OPEN ACCESS Freely available online

PLoS one

Recombination Drives Genetic Diversification of *Streptococcus dysgalactiae* Subspecies *equisimilis* in a Region of Streptococcal Endemicity

David J. McMillan[1,2]*, Santosh Y. Kaul[3], P. V. Bramhachari[4], Pierre R. Smeesters[5], Therese Vu[2], M. G. Karmarkar[3], Melkote S. Shaila[4], Kadaba S. Sriprakash[1]

1 Bacterial Pathogenesis Laboratory, Queensland Institute of Medical Research, Brisbane, Queensland, Australia, **2** Griffith Medical Research College, a joint collaboration between Griffith University and Queensland Institute of Medical Research, Brisbane, Queensland, Australia, **3** Department of Microbiology, KEM Hospital, Mumbai, India, **4** Department of Microbiology and Cell Biology, Indian Institute of Science, Bangalore, India, **5** Laboratoire de Genetique et Physiologie Bacterienne, Institut de Biologie et de Medecine Moleculaires, Faculte des Sciences, Universite Libre de Bruxelles, Brussels, Belgium

Abstract

Infection of the skin or throat by *Streptococcus dysgalactiae* subspecies *equisimilis* (SDSE) may result in a number of human diseases. To understand mechanisms that give rise to new genetic variants in this species, we used multi-locus sequence typing (MLST) to characterise relationships in the SDSE population from India, a country where streptococcal disease is endemic. The study revealed Indian SDSE isolates have sequence types (STs) predominantly different to those reported from other regions of the world. Emm-ST combinations in India are also largely unique. Split decomposition analysis, the presence of emm-types in unrelated clonal complexes, and analysis of phylogenetic trees based on concatenated sequences all reveal an extensive history of recombination within the population. The ratio of recombination to mutation (*r/m*) events (11:1) and per site *r/m* ratio (41:1) in this population is twice as high as reported for SDSE from non-endemic regions. Recombination involving the *emm*-gene is also more frequent than recombination involving housekeeping genes, consistent with diversification of M proteins offering selective advantages to the pathogen. Our data demonstrate that genetic recombination in endemic regions is more frequent than non-endemic regions, and gives rise to novel local SDSE variants, some of which may have increased fitness or pathogenic potential.

Citation: McMillan DJ, Kaul SY, Bramhachari PV, Smeesters PR, Vu T, et al. (2011) Recombination Drives Genetic Diversification of *Streptococcus dysgalactiae* Subspecies *equisimilis* in a Region of Streptococcal Endemicity. PLoS ONE 6(8): e21346. doi:10.1371/journal.pone.0021346

Editor: Frank R. DeLeo, National Institute of Allergy and Infectious Diseases, National Institutes of Health, United States of America

Received January 26, 2011; **Accepted** May 29, 2011; **Published** August 3, 2011

Funding: This work was supported by an Australian Government National Health and Medical Research Council, National and International Alliance Program of Queensland, Australia, and India-Australia Strategic Research Fund and Department of Biotechnology, India. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: david.mcmillan@qimr.edu.au

Introduction

Streptococcus dysgalactiae subspecies *equisimilis* (SDSE) is a β-haemolytic Gram-positive bacterium that typically colonises the oropharynx and skin of humans. The species is closely related to *S. pyogenes*, a significant human pathogen [1]. Although generally a less common cause of human infection, SDSE causes many of the same diseases as *S. pyogenes*, including pharyngitis, pyoderma, post-glomerulonephritis, bacteraemia and other invasive diseases [2,3]. The incidence of SDSE disease is also reported to be increasing, and in some studies exceeds that of *S. pyogenes* disease [4,5,6].

SDSE and *S. pyogenes* share in common many virulence factors that contribute to virulence, including the M-protein [7,8,9]. The M-protein protects bacteria from opsonophagocytosis by blocking deposition of complement on the bacterial surface. Nucleotide variation in *emm*, the gene encoding the M-protein, is used to type both SDSE and *S. pyogenes* at the subspecies level. Currently more than 50 SDSE *emm*-genes are present in the Centre for Disease Control *emm*-gene database (http://www.cdc.gov/ncidod/biotech/strep/types_*emm*103-124.htm). Other typing methods such as vir-typing [10,11] and *emm*-pattern typing [12] also target the locus encoding the *emm*-gene. However several studies have reported the *emm*-gene, and surrounding loci to be subject to lateral gene transfer (LGT) [13,14].

Recovery of SDSE from the throats has been reported to exceed that of *S. pyogenes* in regions where streptococcal disease is endemic [15,16]. The diversity of circulating SDSE and *S. pyogenes emm*-types in endemic regions is high. Multilocus sequence typing (MLST) is a nucleotide based method for characterising genetic relationships amongst isolates of the same bacterial species. Unlike *emm*-typing, MLST utilises multiple housekeeping genes considered to be selectively neutral that are located in different parts of the genome. MLST therefore provides a better tool for determination evolutionary relationships within the SDSE population than typing using *emm* gene, which is under strong diversifying selective pressure. Recent MLST studies of SDSE isolates from Australia, Portugal and USA reported a high degree of genetic diversity in these populations, and revealed LGT of housekeeping alleles was occurring [14,17]. In the current study we have used MLST to assess the genetic diversity of SDSE recovered from India, a country where streptococcal disease is endemic [18,19] and *emm*-type diversity is high [15,20]. Our results demonstrate this geographically confined collection to contain predominantly novel sequence types (STs). The ratio of

recombination to mutation in house-keeping alleles in this endemic region surpasses that reported for non-endemic regions [17]. The data additionally suggests that LGT between SDSE and other streptococcal species occur. Our findings suggest an evolutionary process in which novel genetic variants of SDSE, possibly with altered fitness or pathogenic potential, are more likely to arise in endemic regions than non-endemic regions.

Results

Allelic variation

Details of the ST and *emm*-type of the 181 SDSE isolates from India are provided in Table S1. A summary of nucleotide variation in the seven loci used for MLST in SDSE is provided in Table 1. The total number of alleles present at each locus varied from six for *gtr* to sixteen for *xpt*. Ten of the new SDSE MLST alleles identified in the study were identical to alleles from GAS (Table S2). Another allele, *recP22*, was identical to a nucleotide sequence from *S. agalactiae*, and shares greater than 99% identity with the same sequence in two other *S. agalactiae* genomes. As *recP22* is less than 90% identical to other so far known SDSE and *S. pyogenes recP* alleles, the allele was most likely acquired from *S. agalactiae*. Although evidence for mobile genetic element (MGE) mediated LGT between SDSE and *S. agalactiae* has been reported [21], to our knowledge this is the first evidence suggesting lateral transfer of an *S. agalactiae* housekeeping gene to SDSE.

When MLST alleles predicted to have been acquired by SDSE from non-SDSE sources through recombination (defined below) were disregarded, nucleotide diversity ranged from 0.004 for *gtr* to 0.042 for *atoB*. With the exception of *gtr* and *atoB*, the d_n/d_s ratio was less than 0.4 for all loci, indicating that the variation observed in these alleles is likely constrained by purifying selection. Although the d_n/d_s for *gtr* was relatively high compared to the other MLST alleles, nucleotide variation only occurred at four sites. The number of variable sites for the other MLST alleles ranged from 6 to 47. Phylogenetic analysis of the SDSE *atoB* and its orthologue in GAS (*yqiL*), showed two SDSE alleles, *atoB16* and *atoB17*, which accounted for the majority of non-synonomous mutations, lie at a node midway between the SDSE and GAS clusters (Figure S1).

Clonal relationships in the Indian SDSE population

MLST resolved the 181 isolates into 52 STs. The most common ST, ST84, was found in 43 isolates, all of which were *emm*-type stg4831. The five next most abundant STs (ST44, ST15, ST89, ST81 and ST107) each individually constitute 5 to 10% of the total collection, and together with ST84 account for 53% of isolates. In contrast, thirty two STs (62%) are represented by a single isolate. The overall diversity of *emm*-type ($D = 0.898$, 95% CI: 0.872–0.924) and ST ($D = 0.916$, 95% CI: 0.890–0.942) in the population is similar. The eBURST analysis segregated the 52 STs into seven single locus variant clonal complexes (CC_{slv}s), and twenty singletons (Figure 1). $CC_{slv}44$ contained the greatest number of STs (n = 13). $CC_{slv}84$ contained the greatest number of isolates (n = 43), and together with $CC_{slv}44$ (n = 41), account for 58% of all isolates. No other CC_{slv} contained more than 15 isolates. When clustered at the DLV (CC_{dlv}) level, seven complexes, and six singletons were defined. Together $CC_{dlv}44$ and $CC_{dlv}107$ account for 70% of isolates. Overall, while a large proportion of SLV defined singletons became a part of CC_{dlv}s, the CCs_{lv}s themselves remained largely discrete.

Relationships between STs were examined at the nucleotide level by constructing Minimum Evolution (ME) spanning trees using concatenated sequences of all seven MLST loci. In general, there is a good congruence between relationships of STs within CC_{slv}s, and the clustering of STs within individual branches of the ME tree (Figure 2). The branches of the ME tree were subsequently partitioned at nodes separating STs belonging to different CC_{slv}s The grouping of STs in these partitions were then compared to the grouping of STs by eBURST at the CC_{slv} level using the Wallace co-efficient [17,22]. The correlation between eBURST derived partition and ME derived partitions was calculated to be 0.630 (CI95% 0.388–0.872). There were five instances in the tree where STs did not cluster with the CC_{slv} as predicted by eBURST. As examples, ST117 and ST123, both of which belong to $CC_{slv}44$, possess gki15 and gki16. These alleles are identical to the GAS MLST alleles gki67 and gki38, respectively. ST86, another ST from $CC_{slv}44$ possesses the divergent *ato17* (Figure S1). Genetic relationships consistent with recombination were also apparent in STs that did not cluster as predicted from eBURST analysis.

Recombination and mutation in MLST loci

Estimates of recombination and mutation in SLV pairs were determined using previously described methods [22,23,24]. Of the 25 SLV pairs 22 are predicted to have arisen through recombination (Table S3); three recombination events involved non-SDSE derived alleles. Only two SLV pairs are predicted to have arisen through point mutation. The *r/m* ratio in the population was 11, and per site *r/m* ratio 164. When alleles predicted to be derived from non-SDSE species were excluded the per site *r/m* ratio fell to 41. The standardised Index of Association (I_A) across the population was 0.28, also suggestive of a high rate of recombination.

Table 1. Sequence variation in SDSE MLST loci from India.

Gene	Size of partial gene	Total alleles	New alleles	New non-SDSE alleles	nt variant positions	π[a]	d_n	d_s	d_n/d_s
gki	498	8	3	3	14	0.011	0.011	0.043	0.247
gtr	450	6	2	2	4	0.004	0.004	0.006	0.678
murI	438	9	4	2	15	0.015	0.008	0.036	0.222
mutS	405	9	6	1	6	0.006	0.002	0.019	0.081
recP	459	14	4	2	31	0.034	0.007	0.150	0.041
xpt	450	16	8	0	28	0.020	0.005	0.071	0.067
atoB	434	10	5	1	47	0.042	0.033	0.067	0.486

[a]Nucleotide diversity, d_n and d_s values were determined using alleles unique to SDSE. i.e. alleles likely acquired from *S. pyogenes* and *S. agalactiae* were excluded.
doi:10.1371/journal.pone.0021346.t001

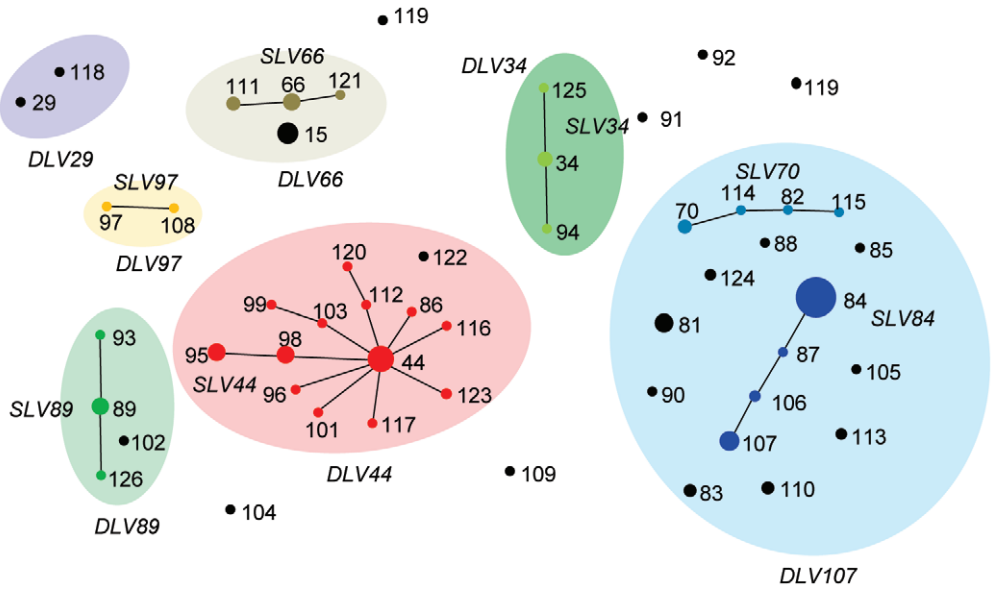

Figure 1. Population structure of Indian SDSE isolates. The 52 STs clustered into seven SLV-defined CCs. STs part of the same SLV defined CC are joined by black lines. When grouped at the less stringent DLV level seven CCs are defined (light shaded circles).
doi:10.1371/journal.pone.0021346.g001

Recombination events can distort or conceal true evolutionary relationships that exist between STs. In these instances, standard phylogenetic trees, which only display single relationships between isolates or clones, do not provide the equal representation for all possible evolutionary relationships. Split decomposition analysis was therefore used to visualise alternative phylogenetic relationships between STs (Figure 3). The reticulated phylogenetic structure of this figure is indicative of extensive recombination of loci [25] providing additional support for the high estimates of recombination and low I_A. The majority of STs were found in the same groupings as determined by eBURST. However STs belonging to DLV107 segregated into two separate groups.

Relationship between *emm*-type and ST

emm-typing is the most common method of typing SDSE. Discrepancies between *emm*-type and ST have previously been reported [17] and offers evidence that LGT of *emm*-genes occurs in nature [26]. The presence of multiple *emm*-types within a single ST that are identical to, or have close similarity to emm-genes found in distant STs is suggestive of LGT of this gene [17,27]. Alternatively, diversifying selection pressure on the emm-gene may also give rise to STs that harbour multiple emm-types. In this study, eleven STs were associated with more than one *emm*-type (Table 2). Two of these, ST44 and ST15, were each associated with five *emm*-types. Together nineteen instances of multiple emm-genes present in individual STs was observed. A complementary method for inference of LGT is the identification of the same *emm*-gene in distantly related or unrelated STs. In total eleven *emm*-genes were found in two or more CC_{dlv}s (Table S4). Four *emm* genes, *stc36*, *stg485*, *stg480* and *stg866* were present in 4, 4, 3 and 3 CC_{dlv}s respectively (Figure 3). Seventeen instances of *emm*-gene LGT were inferred using this method. In contrast to the frequency of recombination of the *emm*-gene, the predicted number of recombination events for individual housekeeping alleles used in MLST ranged from one for *xpt* to six for *murI* based on SLV relationships.

To further investigate the relationships between *emm*-type and ST, we constructed a phylogenetic tree using the 150 nucleotides of the emm-gene used to determine emm-type (Figure 4). The *emm*-gene sequences were aligned using ClustalW prior to construction of the ME tree. When MLST data was overlaid onto this tree, CC_{slv}s and CC_{dlv}s were scattered throughout the tree. When *emm*-types were mapped onto the SplitsTree phylogram (which accommodates effects of recombination), a much clearer association between *emm*-type and MLST data became apparent, with the majority of isolates of the *emm*-types falling into the same cluster (Figure 3). These clusters are therefore the likely progenitors of specific emm-genes.

Indian isolates possess unique STs

Eighty STs were described in a previous study of 178 SDSE isolates from non-endemic regions [17]. Only six of these STs (ST15 ST29 ST34 ST44 ST66 ST70) were found amongst the Indian strains. Of the 46 new STs in this study, nineteen possessed previously reported alleles at all seven loci, but in combinations not previously found. Twenty-seven STs possessed at least one new allelelic sequence. When all known SDSE STs were compared using eBURST, seven CC_{dlv}s were defined (Figure S2). One of these complexes was considerably larger than all others with 87 STs. With the exception of four STs (ST97, ST108, ST100 and ST104), all Indian STs were found in this large complex. In contrast only 43 of the 74 non-Indian STs were found within this complex. Further comparisons using split decomposition (Figure S3) and ME analysis (Figure S4) also segregated Indian isolates into subclusters separate from other SDSE isolates.

Association between disease and ST

Chi-squared goodness-of-fit tests were used to determine if any ST or CC was over-represented in strains recovered from individuals with throat infection,(i.e. presenting with pharyngitis or tonsillitis) when compared to strains recovered from individuals without symptoms of throat infection [15]. The latter control group included all SDSE throat carriage isolates, and SDSE recovered from the skin. Three STs were found to have statistically different distributions between throat-infection and control isolates. ST89 and ST111 were found to be overrepresented in

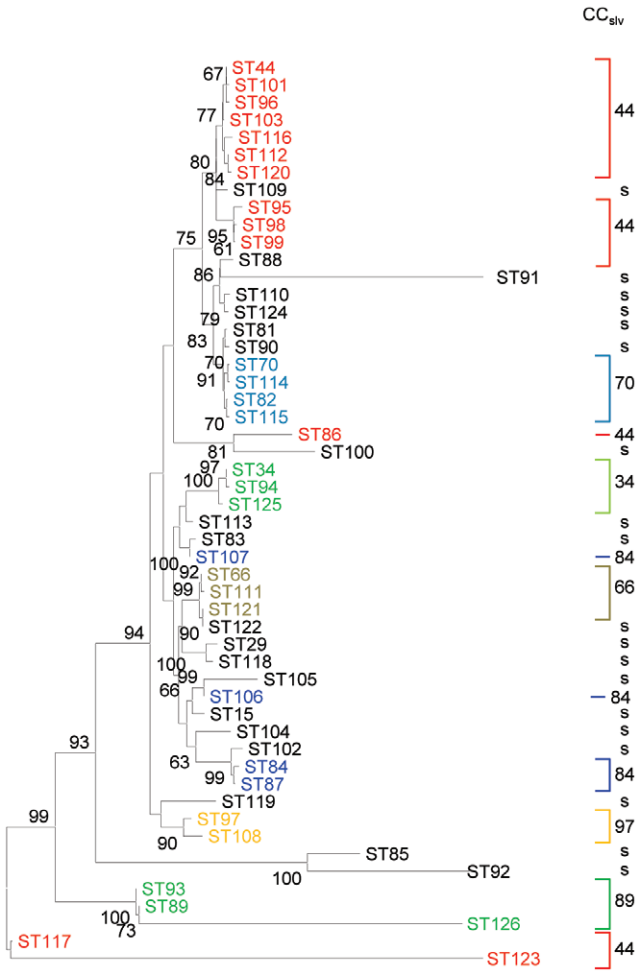

Figure 2. Minimum evolutionary tree of concatenated SDSE MLST loci. The tree was constructed using concatenated sequences of 52 SDSE STs. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Only bootstrap values greater than 80% are shown. Clonal complexes at both the SLV and DLV level, and singletons (s), as determined using eBURST, are also depicted.
doi:10.1371/journal.pone.0021346.g002

the throat infection isolates (relative risk of 2.61 and 2.85 respectively). The third ST, ST107 was underrepresented (relative risk of 1.63 for non-throat infection). No CC had an increased association with throat infection.

Discussion

Streptococcal infection is considered to be endemic in India. As a consequence co-infection with multiple SDSE strains, or SDSE and other streptococcal species, is more likely to occur here than in countries where streptococcal disease is non-endemic. An outcome of increased inter-strain contact is increased opportunity for LGT, resulting in an increase in overall strain diversity. The ratio of recombination to mutation reported here is double that reported for SDSE from non-endemic areas, supporting this scenario [17]. Further the inter-relatedness of Indian isolates and presence of only a few global STs, support a model whereby ongoing recombination between isolates in a single geographic location result in strains that over time become more related to each other than to strains from outside of the population.

In our study, recombination involving the *emm*-gene was more frequent than recombination involving individual housekeeping alleles. The *emm*-gene is part of an ancient pathogenicitiy island that is mobilisable *in vitro* [13]. An active mechanism for LGT involving the emm-gene therefore exists, and may account for the increased LGT of this gene. The acquisition of new emm-genes that assist SDSE in evasion of host immune responses, giving recipients a selective advantage, may also increase the frequency in which LGT involving the emm-gene is observed when compared to selectively neutral house-keeping genes. In spite of this, we observed more *S. pyogenes* housekeeping alleles than *S. pyogenes emm*-genes among our SDSE isolates. This suggests restricted compatibility for the *S. pyogenes emm* genes in SDSE, and may explain distinct evolutionary clades for the genes in the two species. The amino-termini of the mature M-proteins are the major target of type-specific anti-GAS and anti-SDSE antibodies. While population and virulence studies of *S. pyogenes* report associations between *emm*-types and specific disease, associations between *emm*-type and SDSE-mediated disease are not obvious [28,29,30]. One outcome from our study that could account for this observation is that the greater recombination occurring among SDSE isolates in regions where streptococcal infection is endemic results in the weakening of emm-typing as a useful tool for assigning genetic relationships over a meaningful time frame amongst geographically separated SDSE isolates. Of the six STs shared between India and the rest of the world, only five *emm*-STs were common. Twelve *emm*-STs were only found in India, and another nine *emm*-STs found in non-Indian isolates.

One *emm*-type, *stg480*, has been associated with SDSE infection in several studies [5,31]. This *emm*-type was also one of the most commonly recovered *emm*-types in this study, but was not recovered more frequently from individuals with pharyngitis. The apparent increased association with diseases reported in other studies may reflect the relative abundance of this *emm*-type within the population, rather than an increased virulence potential. *stg480* has now been associated with five STs in this and previous studies [14,17], suggesting that this *emm*-locus has propensity to participate in LGT frequently, which indeed further clouds epidemiological findings based on *emm*-type. Larger prospective studies that include characterisation of both the *emm*-gene and ST are required to determine the pathogenic potential of individual SDSE lineages.

In contrast to *stg480*, the majority of *stg4831* isolates were predominantly associated with ST84. All ST84 isolates also possessed the *stg4831 emm*-gene. As all ST84 and *stG4831* isolates were only recovered in Mumbai, it is likely that during the collection period, an outbreak of *stg4831*-ST84 was occurring. The general lack of variation in ST-*emm* combinations suggests this outbreak is relatively new, leaving little time for mutation or recombination with existing strains. Nevertheless, the presence of three stG4831 clones with different ST (ST87, ST85 and ST102), all recovered from Mumbai, once again suggests that recombination was occurring at a local level. ST87 is an SLV of ST84, predicted to have arisen *via* recombination. ST85, a DLV variant of ST84 is more closely related to ST84 than any other ST. Two of the three alleles that differ between ST102 and ST84 are found in ST89. In this instance, the data suggest recombination between an ST84 clone and isolate from CC$_{slv}$89 has occurred.

Frequent recombination is emerging as a paradigm for the β-hemolytic streptococci. Taken together our findings suggest that SDSE ST diversity is high in regions where streptococcal disease is considered endemic, and is driven mainly by recombination. In endemic regions, the opportunity for different streptococcal isolates to come into contact, and share genes is greater than in

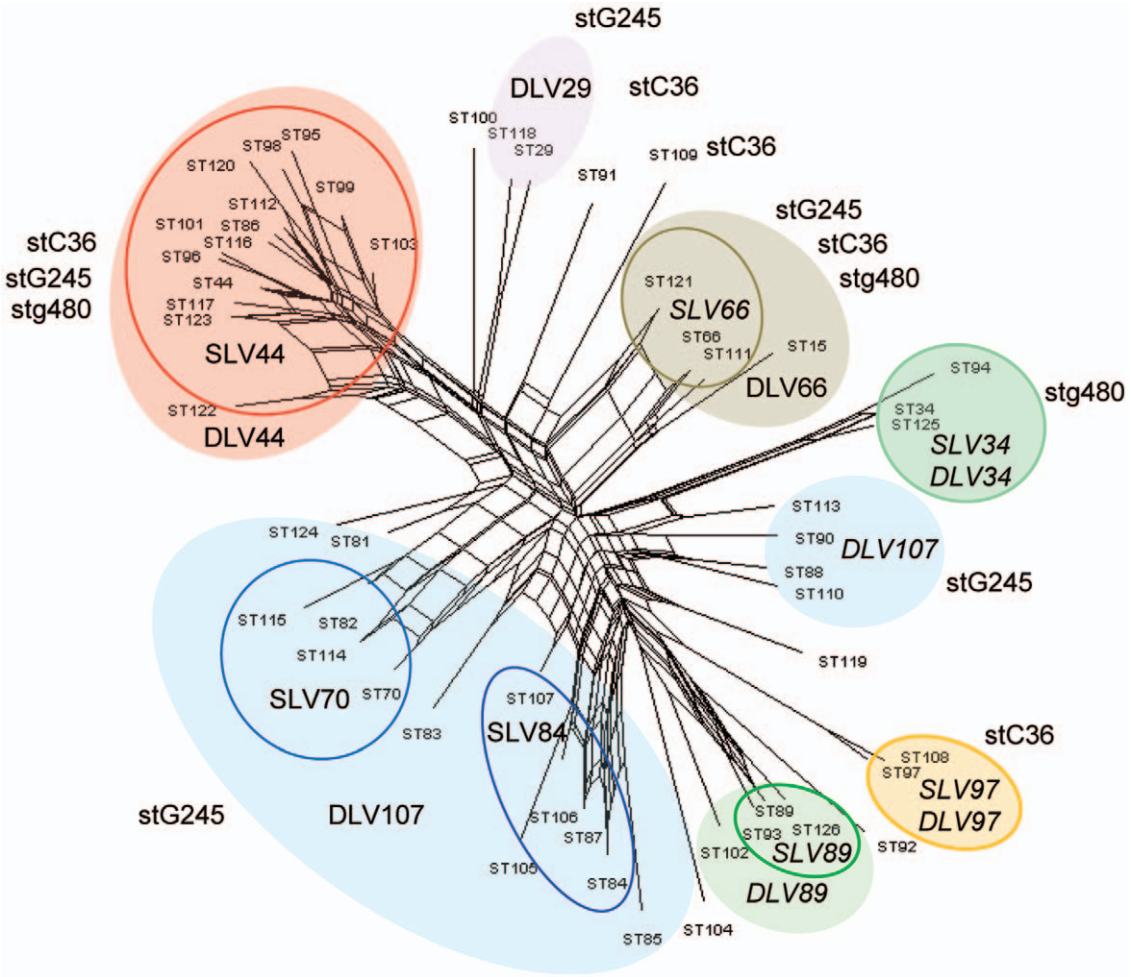

Figure 3. Split decomposition analysis of Indian STs. Shaded circles indicate DLV defined CCs. Solid lines enclose SLV defined CCs. STs associated with *stC36*, *stg245* and *stg480* are also indicated.
doi:10.1371/journal.pone.0021346.g003

regions where streptococcal infection is less prevalent. SDSE, *S. pyogenes* and *S. agalactiae* all possess multiple MGEs, and MGE-related interspecies transfer of virulence genes has been reported by several groups [26,32,33], The transfer of non-MGE housekeeping genes may occur as bystander event during transfer of MGEs [34]. It is noteworthy that transfer of allelic variants of virulence genes not associated with mobile genetic elements is also occurring, possibly having a greater impact on the fitness or virulence of some SDSE isolates. The observed increase in the rate of recovery of new variants in endemic regions suggests rapid emergence of more fit or virulent clones are more likely to occur in these regions.

Table 2. Sequence types associated with multiple *emm*-types.

ST	Emm-type
15	*stc36, stg10, stg245, stg866, stgL265*
29	*stc74a, stg245*
34	*stg480, stgm22*
44	*stc36, stg245, stg480, stg6, stgL265*
66	*stC1400, stg6792*
81	*stg245, stg6*
83	*stc74a, stg245*
95	*stg653, stg6972*
98	*stc36, stg2078, stg480*
107	*emm23, stC5345, stc6979*
110	*stg1750, stg245*

doi:10.1371/journal.pone.0021346.t002

Materials and Methods

Ethics statement

Ethical approval for swabbing of individuals in the study was granted by the Seth G. S. Medical School and KEM Hospital Ethic Committee, India (EC/Gov/-4/2006). Written informed consent for the swabbing of children was obtained from the guardians of all children included in the study.

Bacterial Strains

A total of 181 SDSE isolates collected as part of community and school surveys carried out in Mumbai [15] and Chennai [35] were used in the study. Eighty-five isolates were collected from individuals presenting with pharyngitis (n = 66) or tonsillitis

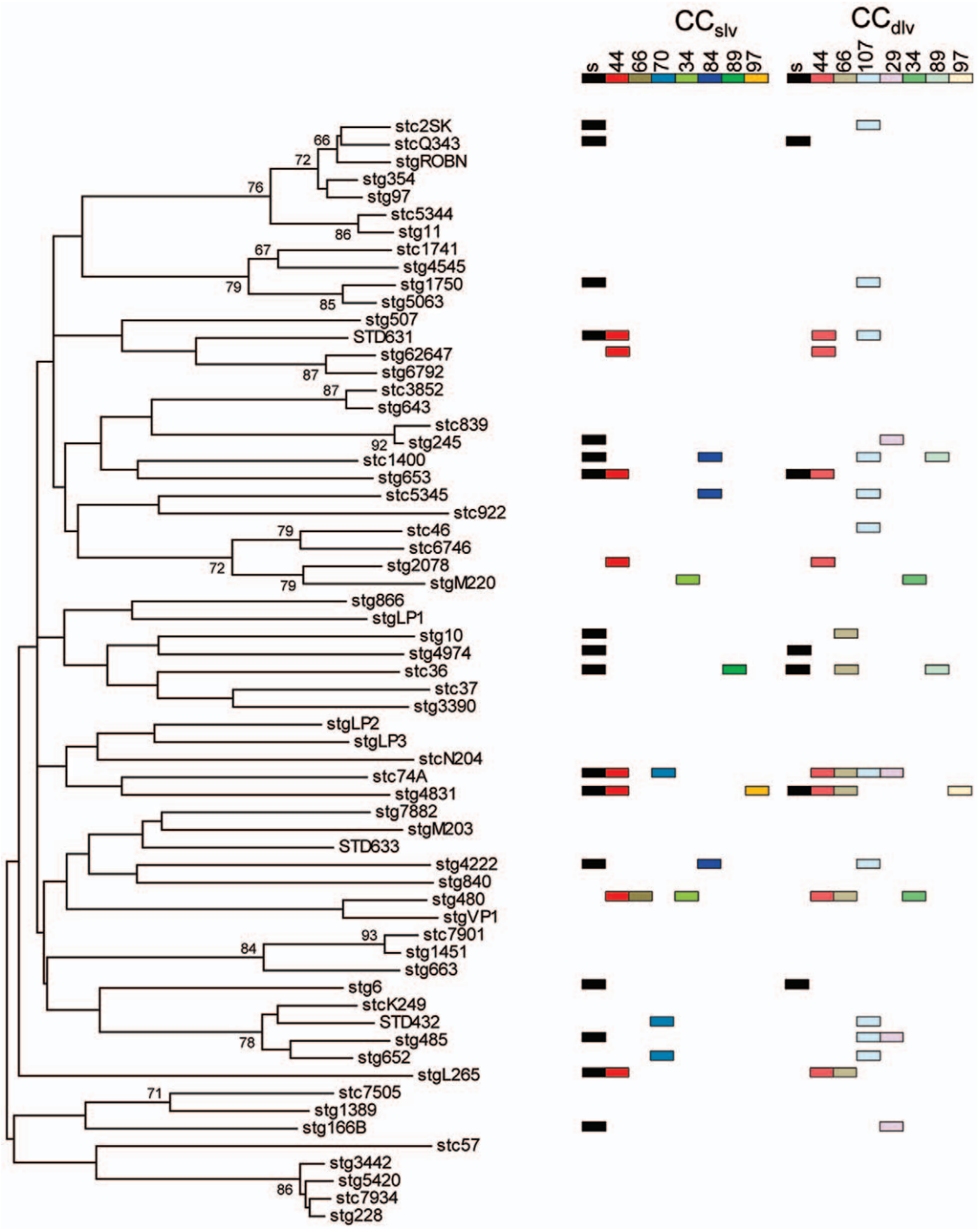

Figure 4. Phylogenetic relationship of SDSE based on *emm* nucleotide sequence. The tree was constructed using the ME algorithm. Bootstrap values in which associated taxa are clustered in greater than 60% of cases are shown next to branches. CC_{slv}s and CC_{dlv}s associated with emm-types are also shown.
doi:10.1371/journal.pone.0021346.g004

(n = 10), and were classified as 'throat-infection' isolates. Another 102 isolates were recovered from the throats of individuals lacking clinical signs of streptococcal infection. Nine SDSE isolates were collected from the skin of individuals presenting with pyoderma. All isolates were classified as SDSE on the basis of β-hemolytic activity, expression of the group C or G carbohydrate, and possession of characteristic *emm*-type and molecular markers characteristic of SDSE [10]. *Emm*-types for these isolates were previously reported [15,35] or determined using standard protocols [20,36]. Full details of strains used in this study are provided in Table S1.

MLST

The SDSE MLST scheme based on the following seven gene targets (glucose kinase (*gki*), glutamine transport protein (*gtr*), glutamate racemase (*murI*), DNA mismatch repair protein (*mutS*),

transketolase (*recP*), xanthine phosphoribosyl transferase (*xpt*) and acetoacetyl-coathioloase (*atoB*)) has been described previously [17]. With the exception of *atoB* these alleles are the same as used in the GAS MLST scheme [27]. DNA was extracted using the QIAGEN DNAeasy kit (QIAGEN, Australia), and 450–500 base-pair internal fragments of these genes were amplified under the following conditions; 2 min denaturation at 95°C, followed by 35 cycles of 95°C (45 s), 50°C (45 s) and 72°C (60 s). PCR products were purified using ExoSAPIT (USB Corp, USA), and sequenced in the forward and reverse directions by Macrogen (Korea) or in-house. Sequencher (Genecodes, USA) was used for initial analysis and trimming of sequences to match reference sequences. All sequences were compared to existing SDSE MLST alleles to determine specific allele number at a given locus. Unique allelic sequences identified in this study were assigned a new allele number. The combination of seven allele numbers was then used to determine the sequence type (ST). goeBURST (http://goeburst.phyloviz.net/) was used to identify related STs [37,38]. In this study Clonal Complex (CCs) were defined as a group of STs that are related to each other at the Single Locus Variant (CC_{slv}) or Double Locus Variant (CC_{dlv}) level.

Recombination and mutation

Rates of recombination and mutation that give rise to SLV pairs were estimated as previously described [22,23,24]. Briefly, SLV pairs that contain greater than one nucleotide difference in the variant alleles were classified as arising through recombination. SLV pairs in which one ST contains a unique single nucleotide polymorphism not found in other STs were classified as a point mutation event. Single nucleotide changes in SLVs pairs giving rise to alleles already present in unrelated STs in different CC_{slv}s were classified as recombination events. Both the ratio of recombination/mutation events (r/m) and per nucleotide site ratio of recombination/mutation are reported.

Phylogenetic Analysis

Nt diversity (π), nonsynonymous (d_n) and synonymous substitution rates (d_s) were calculated using DnaSP (version 5) [39]. Linkage equilibrium, expressed as standardised Index of Association (I_A) was calculated using LIAN 3.5 [40]. Distance matrices for phylogenetic analysis were calculated using START [41]. Phylogenetic networks were constructed using Splitstree 4 [25]. Minimum evolution (ME) spanning trees using concatenated nucleotide sequences from all seven MLST loci were constructed in Mega4 [42], with support for branches provided by bootstrapping (n = 1000). Phylogenetic analysis of the 180 nt variable region of the *emm*-gene used for *emm*-typing was also performed using Mega4. The sequences were aligned using ClustalW, using default parameters, prior to construction of the ME tree. The Wallace co-efficient [23,24] was calculated using www.comparingpartitions.info.

Statistical analysis

Statistically significant associations between ST, Clonal Complex and other epidemiological factors were assessed using the Chi-squared goodness of fit test ($p<0.05$). Confidence intervals for Simpson index of diversity (D) were calculated as previously described [43].

Supporting Information

Figure S1 Phylogenetic analysis of SDSE *atoB* and S. *pyogenes yqiL* alleles. The relationship between alleles was inferred using the Minimum Evolution method, and support for branches provided by bootstrapping (n = 1000). Bootstrap values are only provided for branches with greater than 50% support.
(TIF)

Figure S2 eBURST analysis of all known SDSE STs. Blue circles represent STs only found in India. Red circles represent STs found outside of India (McMillan et al., 2010). Green circles represent STs found in both collections. Dark connecting lines join SLV related pairs. Grey lines connect DLV related pairs.
(TIF)

Figure S3 Split decomposition of all SDSE STs. STs found in India are circled.
(TIF)

Figure S4 ME tree of all SDSE STs. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Only bootstap values greater than 50% are shown. Blue circles represent STs only found in India. Red circles represent STs found outside of India (McMillan et al., 2010). Green circles represent STs found in both collections.
(TIF)

Table S1 Details of SDSE isolates used in this study.
(DOC)

Table S2 Novel MLST alleles in the Indian SDSE population.
(DOC)

Table S3 Recombination and mutation in SDSE.
(DOC)

Table S4 *Emm*-types associated with multiple Sequence Types.
(DOC)

Author Contributions

Conceived and designed the experiments: DJM MGK MSS KSS. Performed the experiments: DJM SYK PVB TV. Analyzed the data: DJM SYK PVB PRS. Contributed reagents/materials/analysis tools: DJM MSS KSS. Wrote the paper: DJM PRS KSS.

References

1. Carapetis JR, Steer AC, Mulholland EK, Weber M (2005) The global burden of group A streptococcal diseases. Lancet Infect Dis 5: 685–694.
2. Efstratiou A (1997) Pyogenic streptococci of Lancefield groups C and G as pathogens in man. Soc Appl Bacteriol Symp Ser 26: 72S–79S.
3. Brandt CM, Spellerberg B (2009) Human infections due to Streptococcus dysgalactiae subspecies equisimilis. Clin Infect Dis 49: 766–772.
4. Broyles LN, Van Beneden C, Beall B, Facklam R, Shewmaker PL, et al. (2009) Population-based study of invasive disease due to beta-hemolytic streptococci of groups other than A and B. Clin Infect Dis 48: 706–712.
5. Cohen-Poradosu R, Jaffe J, Lavi D, Grisariu-Greenzaid S, Nir-Paz R, et al. (2004) Group G streptococcal bacteremia in Jerusalem. Emerg Infect Dis 10: 1455–1460.
6. Skogberg K, Simonen H, Renkonen OV, Valtonen VV (1988) Beta-haemolytic group A, B, C and G streptococcal septicaemia: a clinical study. Scand J Infect Dis 20: 119–125.
7. Chhatwal GS, McMillan DJ, Talay SR (2006) Pathogenicity Factors in Group C and G Streptococci. In: Fischetti VA, Novick RP, Ferretti JJ, Portnoy DA, Rood JI, eds. Gram-Positive Pathogens. 2 ed. Washington, D.C.: ASM Press. pp 213–221.
8. Fischetti VA (1989) Streptococcal M protein: molecular design and biological behavior. Clin Microbiol Rev 2: 285–314.
9. Smeesters PR, McMillan DJ, Sriprakash KS (2010) The streptococcal M protein: a highly versatile molecule. Trends Microbiol 18: 275–282.
10. McMillan DJ, Vu T, Bramhachari PV, Kaul SY, Bouvet A, et al. (2010) Molecular markers for discriminating Streptococcus pyogenes and S. dysgalactiae subspecies equisimilis. Eur J Clin Microbiol Infect Dis 29: 585–589.
11. Hartas J, Hibble M, Sriprakash KS (1998) Simplification of a locus-specific DNA typing method (Vir typing) for Streptococcus pyogenes. J Clin Microbiol 36: 1428–1429.

12. Bessen DE, Sotir CM, Readdy TL, Hollingshead SK (1996) Genetic correlates of throat and skin isolates of group A streptococci. J Infect Dis 173: 896–900.
13. Panchaud A, Guy L, Collyn F, Haenni M, Nakata M, et al. (2009) M-protein and other intrinsic virulence factors of Streptococcus pyogenes are encoded on an ancient pathogenicity island. BMC Genomics 10: 198.
14. Ahmad Y, Gertz RE, Jr., Li Z, Sakota V, Broyles LN, et al. (2009) Genetic relationships deduced from emm and multilocus sequence typing of invasive Streptococcus dysgalactiae subsp. equisimilis and S. canis recovered from isolates collected in the United States. J Clin Microbiol 47: 2046–2054.
15. Bramhachari PV, Kaul SY, McMillan DJ, Shaila MS, Karmarkar MG, et al. (2010) Disease burden due to Streptococcus dysgalactiae subsp. equisimilis (group G and C streptococcus) is higher than that due to Streptococcus pyogenes among Mumbai school children. J Med Microbiol 59: 220–223.
16. McDonald M, Towers RJ, Andrews RM, Carapetis JR, Currie B (2007) Epidemiology of Streptococcus dysgalactiae subsp. equisimilis in Tropical Communities, Northern Australia. Emerg Infect Dis 13: 1694–1700.
17. McMillan DJ, Bessen DE, Pinho M, Ford C, Hall GS, et al. (2010) Population genetics of Streptococcus dysgalactiae subspecies equisimilis reveals widely dispersed clones and extensive recombination. PLoS One 5: e11741.
18. Padmavati S (2001) Rheumatic heart disease: prevalence and preventive measures in the Indian subcontinent. Keywords: rheumatic heart disease; rheumatic fever. Heart 86: 127.
19. Shet A, Kaplan E (2004) Addressing the burden of group A streptococcal disease in India. Indian J Pediatr 71: 41–48.
20. Dey N, McMillan DJ, Yarwood PJ, Joshi RM, Kumar R, et al. (2005) High diversity of group A Streptococcal emm types in an Indian community: the need to tailor multivalent vaccines. Clin Infect Dis 40: 46–51.
21. Davies MR, Shera J, Van Domselaar GH, Sriprakash KS, McMillan DJ (2009) A novel integrative conjugative element mediates genetic transfer from group G Streptococcus to other {beta}-hemolytic Streptococci. J Bacteriol 191: 2257–2265.
22. McGregor KF, Spratt BG, Kalia A, Bennett A, Bilek N, et al. (2004) Multilocus sequence typing of Streptococcus pyogenes representing most known emm types and distinctions among subpopulation genetic structures. J Bacteriol 186: 4285–4294.
23. Feil EJ, Enright MC, Spratt BG (2000) Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between Neisseria meningitidis and Streptococcus pneumoniae. Res Microbiol 151: 465–469.
24. McMillan DJ, Bessen DE, Pinho M, Ford C, Hall GS, et al. Population genetics of Streptococcus dysgalactiae subspecies equisimilis reveals widely dispersed clones and extensive recombination. PLoS One 5: e1741.
25. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23: 254–267.
26. Simpson WJ, Musser JM, Cleary PP (1992) Evidence consistent with horizontal transfer of the gene (emm12) encoding serotype M12 protein between group A and group G pathogenic streptococci. Infect Immun 60: 1890–1893.
27. Enright MC, Spratt BG, Kalia A, Cross JH, Bessen DE (2001) Multilocus sequence typing of Streptococcus pyogenes and the relationships between emm type and clone. Infect Immun 69: 2416–2427.
28. Davies MR, McMillan DJ, Beiko RG, Barroso V, Geffers R, et al. (2007) Virulence profiling of Streptococcus dysgalactiae subspecies equisimilis isolated from infected humans reveals 2 distinct genetic lineages that do not segregate with their phenotypes or propensity to cause diseases. Clin Infect Dis 44: 1442–1454.
29. Pinho MD, Melo-Cristino J, Ramirez M (2006) Clonal relationships between invasive and noninvasive Lancefield group C and G streptococci and emm-specific differences in invasiveness. J Clin Microbiol 44: 841–846.
30. Ikebe T, Murayama S, Saitoh K, Yamai S, Suzuki R, et al. (2004) Surveillance of severe invasive group-G streptococcal infections and molecular typing of the isolates in Japan. Epidemiol Infect 132: 145–149.
31. Lopardo HA, Vidal P, Sparo M, Jeric P, Centron D, et al. (2005) Six-month multicenter study on invasive infections due to Streptococcus pyogenes and Streptococcus dysgalactiae subsp. equisimilis in Argentina. J Clin Microbiol 43: 802–807.
32. Towers RJ, Gal D, McMillan D, Sriprakash KS, Currie BJ, et al. (2004) Fibronectin-binding protein gene recombination and horizontal transfer between group A and G streptococci. J Clin Microbiol 42: 5357–5361.
33. Sriprakash KS, Hartas J (1996) Lateral genetic transfers between group A and G streptococci for M-like genes are ongoing. Microb Pathog 20: 275–285.
34. Brochet M, Rusniok C, Couve E, Dramsi S, Poyart C, et al. (2008) Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of Streptococcus agalactiae. Proc Natl Acad Sci U S A 105: 15961–15966.
35. Menon T, Lloyd C, Malathy B, Sakota V, Jackson D, et al. (2008) emm type diversity of beta-haemolytic streptococci recovered in Chennai, India. J Med Microbiol 57: 540–542.
36. Beall B, Facklam R, Thompson T (1996) Sequencing emm-specific PCR products for routine and accurate typing of group A streptococci. J Clin Microbiol 34: 953–958.
37. Francisco AP, Bugalho M, Ramirez M, Carrico JA (2009) Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinformatics 10: 152.
38. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG (2004) eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J Bacteriol 186: 1518–1530.
39. Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451–1452.
40. Haubold B, Hudson RR (2000) LIAN 3.0: detecting linkage disequilibrium in multilocus data. Linkage Analysis. Bioinformatics 16: 847–848.
41. Jolley KA, Feil EJ, Chan MS, Maiden MC (2001) Sequence type analysis and recombinational tests (START). Bioinformatics 17: 1230–1231.
42. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.
43. Grundmann H, Hori S, Tanner G (2001) Determining confidence intervals when measuring genetic diversity and the discriminatory abilities of typing methods for microorganisms. J Clin Microbiol 39: 4190–4192.