates for GAS, yet at the same time, the known contribution of GCS and GGS to disease remains below that for GAS (36).

GAS, GCS, and GGS share several properties that potentially contribute to human infection (reviewed in references 7 and 30). All of the organisms possess genes (*emm*) encoding M proteins (10, 39), which form surface fibrils and can act as key virulence factors in GAS. The *emm* genes of GCS and GGS display sequence heterogeneity at their 5′ ends, giving rise to at least 30 distinct *emm* sequence types (www.cdc.gov/ncidod/biotech/strep/strains.html). For GAS, >150 *emm* types are recognized, and antibodies directed to M-type-specific epitopes, located at the fibril tips, can confer protective immunity (11, 28). Infections caused by GAS, GCS, and GGS can be associated with a rise in antibody titers to streptolysin O (4, 23). Other shared, putative virulence factors include streptokinase, C5a peptidase, binding proteins for fibronectin, plasminogen and immunoglobulins, and a hyaluronic acid capsule (8, 25, 27, 35, 41). Perhaps the best evidence for interspecies transfer of virulence loci between GAS and GCS or GGS involves the *emm* gene (40, 44) and streptokinase gene (25).

A fundamental objective in the study of bacterial pathogenesis is to understand why related organisms differ in their virulence properties. One approach towards this goal is to compare the gene content of pathogenic and commensal species that evolved along distinct pathways yet remain close genetic relatives. Many adaptive loci, such as *emm* and other virulence genes, frequently undergo diversifying or purifying selection, and as a result, the ancestral relationships between isolates can be difficult to decipher using these loci. Phylogenetic histories are easiest to trace for genes that are both vital to the bacterial cell and selectively neutral, such as the housekeeping loci, which perform general maintenance functions. In this report, phylogenetic relationships between the alleles at multiple housekeeping loci are used to construct a model for the evolution of group A, C, and G streptococci.

MATERIALS AND METHODS

Bacteria. Thirty-four GCS and GGS isolates obtained from human subjects were the focus of this study (Table 1). All were beta-hemolytic and formed large colonies following growth on Todd-Hewitt sheep blood agar. Group carbohydrate was ascertained by a latex agglutination test (Difco, Inc.). Additional phenotypic tests for taxonomic classification measured bacitracin sensitivity, pyrrolidonyl arylamidase production, and utilization of sorbitol, glycogen, and trehalose (16). The Voges-Proskauer test for acetoin production was also performed. All GCS and GGS isolates were kindly provided by Bernard Beall and Richard Facklam (Centers for Disease Control [CDC], Atlanta, Ga.); for many of the isolates, *emm* sequence type was also provided. Several isolates (CDC numbers 5341, 5344, 5345, 5353, 5354, and 5357) originated from Androulla Efstratiou (Central Public Health Laboratory, Colindale, London, England),

* Corresponding author, Mailing address: Yale University School of Medicine, Department of Epidemiology & Public Health, 60 College Street, Box 208034, New Haven, CT 06520-8034. Phone: (203) 785-4480. Fax: (203) 737-4285. E-mail: debra.bessen@yale.edu.

TABLE 1. Epidemiological properties and MLST analysis of GCS and GGS

Strain	Tissue[a]	Disease[a]	Yr	Location[b]	*emm* type	Group carbohydrate	ST	Allele assignments[c]							Code in dendrogram
								gki	*gtr*	*murI*	*mutS*	*recP*	*xpt*	*yqiL*	
4236	Sterile	Invasive	ND	USA	*stG643*	G	1	101	**106**	104	101	***116***	101	**107**	GstG643
4031	Blood	Bacteremia	1998	Argentina	*stG4831*	G	2	**102**	102	103	**106**	104	**114**	106	GstG4831
4949	URT	None	1996	India	*stC839*	C	3	**103**	104	**109**	**106**	103	107	105	Cst839-1
4241	Sterile	Invasive	ND	USA	*stC839*	C	4	**103**	104	**112**	**106**	101	107	**112**	Cst839-2
4231	Sterile	Invasive	ND	USA	*stG653*	C	5	**103**	**105**	**111**	**106**	101	107	**116**	CstG653
5341	URT	ND	1920s	UK	*stG93464*	C	6	**103**	**111**	**109**	**106**	101	107	**110**	Cst93464
SS1069	URT	ND	1974	N.J.	*stC839*	C	7	**103**	**111**	**109**	**106**	101	107	**112**	Cst839-3
MGAS338	Sterile	Invasive	1980s	USA	*stC839*	C	8	**104**	102	**109**	**106**	101	107	104	Cst839-4
4966	URT	Pharyngitis	1996	India	*stC839*	C	9	**104**	**111**	**109**	**106**	101	107	**107**	Cst839-5
4277	Sterile	Invasive	ND	USA	*stC36*	G	10	**105**	**106**	**109**	**103**	104	**111**	**109**	GstC36
5344	Skin	Wound	1980s	UK	*stC5344*	C	11	**105**	**106**	**112**	**106**	**113**	105	**111**	CstC5344
4232	Sterile	Invasive	ND	USA	*stG10*	G	12	**105**	**109**	**109**	**106**	**113**	104	**113**	GstG10-1
1891	Sterile	Cellulitis	1999	Ill.	*stG480*	G	13	**106**	101	101	**106**	**113**	104	102	Gst480-1
D421	Skin	Impetigo	1971	Trinidad	*stG480*	G	14	**106**	103	102	**106**	**113**	104	103	Gst480-2
4030	Blood	Bacteremia	1998	Argentina	*stG11*	G	15	**106**	104	**105**	**106**	**113**	104	106	GstG11-1
4234	Sterile	Invasive	ND	USA	*stG480*	G	16	**106**	**108**	**105**	**104**	**113**	104	111	Gst480-3
4255	Sterile	Invasive	ND	USA	*stG653*	G	17	**106**	**108**	**105**	**106**	114	104	107	GstG653
5353	Skin	Ulcer	1980s	UK	*stG11*	G	18	**106**	**110**	**109**	**106**	113	104	**113**	GstG11-2
5357	Skin	Ulcer	1980s	UK	*stG652*	G	19	**107**	**112**	**108**	**105**	**111**	**109**	**108**	GstG652
1778	URT	ND	1993	Wyo.	*stC36*	C	20	**108**	**106**	**105**	102	102	102	**110**	CstC36-1
4288	Sterile	Invasive	ND	USA	*stG485*	C	21	**108**	**106**	105	**106**	107	104	**114**	Cst485
4265	Sterile	Invasive	ND	USA	*stC36*	C	22	**108**	106	106	**106**	108	106	108	CstC36-2
4242	Sterile	Invasive	ND	USA	*stG485*	G	23	**108**	107	109	**105**	102	104	115	Gst485-1
4276	Sterile	Invasive	ND	USA	*stG10*	G	24	108	**109**	**107**	**105**	105	110	**116**	GstG10-2
3296	ND	Unknown	1994	Ariz.	*stG485*	G	25	108	**109**	**109**	**105**	102	104	**116**	Gst485-2
5354	Blood	Bacteremia	1980s	UK	*stG62467*	G	26	**108**	**109**	**109**	**106**	**107**	**113**	**113**	Gst62467
SS957	ND	ND	1969	ND	*stC957*	C	27	**108**	**111**	**105**	106	**109**	104	**109**	CstSS957
SS188	URT	ND	1941	USA	*stC74a*	C	28	**108**	**113**	**109**	106	**108**	**112**	**111**	CstC74a
4286	Sterile	Invasive	ND	USA	*stG6*	G	29	**108**	**113**	**109**	**106**	**110**	**115**	**109**	GstG6
1775	URT	ND	1993	Wyo.	*stC36*	C	30	**108**	**114**	**110**	102	**107**	101	**108**	CstC36-3
5345	Blood	Bacteremia	1983	UK	*stG5345*	C	31	108	**115**	**109**	**105**	**112**	**108**	**110**	CstG5345
3110	Blood	Invasive	1994	Ariz.	*stG643*	C	32	108	**116**	**109**	**105**	**115**	103	**115**	CstG643
4247	Sterile	Invasive	ND	USA	*stC74a*	G	33	**109**	**109**	**109**	**105**	102	104	**110**	GstC74a
4951	URT	ARF	1996	India	*stG4974*	G	34	**110**	101	**109**	**106**	106	**115**	101	GstG4974

[a] Isolation of a GCS or GGS organism from a subject with noninvasive disease does not necessarily imply that the organism caused that disease. Sterile, a normally sterile site; URT, upper respiratory tract; ARF, acute rheumatic fever. ND, not determined.

[b] UK, United Kingdom.

[c] Allele assignments in bold are assigned to cluster II, whereas the allele in bold italics (*recP116*) is highly divergent from both sequence similarity clusters I and II (see Table 3 and Fig. 2).

who kindly provided accompanying epidemiological information; the strain designations are NCTC 5370, 11552, 11553, 11557, 11566, and 11569, respectively (http://www.phls.co.uk/services/nctc) (13). Two isolates originated from the Lancefield collection (SS957 [B337] and SS188 [C74]). All GAS isolates were previously described (14).

***emm* sequence typing.** The *emm* sequence type was determined for the GCS and GGS isolates in accordance with methods used for GAS with the same universal oligonucleotide primers (1, 15). A unique *emm* type is defined as having <95% sequence identity to any other known *emm* type over 160 bp near the 5′ end; indels of greater than four codons or frameshift mutations relative to the reference *emm* typing strain were not encountered. One newly recognized *emm* type, *emmst*C957, was deposited in GenBank (accession no. AF332809). A listing of *emm* types found in association with GCS and GGS is maintained on the Internet (www.cdc.gov/ncidod/biotech/strep/strains.html).

MLST. For multilocus sequence typing (MLST), chromosomal DNA was prepared from freshly grown GCS and GGS isolates by the mutanolysin procedure, as previously described for GAS (2, 14). Internal fragments (~400 to 500 bp) of seven housekeeping genes, encoding putative glucose kinase (*gki*), glutamine transport protein (*gtr*), glutamate racemase (*murI*), mismatch repair enzyme (*mutS*), transketolase (*recP*), xanthine phosphoribosyltransferase (*xpt*), and acetylcoenzyme A acetyltransferase (*yqiL*), were amplified by PCR using primer pairs designed for GAS loci, as previously described (14).

Primer pairs for two of the loci (*gtr* and *yqiL*) failed to amplify the gene fragments for many of the GCS and GGS isolates. Therefore, alternative sets of primers were designed using sequence information from the *Streptococcus equi* subsp. *equi* genome sequencing project (www.sanger.ac.uk). To confirm that the fragments amplified by these alternative primers were homologous to the *gtr* or *yqiL* locus of GAS, their sequences underwent a TBlastX search analysis (translated nucleotide [nt] comparisons) with both the GenBank and GAS genome sequence (www.genome.ou.edu) databases; the amino acid identity with GAS alleles was ~91 and 77% for the alternative *gtr* and *yqiL* fragments, respectively. Alternative primers used for amplification of *gtr* and *yqiL* from GCS and GGS which failed to amplify with the GAS primers are as follows: *gtr*(CG)-up, 5′-TTT ACT TCG TAC CAT GAA CCT TCT T-3′; *gtr*(CG)-dn, 5′-GAC CAT AGT CAT CCC AGA TTT AGC-3′; *yqiL*(CG)-up, 5′-ACG AAA TTG TCC CTG TCT CTG T-3′; and *yqiL*(CG)-dn, 5′-AAA GTG TTG CTA GTC CTC TGG TTA C-3′. In order to confirm that the alternative primers amplified GCS and GGS loci that were equivalent to the GAS *gtr* and *yqiL* loci, it could be shown that each GCS and GGS isolate was amplified by one primer pair or the other (GAS or alternative primers), but not by both pairs. As a control, *gtr* and *yqiL* loci of GAS strain ATCC 700294 were shown to amplify with the GAS-like primer pairs only.

The PCR amplifications were performed in volumes of 50 μl for 30 cycles, with an initial denaturation at 94°C for 1 min (4 min for first cycle only), annealing at 52°C for 1 min (except for *murI* primers, which were annealed at 46°C), and extension at 72°C for 1 min. The amplified DNA fragments were confirmed for expected size by agarose gel electrophoresis and were subsequently purified and subjected to nt sequence determination as previously described (14), with the same primers used for PCR amplification.

To ensure the accuracy of the nt sequences obtained for the housekeeping loci, verification steps were taken. Of the 238 housekeeping alleles characterized for the 34 GCS and GGS isolates, 54 alleles underwent nt sequencing a second time to rule out possible cross-contamination of DNA template. For these checks, three to five colonies of bacteria, freshly grown on agar plates, were suspended in distilled water and boiled for 10 min to release template DNA, which was used immediately for PCR amplification. For each fresh culture, group carbohydrate

TABLE 2. Housekeeping alleles of GCS and GGS and comparison to GAS

Locus[a]	Length of sequence (bp)	No. of GAS alleles (from 212 isolates)	No. of GCS-GGS alleles (from 34 isolates)	Maximal % divergence		
				Within GAS	Within GCS-GGS	Between GAS and GCS-GGS
gki	498	34	10	2.6	11.7	12.6
gtr	430	32	16	2.6	22.8	24.5
murI	438	24	12	1.4	32.2	32.6
mutS	405	21	6	2.2	4.9	4.9
recP	459	35	16	6.1	6.4	7.6
xpt	450	29	15	2.9	4.9	4.9
yqiL	434	22	16	1.4	26.6	26.9

[a] For GCS and GGS isolates, the *gtr* locus was truncated to 430 bp (20 bp smaller than that originally reported for GAS). The allele *recP116* from GGS strain 4236 formed a third cluster (not assigned to either cluster I or II) and is excluded from the maximal divergence calculations.

composition was also reconfirmed. Loci selected for verification included one representative of each GAS-like cluster I allele; for isolates sharing an *emm* type, many of the cluster II alleles that differed from the majority of strains of that *emm* type were sequenced a second time. Verification data were 100% concordant with the first set of data.

For each locus, every different sequence was assigned a distinct allele number, and each isolate was defined by a series of seven integers (the allelic profile) corresponding to the alleles at the seven loci, in the order *gki-gtr-murI-mutS-recP-xpt-yqiL*. The allelic profile defines the sequence type (ST). Alleles from GCS and GGS isolates were given numeric assignments beginning with 101 irrespective of whether an identical allele had been previously identified in GAS (GAS allele assignments are two-digit, beginning with 01). Because of sequence ambiguities at the ends of the *gtr* fragment, trim points were adjusted to yield a slightly shorter (430 bp) sequence; however, none of the previously described *gtr* alleles from GAS had nt substitutions in the portions of the gene that were trimmed (14).

Computations and phylogenetic analysis. A matrix of pairwise differences in allelic profiles was constructed, and the similarities between the allelic profiles of the isolates were assessed by cluster analysis using the unweighted pair-group method with arithmetic averages (UPGMA) and the percent disagreement distance measure (Statistica version 5.5; StatSoft, Tulsa, Okla.).

The maximum percent nt divergence between pairs of alleles at a given locus was calculated using Mega version 2.0 (http://www.megasoftware.net).

The index of association (32) was used to test for linkage disequilibrium between alleles at the seven housekeeping loci (START version 1.1; http://www.mlst.net). The observed variance in the distribution of allelic mismatches in all pairwise comparisons of the allelic profiles was compared to that expected in a freely recombining population (linkage equilibrium). The significance of the difference in the observed and expected variance was evaluated by computing the maximum variance in the distribution of allelic mismatches obtained using 100 randomizations of the data set. Significant linkage disequilibrium was established if the observed variance obtained with the actual data was greater than that found with any of the 100 randomized data sets; otherwise there was no evidence of a departure from linkage equilibrium.

Phylogenetic trees of the nt sequences from each housekeeping locus were constructed using the maximum likelihood (ML) method available in the PAUP* package (version 4.0; Sinaver Associates, Mass.). Start-up trees were obtained using the neighbor-joining method. ML trees were reconstructed using optimized values for the transition-transversion ratio and α parameter, which describes the extent of rate variation among nt sites assuming a discrete gamma distribution with eight categories; both were estimated from the empirical data during tree constructions. Optimization of the desired evolutionary model of DNA substitution and the parameters was done using hierarchical likelihood ratio tests (21), with the aid of MODELTEST version 3.0 (38). To determine the significance of the observed groupings, bootstrap analysis with 1,000 replicates was performed, using trees reconstructed by the neighbor-joining method to avoid excessive computational time, while incorporating the same ML substitution parameters. Phylogenetic trees were visualized using TreeView (version 1.6; http://www.taxonomy.zoology.gla.ac.uk/rod/rod.html).

Split decomposition analysis was performed using SplitsTree (version 3.1) and uncorrected Hamming distance measures (22).

RESULTS

Taxonomic classification based on phenotypic traits. All GCS and GGS isolates evaluated in this report were obtained from human subjects, and the vast majority were known to be associated with disease (Table 1). They were assigned as GCS or GGS by their reactivity with the group C- or group G-specific carbohydrate antiserum. The 34 GCS and GGS isolates are best classified as *Streptococcus dysgalactiae* subsp. *equisimilis* on the basis of a negative test for pyrrolidonyl arylamidase production (*Streptococcus pyogenes* gives a positive test) and bacitracin resistance (*S. pyogenes* is sensitive), although one GGS strain (3296) was partially sensitive to bacitracin; in addition, all 34 GCS and GGS isolates were negative for acid production following growth in the presence of either sorbitol or glycogen, but positive for acid production with trehalose. Whether human isolates of GCS displaying these phenotypes are best placed in the *S. dysgalactiae* complex along with their GGS counterparts or should be classified as *S. equi* subsp. *equisimilis* has been controversial (46, 47). As will be shown in this report, the group C isolates under study have genotypes similar to those of the group G isolates of *S. dysgalactiae* subsp. *equisimilis*, and herein, the 34 GCS and GGS isolates will be regarded as falling under this classification.

MLST of GCS and GGS. In a recent study, MLST of 212 GAS (*S. pyogenes*) isolates was performed by determining the nt sequence of an internal portion of seven housekeeping loci (14). In an effort to assess the phylogenetic relationships between GAS and it close genetic relatives GCS and GGS, the same seven housekeeping loci (*gki, gtr, murI, mutS, recP, xpt*, and *yqiL*) were used for MLST analysis of GCS and GGS. A total of 16 GCS and 18 GGS isolates of *S. dysgalactiae* subsp. *equisimilis* were selected for study (Table 1). This set of isolates represents 18 of the ~30 known *emm* types that have been found in association with GCS or GGS (www.cdc.gov/ncidod/biotech/strep/strains.html) and, by this measure, constitute a heterogeneous sample. However, the majority of isolates were recovered from cases of invasive disease and therefore may not be representative of strains collected by population-based sampling of all tissue sites.

As with the MLST scheme for GAS, for a given housekeeping locus, sequences differing from all others by 1 nt or more were given a new allele assignment. Among the 34 GCS and GGS isolates studied, the number of alleles per locus ranged from 6 (for *mutS*) to 16 (for *gtr, recP*, and *yqiL*) (Table 2). For the 34 GCS and GGS isolates examined, 34 unique combinations of allelic profiles (STs) were obtained (Table 1). Each allelic profile is regarded as a unique strain or clone.

The average number of alleles per locus was 13, and therefore, MLST is able to distinguish $>10^7$ allelic profiles among

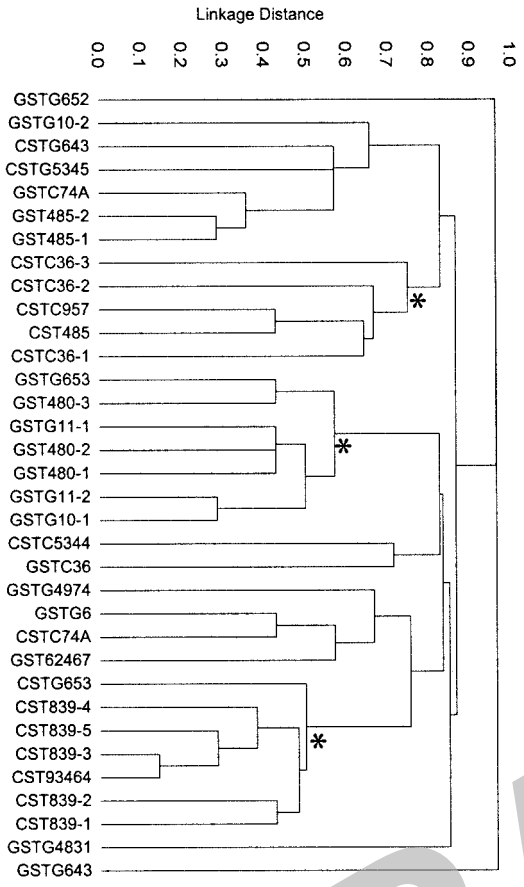

FIG. 1. Dendrogram showing UPGMA cluster analysis of GCS and GGS. Strain designations at branch tips are listed in the Code for dendrogram column in Table 1, showing group carbohydrate (C or G) followed by *emm* type. Three branch points showing large clusters of GCS or GGS isolates are indicated by asterisks.

GCS and GGS. No isolate was found to have the combination of the most frequently occurring allele at each locus. Since only 34 isolates were evaluated for this study, it is anticipated that many more alleles exist in nature and remain to be identified. Thus, the resolving power of the MLST scheme for GCS and GGS should prove to be very considerably higher.

A matrix of pairwise differences in allelic profiles was determined, and a dendrogram displaying the genetic distance between the 34 GCS and GGS isolates was constructed by cluster analysis using UPGMA (Fig. 1). Of the 18 *emm* types, 9 *emm* types are represented by more than one isolate; 25 of the 34 isolates are of an *emm* type having more than one representative within the set. Among these 25 isolates, only 2 of the same *emm* type (isolates 4966 and SS1069, *emmst*C839) were closely related in genotype, sharing alleles at five of seven housekeeping loci. The two most closely related isolates among the set of 34 strains are SS1069 (*emmstC839*) and 5341 (*emmst93464*), which share six of seven housekeeping alleles; both are GCS. These findings are in sharp contrast to the data observed for GAS, where the vast majority of isolates of a single *emm* type are members of a single clone or clonal complex, even when recovered from their human host several decades apart (14).

For each of the seven housekeeping loci, there are examples of alleles shared by both GCS and GGS (Table 1). Furthermore, some *emm* types were observed in association with both GCS and GGS isolates (*emmstC36, emmstC74a, emmstG485, emmstG643*, and *emmstG653*).

The GCS and GGS isolates were not resolved by MLST into two distinct clusters (Fig. 1). However, there were two prominent clusters of GCS strains, formed at genetic linkage distances of ~0.5 and 0.75, and one large cluster of GGS isolates branching from a node at a linkage distance of ~0.65 (indicated by asterisks). A dendrogram was also constructed for the combined GAS (100 STs) and GCS-GGS (34 STs) data sets. With one exception (strain 4236), all GCS-GGS isolates were resolved from all GAS isolates (data not shown). Strain 4236 clustered with the GAS isolates, as the alleles at both *murI* and *mutS* were identical to alleles present among GAS, and the allelic profile of this strain had two of seven loci in common with five GAS STs.

The extent of recombination among the set of 34 GCS and GGS isolates was assessed by the index of association (32). Using all 34 isolates in the calculation, no significant departure from linkage equilibrium among housekeeping loci was observed. The data suggest that the alleles in GCS and GGS undergo high rates of genetic recombination which are sufficient to eliminate any association between the alleles at different loci.

Phylogenetic analysis of housekeeping loci of GAS, GCS, and GGS. In MLST, the alleles of each housekeeping locus are regarded as character states, having unique integer assignments that do not take into account the level of nt sequence divergence between alleles. The relationships between GAS, GCS, and GGS were analyzed further by using the nt sequences of the alleles at the seven loci used in MLST.

In an earlier report on the MLST of 212 GAS isolates, using the same set of housekeeping loci, the maximal divergence in nt sequence between any pair of alleles at each locus ranged from 1.4% (for *yqiL* and *murI*) to 6.1% (for *recP*). In sharp contrast, four of the seven housekeeping loci (*gki, gtr, murI*, and *yqiL*) exhibited high levels of maximal nt sequence divergence within the GCS and GGS set, ranging from ~10 to 30% (Table 2). There is no obvious relationship between the degree of nt sequence divergence within a locus and relative chromosomal position, based on the genome map derived from GAS strain 700294 (14) (www.genome.ou.edu). The maximal percent nt divergence within GCS and GGS was similar in magnitude to the maximal percent divergence between alleles from GAS and GCS or GGS isolates (Table 2). Compared to GAS, the GCS and GGS housekeeping loci display a much greater degree of sequence diversity.

ML trees were constructed for each of the seven housekeeping loci and included all known GAS (14) and GCS-GGS alleles (Fig. 2 for *mutS, xpt,* and *yqiL*; trees for the other four loci are available from the authors upon request). None of the seven gene trees resolved all of the GAS alleles from all of the GCS-GGS alleles. Importantly, GCS alleles were not resolved from GGS alleles; at each locus, there were alleles shared by both GCS and GGS. The trees for *gki, gtr, murI, mutS,* and *yqiL* show two strongly supported clusters representing GAS (sequence similarity cluster I) and GCS-GGS (sequence similarity cluster II) alleles. These two groups of alleles are also evident from visual inspection of the distribution of the polymorphic nt

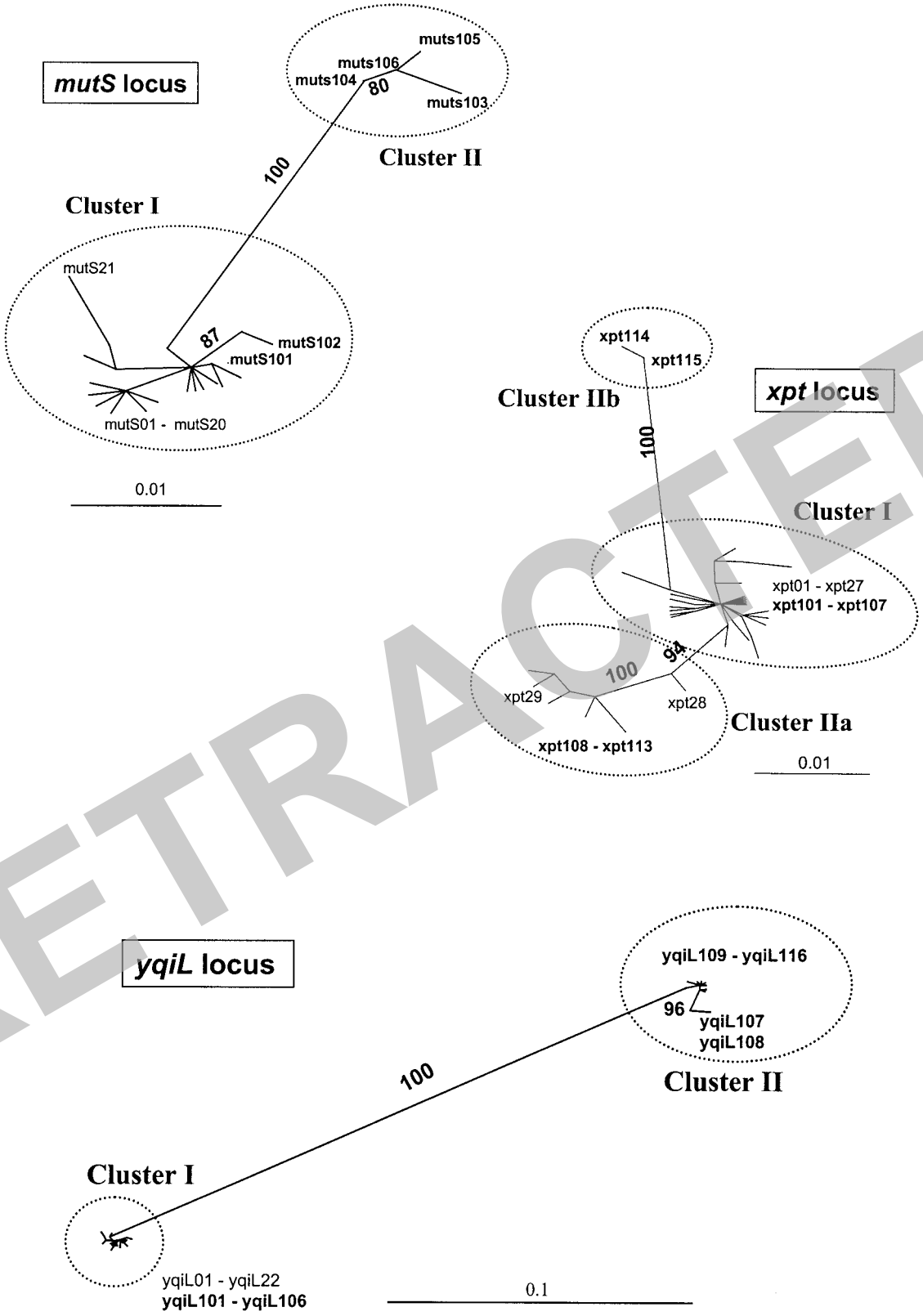

FIG. 2. ML gene trees of housekeeping loci. Unrooted radial gene trees, generated by the ML method, contain all reported GAS alleles (14) and all GCS-GGS alleles listed in Table 1. Bootstrap values of ≥80% are indicated; GCS-GGS alleles are highlighted in bold. One highly divergent allele listed in Table 1 (*recP116*) falls outside of sequence similarity clusters I and II. The most appropriate models for DNA substitution were determined to be as follows: HKY85 (for *mutS, recP*, and *yqiL*), HKY85+G (for *gtr, murI*, and *xpt*), and TrN+I+G (for *gki*). Scale bars indicate the number of nt substitutions per site. Trees are shown for *mutS, xpt*, and *yqiL*; the complete set of gene trees for all housekeeping loci is available from the authors upon request.

sites (Fig. 3 for *yqiL* and *mutS*; alignments of polymorphic nt sites for other loci are available from the authors upon request). However, in each of the gene trees, some GCS-GGS alleles are grouped with the GAS cluster I alleles (ranging from 10% of the total alleles for *gki* to 38% for *yqiL*; Table 3), and in several instances, identical alleles were found within GAS and GCS-GGS isolates (Table 4). In all instances, sequence similarity clusters I and II were supported by bootstrap values of 100%. The maximal percent divergence within cluster I or within cluster II for each of these five housekeeping loci ranged from 0.7 to 4.0% (Table 3).

The ML tree for *xpt* alleles (Fig. 2) and visual examination of the polymorphic sites (Fig. 3) do not show the clear distinction into GAS-like (cluster I) and GCS-GGS-like (cluster II) alleles that was observed for *gki, gtr, murI, mutS*, and *yqiL*. The rather uniform alleles that were typically present in GAS, and which were also present in some GCS-GGS, were considered cluster I alleles (Fig. 2). Alleles *xpt114* and *xpt115* (cluster IIb) were found only in GCS and GGS and were clearly distinct from the cluster I alleles. GCS-GGS alleles *xpt108* through *xpt113* (cluster IIa) were also distinct from cluster I, although two GAS alleles that shared some or all of the distinctive run of polymorphisms between nt 309 and 366 were evident (*xpt28* and *xpt29*). However, when considered separately, cluster IIa and IIb alleles each displayed lower values of maximal nt divergence with cluster I alleles than with each other (Table 3). Intragenic recombination or networked evolution involving *xpt* alleles from GAS and GCS-GGS is supported by split decomposition analysis (Fig. 4). Overall, the divergence between all *xpt* alleles is rather low (Tables 2 and 3), and there was less distinction between the alleles present in GAS and GCS-GGS than for the other loci.

The ML tree for *recP* provides a clear example of some GAS alleles (*recP07, recP15, recP21*, and *recP29*) that are grouped in cluster II. This assignment is supported by high bootstrap values and can be clearly observed by visual inspection of the polymorphic sites (Fig. 3). As with *xpt* alleles, a history of intragenic recombination between *recP* loci of GAS and GCS-GGS is supported by split decomposition analysis. Like *xpt*, the maximal divergence between cluster I and II *recP* alleles is relatively low (7.6%) and not much higher than the maximal divergence within each cluster (Table 3). One highly divergent *recP* allele (*recP116*) was found in a single strain (4236). This allele differed from all the other GCS-GGS alleles at 48 to 50% of nt sites.

Mosaic genomes of GCS and GGS. Housekeeping loci have been shown to be relatively uniform in GAS; unusually divergent alleles are only found at *recP* and were ascribed to the introduction of diverged sequences from closely related species (14). The uniformity of GAS housekeeping genes suggests that interspecies recombinational exchanges have been rare in this species. In sharp contrast, the 34 GCS and GGS isolates under evaluation exhibit a wide range in their content of sequence similarity cluster I and II alleles. Six strains have four cluster I alleles, involving four different combinations of loci; both GCS and GGS isolates are included in this group, and four *emm* types are represented (Table 1). Four isolates have three cluster I alleles and include both GCS and GGS strains. Only five strains were completely devoid of cluster I alleles at all seven housekeeping loci. The most likely explanation for the mosaic genomes of GCS and GGS isolates is that some loci received recombinational replacements from different species.

The distribution of cluster I and II alleles among GCS and GGS isolates fails to provide evidence that they represent two genetically distant populations. Cluster I alleles of the *gtr, mutS, recP, xpt,* and *yqiL* housekeeping loci were found among both GCS and GGS (Table 1, Fig. 2). The four cluster I alleles of *murI* were restricted to four GGS isolates; however, two were found in GGS isolates that were outside the principal GGS cluster constructed by pairwise differences in the allelic profiles assigned by MLST (Fig. 1). For all loci, sequence similarity cluster II alleles were distributed among both GCS and GGS strains (Table 1). In many instances, identical alleles were found in both GCS and GGS isolates. These include six cluster I and 17 cluster II alleles; all seven housekeeping loci were represented. Based on these data, neither the nature nor the extent of genome mosaicism appears to be different for GCS versus GGS isolates.

A total of 16 housekeeping alleles were identical among GAS and GCS-GGS (Table 4). Many of these shared alleles were highly abundant within the set of 212 GAS isolates, present in 10 to >25% of the total GAS isolates evaluated (*gtr06*/*gtr102, murI02*/*murI104, murI04*/*murI101, murI08*/*murI102, recP02*/*recP105, recP04*/*recP106, xpt02*/*xpt103, yqiL01*/*yqiL105*, and *yqiL04*/*yqiL102*). Of the 16 alleles shared by GCS-GGS and GAS, 11 were represented by three or more distinct *emm* types among GAS, and therefore, the high allelic frequencies observed in GAS are not strictly a reflection of a single dominant clone. Among the set of 34 GCS-GGS isolates, none of the shared alleles exceeded a frequency of 0.09.

Besides the 16 identical cluster I alleles, there were also cluster I alleles found in GCS-GGS isolates that were almost identical to known GAS alleles (Fig. 2 and 3). The set of 212 GAS isolates represent only half of the known *emm* types, and some of the GCS-GGS cluster I alleles may correspond to GAS alleles that have not yet been sampled. If this is the case, many of the variable sites in the GAS-like alleles should be found in the known GAS alleles, albeit in different combinations. Among the 14 cluster I alleles in GCS-GGS isolates that were similar but not identical to known GAS alleles, 64 of the 75 variable sites corresponded to polymorphisms present within GAS alleles. For the *recP* locus, it is likely that the four divergent alleles in the GAS isolates have been introduced from GCS or GGS. None of these divergent GAS alleles were identical to alleles in GCS or GGS, but all were very similar; of the 88 variable sites within the four GAS divergent alleles (*recP7, recP15, recP21*, and *recP29*), 77 were polymorphisms that are present in GCS-GGS alleles.

DISCUSSION

Historically, it has been suggested that human isolates of GCS and GGS should be separately classified as *S. equi* subsp. *equisimilis* and *S. dysgalactiae* subsp. *equisimilis*, respectively, based on differences in group carbohydrate. However, their homogeneity in other phenotypic traits supports their placement in a single taxonomic category (46, 47). The findings of this report show that the genetic difference between GCS and GGS isolates is roughly the same as the differences within the GCS or GGS subsets of isolates. Thus, there is no known

mutS Locus

```
                     11111222222222233333333333
          2347777889922356011122479901222344579
         25581238136918005516956966765457605459
MUTS01   CCCGTCGTACCGACAGTGCCTGCTGCCTGCCTATTTTA
MUTS02   ..TA........G.G.......................
MUTS06   ..........T.G..C.............T........
MUTS07   ............G.........................
MUTS101  ............G.........................[Same as MUTS07]
MUTS102  ........G..AG...C.....................
MUTS103  ....C..C....GA...AGACATC.G.CA.T.GC....
MUTS104  ....C..C....GA...AGATATC...CA...GC....
MUTS105  ....C..C....GA...AGACATC..TCA...GC....
MUTS106  ....C..C....GA...AGACATC...CA...GC....
```

yqiL Locus

```
                                     111111111111111111111111111111122222222222222222222333333333333333333333444444444444444
         1111222222333444666667778889000011111233344455666677788999901122334444455688899001223333566667888999000001111112222
        92589012789067567034892691799568923456937817903028907869125872325171346946702367462490236403492178078235681247890349
YQIL01  TTCACGTACCAACAAGGAGACATAATGGAACCGGATCATTGAGGATGAGCCCTATGGGAATCTCTAAATTAGGATCCCAAATAACAGTTATGAGTGGCGGACATGTGATAGCACGT
YQIL02  ..................................................T...............G.................................................
YQIL03  .............................G..A.................................G.................................................
YQIL04  ..................................................................G.................................................
YQIL05  ...........................A........................................................................................
YQIL101 ...G................................................................................................................[Same as YQIL12]
YQIL102 ..................................................................G.................................................[Same as YQIL04]
YQIL103 ................................A.................................G.................................................[Same as YQIL15]
YQIL104 ..................................................................G...............................................C.
YQIL107 AA..GAA.TGCTATTCACA.TC.TGGA.C.AT.ACAACAGATAT.AACAT..CCAA.ACTG.CACT.TGAGACG.TAACT.CTTTGACGTCTCACA..ATCTTG.CAGGGT.CT.A
YQIL108 AA.....TGCTATTCACA.TC.TGGA.C.AT.ACAACAGATAT.AACAT..CCAA.ACTG.CACT.TGAGACG.TAACT.CTTTGACGTCTCACA..ATCTTG.CAGGGT.CT.A
YQIL109 AA..........ATT..CA.TC.TGGA.C.AT.ACAACAGATCT.AACAT..CCAA.ACTG.CACT.TGAGACG.TAACT.CTTTGACGTCTCA.A..ATCTTG.CAGGGT.CT.A
YQIL110 AA..........ATT..CA.TC.TGGA.C.AT.ACAACAGATAT.AACAT..CCAA.ACTG.CACT.TGAGACG.TAACT.CTCTGACGTCTCACA..ATCTTG.CAGGGT.CT.A
```

recP Locus

```
                   1111111111112222222222222222222233333333333344444444
           133449990022345668880111122233344667778902344455889901123445
         19269253695869586251239036912802739140368774725849470731796479
RECP01   CCTCCTCATTTCGTATAAGGGTTGCCAACCAGCATCACTTACACGCCATAGCCGAATGGATC
RECP02   ............ACT..........................T.....C..............
RECP03   T.............................................................
RECP04   ............ACT........................G.T....................
RECP05   .....................A........................................
RECP06   ....................C.........................................
RECP07   ..C.T.AGAC.TACT.GGA..A.........A.........T...T....T.T...CC.TCT
RECP15   ..C.T.AG......TA.......TTTC.TG..T..TGT........TC..............
RECP21   ..C.T.AGAC.TACT.GGA..A.......................T....T.T...CC.TCT
RECP29   ..C.T.AGAC.TACTA.......T..CGTG..T..TGT..C.....TC......T.......
RECP101  ............ACT........................G.T........TCT...CC.TCT
RECP102  ............ACT..........................T.....C..T.T...CC.TCT
RECP103  ...........TACT........................G.T..........T...CC.TCT
RECP104  .......G....ACT..........................T.....C..............
RECP105  ............ACT..........................T.....C..............[Same as RECP02]
RECP106  ............ACT........................G.T....................[Same as RECP04]
RECP107  ..C.T.AGAC.TACT.GGA....TTTC.TG..T....T...T...T....T.T...CC.TCT
RECP108  ..C.T.AGAC.TACT.GGA....TTTC.TG..T....T...T...T....T.T...CT.TCT
RECP109  ..C.T.AGAC.TACT.GGA....TTTC.TG..T....T...T...T......T...CCTTCT
RECP110  ..C.T.AGAC.TACT.GGA....TTTC.TG..T....T...T...T......T...CC.TCT
```

xpt Locus

```
                         11111111222222222233333333333344
         133345556667778900135789256667899901355666789944
         356981463495682902940592454671845792947069096979
XPT01    CCTTTCGTAGTAGGGGGTTCTGTGGTGCGATCCACGTCACCGACACGT
XPT02    ....................C...................T...C...
XPT03    ....................C.......................C...
XPT04    T...................C.....T.................C...
XPT05    ...C................C.......................C...
XPT28    ....................C.........C......AGTT...C...
XPT29    ............A............G........T.CAGTT...CT..
XPT101   ............A...............................C...
XPT102   ............A.A.............................C...
XPT103   ....................C...................T...C...[Same as XPT02]
XPT111   ....A.......A............G........T.CAGTT...C...
XPT112   ............A.......C....G........T.CAGTT...C...
XPT113   ............A.....C.C....G........T.CAGTT...C...
XPT114   T.CC.TA.G.CG.A...A..C..A.G..................C..G
XPT115   T.CC.TA.G.CG.A...A..C..A.G..................C...
```

FIG. 3. Polymorphic nt sites among housekeeping alleles of GAS, GCS, and GGS. Polymorphic nt sites are shown for selected GCS and GGS alleles (three digits) and GAS alleles (two digits) (14). GCS-GGS alleles identical to a GAS allele are indicated. GenBank accession numbers for all GCS-GGS alleles listed in Table 1 are as follows: AF332626 to AF3326351 (*gki*), AF332636 to AF332651 (*gtr*), AF332797 to AF332808 (*murI*), AF330232 to AF330237 (*mutS*), AF332810 to AF332825 (*recP*), AF332826 to AF33284 (*xpt*), and AF332841 to AF332856 (*yqiL*). The *gtr* alleles are truncated 7 bp at the 5′ end (and 13 bp at the 3′ end) relative to the *gtr* alleles from GAS, as reported previously. Alignments are shown for some alleles of *mutS, recP, xpt*, and *yqiL*; the complete set of alignments for polymorphic nt sites among all GAS and GCS-GGS alleles is available from the authors upon request.

TABLE 3. Summary of sequence clusters for the housekeeping loci[a]

Locus	% of GAS alleles represented		% of GCS-GGS alleles represented		Maximal % divergence		
	Sequence cluster I	Sequence cluster II	Sequence cluster I	Sequence cluster II	Within cluster I	Within cluster II	Between clusters I and II
gki	100	0	10	90	2.7	2.5	12.6
gtr	100	0	25	75	2.6	2.4	24.5
murI	100	0	33	67	1.4	2.1	32.6
mutS	100	0	33	67	2.2	0.7	4.9
recP	89	11	38	56	3.1	6.4	7.6
xpt	93	7	47	53	1.8	4.9	4.9
yqiL	100	0	38	62	1.4	4	26.9

[a] Includes all alleles assigned to cluster I or II. The allele *recP116* from GGS strain 4236 formed a third cluster (not assigned to either cluster I or II) and is excluded from this analysis. Cluster II alleles of *xpt* form two distinct groups (cluster IIa and IIb); the maximal divergence between clusters I and IIa is 3.0%, and that between clusters I and IIb is 3.7%.

genetic basis for the separate classification of these particular GCS and GGS strains with the exception of genes encoding the group carbohydrate biosynthesis enzymes themselves, which remain to be characterized. The combined genotypic and phenotypic data support the singular grouping of the 34 GCS-GGS isolates as *S. dysgalactiae* subsp. *equisimilis*.

Despite their shared classification as *S. dysgalactiae* subsp. *equisimilis*, the 34 GCS and GGS isolates display wide variation in their phylogenetic relationships at the seven housekeeping loci. The maximal nt sequence divergence among alleles (excluding *recP116*) at a given locus ranges from ~5 to 32%. At most loci, GCS-GGS alleles segregate neatly into two sequence similarity clusters, whereby the maximal nt sequence divergence within a cluster ranges from 0.7 to 2.2% for cluster I alleles (excluding *gki*, which has only one cluster I allele) and from 0.7 to 4.9% for cluster II alleles. Furthermore, the 34 GCS and GGS isolates are highly heterogeneous in their overall content of cluster I and II alleles. The mosaic nature of the genomes of individual GCG and GGS isolates indicates that different loci have different evolutionary histories, and this can be most readily explained by interspecies recombinational exchanges.

In sharp contrast to the 34 GCS and GGS isolates, the diverse set of 212 GAS (*S. pyogenes*) isolates is nearly uniform in nt sequence at each housekeeping locus, with maximal divergence of <3.0% at all loci except *recP* (6.1%) (14). The relatively low levels of sequence diversity within each housekeeping locus of the GAS isolates suggest that introduction of alleles by interspecies recombination is rare. It was previously suggested that the diverged *recP* alleles found among a few GAS isolates arose via importation of DNA from a closely related species (14). The similarity of the diverged *recP* alleles of GAS to the most typical alleles found among GCS and GGS suggests that they were introduced into GAS from GCS-GGS strains.

At most housekeeping loci, the alleles that are most prevalent in the GCS-GGS isolates (i.e., cluster II alleles) are highly

TABLE 4. Housekeeping alleles shared by GAS and GCS-GGS[a]

Locus	Shared GCS-GGS allele	GCS-GGS *emm* types represented	Group carbohydrate represented	No. of GCS-GGS isolates represented	GCS-GGS allelic frequency	Shared GAS allele	GAS *emm* types represented	No. of GAS isolates represented	GAS allelic frequency
gtr	101	*stG4974, stG480*	G	2	0.0588	03	1, 87, *st4935*	18	0.0849
	102	*stC839, stG4831*	C, G	2	0.0588	06	3, 28, 49, 53, 60, 76, 77, 91, *st4973, st64/14, stD633, stNS5*	40	0.1887
	103	*stG480*	G	1	0.0294	13	93	1	0.0047
	104	*stC839, stG11*	C, G	3	0.0882	26	57	1	0.0047
murI	101	*stG480*	G	1	0.0294	04	1	22	0.1038
	102	*stG480*	G	1	0.0294	08	3, 4, 8, 19, 24, 26, 29, 34, 52, 66, 78, 89, 93, *st1RP31, st2370.1, st4973, st64/14*	55	0.2594
	103	*stG4831*	G	1	0.0294	24	70	1	0.0047
	104	*stG643*	G	1	0.0294	02	12, 14, 39, 49, 51, 55, 63, 77, 81, 83, 90, 91, *st3365, st4529, st4935, st833, stD432, stD633*	30	0.1415
mutS	101	*stG643*	G	1	0.0294	07	9, 11, 25, 42, 49, 57, 60, 81, 89, *stD432*	14	0.0660
recP	105	*stG10*	G	1	0.0294	02	2, 3, 33, 43, 53, 83, 90, 91, 92, 93, *st2370.1, st4935, st4973, stNS5*	46	0.2170
	106	*stG4974*	G	1	0.0294	04	1, 70, *st4529*	22	0.1038
xpt	103	*stG643*	C	1	0.0294	02	1, 4, 12, 33, 42, 52, 53, 60, 76, 87, 91, 92, *stNS5*	55	0.2594
yqiL	101	*stG4974*	G	1	0.0294	12	51, *st2370.1, st3365*	3	0.0142
	102	*stG480*	G	1	0.0294	04	1, 2, 4, 8, 11, 22, 25, 49, 50, 62, 68, 73, 76, 44/61, 90, *st833*	43	0.2029
	103	*stG480*	G	1	0.0294	15	88, 92	3	0.0142
	105	*stC839*	C	1	0.0294	01	1, 4, 9, 11, 13, 32, 39, 49, 66, 67, 78, 44/61, 82, 89, *st2346, st4592, st4935, st64/14*	48	0.2264

[a] Analysis includes 34 isolates of GCS and GGS (*S. dysgalactiae* subsp. *equisimilis*) and 212 isolates of GAS. Minimum possible allelic frequencies are 0.0294 and 0.0047 for GCS-GGS and GAS, respectively.

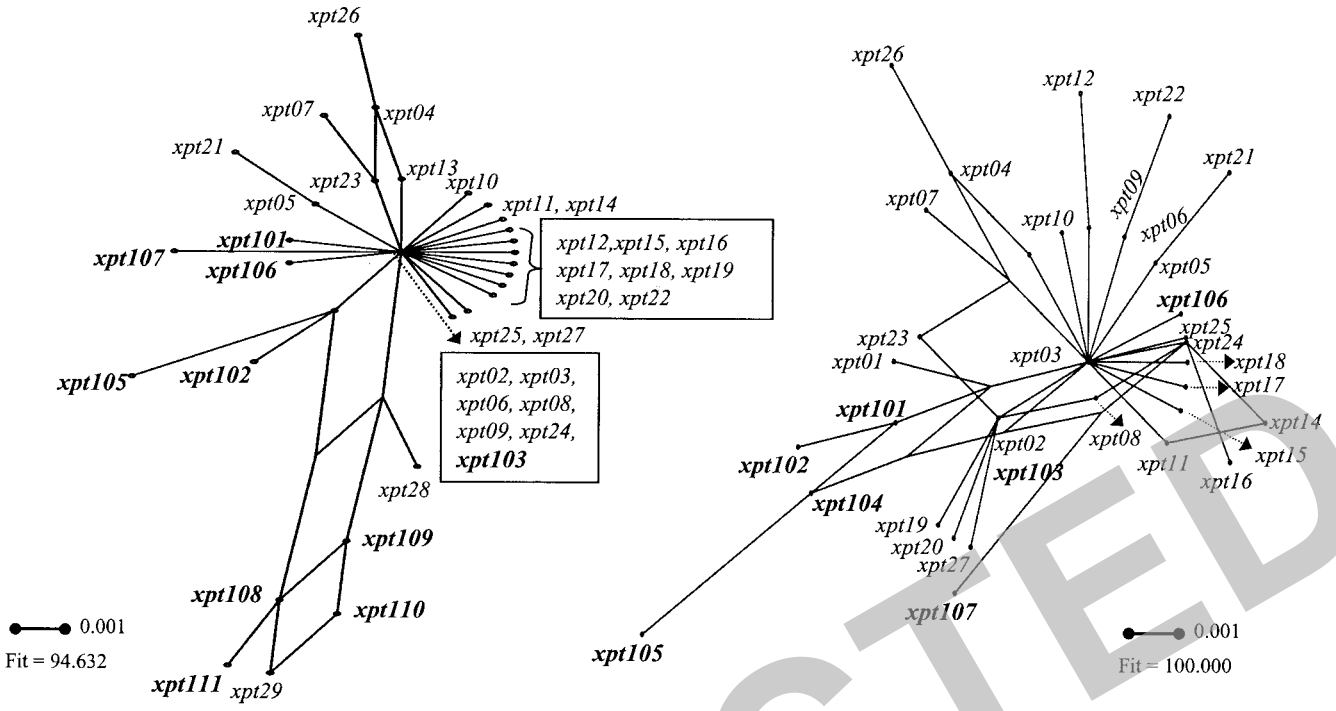

FIG. 4. Split decomposition analysis. The effect of recombination on evolutionary relationships was assessed by split decomposition (19, 22, 42). This method depicts multiple pathways linking sequences and allows visualization of the extent of conflicting phylogenetic signals. Thus, recombinational events are depicted as an interconnected network of phylogenetic relationships. A lack of treelike structure was evident for both *recP* and *xpt* alleles; splits graphs for *recP* are available upon request fron the authors. This finding was consistent with the visual inspection of polymorphic nt sites (Fig. 3). Annotated split decomposition graphs are shown for *xpt*. Graph on the left includes all alleles except *xpt112* through *xpt115* (all cluster II); these alleles were removed because their long branch lengths prevented resolution of central networks. Uncorrected Hamming distances were used; similar results were obtained with other estimated distance measures (Kimura 3-ST model and Jukes-Cantor). Branch lengths are drawn to scale. The fit parameters improved after the additional removal of alleles representing the longer branches, resulting in further resolution of inner branches (splits graph shown to the right). For *xpt* (right graph), removed alleles are the remaining cluster II alleles (*xpt108* through *xpt111, xpt28*, and *xpt29*). A fit parameter of 100% indicates that all phylogenetic information is represented by the graph. Splits graphs containing cluster I and II alleles for each of the *gki, gtr, murI, mutS*, and *yqiL* loci showed no evidence of networking.

divergent from those in GAS. However, at each of the seven housekeeping loci, there are one or more GCS-GGS alleles that are highly similar or identical to alleles found in GAS isolates; most GCS and GGS isolates possess GAS-like alleles (i.e., cluster I) at one or more loci. Since cluster II alleles are far more prevalent among GCS and GGS, we assume that they represent the clonal frame DNA (33) for this species, whereas the GCS-GGS alleles that are highly similar or identical to GAS alleles were imported from GAS donors during cocolonization or coinfection. Homologous recombination between genes having overall sequence divergence of 20 to 30%, as in the case of *gtr, murI*, and *yqiL* cluster I alleles (from GAS donors) and cluster II alleles (in GCS-GGS recipients), is not unprecedented in bacteria (29, 43).

The identification of 16 alleles (at six different loci) in GCS and GGS that are identical in sequence to GAS alleles suggests that many of the interspecies recombinational exchanges occurred recently, since there were no additional nt changes at the locus since their transfer. The proposed direction for most horizontal gene movements—from GAS donors to GCS and GGS recipients—is supported by the observed prevalence of shared alleles among GAS and GCS-GGS; since these transfers must have occurred recently, the population with the highest prevalence of the allele is most likely to be the donor. Of the 16 shared alleles, 9 are highly prevalent within GAS, being present in ~10 to 25% of the 212 GAS isolates sampled, whereas 13 of the shared alleles were present in only a single GCS or GGS isolate. The finding of GCS-GGS-like (cluster II) alleles of *recP* in a few GAS isolates (representing *emm* types 4 and 49 only) provides evidence that gene flow in the opposite direction, from GCS-GGS to GAS, can also occur.

The ability of some bacteria to acquire DNA by homologous recombination from other bacterial species via horizontal transfer can blur the definition of what constitutes a species. Clearly, no single gene tree provides an accurate indicator of the phylogenetic relationships between GAS, GCS, and GGS. Depending on the isolates and housekeeping locus, GAS, GCS, and GGS could appear to be indistinguishable or as much as 32% diverged in sequence. A blurring of taxonomic boundaries has been noted for human-commensal *Neisseria* species, also as a consequence of interspecies recombination (42). Sequences characteristic of *Neisseria lactamica* and *Neisseria cinerea* (and other human-commensal *Neisseria* spp.) are commonly found within *Neisseria meningitidis* housekeeping genes, and vice versa (50; E. J. Feil, J. Zhou, and B. G. Spratt, unpublished data). The phylogenetic relationships inferred from analysis of housekeeping genes of some viridans streptococci are also suggestive of a history of interspecies recombination (49). The difficulties encountered in unambiguously defining the taxonomy and phylogenetic relationships among

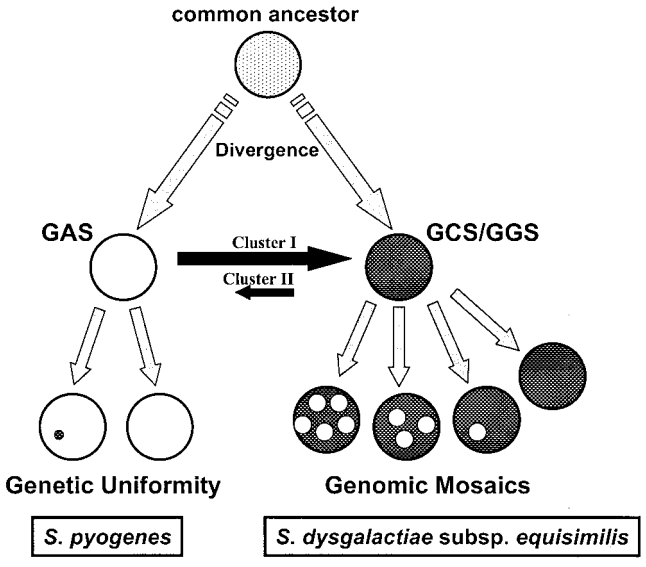

FIG. 5. Model for evolution of GAS, GCS, and GGS. Model for evolution based on selectively neutral housekeeping loci shows divergence of *S. pyogenes* and *S. dysgalactiae* subsp. *equisimilis* into cluster I and II alleles, respectively, followed by more recent interspecies gene flow that is dominated by movement of cluster I alleles from GAS to GCS-GGS. This model also shows that neutral gene flow from GAS donors to GCS-GGS recipients tends to involve larger blocks of DNA (allelic replacements) than the genetic material transferred from the reverse direction (intragenic recombination).

closely related *Neisseria* and viridans streptococci may, as in the case of *S. dysgalactaie* subsp. *equisimilis*, be rooted in the mosaic structure of the genomes of these organisms.

Despite close parallels between GCS-GGS and *Neisseria* spp. in terms of a key role for interspecies recombination in their evolution, there is also an important distinction. *N. meningitidis* has a broad range of sequence divergence between the alleles at each housekeeping locus and very clear evidence of mosaic gene structure, suggesting a long history of highly localized, intragenic recombinational events. In contrast, there is very sharp divergence between cluster I and II alleles at most loci of GCS and GGS, with an absence of alleles having intermediate levels of divergence, indicating that the importation of GAS alleles into GCS and GGS is a recent event.

The evolutionary model best supported by the findings of this report is that GAS (*S. pyogenes*) and GCS and GGS (*S. dysgalactiae* subsp. *equisimilis*) diverged in the relatively distant past and evolved along separate paths, which is reflected in the observed differences between the cluster I and cluster II housekeeping alleles (Fig. 5). Recently, genetic exchange with GAS isolates resulted in the replacement of the GCS-GGS alleles at some loci with those from GAS, to produce the mosaic genomes observed among the human isolates of GCS-GGS. The triggering event for the proposed recent gene movement between GAS and GCS-GGS is unknown. Humans are believed to be the sole biological host for *S. pyogenes*, whereas many beta-hemolytic streptococcal species bearing group C or G carbohydrate and forming large colonies, such as the *S. equi* complex, *S. dysgalactiae* subsp. *dysgalactiae*, and *S. canis*, are usually found in animals, including horses, dogs, and cattle. The increased exposure of humans to animals, which followed their domestication ~10,000 years ago, may have led to an expanded ecological niche for a few clones from streptococcal species previously associated only with animals. This, in turn, may have provided the human-adapted lineages with new opportunities for interspecies gene exchange with closely related streptococcal species already residing in humans. Future studies on comparative genomics of the various species of group A, C, and G streptococci should provide deeper insight into the origins of *S. dysgalactiae* subsp. *equisimilis*. That the domestication of animals led to the emergence of new human pathogens has been proposed for the origins of *Mycobacterium tuberculosis*, the causative agent of human tuberculosis, which is very closely related to the cattle pathogen *Mycobacterium bovis* (34).

The genetic uniformity at GAS housekeeping loci suggests that GAS have not frequently received recombinational replacements from GCS or GGS or from other related streptococci and that they may be poor recipients of genetic material from other species. It is unlikely that this is due to the lack of a mechanism for genetic exchange among GAS, since there is strong evidence for recombination from the disruption of deep phylogenetic relationships among GAS and from the occasional noncongruence between *emm* type and clone (14, 17, 26). Neither GAS or GCS-GGS are known to be naturally transformable, and generalized transduction by bacteriophage might conceivably be the primary mechanism for gene transfer (48); however, the vehicle(s) for lateral transfer of housekeeping loci remains undefined and cannot be deduced from our data.

The finding of identical alleles shared by GAS and GCS-GGS indicates that horizontal gene transfer occurs between these two populations, and therefore, the general failure of GAS to acquire cluster II alleles from GCS-GGS cannot be readily explained by ecological isolation. One possibility is that GAS are inherently less capable than GCS or GGS of acquiring DNA from related species by homologous recombination. Differences in the impact of restriction-modification, heteroduplex formation, and/or mismatch repair systems (29) or in the host cell preference for transducing phage might account for the observed polarity of gene transfer. In fact, such molecular restraints could explain the failure of penicillin resistance genes to spread into GAS from other streptococcal species (20). As an alternative possibility, the observed polarity could be due to stochastic effects arising from large differences in the population size of GAS versus GCS-GGS, with GAS being larger than GCS-GGS. However, in at least some parts of the world, the prevalence rates for GCS-GGS colonization exceed the rates for GAS (36).

An important distinguishing feature of GAS and GCS-GGS lies in the relationship between *emm* type and clone. Among GAS, most isolates of a given *emm* type are clones or form a clonal complex, as defined by STs with ≥5 housekeeping alleles in common (14). In sharp contrast, all 34 GCS and GGS isolates have distinct allelic profiles, and only one pair of isolates bearing the same *emm* type have similar genotypes. A critical factor may be differential rates of genetic recombination for GAS versus GCS-GGS. For GAS, the lack of congruency between gene trees of the different housekeeping loci suggests that over the long term, recombination has been sufficient to eliminate the phylogenetic signal in these trees (17). However, the index of association (32) shows significant link-

age disequilibrium among housekeeping loci of GAS when all 100 defined clones are included in the measure. Only when one representative of each clonal complex is considered—selected by truncation of the dendrogram at a genetic distance of 0.3, to yield 72 STs—is there an absence of significant linkage disequilibrium (14). Thus, recombination rates within GAS, though relatively high, might not be sufficient to break up associations between allelic profiles of housekeeping genes and *emm* type. The GCS and GGS of *S. dysgalactiae* subsp. *equisimilis* show no departure from linkage equilibrium when all 34 STs are compared, and therefore, these organisms could be subject to higher rates of recombination than GAS.

A second factor that may contribute to the strength of the association between *emm* type and ST is the intensity of host immune selection on the *emm* gene product. Protective immunity against GAS infection is due to antibodies directed against the type-specific portion of *emm* gene products (M proteins) (11, 28). Horizontal movement of a given *emm* type to a new genetic background could lead to competition between the donor and recipient GAS strains, mediated through a shared immunity; the ultimate result would be exclusion of the "least fit" strain (18). However, for GCS and GGS, it remains unproven that M-type-specific immunity is protective, and there is little evidence that patients produce type-specific antibodies during the course of infection. In general terms, organisms that cause asymptomatic carriage tend not to provoke a robust host immune response (24). If the *emm* gene of one GCS or GGS strain moves to another GCS or GGS strain in an environment where selective immune pressures are absent, competition between the donor and recipient strains will be weak and allow coexistence of both strains.

The vast majority of GCS and GGS isolates chosen for this study were found in association with diseased states in humans, and many were recovered from normally sterile tissue sites. Invasive disease is a rare outcome of carriage or infection with either GAS or GCS-GGS. For at least some clones of GAS, invasive disease is highly dependent on host susceptibility factors, and the most virulent clone can also be the most prevalent clone in the community (9). It remains unknown whether there are intrinsic differences, independent of transmission rates, in the ability of individual clones to cause invasive disease. Future studies of value might measure the prevalence of particular clones of GCS-GGS in association with asymptomatic carriage versus disease and thereby determine whether biologically defined clones differ in their genetic relationships with GAS. It may be that GCS-GGS clones which harbor a high proportion of cluster I alleles are also the most similar to GAS in their disease-causing capacity.

Whether or not the observed directionality of neutral gene flow, from GAS to GCS-GGS, has important implications for the evolution of virulence in streptococci remains to be established. Loci encoding several pathogenicity factors of GAS are present in at least some (if not all) GCS-GGS isolates, including M proteins, fibronectin-binding protein F, streptokinase, C5a peptidase, and streptolysin O; several of the shared loci are highly polymorphic in GAS, and the structural variants often confer unique biological properties. However, the dynamics of neutral versus adaptive gene flow are not necessarily the same, since the effect of natural selection on adaptive loci will be more profound. Also, the recombinational mechanism (site specific versus homologous) associated with the successful movement of a particular gene, as well as the host cell range of its vehicle, can influence the direction of gene movement. For example, the *speA* and *speC* genes of GAS, which encode exotoxins that contribute to virulence, are present on bacteriophage and can move into a recipient GAS (and possibly into GCS-GGS) by specialized transduction. While it might be expected that specialized transduction is an efficient way of spreading *speA* and *speC* genes among GAS, they have restricted distributions within GAS for reasons that are not yet known (3). Other virulence genes of GAS that lack homologs in GCS and GGS might be acquired by generalized transduction, followed by recombination that is promoted by homologous flanking loci. Although it is unlikely that we will fully understand the events which influenced the evolutionary history of GAS and GCS-GGS, it would nonetheless be important to know whether a gradual increase in the virulence of GCS and GGS for humans resulting from the horizontal movement of GAS virulence genes is occurring.

ACKNOWLEDGMENTS

We thank Yury Nunez for expert technical assistance; Bernard Beall, Richard Facklam, Androulla Efstratiou, and Vincent Fischetti for providing strains and/or epidemiological information; and Ed Feil, Eddie Holmes, and Ed Kaplan for helpful insights and comments.

This work was supported by grants from the Wellcome Trust (to B.G.S.), the National Institutes of Health (AI-28944 to D.E.B. and GM-60793 to D.E.B. and B.G.S.), the American Heart Association (Grant-in-Aid to D.E.B.), and a Brown-Coxe Postdoctoral Fellowship (to A.K.). M.C.E. is a Royal Society University Research Fellow. D.E.B. is an Established Investigator of the American Heart Association.

REFERENCES

1. **Beall, B., R. Facklam, and T. Thompson.** 1996. Sequencing *emm*-specific PCR products for routine and accurate typing of group A streptococci. J. Clin. Microbiol. **34:**953–958.
2. **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.
3. **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.
4. **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.
5. **Bradley, S. F., J. J. Gordon, D. D. Baumgartner, W. A. Marasco, and C. A. Kauffman.** 1991. Group C streptococcal bacteremia: analysis of 88 cases. Rev. Infect. Dis. **13:**270–280.
6. **Carmeli, Y., and K. L. Ruoff.** 1995. Report of cases and taxonomic considerations for large-colony-foming Lancefield group C streptococcal bacteremia. J. Clin. Microbiol. **33:**2114–2117.
7. **Chhatwal, G. S., and S. R. Talay.** 2000. Pathogenicity factors in C and G streptococci, p. 177– 183. *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.
8. **Cleary, P. P., J. Peterson, C. Chen, and C. Nelson.** 1991. Virulent human strains of group G streptococci express a C5a peptidase enzyme similar to that produced by group A streptococci. Infect. Immun. **59:**2305–2310.
9. **Cockerill, F. R., K. L. MacDonald, R. L. Thompson, F. Roberson, P. C. Kohner, J. Besser-Wiek, J. M. Manahan, J. M. Musser, P. M. Schlievert, J. Talbot, B. Frankfort, J. M. Steckelberg, W. R. Wilson, and M. T. Osterholm.** 1997. An outbreak of invasive group A streptococcal disease associated with high carriage rates of the invasive clone among school-aged children. J. Am. Med. Assoc. **277:**38–43.
10. **Collins, C. M., A. Kimura, and A. L. Bisno.** 1992. Group G streptococcal M protein exhibits structural features analogous to those of class I M protein of group A streptococci. Infect. Immun. **60:**3689–3696.

11. **Cunningham, M. W.** 2000. Pathogenesis of group A streptococcal infections. Clin. Microbiol. Rev. **13:**470–511.
12. **Efstratiou, A.** 1997. Pyogenic streptococci of Lancefield groups C and G as pathogens in man. J. Appl. Microbiol. Symp. Suppl. **83:**72S–79S.
13. **Efstratiou, A.** 1983. The serotyping of hospital strains of streptococci belonging to Lancefield group C and G. J. Hyg. (Cambridge) **90:**71–80.
14. **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.
15. **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.
16. **Facklam, R., and J. Washington.** 1991. Streptococci and related catalase negative gram positive cocci, p. 238–257. *In* A. Balows, W. Hausler, K. Hermann, H. Isenberg, and H. Shadony (ed.), Manual of clinical microbiology, 5th ed. ASM Press, Washington, D.C.
17. **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.
18. **Gupta, S., and R. Anderson.** 1999. Population structure of pathogens: the role of immune selection. Parasitol. Today **15:**497–501.
19. **Holmes, E. C., R. Urwin, and M. C. J. Maiden.** 1999. The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. Mol. Biol. Evol. **16:**741–749.
20. **Horn, D., J. Zabriskie, R. Austrian, P. Cleary, J. Ferretti, V. Fischetti, E. Gotschlich, E. Kaplan, M. McCarty, S. Opal, R. Roberts, A. Tomasz, and Y. Wachtfogel.** 1998. Why have group A streptococci remained susceptible to penicillin? Report on a symposium. Clin. Infect. Dis. **26:**1341–1345.
21. **Huelsenbeck, J., and B. Rannala.** 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science **276:**227–232.
22. **Huson, D.** 1998. SplitsTree: a program for analyzing and visualizing evolutionary data. Bioinformatics **14:**68–73.
23. **Johnson, C., and A. Tunkel.** 2000. Viridans streptococci and groups C and G streptococci, p. 2167–2182. *In* G. Mandell, J. Bennett, and R. Dolin (ed.), Principles and practice of infectious diseases. Churchill Livingstone, Philadelphia, Pa.
24. **Kaplan, E. L.** 1980. The group A streptococcal upper respiratory tract carrier state: an enigma. J. Pediatr. **97:**337–345.
25. **Kapur, V., S. Kanjilal, M. R. Hamrick, L.-L. Li, T. S. Whittam, S. A. Sawyer, and J. M. Musser.** 1995. Molecular population genetic analysis of the streptokinase gene of *Streptococcus pyogenes* mosaic alleles generated by recombination. Mol. Microbiol. **16:**509–519.
26. **Kehoe, M. A., V. Kapur, A. M. Whatmore, and J. M. Musser.** 1996. Horizontal gene transfer among group A streptococci: implications for pathogenesis and epidemiology. Trends Microbiol. **4:**436–443.
27. **Kline, J. B., S. Xu, A. L. Bisno, and C. M. Collins.** 1996. Identification of a fibronectin-binding protein (GfbA) in pathogenic group G streptococci. Infect. Immun. **64:**2122–2129.
28. **Lancefield, R. C.** 1962. Current knowledge of the type specific M antigens of group A streptococci. J. Immunol. **89:**307–313.
29. **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.
30. **Malke, H.** 2000. Genetics and pathogenicity factors of group C and G streptococci, p. 163–176. *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.
31. **Martin, N. J., E. L. Kaplan, M. A. Gerber, M. A. Menegus, M. Randolph, K. Bell, and P. P. Cleary.** 1990. Comparison of epidemic and endemic group G streptococci by restriction enzyme analysis. J. Clin. Microbiol. **28:**1881–1886.
32. **Maynard Smith, J., N. Smith, M. O'Rourke, and B. Spratt.** 1993. How clonal are bacteria? Proc. Natl. Acad. Sci. USA **90:**4384–4388.
33. **Milkman, R., and M. M. Bridges.** 1990. Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. Genetics **126:**505–517.
34. **Musser, J. M.** 1996. Molecular population genetic analysis of emerged bacterial pathogens: selected insights. Emerg. Infect. Dis. **2:**1–17.
35. **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.
36. **Ogunbi, O., Q. Lasi, and S. F. Lawal.** 1974. An epidemiological study of beta-hemolytic streptococcal infections in a Nigerian (Lagos) urban population, p. 281–284. *In* M. J. Haverkorn (ed.), Streptococcal disease and the community. Excerpta Medica, Amsterdam, The Netherlands.
37. **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.
38. **Posada, D., and K. Crandall.** 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14:**817–818.
39. **Schnitzler, N., A. Podbielski, G. Baumgarten, M. Mignon, and A. Kaufhold.** 1995. M or M-like protein gene polymorphisms in human group G streptococci. J. Clin. Microbiol. **33:**356–363.
40. **Simpson, W. J., J. M. Musser, and P. P. Cleary.** 1992. Evidence consistent with horizontal transfer of the gene (*emm12*) encoding serotype M12 protein between group A and group G pathogenic streptococci. Infect. Immun. **60:**1890–1893.
41. **Sjobring, U., L. Bjorck, and W. Kastern.** 1989. Protein G genes: structure and distribution of IgG-binding and albumin-binding domains. Mol. Microbiol. **3:**319–327.
42. **Smith, N., E. Holmes, G. Donovan, G. Carpenter, and B. Spratt.** 1999. Networks and groups within the genus *Neisseria*: analysis of *argF, recA, rho*, and 16S rRNA sequences from human *Neisseria* species. Mol. Biol. Evol. **16:**773–783.
43. **Spratt, B. G., Q. Zhang, D. M. Jones, A. Hutchison, and J. A. Brannigan.** 1989. Recruitment of a penicillin-binding protein gene from *Neisseria flavescens* during the emergence of penicillin resistance in *Neisseria meningitidis*. Proc. Natl. Acad. Sci. USA **86:**8988–8992.
44. **Sriprakash, K. S., and J. Hartas.** 1996. Lateral genetic transfers between group A and G streptococci for M-like genes are ongoing. Microb. Pathog. **20:**275–285.
45. **Turner, J. C., F. G. Hayden, M. C. Lobo, C. E. Ramirez, and D. Murren.** 1997. Epidemiologic evidence for Lancefield group C beta-hemolytic streptococci as a cause of exudative pharyngitis in college students. J. Clin. Microbiol. **35:**1–4.
46. **Vandamme, P., B. Pot, E. Falsen, K. Kersters, and L. A. Devriese.** 1996. Taxonomic study of Lancefield groups C, G, and L (*Streptococcus dysgalactiae*) and proposal of *S. dysgalactiae* subsp. *equisimilis* subsp. nov. Int. J. Syst. Bacteriol. **46:**774–781.
47. **Vieira, V. V., L. M. Teixeira, V. Zahner, H. Momen, R. R. Facklam, A. G. Steigerwalt, D. J. Brenner, and A. C. D. Castro.** 1998. Genetic relationships among the different phenotypes of *Streptococcus dysgalactiae* strains. Int. J. Syst. Bacteriol. **48:**1231–1243.
48. **Wannamaker, L. W., S. Almquist, and S. Skjold.** 1973. Intergroup phage reactions and transduction between group C and group A streptococci. J. Exp. Med. **137:**1338–1353.
49. **Whatmore, E., A. Efstratiou, A. Pickerill, K. Broughton, G. Woodard, D. Sturgeon, R. George, and C. Dowson.** 2000. Genetic relationships between clinical isolates of *Streptococcus pneumoniae, Streptococcus oralis*, and *Streptococcus mitis*: characterization of "atypical" pneumococci and organisms allied to *S. mitis* harboring *S. pneumoniae* virulence factor-encoding genes. Infect. Immun. **68:**1374–1382.
50. **Zhou, J., L. D. Bowler, and B. G. Spratt.** 1997. Interspecies recombination, and phylogenetic distortions, within the glutamine synthetase and shikimate dehydrogenase genes of *Neisseria meningitidis* and commensal *Neisseria* species. Mol. Microbiol. **23:**799–812.

Editor: E. I. Tuomanen

INFECTION AND IMMUNITY, Oct. 2009, p. 4688 Vol. 77, No. 10
0019-9567/09/$08.00+0 doi:10.1128/IAI.00966-09

RETRACTION

Directional Gene Movement from Human-Pathogenic to Commensal-Like Streptococci

Awdhesh Kalia, Mark C. Enright, Brian G. Spratt, and Debra E. Bessen

Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut, and Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, and Department of Infectious Disease Epidemiology, Imperial College School of Medicine, University of London, St. Mary's Campus, London W2 IPG, United Kingdom

Volume 69, no. 8, p. 4858–4869, 2001: We retract the article.

We have uncertainties as to the overall strength of the main conclusion on interspecies gene transfer of highly divergent housekeeping gene alleles from group A streptococci (GAS) to group C and G streptococci (GCS/GGS), due to an inability to replicate some of the original GCS/GGS-derived sequence data. The reasons may be that (i) the GCS/GGS DNA template used for PCR amplification was contaminated with trace amounts of GAS DNA and/or (ii) there are subtleties with the PCR amplification conditions that could not be replicated due to changes in commercial reagents, thermal cyclers, etc. A clear-cut resolution of this matter may require whole-genome sequencing and assembly for at least a few GCS/GGS strains.