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. Bessen3,5*

Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom1; Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, Missouri2; Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut3; Respiratory Diseases Branch, Centers for Disease Control and Prevention, Atlanta, Georgia4; and Department of Microbiology and Immunology, New York Medical College, Valhalla, New York5

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...
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. Siri-prakash 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/emmssequ/ 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, murF, mutS, recP, ski, and ypgL) 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 *cpx* locus, 5'-GGA TAT GAG ATT GCC GAA CCT ATT ACT TTT AAA G-3' (forward) and 5'-GGG GCC TGT TTA TCT TCC ATT CGA ATA ATC AC-3' (reverse) (product size, ~600 bp); for the *aprF* locus, 5'-TCC GCC GGT TTC ATC GGT TTT GCC CAA GTA-3' (forward) and 5'-TAC TAG TTT TTT CCT TCAAG(A)/G(T/C/T/C TAC GT CAA TA TAA-3' (reverse) (product size, ~360 bp). The same primers were used for nucleotide sequence determination.

**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/emmssequ/ 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 (*s4* and *s854*) 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 (*s1815* and *s2111*) 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

<table>
<thead>
<tr>
<th>Original strain name</th>
<th>emm type(s)</th>
<th>ST&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Housekeeping allele assignments</th>
<th>Country of origin&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>gki, gtr, murI, mutS, recP, xpt, yqiL</td>
<td></td>
</tr>
<tr>
<td>3934–98</td>
<td>1</td>
<td>106*</td>
<td>4, 25</td>
<td>Brazil</td>
</tr>
<tr>
<td>4650–97</td>
<td>1</td>
<td>107*</td>
<td>4, 25</td>
<td>Brazil</td>
</tr>
<tr>
<td>6745–99</td>
<td>4</td>
<td>105*</td>
<td>2, 5</td>
<td>Brazil</td>
</tr>
<tr>
<td>A410</td>
<td>11</td>
<td>195</td>
<td>3</td>
<td>Unknown</td>
</tr>
<tr>
<td>T14/46</td>
<td>14</td>
<td>84</td>
<td>20, 22</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>1269–98</td>
<td>15</td>
<td>121</td>
<td>47</td>
<td>USA</td>
</tr>
<tr>
<td>SS8080</td>
<td>15</td>
<td>163</td>
<td>13</td>
<td>Unknown</td>
</tr>
<tr>
<td>J17E/165</td>
<td>17</td>
<td>196</td>
<td>19, 50</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS580</td>
<td>23</td>
<td>160</td>
<td>10</td>
<td>Brazil</td>
</tr>
<tr>
<td>SS730</td>
<td>23</td>
<td>160</td>
<td>10</td>
<td>Brazil</td>
</tr>
<tr>
<td>SS109</td>
<td>30</td>
<td>65</td>
<td>13</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS901</td>
<td>31</td>
<td>316</td>
<td>13</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS53</td>
<td>37</td>
<td>155</td>
<td>10, 19</td>
<td>USA</td>
</tr>
<tr>
<td>SS795</td>
<td>41</td>
<td>162</td>
<td>3, 2</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS642</td>
<td>46</td>
<td>158</td>
<td>13, 7</td>
<td>USA</td>
</tr>
<tr>
<td>SS116</td>
<td>47</td>
<td>83</td>
<td>19, 21</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS737</td>
<td>48</td>
<td>161</td>
<td>4, 1</td>
<td>USA</td>
</tr>
<tr>
<td>A889</td>
<td>52</td>
<td>4</td>
<td>2, 1</td>
<td>USA</td>
</tr>
<tr>
<td>SS686</td>
<td>52</td>
<td>4</td>
<td>2, 1</td>
<td>USA</td>
</tr>
<tr>
<td>NS342</td>
<td>53</td>
<td>299</td>
<td>2, 1</td>
<td>Brazil</td>
</tr>
<tr>
<td>2470–93</td>
<td>54</td>
<td>4</td>
<td>6</td>
<td>USA</td>
</tr>
<tr>
<td>SS725</td>
<td>54</td>
<td>159</td>
<td>13</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS934</td>
<td>55</td>
<td>100</td>
<td>34</td>
<td>Trinidad</td>
</tr>
<tr>
<td>546–96</td>
<td>56</td>
<td>126</td>
<td>2, 8</td>
<td>Unknown</td>
</tr>
<tr>
<td>4756–01</td>
<td>59</td>
<td>172</td>
<td>56</td>
<td>Unknown</td>
</tr>
<tr>
<td>CT95–201</td>
<td>59</td>
<td>172</td>
<td>56, 39</td>
<td>Unknown</td>
</tr>
<tr>
<td>2954–97</td>
<td>64</td>
<td>124</td>
<td>2, 8</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS1451</td>
<td>64</td>
<td>124</td>
<td>2, 8</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS989</td>
<td>64</td>
<td>164</td>
<td>2, 8</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS1098</td>
<td>71</td>
<td>130</td>
<td>13</td>
<td>Egypt</td>
</tr>
<tr>
<td>SS1144</td>
<td>72</td>
<td>131</td>
<td>2, 36</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS132</td>
<td>77</td>
<td>133</td>
<td>4, 31</td>
<td>Trinidad</td>
</tr>
<tr>
<td>SS1151</td>
<td>79</td>
<td>132</td>
<td>50, 8, 30</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS1401</td>
<td>81</td>
<td>171, 13</td>
<td>23, 7</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS1402</td>
<td>82</td>
<td>122</td>
<td>4, 21</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>SS1400</td>
<td>83</td>
<td>103*</td>
<td>2, 2</td>
<td>Brazil</td>
</tr>
<tr>
<td>SS1447</td>
<td>85</td>
<td>109</td>
<td>42</td>
<td>Unknown</td>
</tr>
<tr>
<td>4539–96</td>
<td>87</td>
<td>102*</td>
<td>13, 34</td>
<td>Malaysia</td>
</tr>
<tr>
<td>292–00</td>
<td>89</td>
<td>101*</td>
<td>16, 2</td>
<td>Italy</td>
</tr>
<tr>
<td>SS1475</td>
<td>94</td>
<td>89</td>
<td>24</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS1343</td>
<td>95</td>
<td>14</td>
<td>2, 6</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS1432</td>
<td>97</td>
<td>197</td>
<td>11, 6</td>
<td>Australia</td>
</tr>
<tr>
<td>SS1434</td>
<td>98</td>
<td>10</td>
<td>2, 9</td>
<td>Australia</td>
</tr>
<tr>
<td>SS1433</td>
<td>99</td>
<td>141</td>
<td>52, 31</td>
<td>Australia</td>
</tr>
<tr>
<td>1523–00</td>
<td>102</td>
<td>60</td>
<td>13</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS1437</td>
<td>102</td>
<td>185</td>
<td>13, 45</td>
<td>Australia</td>
</tr>
<tr>
<td>SS1370</td>
<td>103</td>
<td>327</td>
<td>83</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS1371</td>
<td>104</td>
<td>137</td>
<td>4, 2</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS1482</td>
<td>105</td>
<td>715</td>
<td>25</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS1416</td>
<td>106</td>
<td>140</td>
<td>51, 2</td>
<td>Malaysia</td>
</tr>
<tr>
<td>SS1551</td>
<td>107</td>
<td>173</td>
<td>57</td>
<td>Malaysia</td>
</tr>
<tr>
<td>SS1456</td>
<td>108</td>
<td>10</td>
<td>2, 9</td>
<td>Malaysia</td>
</tr>
<tr>
<td>NS030</td>
<td>108</td>
<td>304</td>
<td>2, 6</td>
<td>Malaysia</td>
</tr>
<tr>
<td>SS1468</td>
<td>109</td>
<td>198</td>
<td>58, 48</td>
<td>Malaysia</td>
</tr>
<tr>
<td>NS182</td>
<td>109</td>
<td>305</td>
<td>58, 17</td>
<td>Malaysia</td>
</tr>
<tr>
<td>SS1550</td>
<td>112</td>
<td>194</td>
<td>13, 46</td>
<td>Thailand</td>
</tr>
<tr>
<td>SS1470</td>
<td>113</td>
<td>148</td>
<td>3, 2</td>
<td>New Zealand</td>
</tr>
<tr>
<td>3018–01</td>
<td>114</td>
<td>188</td>
<td>2, 31</td>
<td>USA</td>
</tr>
<tr>
<td>SS1537</td>
<td>114</td>
<td>188</td>
<td>2, 31</td>
<td>USA</td>
</tr>
<tr>
<td>2907–97</td>
<td>115</td>
<td>123</td>
<td>43</td>
<td>Brazil</td>
</tr>
<tr>
<td>SS1366</td>
<td>115</td>
<td>135</td>
<td>2, 6</td>
<td>Brazil</td>
</tr>
<tr>
<td>SS1363</td>
<td>117</td>
<td>134</td>
<td>54</td>
<td>Brazil</td>
</tr>
<tr>
<td>3360–01</td>
<td>118</td>
<td>167</td>
<td>29</td>
<td>Brazil</td>
</tr>
<tr>
<td>SS1499</td>
<td>119</td>
<td>154</td>
<td>31</td>
<td>Brazil</td>
</tr>
<tr>
<td>SS1535</td>
<td>120</td>
<td>168</td>
<td>2, 14</td>
<td>Egypt</td>
</tr>
<tr>
<td>SS1537</td>
<td>121</td>
<td>10</td>
<td>2, 9</td>
<td>Egypt</td>
</tr>
<tr>
<td>SS1542</td>
<td>122</td>
<td>200</td>
<td>62</td>
<td>Egypt</td>
</tr>
<tr>
<td>6949–09</td>
<td>123</td>
<td>123</td>
<td>43</td>
<td>Egypt</td>
</tr>
<tr>
<td>SS1536</td>
<td>124</td>
<td>199</td>
<td>51, 49</td>
<td>Egypt</td>
</tr>
<tr>
<td>SS636</td>
<td>13L</td>
<td>157</td>
<td>44</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>SS582</td>
<td>27G</td>
<td>156</td>
<td>45</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>SS875</td>
<td>44 and 61</td>
<td>31</td>
<td>4, 10</td>
<td>USA</td>
</tr>
<tr>
<td>SS951</td>
<td>50 and 62</td>
<td>2</td>
<td>1, 2</td>
<td>USA</td>
</tr>
</tbody>
</table>

Continued on following page
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 recombination was determined by comparing the allele sequences at the single altered locus with the sequence of the reference allele (14, 16).

### Table 1—Continued

<table>
<thead>
<tr>
<th>Original strain name</th>
<th>emm type(s)</th>
<th>ST&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Housekeeping allele assignments</th>
<th>Country of origin&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS822</td>
<td>65 and 69</td>
<td>127</td>
<td>gki 1 gtr 1 murel 1 mutS 1 recP 8 spt 24</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS1042</td>
<td>65 and 69</td>
<td>127</td>
<td>gki 1 gtr 1 murel 1 mutS 1 recP 8 spt 24</td>
<td>USA</td>
</tr>
<tr>
<td>SS1465</td>
<td>65 and 69</td>
<td>127</td>
<td>gki 1 gtr 1 murel 1 mutS 1 recP 8 spt 24</td>
<td>Unknown</td>
</tr>
<tr>
<td>SS1096</td>
<td>65 and 69</td>
<td>129</td>
<td>gki 14 gtr 17 mutS 2 recP 1 spt 8</td>
<td>Egypt</td>
</tr>
<tr>
<td>6102–99</td>
<td>st11014</td>
<td>277</td>
<td>gki 70 gtr 31 mutS 2 recP 34</td>
<td>Taiwan</td>
</tr>
<tr>
<td>SS1457</td>
<td>st1207</td>
<td>146</td>
<td>gki 40 gtr 29 mutS 5 recP 1</td>
<td>USA</td>
</tr>
<tr>
<td>SS1445</td>
<td>st1389</td>
<td>152</td>
<td>gki 42 gtr 14 mutS 26 recP 43</td>
<td>USA</td>
</tr>
<tr>
<td>SS1683</td>
<td>st1731</td>
<td>303</td>
<td>gki 51 gtr 25 mutS 2 recP 23</td>
<td>Nepal</td>
</tr>
<tr>
<td>SS1479</td>
<td>st1815</td>
<td>150</td>
<td>gki 11 gtr 1 mutS 35 recP 8</td>
<td>USA</td>
</tr>
<tr>
<td>SS1369</td>
<td>st1967</td>
<td>136</td>
<td>gki 40 gtr 9 mutS 5 recP 8</td>
<td>12 Ethiopia</td>
</tr>
<tr>
<td>SS1404</td>
<td>st1969</td>
<td>104*</td>
<td>gki 36 gtr 3 mutS 5 recP 19</td>
<td>1 Ethiopia</td>
</tr>
<tr>
<td>SS1373</td>
<td>st2037</td>
<td>301</td>
<td>gki 77 gtr 8 mutS 7</td>
<td>1 Papua New Guinea</td>
</tr>
<tr>
<td>SS1378</td>
<td>st204</td>
<td>139</td>
<td>gki 53 gtr 1 mutS 7</td>
<td>3 Brazil</td>
</tr>
<tr>
<td>SS1379</td>
<td>st211</td>
<td>3</td>
<td>gki 2 gtr 1 mutS 2</td>
<td>2 Brazil</td>
</tr>
<tr>
<td>SS1408</td>
<td>st213</td>
<td>187</td>
<td>gki 59 gtr 14 mutS 5</td>
<td>2 Brazil</td>
</tr>
<tr>
<td>SS1541</td>
<td>st2147</td>
<td>169</td>
<td>gki 51 gtr 8 mutS 7</td>
<td>3 Egypt</td>
</tr>
<tr>
<td>SS1376</td>
<td>st2460</td>
<td>138</td>
<td>gki 3 gtr 2 mutS 5</td>
<td>4 The Gambia</td>
</tr>
<tr>
<td>SS1405</td>
<td>st2461</td>
<td>128</td>
<td>gki 49 gtr 2 mutS 5</td>
<td>14 The Gambia</td>
</tr>
<tr>
<td>SS1377</td>
<td>st2463</td>
<td>166</td>
<td>gki 4 gtr 11 mutS 34</td>
<td>4 The Gambia</td>
</tr>
<tr>
<td>SS1471A</td>
<td>st2904</td>
<td>186</td>
<td>gki 13 gtr 14 mutS 1</td>
<td>18 Brazil</td>
</tr>
<tr>
<td>2911–97</td>
<td>st2911</td>
<td>174</td>
<td>gki 31 gtr 2 mutS 5</td>
<td>14 12 Brazil</td>
</tr>
<tr>
<td>SS1472</td>
<td>st2917</td>
<td>149</td>
<td>gki 55 gtr 8 mutS 5</td>
<td>3 10 Brazil</td>
</tr>
<tr>
<td>SS1473</td>
<td>st2926</td>
<td>170</td>
<td>gki 2 gtr 3 mutS 18</td>
<td>14 12 Brazil</td>
</tr>
<tr>
<td>SS1469</td>
<td>st2940</td>
<td>147</td>
<td>gki 11 gtr 8 mutS 33</td>
<td>13 18 Brazil</td>
</tr>
<tr>
<td>4770–01</td>
<td>st369</td>
<td>208</td>
<td>gki 60 gtr 35 mutS 7</td>
<td>39 USA</td>
</tr>
<tr>
<td>SS1497</td>
<td>st3757</td>
<td>153</td>
<td>gki 4 gtr 13 mutS 2</td>
<td>24 4 Brazil</td>
</tr>
<tr>
<td>3850–01</td>
<td>st3850</td>
<td>174</td>
<td>gki 31 gtr 2 mutS 5</td>
<td>14 12 Brazil</td>
</tr>
<tr>
<td>5282–00</td>
<td>st5282</td>
<td>182</td>
<td>gki 2 gtr 37 mutS 2</td>
<td>13 1 USA</td>
</tr>
<tr>
<td>6030–00</td>
<td>st6030</td>
<td>294</td>
<td>gki 43 gtr 2 mutS 7</td>
<td>5 1 USA</td>
</tr>
<tr>
<td>6735–99</td>
<td>st6735</td>
<td>165</td>
<td>gki 51 gtr 3 mutS 25</td>
<td>13 27 Brazil</td>
</tr>
<tr>
<td>77–00</td>
<td>st7700</td>
<td>207</td>
<td>gki 2 gtr 13 mutS 2</td>
<td>14 2 USA</td>
</tr>
<tr>
<td>SS1443</td>
<td>st809</td>
<td>143</td>
<td>gki 15 gtr 38 mutS 6</td>
<td>4 23 Brazil</td>
</tr>
<tr>
<td>SS1444</td>
<td>st833</td>
<td>144</td>
<td>gki 29 gtr 2 mutS 5</td>
<td>48 9 4 Brazil</td>
</tr>
<tr>
<td>SS1445</td>
<td>st854</td>
<td>145</td>
<td>gki 46 gtr 17 mutS 5</td>
<td>1 25 1 Brazil</td>
</tr>
<tr>
<td>1165–99</td>
<td>st854</td>
<td>292</td>
<td>gki 4 54 mutS 34 8</td>
<td>44 Egypt</td>
</tr>
<tr>
<td>SS1681</td>
<td>st908584</td>
<td>300</td>
<td>gki 4 6 mutS 4</td>
<td>52 2 1 China</td>
</tr>
<tr>
<td>MTH81</td>
<td>stMTH81</td>
<td>112</td>
<td>gki 16 gtr 35 mutS 24</td>
<td>33 1 Australia</td>
</tr>
<tr>
<td>SS1680</td>
<td>stNS1033</td>
<td>205</td>
<td>gki 16 gtr 2 mutS 3</td>
<td>2 3 2 Australia</td>
</tr>
<tr>
<td>NS292</td>
<td>stNS292</td>
<td>114</td>
<td>gki 4 36 mutS 18</td>
<td>2 3 2 Australia</td>
</tr>
<tr>
<td>SS1676</td>
<td>stNS90</td>
<td>206</td>
<td>gki 11 gtr 1 mutS 7</td>
<td>3 6 Australia</td>
</tr>
<tr>
<td>SS1578</td>
<td>st104</td>
<td>39</td>
<td>gki 5 11 mutS 15</td>
<td>2 1 Taiwan</td>
</tr>
</tbody>
</table>

<sup>a</sup> Asterisks indicated newly analyzed STs previously posted at www.mlst.net by B. Beall.

<sup>b</sup> USA, United States; Trinidad, Trinidad and Tobago.
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 distinctly 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 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 st104 appears to have arisen via fusion of

### Table 2. emm types according to emm pattern marker for tissue site preference

<table>
<thead>
<tr>
<th>emm pattern</th>
<th>Tissue site preference</th>
<th>No. of emm types represented</th>
<th>emm type(s)a</th>
</tr>
</thead>
<tbody>
<tr>
<td>A–C</td>
<td>Throat (pharyngitis)</td>
<td>29</td>
<td>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</td>
</tr>
<tr>
<td>Rearranged</td>
<td></td>
<td>2</td>
<td>st1815, st211</td>
</tr>
<tr>
<td>Uncertainb</td>
<td></td>
<td>1</td>
<td>31</td>
</tr>
<tr>
<td>Not donec</td>
<td></td>
<td>1</td>
<td>st1969</td>
</tr>
</tbody>
</table>

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

\textbf{Analysis of other adaptive loci in \textit{emm}-variable STs.} If recombinational replacement of \textit{emm} type is a recent event, then other loci distant from \textit{emm} on the genome should display little or no sequence variation among isolates that have the same ST but which differ in \textit{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 \textsim 300 \text{kb} from the \textit{emm} region (5, 17, 27, 32). Two FCT region genes, \textit{prtF1} and \textit{cpa}, were examined for sequence diversity in GAS isolates sharing the same ST but differing in \textit{emm} type (Table 4).

For 10 isolates possessing the \textit{prtF1} locus (\textit{emm} patterns A-C and E) and representing four STs and 10 \textit{emm} types, four major sequence clusters of partial \textit{prtF1} genes, corresponding to the 5' end region, were identified (Table 4). The percent nucleotide sequence identity among alleles belonging to different \textit{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 \textit{prtF1} locus. However, in each case, all isolates with the same ST but different \textit{emm} types also had identical \textit{prtF1} alleles. The lack of variation among \textit{prtF1} alleles within \textit{emm}-variable STs adds further support to the idea that recombinational replacements that lead to variation in the \textit{emm} type of isolates of a single ST are recent evolutionary events in \textit{emm} pattern A-C strains.

Many \textit{emm} pattern D strains harbor a \textit{cpa} gene, rather than the \textit{prtF1} gene, within their FCT regions. Of 20 pattern D isolates, belonging to eight STs and representing 20 different \textit{emm} types, the nucleotide sequence was determined for an internal portion (5' end region) of the \textit{cpa} gene for 18 strains (Table 4). The partial \textit{cpa} genes formed three discrete sequence clusters, with two alleles in each major cluster. The percent nucleotide sequence identity among \textit{cpa} alleles belonging to the same sequence cluster was high (>99%). However, the percent nucleotide sequence identity among alleles belonging to different \textit{cpa}
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 **pattern**) E (i.e., evidence for three tandem emm loci). Also of potential relevance, the emm pattern D subpopulation has a more complex genetic background:

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...
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

<table>
<thead>
<tr>
<th>Strain</th>
<th>emm pattern</th>
<th>ST</th>
<th>emm type</th>
<th>prtF1 Cluster</th>
<th>prtF1 Allele</th>
<th>cpa Cluster</th>
<th>cpa Allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>1RP118</td>
<td>A-C</td>
<td>65</td>
<td>19</td>
<td>A</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>3RP70</td>
<td>A-C</td>
<td>65</td>
<td>29</td>
<td>A</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>SS109</td>
<td>A-C</td>
<td>65</td>
<td>30</td>
<td>A</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>2GL32</td>
<td>A-C</td>
<td>65</td>
<td>st1RP31</td>
<td>B</td>
<td>2</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1GL217</td>
<td>A-C</td>
<td>83</td>
<td>17</td>
<td>B</td>
<td>2</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>SS116</td>
<td>A-C</td>
<td>83</td>
<td>47</td>
<td>C</td>
<td>3</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>T14/46</td>
<td>A-C</td>
<td>84</td>
<td>14</td>
<td>C</td>
<td>3</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>A291</td>
<td>A-C</td>
<td>84</td>
<td>51</td>
<td>C</td>
<td>3</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>29487</td>
<td>D</td>
<td>3</td>
<td>33</td>
<td>NA</td>
<td>NA</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>SS1379</td>
<td>Indeterminant</td>
<td>3</td>
<td>st211</td>
<td>NA</td>
<td>NA</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>D407</td>
<td>D</td>
<td>4</td>
<td>43</td>
<td>NA</td>
<td>NA</td>
<td>PCR negative</td>
<td>Null</td>
</tr>
<tr>
<td>A889</td>
<td>D</td>
<td>4</td>
<td>52</td>
<td>NA</td>
<td>NA</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>D964</td>
<td>D</td>
<td>9</td>
<td>86</td>
<td>NA</td>
<td>NA</td>
<td>B</td>
<td>4</td>
</tr>
<tr>
<td>D626</td>
<td>D</td>
<td>9</td>
<td>97.1</td>
<td>NA</td>
<td>NA</td>
<td>C</td>
<td>5</td>
</tr>
<tr>
<td>CK100</td>
<td>D</td>
<td>10</td>
<td>70</td>
<td>NA</td>
<td>NA</td>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>CK344</td>
<td>D</td>
<td>10</td>
<td>80</td>
<td>NA</td>
<td>NA</td>
<td>C</td>
<td>5</td>
</tr>
<tr>
<td>D466</td>
<td>D</td>
<td>10</td>
<td>93</td>
<td>NA</td>
<td>NA</td>
<td>C</td>
<td>5</td>
</tr>
<tr>
<td>SS1434</td>
<td>D</td>
<td>10</td>
<td>98</td>
<td>NA</td>
<td>NA</td>
<td>C</td>
<td>5</td>
</tr>
<tr>
<td>SS1456</td>
<td>D</td>
<td>10</td>
<td>108</td>
<td>NA</td>
<td>NA</td>
<td>C</td>
<td>5</td>
</tr>
<tr>
<td>SS1537</td>
<td>D</td>
<td>10</td>
<td>121</td>
<td>NA</td>
<td>NA</td>
<td>C</td>
<td>5</td>
</tr>
<tr>
<td>ALAB49</td>
<td>D</td>
<td>11</td>
<td>53</td>
<td>NA</td>
<td>NA</td>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>D641</td>
<td>D</td>
<td>11</td>
<td>101</td>
<td>NA</td>
<td>NA</td>
<td>ND</td>
<td>NA</td>
</tr>
<tr>
<td>2907–97</td>
<td>D</td>
<td>123</td>
<td>115</td>
<td>NA</td>
<td>NA</td>
<td>C</td>
<td>6</td>
</tr>
<tr>
<td>6949–99</td>
<td>D</td>
<td>123</td>
<td>123</td>
<td>NA</td>
<td>NA</td>
<td>C</td>
<td>6</td>
</tr>
<tr>
<td>2911–97</td>
<td>D</td>
<td>174</td>
<td>st2911</td>
<td>NA</td>
<td>NA</td>
<td>B</td>
<td>4</td>
</tr>
<tr>
<td>3850–01</td>
<td>D</td>
<td>174</td>
<td>st3850</td>
<td>NA</td>
<td>NA</td>
<td>B</td>
<td>4</td>
</tr>
<tr>
<td>CK416</td>
<td>D</td>
<td>182</td>
<td>101</td>
<td>NA</td>
<td>NA</td>
<td>B</td>
<td>3</td>
</tr>
<tr>
<td>5282–00</td>
<td>D</td>
<td>182</td>
<td>st5282</td>
<td>NA</td>
<td>NA</td>
<td>B</td>
<td>3</td>
</tr>
<tr>
<td>87–231</td>
<td>E</td>
<td>39</td>
<td>4</td>
<td>D</td>
<td>4</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>SS1578</td>
<td>E</td>
<td>39</td>
<td>st104</td>
<td>D</td>
<td>4</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

*a ND, could not obtain a clean sequence after multiple attempts.
*b NA, not applicable.
locus may have occurred subsequent to the recombinational replacement of the \textit{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 \textit{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 \textit{emm} type of a parent (recipient) strain is replaced with a new \textit{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 \textit{emm} type of the donor strain and immune to the \textit{emm} type of the parent strain. The ability to recover multiple \textit{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 \textit{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., \textit{emm}, \textit{ptf1}, and \textit{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 \textit{emm} type. In fact, all except 1 of the 12 \textit{emm}-variable STs (ST182) were represented among the clonal complexes identified by eBURST. Among the 17 SLV pairs represented by an \textit{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 \textit{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 \textit{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 \textit{S. pyogenes}.

A comprehensive catalogue of STs and \textit{emm} patterns for the majority of known \textit{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. \textit{emm} pattern D and E strains account for >80% of \textit{emm} types, and therefore, from a global standpoint, these strains are of medical importance. The STs of \textit{emm} pattern E isolates are rarely associated with one or more \textit{emm} types; divergent \textit{emm} types associated with the same ST were a far more common feature of pattern A-C and D \textit{emm} types; however, the genetic mechanisms underlying the emergence of population structures of the \textit{emm} pattern A-C versus \textit{emm} pattern D subpopulations seem to be distinct. Genetic diversification by recombination appeared to be the dominant mechanism in \textit{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 \textit{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


