JOURNAL OF BACTERIOLOGY, July 2004, p. 4285–4294 Vol. 186, No. 13
0021-9193/04/$08.00+0 DOI: 10.1128/JB.186.13.4285–4294.2004

Multilocus Sequence Typing of *Streptococcus pyogenes* Representing Most Known *emm* Types and Distinctions among Subpopulation Genetic Structures

Karen F. McGregor,[1] Brian G. Spratt,[1] Awdhesh Kalia,[2,3] Alicia Bennett,[3] Nicole Bilek,[1] Bernard Beall,[4] and Debra E. Bessen[3,5]*

Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom[1]; Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, Missouri[2]; Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut[3]; Respiratory Diseases Branch, Centers for Disease Control and Prevention, Atlanta, Georgia[4]; and Department of Microbiology and Immunology, New York Medical College, Valhalla, New York[5]

Received 17 December 2003/Accepted 1 April 2004

A long-term goal is to characterize the full range of genetic diversity within *Streptococcus pyogenes* as it exists in the world today. Since the *emm* locus is subject to strong diversifying selection, *emm* type was used as a guide for identifying a genetically diverse set of strains. This report contains a description of multilocus sequence typing based on seven housekeeping loci for 495 isolates representing 158 *emm* types, yielding 238 unique combinations of sequence type and *emm* type. A genotypic marker for tissue site preference (*emm* pattern) revealed that only 17% of the *emm* types displayed the marker representing strong preference for infection at the throat and that 39% of *emm* types had the marker for skin tropism, whereas 41% of *emm* types harbored the marker for no obvious tissue site preference. As a group, the *emm* types bearing the *emm* pattern marker indicative of no obvious tissue site preference were far less likely to have two distinct *emm* types associated with the same sequence type than either of the two subpopulations having markers for strong tissue tropisms ($P < 0.002$). In addition, all genetic diversification events clearly ascribed to a recombinational mechanism involved strains of only two of the *emm* pattern-defined subpopulations, those representing skin specialists and generalists. The findings suggest that the population genetic structure differs for the tissue-defined subpopulations of *S. pyogenes*. The observed differences may partly reflect differential host immune selection pressures.

Streptococcus pyogenes, also known as group A beta-hemolytic streptococcus (GAS), is among the most highly prevalent bacterial pathogens and has a worldwide distribution. Humans are the only known biological host. Although these organisms can cause severe invasive disease or give rise to an asymptomatic carrier state, most often they cause mild disease by infecting the upper respiratory tract or skin, resulting in pharyngitis or impetigo, respectively (7). The relative incidence of GAS disease varies throughout the world, in accordance with both season and locale. In the temperate regions of North America and Europe, pharyngitis is highly prevalent during the winter months and impetigo (although less common) is most often encountered during warmer weather. In many tropical regions, GAS impetigo is far more common than pharyngeal infection (4), and there may be no discrete seasonal peaks in incidence of disease (30).

Numerous typing schemes have been used to characterize and measure the genetic diversity among isolates of *S. pyogenes*. Perhaps the most common tool used today is *emm* typing (3, 13), which is based on sequence at the 5′ end of a locus (*emm*) that is present in all isolates. The targeted region of *emm* displays the highest level of sequence polymorphism known for a widely distributed *S. pyogenes* gene; >150 *emm* types have been described to date (B. Beall, http://www.cdc.gov/ncidod/biotech/strep/emmtypes.htm). *emm* encodes the M protein, which forms the basis of a serological typing scheme (28). For many M proteins, the type-specific epitopes elicit strong host protective immunity (23).

There are four major subfamilies of *emm* genes, which are defined by sequence differences within the 3′ end, encoding the peptidoglycan-spanning domain (22). The chromosomal arrangement of *emm* subfamily genes reveals five major *emm* patterns, denoted *emm* patterns A through E (6); strains with patterns B and C are rare and are currently grouped with *emm* pattern A strains (referred to as pattern A-C strains). A given isolate of *S. pyogenes* has one, two, or three *emm* genes lying in tandem on the chromosome, and each gene differs in sequence from the others. In strains having three *emm* genes, the determinants of *emm* type lie within the central *emm* locus.

The *emm* pattern A-C strains are usually recovered from cases of pharyngitis, whereas *emm* pattern D strains are most often isolated from impetigo lesions (4, 6, 10). As a group, *emm* pattern E strains are readily found at both primary tissue sites. For example, in tropical Australia, 84% of isolates recovered by population-based surveillance of an aboriginal community experiencing high rates of streptococcal impetigo and no cases of pharyngitis were either *emm* pattern D or E (4). In Rome, 98% of pharyngitis isolates were of *emm* types associated with *emm* pattern A-C or E (10). Thus, *emm* pattern can

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

serve as a genotypic marker for tissue site preferences among *S. pyogenes* strains.

Multilocus sequence typing (MLST) is a relatively new tool for molecular typing of bacteria (8, 33). A principal advantage of MLST over gel-based methods is that the sequence data, which are generated for several neutral housekeeping loci, are unambiguous, electronically portable, and readily queried via the Internet (www.mlst.net). In this report, MLST and *emm* pattern determination are performed for many previously untested *emm* types of *S. pyogenes*. When these data are combined with data from previous reports (10, 12, 31), it is found that the large majority of known *emm* types (http://www.cdc.gov/ncidod/biotech/strep/emmtypes.htm) are represented. An analysis of the relationships among *emm* type, *emm* pattern, and the genetic relatedness defined by MLST is presented.

MATERIALS AND METHODS

Bacteria. The 107 new GAS isolates under study are listed in Table 1. Selection of bacterial isolates for this study was largely guided by knowledge of previously determined *emm* types, for the purpose of assembling a strain set having maximal diversity in *emm* type. Some new isolates were selected in order to ascertain *emm* pattern for more than one isolate of a given *emm* type. Strains designated SS followed by a string of three or four numerals are part of the Centers for Disease Control and Prevention (CDC) strain collection; strains designated with two, three, or four numerals followed by a hyphen and then two additional numerals, where the last two numerals in the series represent year of acquisition, are also part of the CDC strain collection; additional epidemiological information is posted at http://www.cdc.gov/ncidod/biotech/strep/emmtypes.htm for many strains. All other isolates from Australia were provided by K. Sriprakash and B. Currie. Strain CT95-201 was obtained from the State of Connecticut Department of Health (18). All other strains were obtained from the Lancefield collection (The Rockefeller University, New York, N.Y.).

Of the 495 isolates, the tissue site of isolation was unknown for 73 (15%), 183 (37%) were derived from impetigo lesions, 136 (27%) were recovered from normally sterile tissue sites, and 103 (21%) were from the upper respiratory tract. Of the upper respiratory tract isolates, 23 (22%) were known to be recovered from subjects with no disease symptoms, whereas 54 (52%) were definitively associated with disease. Isolates from impetigo lesions and normally sterile sites are, by definition, disease associated.

***emm* sequence typing.** *emm* type, which closely corresponds to M serotype, was ascertained by nucleotide sequence determination as previously described (3, 12, 29); a unique *emm* type is defined as having <95% sequence identity to any other known type over the first 160 bp of sequence, allowing for small indels. A complete and current listing of GAS *emm* types is posted at ftp://ftp.cdc.gov/pub/infectious_diseases/biotech/emmsequ/ and http://www.cdc.gov/ncidod/biotech/strep/emmtypes.htm). *emm* pattern was determined by a PCR-based method, as previously described (4).

MLST. Internal fragments of seven housekeeping genes (*gki*, *gtr*, *murI*, *mutS*, *recP*, *xpt*, and *yqiL*) were amplified and sequenced with primers and under conditions described previously (12). For each locus, distinct allele numbers were assigned to each unique sequence, generating a seven-integer allelic profile for each isolate. Isolates with identical allelic profiles were assigned to the same sequence type (ST). A complete database of alleles, allele sequences, and STs is maintained on the Internet at www.mlst.net.

Additional nucleotide sequence determination. Using bacterial DNA as a template, PCR amplification products were generated (annealing temperature, 50 or 55°C) with the following oligonucleotide primers: for the *cpa* locus, 5′-GGA TAT GAG ATT GCC GAA CCT ATT ACT TTT AAA G-3′ (forward) and 5′-GGA GCC TGT TTA TCT TCC ATT CGA ATA ATA TCC AC-3′ (reverse) (product size, ~600 bp); for the *prtF1* locus, 5′-TGC GCG GGT TCT ATC GGT TTT GGT CAA GTA-3′ (forward) and 5′-AAT TAG TTT T(T/C)T CA(G/A) (T/A)GC (T/C)TC ACG CAT TAA-3′ (reverse) (product size, ~360 bp). The same primers were used for nucleotide sequence determination.

Computational analysis. Sequence (nucleotide and amino acid) alignments and percent sequence identity calculations were performed with Clustal W (DNAStar; version 5.05). The eBURST algorithm was applied with software available at http://eburst.mlst.net (15). Average distances between STs was calculated by the START-distance matrix method (www.mlst.net). For tests for independence, Fisher's two-tailed exact test was used (DnaSP; version 3.99).

Nucleotide sequence accession numbers. The new housekeeping allele sequences generated as part of this report were submitted to GenBank and assigned accession numbers AY520918 through AY521006. The new allele sequences associated with the *cpa* and *prtF1* loci were submitted to GenBank and assigned accession numbers AY579608 through AY579635.

RESULTS

MLST of GAS. Allelic profiles at seven housekeeping loci were determined for 107 isolates of GAS (Table 1). The majority of isolates (75%) represent *emm* types not previously reported on for MLST. When these data were combined with previously reported data on 388 GAS isolates, 220 STs were recognized (10, 12, 31) (www.mlst.net). The total number of alleles at each housekeeping locus ranged from 36 (*mutS*) to 66 (*gki*). Collectively, 158 distinct *emm* types are included in the set of 495 isolates and 238 unique *emm* type-ST combinations can be defined. These 158 *emm* types represent the large majority of known *emm* types found in GAS, as defined by ftp://ftp.cdc.gov/pub/infectious_diseases/biotech/emmsequ/ and http://www.cdc.gov/ncidod/biotech/strep/emmtypes.htm.

Markers for tissue site preference. *emm* pattern serves as a useful genotypic marker for tissue site preferences of individual strains and clones. Of the 158 *emm* types represented within the complete set of 495 isolates, *emm* pattern was established for one or more isolates of 156 *emm* types (Table 2). Of the 76 *emm* types for which *emm* pattern was determined for two or more isolates, 74 (97%) of the *emm* types included isolates belonging to a single *emm* pattern group (i.e., A-C, D, or E). Only two *emm* types (54 and st854) were found in association with two *emm* pattern groups. Therefore, isolates of a given *emm* type usually have the same *emm* pattern grouping.

The classical throat strains (*emm* pattern A-C) displayed the least diversity in *emm* type, accounting for only 17% of the 156 *emm* types that could be assigned an *emm* pattern (Table 2). *emm* types associated with patterns D and E were most abundant, representing 39 and 41%, respectively, of the total *emm* types. Two of the *emm* types (st1815 and st211) had a rearranged *emm* region. The data show that *emm* pattern D and E strains display the most diversity in *emm* type, whereas pattern A-C strains display the least.

The relationship between *emm* pattern subpopulations and genetic diversity, as defined by MLST, was also evaluated. Of the 220 STs resolved by MLST, *emm* pattern was determined for at least one representative of 202 STs. The classical throat strains (*emm* pattern A-C) displayed the least genetic diversity in their allelic profiles, accounting for only 18% of the 202 STs examined. STs associated with patterns D and E were most abundant, representing 36 and 47%, respectively, of the total number of STs. The data show that *emm* pattern E strains, as a group, display the most diversity in ST, whereas pattern A-C strains display the least. Pattern D strains are intermediate in their overall diversity of STs.

Relationships among STs. Of the 220 STs of GAS, the average distance from an ST to all other STs was 6.21 housekeeping alleles, calculated by the START-distance matrix method. The mean distance of an ST to the ST with the most similar allelic profile was 2.35 housekeeping alleles. Thus, many STs are distally related to all others.

eBURST is an algorithm that can be used to subdivide

TABLE 1. New GAS isolates associated with this report

Original strain name	emm type(s)	ST[a]	Housekeeping allele assignments							Country of origin[b]
			gki	*gtr*	*murI*	*mutS*	*recP*	*xpt*	*yqiL*	
3934–98	1	106*	4	25	3	4	4	2	1	Brazil
4650–97	1	107*	4	25	26	4	4	2	1	Brazil
6745–99	4	105*	35	2	2	5	13	30	1	Brazil
A410	11	195	3	4	6	7	1	5	36	Unknown
T14/46	14	84	20	22	2	3	27	18	12	United Kingdom
1269–98	15	121	47	3	2	5	10	34	12	USA
SS800	15	163	13	2	33	1	13	13	28	Unknown
J17E/165	17	196	19	50	1	1	5	3	3	Unknown
SS580	23	160	10	2	7	27	5	37	4	Unknown
SS730	23	160	10	2	7	27	5	37	4	United Kingdom
SS109	30	65	13	7	8	1	13	9	8	Unknown
SS901	31	316	13	7	57	1	13	9	8	Unknown
SS53	37	155	10	19	30	27	5	2	3	USA
SS795	41	162	3	2	32	2	2	2	2	Unknown
SS642	46	158	13	7	8	1	13	3	13	USA
SS116	47	83	19	21	1	1	5	3	3	Unknown
SS737	48	161	16	41	2	29	2	26	7	Unknown
A889	52	4	2	2	1	2	2	3	2	USA
SS686	52	4	2	2	1	2	2	3	2	USA
NS342	53	299	2	7	1	3	2	14	2	Australia
2470–93	54	122	4	6	1	28	2	2	2	Papua New Guinea
SS725	54	159	13	7	8	1	42	13	1	Unknown
SS934	55	100	34	2	2	21	1	29	16	Trinidad
546–96	56	126	2	2	8	13	18	14	2	Unknown
4756–01	59	172	56	24	39	7	30	2	38	Unknown
CT95–201	59	172	56	24	39	7	30	2	33	USA
2954–97	64	124	2	2	8	3	5	2	2	Chile
SS1451	64	124	2	2	8	3	5	2	2	Unknown
SS989	64	164	2	2	8	3	5	2	29	Israel
SS1098	71	130	13	2	35	7	2	38	1	Egypt
SS1144	72	131	2	2	36	3	18	14	2	Trinidad
SS132	77	133	4	31	2	31	2	8	7	Unknown
SS1151	79	132	50	2	8	30	8	3	30	Unknown
SS1401	81	171	11	23	2	7	5	3	2	Unknown
SS1402	82	26	4	2	21	16	17	3	1	United Kingdom
SS1400	83	103*	2	2	2	3	37	3	2	Brazil
SS1447	85	109	42	37	8	22	33	3	4	Unknown
4539–96	87	102*	13	34	25	6	36	3	1	Malaysia
292–00	89	101*	16	2	8	3	1	19	3	Italy
SS1475	94	89	24	2	3	5	1	3	1	Unknown
SS1343	95	14	2	6	8	3	9	3	1	Unknown
SS1432	97	197	11	6	42	5	54	3	6	Australia
SS1434	98	10	2	2	9	13	2	14	2	Australia
SS1433	99	141	52	31	9	6	35	2	12	Australia
1523–00	102	60	13	2	14	1	9	3	1	USA
SS1437	102	185	13	45	8	4	34	3	6	Australia
SS1370	103	327	83	2	8	6	74	3	4	Papua New Guinea
SS1371	104	137	4	2	2	2	46	3	4	Papua New Guinea
SS1482	105	151	2	28	2	17	4	5	1	Malaysia
SS1416	106	140	51	2	14	26	19	35	31	Malaysia
SS1551	107	173	57	3	40	25	1	3	10	Malaysia
SS1456	108	10	2	2	9	13	2	14	2	Malaysia
NS030	108	304	2	6	3	2	2	2	2	Australia
SS1468	109	198	58	48	2	5	53	23	25	Malaysia
NS182	109	305	58	17	2	5	2	3	25	Australia
SS1550	112	194	13	46	14	32	1	13	1	Thailand
SS1470	113	148	43	2	3	6	6	2	2	New Zealand
3018–01	114	188	2	31	8	25	52	2	27	USA
SS1357	114	188	2	31	8	25	52	2	27	USA
2907–97	115	123	43	2	8	7	1	3	1	Brazil
SS1366	115	135	2	6	37	5	2	3	28	USA
SS1363	117	134	54	24	14	4	9	2	2	USA
3360–01	118	167	29	32	2	5	48	5	21	USA
SS1499	119	154	31	2	2	5	18	24	12	USA
SS1535	120	168	2	2	14	3	22	3	2	Egypt
SS1537	121	10	2	2	9	13	2	14	2	Egypt
SS1542	122	200	62	2	2	7	1	3	1	Egypt
6949–99	123	123	43	2	8	7	1	3	1	Argentina
SS1536	124	199	51	49	8	6	33	3	35	Egypt
SS636	13L	157	44	40	2	5	19	5	4	United Kingdom
SS582	27G	156	45	6	31	7	1	36	1	United Kingdom
SS875	44 and 61	31	4	10	3	6	14	10	4	USA
SS951	50 and 62	2	1	2	20	5	35	3	4	USA

Continued on following page

TABLE 1—*Continued*

Original strain name	emm type(s)	ST[a]	Housekeeping allele assignments							Country of origin[b]
			gki	*gtr*	*murI*	*mutS*	*recP*	*xpt*	*yqiL*	
SS822	65 and 69	127	3	39	1	7	1	8	24	Unknown
SS1042	65 and 69	127	3	39	1	7	1	8	24	USA
SS1465	65 and 69	127	3	39	1	7	1	8	24	Unknown
SS1096	65 and 69	129	16	42	34	7	45	2	17	Egypt
6102–99	st11014	277	70	31	2	35	1	3	34	Taiwan
SS1457	st1207	146	40	2	29	5	19	3	1	USA
SS1485	st1389	152	42	6	14	26	43	35	4	USA
SS1683	st1731	303	51	31	2	25	2	23	31	Nepal
SS1479	st1815	150	11	2	1	3	50	8	7	USA
SS1369	st1967	136	40	2	9	5	9	8	12	Ethiopia
SS1404	st1969	104*	36	33	2	5	19	24	1	Ethiopia
SS1373	st2037	301	77	6	8	7	1	3	50	Papua New Guinea
SS1378	st204	139	53	39	1	7	1	3	24	Brazil
SS1379	st211	3	2	2	1	2	2	2	2	Brazil
SS1408	st213	187	59	43	14	5	47	2	1	Brazil
SS1541	st2147	169	51	3	8	7	33	42	1	Egypt
SS1376	st2460	138	3	2	2	5	19	34	4	The Gambia
SS1405	st2461	128	49	2	2	5	44	14	12	The Gambia
SS1377	st2463	166	4	31	2	11	34	3	4	The Gambia
SS1471A	st2904	186	13	2	14	1	9	18	1	Brazil
2911–97	st2911	174	31	2	2	5	35	14	12	Brazil
SS1472	st2917	149	55	44	8	5	4	3	10	Brazil
SS1473	st2926	170	2	2	2	3	18	14	12	Brazil
SS1469	st2940	147	11	2	8	33	1	13	18	Chile
4770–01	st369	208	60	47	35	7	2	39	3	USA
SS1497	st3757	153	4	2	2	13	2	24	4	Brazil
3850–01	st3850	174	31	2	2	5	35	14	12	USA
5282–00	st5282	182	2	2	37	2	2	13	1	USA
6030–00	st6030	294	43	2	2	7	5	51	1	USA
6735–99	st6735	165	51	3	1	25	1	3	27	Brazil
77–00	st7700	207	2	2	8	13	2	14	2	USA
SS1443	st809	143	15	12	38	6	4	23	6	Brazil
SS1444	st833	144	29	32	2	5	48	9	4	Brazil
SS1445	st854	145	46	17	2	5	1	25	1	Brazil
1165–99	st854	292	4	3	54	5	34	8	44	Egypt
SS1681	st980584	300	4	6	4	4	52	2	1	China
MTH81	stMTH81	112	16	35	27	24	35	33	1	Australia
SS1680	stNS1033	205	16	2	8	3	2	3	2	Australia
NS292	stNS292	114	4	36	2	18	2	3	4	Australia
SS1676	stNS90	206	11	6	1	7	2	3	6	Australia
SS1578	sts104	39	5	11	8	5	15	2	1	Taiwan

[a] Asterisks indicated newly analyzed STs previously posted at www.mlst.net by B. Beall.
[b] USA, United States; Trinidad, Trinidad and Tobago.

MLST data into nonoverlapping groups of STs with a user-defined level of similarity in their allelic profiles (15). The most stringent definition of an eBURST group, where all STs assigned to the same group must share alleles at at least six of the seven MLST loci with at least one other ST in the group, identifies clusters of closely related genotypes that are considered to be descended from the same founder and that are defined as clonal complexes (15). To obtain a population snapshot, the group definition is set at zero of seven shared housekeeping alleles. Thirty-one clonal complexes were observed among the 220 STs with eBURST, and most of these were small clusters of two or three linked STs (Fig. 1). eBURST identifies the most likely founder of a clonal complex and provides bootstrap support for the assignment. For the 220 GAS STs, a founder ST was assigned in only 11 of the 31 clonal complexes; 65% of the clonal complexes were doublets where the direction of evolution is unknown. However, the bootstrap support was <70% for all founder STs, except for ST65 (99% confidence). For each of the 31 clonal complexes, all STs had *emm* types belonging to the same *emm* pattern group (Table 2).

The relative contributions of point mutation and recombination to the initial stages of clonal diversification can be assessed from MLST data, by identifying those STs that are very closely related, differing at only one of the seven MLST loci (single-locus variants [SLVs]). The sequences of the alleles at the single altered locus are then analyzed to distinguish whether the change in the housekeeping gene has occurred by recombination or by mutation (14, 16). The criteria for assigning an allelic change as resulting from mutation or recombination used by Feil et al. (16) are based on the identification of the founder ST and its associated SLVs within a clonal complex. Within the GAS data set there are 48 pairs of STs that differ at a single locus, but founders cannot be confidently predicted in the majority of cases. In this study, the assignment of the allelic change in SLV pairs as the result of mutation or

TABLE 2. *emm* types according to *emm* pattern marker for tissue site preference

emm pattern	Tissue site preference	No. of *emm* types represented	*emm* type(s)[a]
A–C	Throat (pharyngitis)	29	1, 3, 5, 6, 12, 14, 17, 18, 19, 23, 24, 26, 29, 30, 37, 38/40, 39, 46, 47, 51, 54, 55, 57, st1RP31, st3765, st854, st980584 (stHK), stCK401, stNS90
D	Skin (impetigo)	63	32, 33, 34, 36, 41, 42, 43, 52, 53, 54, 56, 59, 64, 65/69, 67, 70, 71, 72, 74, 80, 81, 83, 85, 86, 91, 93, 95, 97, 98, 99, 100, 101, 105, 108, 111, 115, 116, 119, 120, 121, 122, 123, st1967, st2037, st204, st2461, st2911, st2917, st2926, st2940, st369, st3757, st3850, st5282, st6030, st7700, st809, st854, stCK249, stD432, stD631, stD633, stNS1033
E	No obvious preference	64	2, 4, 8, 9, 11, 15, 22, 25, 28, 44/61, 48, 49, 50/62, 58, 60, 63, 66, 68, 73, 75, 76, 77, 78, 79, 82, 84, 87, 88, 89, 90, 92, 94, 96, 102, 103, 104, 106, 107, 109, 110, 112, 113, 114, 117, 118, 124, 13L, 27G, st11041, st1207, st1389, st1731, st213, st2147, st2460, st2463, st2904, st6735, st833, stDYD, stMTH81, stNS292, stNS554, sts104
Rearranged		2	st1815, st211
Uncertain[b]		1	31
Not done[c]		1	st1969

[a] 54 and st854 (underlined) are associated with A–C and D, *sof*-positive pattern D strains were observed for *emm* types 59, 65/69, 81, and 85.
[b] Difficulty distinguishing between patterns D and E, which may reflect rearranged *emm* region.
[c] st1969 underwent MLST, but not *emm* pattern determination.

recombination was therefore based on the following assumptions. If there are multiple (more than one) nucleotide differences among the alleles at the locus that differs among SLVs, a recombinational event is assumed to have occurred, because the probability of multiple independent point mutations at one locus with none at any of the other six loci is low. If there is only a single nucleotide difference, assignment of the variant as the result of mutation or recombination is more complicated (16). A random point mutation is expected to produce a novel allele restricted to the SLV in which it arises; however, alleles introduced by recombination should be present in other strains in the population and most may be present in a large MLST database. In this study, where both alleles at the variant locus of the SLV pair were found in distantly related STs, we assume that the allelic change was due to recombination. The remaining alleles that differ at a single site will probably still include some that arose by recombination from a donor allele that is absent from the data set, and thus this procedure provides a minimum estimate of the extent of recombination compared to point mutation.

Among the 48 SLV pairs identified by eBURST among the 220 STs (Fig. 1), 20 allelic changes were designated recombination events, based on multiple nucleotide differences among alleles at the variant locus (data not shown). The remaining 28 SLV pairs had a single nucleotide difference among the alleles. Of these, in eight cases both alleles of the SLV pair were present in one or more distantly related STs. Thus, 28 of the allelic changes were considered to be due to recombination and ≤20 were considered to be due to point mutation, and housekeeping loci in GAS are estimated to change by recombination at least 1.4 times more frequently than by point mutation.

Of the 28 allelic changes classified as recombination events, all involved *emm* pattern D or E strains (16 and 12 genetic events, respectively). Further studies are required to obtain a more precise estimate of the ratio of recombination to mutation and to firmly establish whether recombination is a more common mode of evolutionary change at housekeeping loci in *emm* pattern D and E strains than in pattern A-C strains.

Association of multiple *emm* types with a single ST. The great majority of STs were found in association with a single *emm* type (208 of 220; 95%). Only 12 STs included isolates of two or more different *emm* types (Table 3); these are referred to as *emm*-variable STs. However, the 12 *emm*-variable STs involved a disproportionately large fraction of the total number of *emm* types (30 of 158, 19%). Three *emm*-variable STs were associated with *emm* pattern A-C strains, eight were associated with pattern D strains, and only one was associated with pattern E strains. None of the STs were associated with *emm* types corresponding to different *emm* pattern groups.

Eight (28%) of the 29 *emm* types associated with *emm* pattern A-C strains (Table 2) were found among *emm*-variable STs (ST65, -83, and -84; Table 3). Similarly, for *emm* pattern D strains, 20 of the 63 *emm* types (32%) were associated with *emm*-variable STs (ST3, -4, -9, -10, -11, -123, -174, and -182). In sharp contrast, only 2 of the 64 pattern E *emm* types (3%) were associated with an *emm*-variable ST (ST39). Thus, unlike the *emm* types characteristic of *emm* pattern A-C and D strains, the STs of *emm* pattern E isolates rarely include isolates with more than one *emm* type ($P < 0.002$; Fisher's exact test, two-tailed).

The extent of similarity between the *emm* sequences of those isolates that have the same ST but different *emm* types was examined, as this may distinguish variation in *emm* type that has arisen by the accumulation of point mutations from that arising by horizontal gene transfer. For many of these *emm*-variable STs, the different *emm* types have <50% nucleotide sequence identity, and, for all *emm* types associated with the same ST, the *emm* type sequences were ≤91% identical in nucleotide sequence and ≤84% identical in the corresponding amino acid sequence of the M protein (Table 3). However, close examination of sequence alignments suggests that *emm* type st1RP31 arose from *emm* type 30 via intragenic recombination resulting in small deletions; both strains are ST65. Furthermore, *emm* type sts104 appears to have arisen via fusion of

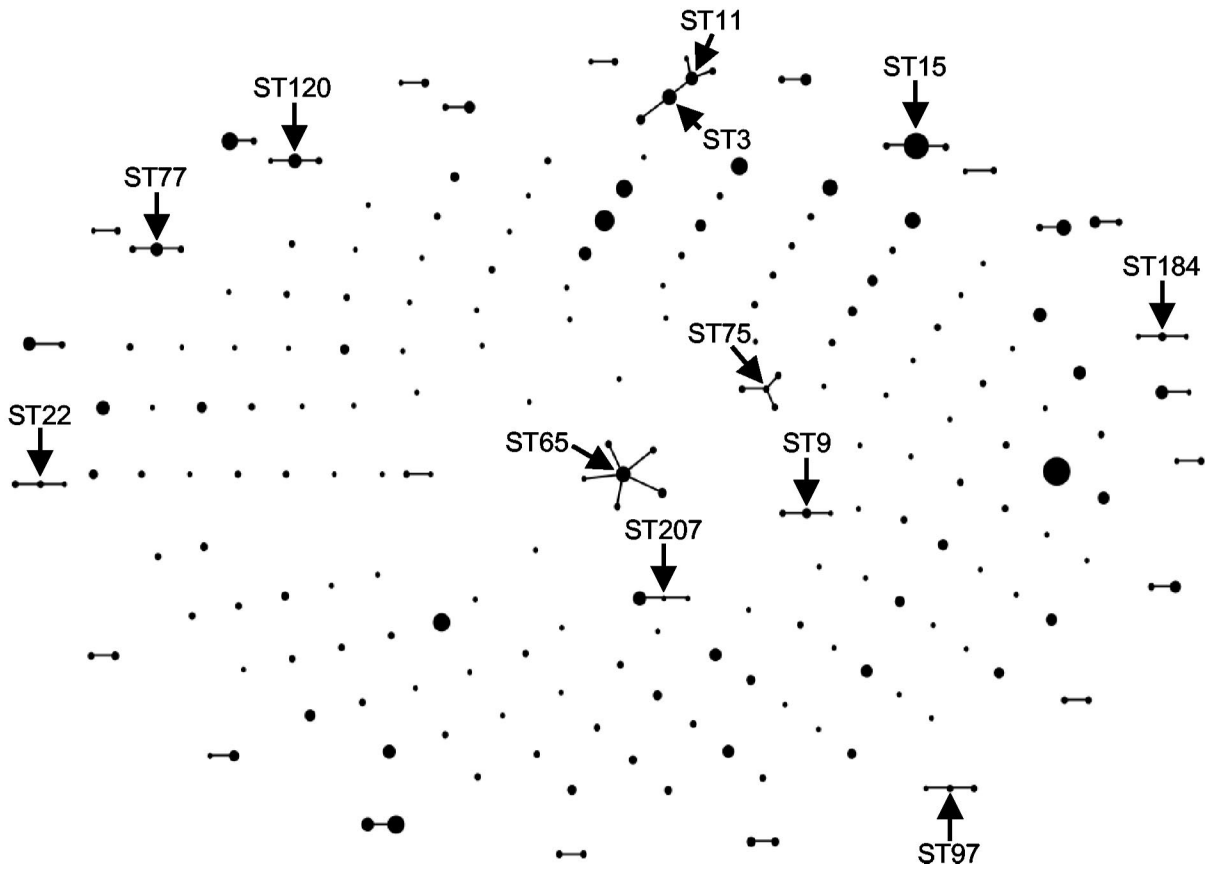

FIG. 1. Population snapshot by eBURST. The entire *S. pyogenes* database of 495 isolates is displayed as a single eBURST diagram, by setting the group definition to zero of seven shared alleles, which places all isolates in a single group. Each dot represents an ST, and the size of the dot reflects the number of GAS isolates in each ST for the set of 495 isolates under study. STs that differ by a single locus are linked with a solid line; clusters of linked isolates correspond to clonal complexes. Founder STs are labeled (arrows), although, except for ST65, the bootstrap support for the founders was low. The distribution and spacing of unlinked STs and clonal complexes in a population snapshot are not relational and provide no information about the genetic distances between them.

the leader-coding region of *emm4* with a downstream *emm* gene (*enn4*), on an ST39 genetic background, although the *emm* type sts104 strain was successfully mapped as *emm* pattern E. Aside from the two exceptions noted, the large number of sequence differences between *emm* types strongly suggests that horizontal transfer of *emm* followed by intergenomic recombination is the primary mechanism underlying *emm*-variable STs, rather than intragenomic recombination or divergence by point mutation.

Analysis of other adaptive loci in *emm*-variable STs. If recombinational replacement of *emm* type is a recent event, then other loci distant from *emm* on the genome should display little or no sequence variation among isolates that have the same ST but which differ in *emm* type. The FCT (fibronectin-collagen-T antigen) region of the GAS genome encodes surface proteins that bind host extracellular matrix proteins (fibronectin and collagen). The FCT region displays high overall levels of genetic diversity and lies ~300 kb from the *emm* region (5, 17, 27, 32). Two FCT region genes, *prtF1* and *cpa*, were examined for sequence diversity in GAS isolates sharing the same ST but differing in *emm* type (Table 4).

For 10 isolates possessing the *prtF1* locus (*emm* patterns A-C and E) and representing four STs and 10 *emm* types, four major sequence clusters of partial *prtF1* genes, corresponding to the 5′ end region, were identified (Table 4). The percent nucleotide sequence identity among alleles belonging to different *prtF1* sequence clusters ranged from 62 to 69%, whereas the amino acid sequence identity among different clusters ranged from 46 to 58%, suggestive of a history of strong diversifying selection at the *prtF1* locus. However, in each case, all isolates with the same ST but different *emm* types also had identical *prtF1* alleles. The lack of variation among *prtF1* alleles within *emm*-variable STs adds further support to the idea that recombinational replacements that lead to variation in the *emm* type of isolates of a single ST are recent evolutionary events in *emm* pattern A-C strains.

Many *emm* pattern D strains harbor a *cpa* gene, rather than the *prtF1* gene, within their FCT regions. Of 20 pattern D isolates, belonging to eight STs and representing 20 different *emm* types, the nucleotide sequence was determined for an internal portion (5′ end region) of the *cpa* gene for 18 strains (Table 4). The partial *cpa* genes formed three discrete sequence clusters, with two alleles in each major cluster. The percent nucleotide sequence identity among *cpa* alleles belonging to the same sequence cluster was high (>99%). However, the percent nucleotide sequence identity among alleles belonging to different *cpa*

TABLE 3. STs associated with more than one *emm* type

ST	*emm* types	*emm* pattern	Pairwise comparisons of *emm* types sharing the same ST (Clustal W)[a]			Pairwise comparisons of *emm* types that are SLVs (Clustal W)[a]	
			emm types compared	Max[c] nucleotide identity between any pair (%)	Max aa[d] identity between any pair (%)	SLV(s) (*emm* types compared)	Max aa identity (%)
65	19, 29, 30, st1RP31	A–C	19 and 29	75	70	ST66 and ST70 (*emm*st1RP31 and *emm*24)	58
			30 and st1RP31	86		ST68 (both *emm*st1RP31)	>95
			29 and 30	55		ST69 (both *emm*19)	>95
			29 and st1RP31	53			
			All other pairs	<50			
83	17, 47	A–C	17 and 47	<50	<50	ST196 (both *emm*17)	>95
84	14, 51	A–C	14 and 51	58	<50	ST85 (both *emm*14)	>95
3	33, st211	D	33 and st211	<50	<50	ST4 (*emm*43, *emm*52, and others)	<50
						ST11 (*emm*53, *emm*101, and others)	<50
4	43, 52	D	43 and 52	69	<50	ST3 (*emm*33, *emm*st211, and others)	<50
9	86, 97	D	86 and 97	<50	<50	ST8 (*emm*80 and others)	<50
						ST131 (*emm*72 and others)	<50
10	70, 80, 93, 98, 108, 121	D	70 and 108	91	84	ST207 (*emm*st7700 and others)	<50
			80 and 93	62			
			80 and 98	77			
			80 and 121	78			
			93 and 98	66			
			93 and 98	66			
			98 and 121	82			
			All other pairs	<50			
11	53, 101	D	53 and 101	73	54	ST3 (*emm*33, *emm*st211, and others)	<50
						ST12 (*emm*91 and *emm*101)	65
						ST304 (*emm*108 and others)	<50
123	115, 123	D	115 and 123	<50	<50	ST120 (*emm*74 and others)	<50
174	st2911, st3850	D	st2911 and st3850	<50	<50	ST97 (*emm*119 and *emm*st3850)	59
182	101, st5282	D	101 and st5282	<50	<50	None in set	Not applicable
39	4, st104[b]	E	4 and st104	<50	<50	ST38 (both *emm*4)	>95

[a] Alignments of 240 bp, whereby the (putative) signal peptidase cleavage site lies between amino acid residues 21 and 22 (nucleotide positions 63 and 64) for all alignments, except for ST65 (between amino acid residues 19 and 20) and ST83 and ST84 (between amino acid residues 22 and 23).

[b] st104 appears to be a fusion between an *emm*4-like leader-coding region and the downstream *enn4* (SF3 *emm*) gene, although this strain was scored as *emm* pattern E (i.e., evidence for three tandem *emm* loci). Also of potential relevance, the *emm* st104 isolate has a *sof4* allele, which is often characteristic of an *emm* type 4 genetic background.

[c] Max, maximal.

[d] aa, amino acid.

sequence clusters was much lower, ranging from 62 to 68%; the amino acid sequence identity among different clusters ranged from 49 to 60%. The sequence data suggest that, like *prtF1*, the *cpa* locus has a history of being subject to strong diversifying selection.

In contrast to what was found for *prtF1*, strains having distinct *emm* types but the same ST were not necessarily uniform in their *cpa* genes (Table 4). Although four of the seven *emm*-variable STs examined had identical *cpa* alleles in strains with different *emm* types (ST3, -123, -174, and -182), two *emm*-variable STs had *cpa* genes belonging to distant sequence clusters (ST9 and -11); in a third (ST4), the *cpa* fragment could not be amplified from one of the two strains with the *cpa* primers. These findings suggest that, for the *emm* pattern D subpopulation, the emergence of strains of the same ST, but with different *emm* types, may in some cases be more complex than a one-step recombinational replacement of *emm*.

DISCUSSION

The findings presented in this report are part of a long-term effort to gain a comprehensive understanding of the genetic diversity present within this medically important bacterial species. *emm* types are commonly used to characterize GAS, and at least one representative strain of the large majority of currently known *emm* types was examined for both ST (based on housekeeping genes) and a genetic marker for tissue site preference (based on *emm* pattern). The lower level of genetic diversity, for both *emm* type and ST, that was observed among the throat strain group (*emm* pattern A-C) may reflect its lower prevalence among the world's human host population: The majority of human hosts inhabit tropical and semitropical regions of the developing world, in which the incidence of streptococcal skin infection is generally high and the incidence of streptococcal pharyngitis is often moderate to low.

At least 58% (28 of 48) of the recent changes at housekeeping loci in GAS appear to be due to recombination, and this value may be substantially greater, since many alleles among SLVs that differ at a single nucleotide site may have arisen by recombination involving a very similar donor allele rather than by point mutation. The best estimate at present is that recombination changes alleles of housekeeping loci at least 1.4 times more commonly than point mutation. The major contribution

TABLE 4. Genetic diversity at other adapative loci for isolates of differing *emm* types sharing an ST

Strain	*emm* pattern	ST	*emm* type	*prtF1*		*cpa*	
				Cluster	Allele	Cluster	Allele
1RP118	A–C	65	19	A	1	NA	NA
3RP70	A–C	65	29	A	1	NA	NA
SS109	A–C	65	30	A	1	NA	NA
2GL32	A–C	65	st1RP31	A	1	NA	NA
1GL217	A–C	83	17	B	2	NA	NA
SS116	A–C	83	47	B	2	NA	NA
T14/46	A–C	84	14	C	3	NA	NA
A291	A–C	84	51	C	3	NA	NA
29487	D	3	33	NA[b]	NA	A	1
SS1379	Indeterminant	3	st211	NA	NA	A	1
D407	D	4	43	NA	NA	PCR negative	Null
A889	D	4	52	NA	NA	A	1
D964	D	9	86	NA	NA	B	4
D626	D	9	97.1	NA	NA	C	5
CK100	D	10	70	NA	NA	A	2
CK344	D	10	80	NA	NA	C	5
D466	D	10	93	NA	NA	C	5
SS1434	D	10	98	NA	NA	C	5
SS1456	D	10	108	NA	NA	C	5
SS1537	D	10	121	NA	NA	C	5
ALAB49	D	11	53	NA	NA	A	2
D641	D	11	101	NA	NA	ND[a]	NA
2907–97	D	123	115	NA	NA	C	6
6949–99	D	123	123	NA	NA	C	6
2911–97	D	174	st2911	NA	NA	B	4
3850–01	D	174	st3850	NA	NA	B	4
CK416	D	182	101	NA	NA	B	3
5282–00	D	182	st5282	NA	NA	B	3
87–231	E	39	4	D	4	NA	NA
SS1578	E	39	sts104	D	4	NA	NA

[a] ND, could not obtain a clean sequence after multiple attempts.
[b] NA, not applicable.

of recombination to allelic change is consistent with previous findings that demonstrated a complete absence of congruence among the gene tree topologies for the seven MLST loci, for GAS genotypes representing all *emm* pattern groups (14). The lack of congruency between loci suggests that, in the long term, recombination has eliminated all phylogenetic signal from gene trees. This finding is further supported by a lack of strong bootstrap support in a phylogenetic tree based on concatenated housekeeping alleles (25).

The *emm* pattern A-C subpopulation (throat specialists) of *S. pyogenes* may differ from the skin specialists (*emm* pattern D) and generalists (*emm* pattern E) in the relative impact of recombination compared to point mutation in genetic diversification at housekeeping loci. Those recombinational changes at MLST loci that can clearly be discerned appear to have been much more common in *emm* pattern group D and E strains than in pattern A-C strains. This trend was also observed in an analysis of congruence among housekeeping gene tree topologies, where 5, 0, and 1 of the 42 possible pairwise tree comparisons were significantly congruent for the *emm* pattern A-C, D, and E subpopulations, respectively (25). Although allelic changes by recombination were less readily detected among the *emm* pattern A-C strains using eBURST, it is important to emphasize that recombination was observed in all of the *emm* pattern-defined subpopulations according to several analytic methods (25).

The total number of STs within each clonal complex identified by eBURST was rather low and probably reflects our sampling strategy. In general, eBURST may identify few clonal complexes, and few large clonal complexes, in populations where sampling has largely been designed to uncover the genetic diversity within the species (11, 16, 34), as in this work, where a small number of isolates or a single isolate of most *emm* types was examined. Thus, a more optimal sampling of GAS will be required for identifying many additional clonal complexes, for defining their founding genotypes, and for exploring the patterns of descent, in order to provide a better assessment of the impact of recombination and mutation.

The data suggest that a significant proportion of *emm* pattern A-C and D strains, but not pattern E strains, have a recent history of recombinational replacement of *emm* type, yielding STs that are associated with multiple, divergent *emm* types. These events may be relatively recent, as no variation in a gene that is believed to be under diversifying selection (*prtF1*) was detected in isolates of the *emm*-variable STs of *emm* pattern A-C. Pattern D strains generally lack *prtF1* but instead harbor *cpa*, which is located at the same approximate position within the genome (5, 32). Not all pattern D strains sharing the same ST and harboring divergent *emm* types had the same *cpa* allele; distant sequence clusters of *cpa* genes were observed on the same ST background in association with different *emm* types. Thus, in some cases, diversification at the rapidly evolving *cpa*

locus may have occurred subsequent to the recombinational replacement of the *emm* gene. Analysis of additional loci may aid in obtaining a more complete understanding of the recent evolutionary history of these strains.

Recombinational replacement of *emm* type, which may occur during coinfection of a single host tissue site by multiple GAS strains, can potentially provide an avenue for immune escape. The ability of a strain to successfully be transmitted to a new human host diminishes as protective immunity arising from infection gradually builds among the host population (1, 19, 20). For many GAS strains, the type-specific epitopes of the M protein elicit strong protective immunity (2, 9, 23, 24, 28). If the *emm* type of a parent (recipient) strain is replaced with a new *emm* type from an unrelated donor strain, the new genotype may have a strong selective advantage if the host population is largely nonimmune to the *emm* type of the donor strain and immune to the *emm* type of the parent strain. The ability to recover multiple *emm* types in association with a single ST through epidemiologic sampling, as shown in this report, may reflect past strain-to-strain competition mediated through herd immunity. Patients with impetigo often differ from those with pharyngitis in their immune response to specific *S. pyogenes* antigens (26). This may be the result of fundamental differences in the host immune response to infection at these two tissue sites, which in turn, may provide a basis for differential selection pressures on the subpopulations of strains. Examination of the relationships between alleles at neutral (housekeeping) and adaptive (e.g., *emm*, *prtF1*, and *cpa*) loci of GAS may allow one to make reasonable predictions on the strength of host immune selection acting on each adaptive locus.

Of the 48 SLV pairs identified by eBURST, 17 pairs were represented by an ST that was also a recipient for recombinational replacement of *emm* type. In fact, all except 1 of the 12 *emm*-variable STs (ST182) were represented among the clonal complexes identified by eBURST. Among the 17 SLV pairs represented by an *emm*-variable ST, nine (53%) of the genetic diversification events at housekeeping loci were attributed to recombination; however, in most cases, it remained unclear as to whether the *emm*-variable ST was the likely ancestral ST. Frequent acquisition of genes via horizontal transfer could be due to high prevalence of the recipient strain within the human host population, with increased opportunities to be present within mixed infections, or, alternatively, could be due to intrinsic properties that render certain STs highly efficient as recipients of recombinational and/or lateral gene transfer events. It is perhaps of relevance here that some strains of *S. pyogenes* appear to be naturally transformable and that, furthermore, the locus (*sil*) that confers the competence phenotype has a limited distribution among strains (21). Generalized transduction may be an important mechanism for horizontal transfer leading to homologous recombination in *S. pyogenes*.

A comprehensive catalogue of STs and *emm* patterns for the majority of known *emm* types of GAS, as presented in this report, provides a foundation for addressing questions on the population substructure of this biologically diverse bacterial pathogen. *emm* pattern D and E strains account for >80% of *emm* types, and therefore, from a global standpoint, these strains are of medical importance. The STs of *emm* pattern E isolates are rarely associated with one or more *emm* types; divergent *emm* types associated with the same ST were a far more common feature of pattern A-C and D *emm* types; however, the genetic mechanisms underlying the emergence of population structures of the *emm* pattern A-C versus *emm* pattern D subpopulations seem to be distinct. Genetic diversification by recombination appeared to be the dominant mechanism in *emm* pattern D and E strains but was less readily detectable among pattern A-C strains. When genetic diversification is combined with differential effects of host immune selection on each of the *emm* pattern-defined subpopulations, distinct population substructures can emerge.

ACKNOWLEDGMENTS

We thank the many investigators worldwide who provided GAS strains to the CDC and the investigators who provided strains to D.E.B.

This work was supported by the National Institutes of Health (GM60793, to D.E.B. and B.G.S.; AI053826, to D.E.B.), the American Heart Association (grant-in-aid, to D.E.B.), and the Wellcome Trust (to B.G.S.). B.G.S. is a Wellcome Trust Principal Research Fellow.

REFERENCES

1. **Anderson, R. M.** 1998. Analytic theory of epidemics, p. 23–50. *In* R. M. Krause (ed.), Emerging infections. Academic Press, New York, N.Y.
2. **Beachey, E. H., J. M. Seyer, J. B. Dale, W. A. Simpson, and A. H. Kang.** 1981. Type-specific protective immunity evoked by synthetic peptide of *Streptococcus pyogenes* M protein. Nature **292:**457–459.
3. **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.
4. **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.
5. **Bessen, D. E., and A. Kalia.** 2002. Genomic localization of a T-serotype locus to a recombinatorial zone encoding extracellular matrix-binding proteins in *Streptococcus pyogenes*. Infect. Immun. **70:**1159–1167.
6. **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.
7. **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.
8. **Chan, M. S., M. C. Maiden, and B. G. Spratt.** 2001. Database-driven multi locus sequence typing (MLST) of bacterial pathogens. Bioinformatics **17:** 1077–1083.
9. **Dale, J. B., and E. H. Beachey.** 1986. Localization of protective epitopes of the amino terminus of type 5 streptococcal M protein. J. Exp. Med. **163:** 1191–1202.
10. **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.
11. **Enright, M. C., D. A. Robinson, G. Randle, E. J. Feil, H. Grundmann, and B. G. Spratt.** 2002. The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). Proc. Natl. Acad. Sci. USA **99:**7687–7692.
12. **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.
13. **Facklam, R. F., D. R. Martin, M. Lovgren, D. R. Johnson, A. Efstratiou, T. A. Thompson, S. Gowan, P. Kriz, G. J. Tyrrell, E. Kaplan, and B. Beall.** 2002. Extension of the Lancefield classification for group A streptococci by addition of 22 new M protein gene sequence types from clinical isolates: emm103 to emm124. Clin. Infect. Dis. **34:**28–38.
14. **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.
15. **Feil, E. J., B. C. Li, D. M. Aanensen, W. P. Hanage, and B. G. Spratt.** 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J. Bacteriol. **186:**1518–1530.

16. **Feil, E. J., J. M. Smith, M. C. Enright, and B. G. Spratt.** 2000. Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. Genetics **154:**1439–1450.
17. **Ferretti, J. J., W. M. McShan, D. Ajdic, D. J. Savic, G. Savic, K. Lyon, C. Primeaux, S. Sezate, A. N. Suvorov, S. Kenton, H. S. Lai, S. P. Lin, Y. Qian, H. G. Jia, F. Z. Najar, Q. Ren, H. Zhu, L. Song, J. White, X. Yuan, S. W. Clifton, B. A. Roe, and R. McLaughlin.** 2001. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc. Natl. Acad. Sci. USA **98:**4658–4663.
18. **Fiorentino, T. R., B. Beall, P. Mshar, and D. E. Bessen.** 1997. A genetic-based evaluation of principal tissue reservoir for group A streptococci isolated from normally sterile sites. J. Infect. Dis. **176:**177–182.
19. **Gupta, S., and R. Anderson.** 1999. Population structure of pathogens: the role of immune selection. Parasitol. Today **15:**497–501.
20. **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.
21. **Hidalgo-Grass, C., M. Ravins, M. Dan-Goor, J. Jaffe, A. E. Moses, and E. Hanski.** 2002. A locus of group A *Streptococcus* involved in invasive disease and DNA transfer. Mol. Microbiol. **46:**87–99.
22. **Hollingshead, S. K., T. L. Readdy, D. L. Yung, and D. E. Bessen.** 1993. Structural heterogeneity of the *emm* gene cluster in group A streptococci. Mol. Microbiol. **8:**707–717.
23. **Hu, M. C., M. A. Walls, S. D. Stroop, M. A. Reddish, B. Beall, and J. B. Dale.** 2002. Immunogenicity of a 26-valent group A streptococcal vaccine. Infect. Immun. **70:**2171–2177.
24. **Jones, K. F., B. N. Manjula, K. H. Johnston, S. K. Hollingshead, J. R. Scott, and V. A. Fischetti.** 1985. Location of variable and conserved epitopes among the multiple serotypes of streptococcal M protein. J. Exp. Med. **161:**623–628.
25. **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.
26. **Kaplan, E., B. Anthony, S. Chapman, E. Ayoub, and L. Wannamaker.** 1970. The influence of the site of infection on the immune response to group A streptococci. J. Clin. Investig. **49:**1405–1414.
27. **Kreikemeyer, B., K. S. McIver, and A. Podbielski.** 2003. Virulence factor regulation and regulatory networks in *Streptococcus pyogenes* and their impact on pathogen-host interactions. Trends Microbiol. **11:**224–232.
28. **Lancefield, R. C.** 1962. Current knowledge of the type specific M antigens of group A streptococci. J. Immunol. **89:**307–313.
29. **Li, Z. Y., V. Sakota, D. Jackson, A. R. Franklin, B. Beall, and the Active Bacterial Core Surveillance/Emerging Infections Program Network.** 2003. Array of M protein gene subtypes in 1064 recent invasive group A streptococcus isolates recovered from the active bacterial core surveillance. J. Infect. Dis. **188:**1587–1592.
30. **Martin, D. R., and K. S. Sriprakash.** 1996. Epidemiology of group A streptococcal disease in Australia and New Zealand. Rec. Adv. Microbiol. **4:**1–40.
31. **McGregor, K., N. Bilek, A. Bennett, A. Kalia, B. Beall, J. Carapetis, B. Currie, K. Sriprakash, B. Spratt, and D. Bessen.** 2004. Group A streptococci from a remote community have novel multilocus genotypes but share emm-types and housekeeping alleles. J. Infect. Dis. **189:**717–723.
32. **Podbielski, A., M. Woischnik, B. A. B. Leonard, and K.-H. Schmidt.** 1999. Characterization of *nra*, a global negative regulator gene in group A streptococci. Mol. Microbiol. **31:**1051–1064.
33. **Spratt, B. G.** 1999. Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the Internet. Curr. Opin. Microbiol. **2:**312–316.
34. **Spratt, B. G., W. P. Hanage, and E. J. Feil.** 2001. The relative contributions of recombination and point mutation to the diversification of bacterial clones. Curr. Opin. Microbiol. **4:**602–606.