Transcriptome Remodeling Contributes to Epidemic Disease Caused by the Human Pathogen *Streptococcus pyogenes*

Stephen B. Beres, a Priyanka Kachroo, a Waleed Nasser, a Randall J. Olsen, a,b Luchang Zhu, a Anthony R. Flores, a,c Ivan de la Riva, a Jesus Paez-Mayorga, a Francisco E. Jimenez, a Concepcion Cantu, a Jaana Vuopio, a,d Jari Jalava, a Karl G. Kristinsson, a,d Magnus Gottfredsson, a Jukka Corander, a Nahuel Fittipaldi, a Maria Chiara Di Luca, j Dezemona Petrelli, a Luca A. Vitali, j Annessa Raiford, j Leslie Jenkins, j James M. Musser a,b

Center for Molecular and Translational Human Infectious Diseases Research, Department of Pathology and Genomic Medicine, Houston Methodist Research Institute, and Houston Methodist Hospital, Houston, Texas, USA; Departments of Pathology and Laboratory Medicine and Microbiology and Immunology, Weill Cornell Medical College, New York, New York, USA; Section of Infectious Diseases, Department of Pediatrics, Texas Children’s Hospital and Baylor College of Medicine, Houston, Texas, USA; Department of Medical Microbiology and Immunology, Medical Faculty, University of Turku, Turku, Finland; Department of Infectious Diseases, National Institute for Health and Welfare, Turku, Finland; Departments of Clinical Microbiology and Infectious Diseases, Landspitali University Hospital, Reykjavik, Iceland; Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland; Department of Biostatistics, University of Oslo, Oslo, Norway, and Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland; Public Health Ontario, and Department of Laboratory Medicine and Pathobiology, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada; School of Pharmacy, University of Camerino, Camerino, Italy; School of Biosciences and Veterinary Medicine, University of Camerino, Camerino, Italy; Comparative Medicine Program, Houston Methodist Research Institute, Houston, Texas, USA

S.B.B., P.K., W.N., R.J.O., and L.Z. contributed equally to this article.

ABSTRACT For over a century, a fundamental objective in infection biology research has been to understand the molecular processes contributing to the origin and perpetuation of epidemics. Divergent hypotheses have emerged concerning the extent to which environmental events or pathogen evolution dominates in these processes. Remarkably few studies bear on this important issue. Based on population pathogenomic analysis of 1,200 *Streptococcus pyogenes* type emm89 infection isolates, we report that a series of horizontal gene transfer events produced a new pathogenic genotype with increased ability to cause infection, leading to an epidemic wave of disease on at least two continents. In the aggregate, these and other genetic changes substantially remodeled the transcriptomes of the evolved progeny, causing extensive differential expression of virulence genes and altered pathogen-host interaction, including enhanced immune evasion. Our findings delineate the precise molecular genetic changes that occurred and enhance our understanding of the evolutionary processes that contribute to the emergence and persistence of epidemically successful pathogen clones. The data have significant implications for understanding bacterial epidemics and for translational research efforts to blunt their detrimental effects.

IMPORTANCE The confluence of studies of molecular events underlying pathogen strain emergence, evolutionary genetic processes mediating altered virulence, and epidemics is in its infancy. Although understanding these events is necessary to develop new or improved strategies to protect health, surprisingly few studies have addressed this issue, in particular, at the comprehensive population genomic level. Herein we establish that substantial remodeling of the transcriptome of the human-specific pathogen *Streptococcus pyogenes* by horizontal gene flow and other evolutionary genetic changes is a central factor in precipitating and perpetuating epidemic disease. The data unambiguously show that the key outcome of these molecular events is evolution of a new, more virulent pathogenic genotype. Our findings provide new understanding of epidemic disease.

Received 4 March 2016 Accepted 4 May 2016 Published 31 May 2016


Editor Richard P. Novick, Skirball Institute of Biomolecular Medicine, New York University Medical Center

Copyright © 2016 Beres et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to James M. Musser, jmmusser@houstonmethodist.org.

This article is a direct contribution from a Fellow of the American Academy of Microbiology. External solicited reviewers: Eldie Ghe din, New York University; Alan R. Hauser, Northwestern University; Arnold L. Smith, Children’s Hospital Seattle.

Genetic diversity begets phenotype variation and with it the possibility of a different life. Considerable effort has been expended in the last 40 years to understand the genetic diversity and population structure of many bacterial pathogens, especially those that detrimentally affect human and livestock health and cause epidemics (1–27). These studies have led to the general concept that some bacterial species are clonal, with relatively little evidence that horizontal gene transfer (HGT) and recombination shape species diversity, whereas other bacterial pathogens are highly recombinogenic, with species diversity mediated by extensive HGT events (1–27). Genetic studies have been greatly facilitated in recent years by relatively inexpensive large-scale comparative DNA sequencing, which now makes it possible to precisely delineate the nature and extent of genomic variation present in...
large populations (hundreds to many thousands) of individual pathogenic bacterial species (4, 5, 10–18, 23–26). For example, analyses of important pathogens such as *Staphylococcus aureus*, *Streptococcus pyogenes*, *Streptococcus pneumoniae*, *Escherichia coli*, *Salmonella enterica* serovars, and *Legionella pneumophila* have been conducted, resulting in much new information about genetic variation in these and other species (4, 5, 10–18, 23–26, 28–35).

In parallel with studies of bacterial population genetic structure, there has been interest in identifying the precise genomic changes that contribute to the emergence, numerical success, and epidemic behavior of members of some bacterial species. A major effort has been devoted to analysis of comprehensive, population-based samples of the strict human pathogen *S. pyogenes* (commonly, group A streptococcus [GAS]) as a model pathogen (28–35). *S. pyogenes* is endemic in humans worldwide and periodically causes epidemics of superficial (e.g., pharyngitis and impetigo) and invasive (e.g., necrotizing fasciitis, pneumonia, myositis) infections. Globally, the organism causes an estimated 711 million human infections and over 500,000 deaths annually (36). The species is genetically diverse, with more than 240 *emm*-types (typing based on sequence differences in the hypervariable amino-terminal portion of the *emm* gene encoding the antiphagocytic Emm virulence protein; http://www.cdc.gov/abcs/index.html), and approximately 650 multilocus sequence types (MLSTs) (http://sppyogenes.mlst.net) described.

In the early 1980s, a dramatic increase in the frequency and severity of infections caused by *S. pyogenes* led to the recognition of a global pandemic caused by *emm*1 strains (37–44). This pandemic afforded the opportunity to compare preepidemic and epidemic strains for potential bacterial factors contributing to this global health problem. To gain insight into the emergence, dissemination, and diversification of *emm*1 strains causing this pandemic, we sequenced the genome of 3,615 *emm* infection isolates (32). Phylogenetic analyses revealed that the pandemic *emm*1 strains that emerged are a genetically closely related clonal population that evolved from its most recent preepidemic progenitor in the early 1980s. The key genetic event underpinning the pandemic was acquisition by HGT and recombinational replacement of a 36 kb segment of the *S. pyogenes* core chromosome (i.e., that portion of the chromosome/genome that is largely conserved across *emm*-types and not present on obvious mobile genetic elements such as plasmids and integrative-conjugative elements) that mediated enhanced production of toxins NAD⁺ glycohydrolase (SPN [S. pyogenes NADase]) and streptolysin O (SLO) (32). A subsequent study (35) showed that the striking upregulation of SPN and SLO production and altered virulence phenotype by members of the pandemic clone occurred as a consequence of only three single nucleotide polymorphisms (SNPs). Two are located in the −35 to −10 spacer region of the promoter sequence upstream of the *rnga*-ifs-slo transcriptional unit and resulted in increased gene expression. The third, a nonsynonymous SNP in the *rnga* gene, increases the activity of SPN, a secreted cytotoxic virulence factor (45). Additional evidence supporting the notion of upregulation of SPN and SLO as a contributing cause of *S. pyogenes* epidemic disease was found by sequence analysis of 1,125 *emm*89 genomes (35) obtained in comprehensive population-based surveillance studies conducted in the United States, Finland, and Iceland between 1995 and 2013. Among these *emm*89 strains, we identified three distinct phylogenetic clades (designated clade 1, clade 2, and clade 3). The current worldwide recent increase in the incidence of *emm*89 invasive infections corresponded temporally with the emergence and expansion of clade 3 strains upregulated in SPN and SLO production (35, 46).

Thus, progress is being made in understanding genomic alterations that are linked with increases in disease frequency and severity in some human pathogens. However, despite these advances, very little analogous work has been conducted to investigate global changes in gene expression that may contribute to the origin and perpetuation of bacterial epidemics. Similarly, there is a general lack of studies linking genome variation, transcriptional changes, and altered virulence in epidemic forms. The primary goal of this investigation was to study how genome variation linked with changes in transcriptome and altered virulence might contribute to the origin and perpetuation of bacterial epidemics, using the ongoing *S. pyogenes emm*89 epidemic as a convenient model system. We used comparative pathogenomics to dissect the precise molecular genetic events that have mediated the evolutionary origin and diversification of the epidemic *emm*89 strains. Unexpectedly, we found that a high frequency of HGT events has shaped the *emm*89 population genetic structure to a far greater extent than vertically inherited SNPs and short insertions and deletions (indels). Three main mechanisms that mediate HGT in bacteria have been described: conjugation, transduction, and transformation. Although *S. pyogenes* is not considered to be naturally competent, analysis of MLST data found *S. pyogenes* to have a level of recombination comparable to that of *Streptococcus pneumoniae*, a species that is naturally competent (47). The mechanism mediating the relatively high level of recombination detected in GAS is not known, but, given the prevalence of phage in GAS genomes, generalized transduction may play an important role. Global transcriptome sequencing (RNAseq) analysis was conducted on genetically representative preepidemic and epidemic *emm*89 strains to determine the extent to which the genomic changes causing altered gene expression may have contributed to the epidemic. We found that HGT is extensive in the *emm*89 population and has contributed disproportionately to the diversification of virulence factors and their expression. Nonsynonymous SNPs in major regulatory genes and other modest genetic changes have also led to transcriptome remodeling intimately linked with the origination and perpetuation of the epidemic. The results have significant implications for understanding epidemic bacterial disease and for translational research efforts designed to control or limit the detrimental effect of infectious agents. The overall strategy used as described here is of general utility and pertinence to the investigation of other pathogens.

**RESULTS AND DISCUSSION**

Population genetic structure and contribution of horizontal gene transfer (HGT). We studied 1,200 *emm*89 *S. pyogenes* strains, virtually all (*n* = 1,198) cultured from patients with invasive infections that occurred between 1995 and 2014 (Fig. 1; see also Table S1 in the supplemental material). The great majority of strains (*n* = 1,180) were collected as part of comprehensive population-based studies conducted in the United States, Finland, and Iceland. The genomes of all 1,200 strains were sequenced to a mean 60-fold depth of coverage (range, 13-fold to 440-fold) using an Illumina paired-end strategy, and polymorphisms were identified. Inference of genetic relationships using core chromosomal SNPs revealed that these *emm*89 strains have a major population of 1,193 strains and a minor population of 7 substantially diver-
89 strains are the only (see Table S2 in the supplemental material) showed that the minor population branches in the phylogenetic network (Fig. 3). Although we found S. pyogenes data available for 37 -type. Comparison of the emm S. pyogenes dent level of genetic diversity for strains of a single that we recently reported (35).

The epidemiological information available for the 1,200 strains MGAS27061 of clade 3) were closed and annotated (see Fig. S1).

The emm population genomic data revealed an unprecedented level of genetic diversity for strains of a single S. pyogenes emm-type. Comparison of the emm98 genome sequences with data available for 37 S. pyogenes genomes of 18 other emm-types (see Table S2 in the supplemental material) showed that the emm98 strains are the only emm-type to have two deeply rooted branches in the phylogenetic network (Fig. 3). Although we found evidence of recombination within the emm98 population, the random distribution of SNPs and the lack of sequence identity of the 7 minor population emm98 outlier strains with sequences of another GAS emm-type or MLST argue that these strains have not arisen through emm-type switching.

We identified extensive genomic diversity between and within the three primary emm98 clades. The mean genetic distance (MGD) among the 1,193 strains of the 3 clades was 610 SNPs in the core genome (Fig. 2B). In striking contrast, among 3,615 emm1 strains collected in 8 countries on two continents over 45 years (i.e., a collection 3 times larger, from a broader geographical region, and a period 2.5 times longer than those used for analysis of the emm98 sample), the MGD was only 106 core SNPs (32).

There was a nonrandom distribution of SNPs throughout the emm98 core genomes. Multiple regions had elevated SNP density, indicating HGT and core genomes with a mosaic evolutionary history (Fig. 4; see also Fig. S1 in the supplemental material). Gubbins (geneologies unbiased by recombination in nucleotide sequences) statistical analysis of SNP distribution (48) identified 2,316 regions of putative HGT with a mean size of 3,695 bp (range, 4 bp to 71,774 bp) at 526 loci around the genome. Because HGT can distort inferences of genetic relationships and evolutionary history, the phylogeny of the strains was reassessed using sequence filters to exclude regions of recombination (Fig. 2C).

This analysis greatly reduced the MGD (i.e., average pairwise core SNPs) among the 1,193 strains by 78% (from 610 to 134), a level similar to that found in 3,615 emm1 strains (32). The MGD from clade 1 to clade 2 and the MGD from clade 2 to clade 3 were reduced by 87% and 76%, respectively. The MGD between strains within each of the clades also was substantially reduced. The MGD strain-to-strain within clade 1 went from 226 to 100 (−56%), within clade 2 from 83 to 63 (−24%), and within clade 3 from 244 to 45 (−82%). Importantly, however, after exclusion of SNPs present in chromosomal segments associated with HGT events, 3 primary clades still remained among the 1,193 strains.

Outgroup rooting with the genome of emm1 reference strain SF370 showed that the evolutionary pathway leading to the current emm98 epidemic lineage had clades branching in the sequence of clade 1 followed by clade 2 and then clade 3 (see Fig. S2 in the supplemental material). Clade 1 and clade 2 strains differed by 8 regions of HGT encompassing 171.1 kb or 10% of the genome, and clade 2 and clade 3 strains differed by 6 regions of HGT encompassing 15.3 kb or 0.9% of the genome (Fig. 4 and Table 1). Seven of the 8 HGT regions differentiating clade 1 and clade 2 are most similar in sequence to regions in emm2 reference genome MGAS10270 (see Fig. S3). Of special note, 33 isolates in clade 3 differed from the 725 other clade 3 strains by one additional HGT. These strains, designated subclade 3D (SC-3D) (Fig. 2B), first occurred in the Finland sample in 2009 and have disproportionately increased in prevalence in recent years as a cause of bloodstream infections in that country (see Table S1 and Fig. S4).

HGT events are responsible for the bulk of the core sequence differences between the clades. The transferred sequences encompass multiple genes encoding many known secreted and cell surface-associated virulence factors, including the pilus/T-antigen adhesin, fibronectin-binding protein FbaB, the toxin pair NGA and SLO, internalin InLA, C5a peptidase ScpA, antiphagocytic M-like proteins Enn and Mrp, virulence regulators Mga and Ihk-Irr, immunogenic secreted protein Ips1, and the HasABC capsule synthesis enzymes (49). These HGT events have had important consequences. For example, clade 1 strains differ from clade 2 and 3 strains in pilus/T-antigen, and the clade 3 strains cannot produce capsule due to loss of the hasABC genes. Of note, different pilus types have been shown to vary in cell adherence and tissue tropism, and differences in the levels of production of capsule and SPN and SLO cytotoxins can alter virulence (35, 49, 50).

Consistent with SPN and SLO playing a key role in S. pyogenes strain emergence and enhanced fitness, each of the three clades has a distinct nga-ifs-slo region resulting from two independent HGT events. In addition, SC-3D strains differ from the other clade 3 strains due to HGT of a region encoding the SpyA and SpeJ virulence factors (49, 51–53). Inasmuch as these multiple HGT events involve regions encoding virulence factors, it is reasonable to hy-
1.2. Genetic relationships among S. pyogenes primary clades. Genetic relationships were inferred by the neighbor-joining method based on concatenated core chromosomal SNP data using SplitsTree. (A) Genetic relationships based on 28,425 SNPs identified among the members of the entire population of 1,200 S. pyogenes strains (see Tables S1 and S2 in the supplemental material). This analysis identified 64 different profiles of MGE content (>53,000 genes) of 30 GAS genomes of 18 emm-types (see Tables S1 and S2 in the supplemental material). This analysis identified 64 different profiles of MGE content (Fig. S5). ICEs were infrequent in the strain sample. The three most prevalent MGE content profiles, or phage genotypes (PGs), accounted for 72% of the strains (Fig. 5). These three phage genotypes (PG01, PG02, and PG03) correspond to the phage content of the reference genomes for each of the three primary clades (Fig. S1 in the supplemental material). With the exception of PG02 (defined as lack of prophages), most phage genotypes were confined to a single clade. The most prevalent (43%) PG in clade 1 was PG03 (phage 11027.1 encoding SpeC and Spd1 and phage 11027.2 encoding Sdn). Also prevalent were PG05 (13%) and PG06 (11%) strains, potentially derived from PG03 strains by phage loss. Most clade 2 strains are distant outliers relative to the combined gene content (>53,000 genes) of 30 GAS genomes of 18 emm-types (see Tables S1 and S2 in the supplemental material). This analysis identified 64 different profiles of MGE content (>53,000 genes) of 30 GAS genomes of 18 emm-types (see Tables S1 and S2 in the supplemental material). This analysis identified 64 different profiles of MGE content (Fig. S5). ICEs were infrequent in the strain sample. The three most prevalent MGE content profiles, or phage genotypes (PGs), accounted for 72% of the strains (Fig. 5). These three phage genotypes (PG01, PG02, and PG03) correspond to the phage content of the reference genomes for each of the three primary clades (Fig. S1 in the supplemental material). With the exception of PG02 (defined as lack of prophages), most phage genotypes were confined to a single clade. The most prevalent (43%) PG in clade 1 was PG03 (phage 11027.1 encoding SpeC and Spd1 and phage 11027.2 encoding Sdn). Also prevalent were PG05 (13%) and PG06 (11%) strains, potentially derived from PG03 strains by phage loss. Most clade 2 strains are distant outliers relative to the combined gene content (>53,000 genes) of 30 GAS genomes of 18 emm-types (see Tables S1 and S2 in the supplemental material). This analysis identified 64 different profiles of MGE content (Fig. S5). ICEs were infrequent in the strain sample. The three most prevalent MGE content profiles, or phage genotypes (PGs), accounted for 72% of the strains (Fig. 5). These three phage genotypes (PG01, PG02, and PG03) correspond to the phage content of the reference genomes for each of the three primary clades (Fig. S1 in the supplemental material). With the exception of PG02 (defined as lack of prophages), most phage genotypes were confined to a single clade. The most prevalent (43%) PG in clade 1 was PG03 (phage 11027.1 encoding SpeC and Spd1 and phage 11027.2 encoding Sdn). Also prevalent were PG05 (13%) and PG06 (11%) strains, potentially derived from PG03 strains by phage loss. Most clade 2 strains are
PG02 (72%), having no phages. The abundance of PG02 strains representing 20% of the entire emm89 cohort is unusual in that, prior to our investigation, nearly all S. pyogenes genomes had been found to be polylysogenic (35). Most clade 3 strains are PG01 (62%), having phage 27061.1 encoding SpeC and Spd1, followed next in prevalence by PG02 (22%). Of note, although phages 11027.1 and 27061.1 are integrated at the same genomic locus and encode the same two secreted virulence factors, they are different phages (see Fig. S5). PG01 (presence of 27061.1) first occurred in our strain samples in 2003, a time that corresponds to the emergence of the global transcriptome. As a consequence of the greater technical difficulty and expense involved, global transcriptional variation has been far less studied than genomic variation in bacterial pathogens. Moreover, since the data corresponding to the groups of samples studied here were population based and comprehensive and included temporal-spatial information, we had the additional opportunity to assess the potential effect of transcriptome remodeling on strain emergence and dissemination. We used RNAseq to compare transcript variations at two growth points among genetically representative strains of clades 1, 2, and 3 (Fig. 6). These strains have the allelic variant of the major virulence regulators covRS, mga, and ropB that is most common to the clades they represent. These regulators lack known function-altering polymorphisms that influence S. pyogenes gene expression and virulence (49, 56–60). The number of genes differentially expressed in stationary-phase growth exceeded the number in exponential-phase growth by approximately 3-fold in all of the clade-to-clade comparisons (Fig. 7A). A general finding was that the greater the genetic distance between strains was, the greater the number of differentially expressed genes was recorded between MGAS11027 (clade 1) and MGAS23530 (clade 2), consistent with strains in these clades being separated by the greatest MGD (Fig. 2). Genes altered in transcript level by 1.5-fold or greater accounted for 14% and 36% of the genome at the exponential and stationary growth phases, respectively, in comparisons of MGAS11027 (clade 1) and MGAS23530 (clade 2) (see Table S3, S. pyogenes emm-types) provided a unique opportunity to test the hypothesis that these HGT events have enhanced the virulence of the epidemic emm89 strains by remodeling of the global transcriptome. As a consequence of the greater

FIG 3  Genetic relationships between strains of various Emm/M protein serotypes. Genetic relationships were inferred among 49 GAS strains of 20 M types based on 75,184 concatenated core chromosomal SNPs by the neighbor network method. The analysis is based on 42 closed genomes and 7 whole-genome-sequenced strains. The MGD interserotype consists of 16,340 SNPs. The MGD interserotype consists of 16,340 SNPs. emm89 strains are the only emm-type strains with two distinct lineages (L1 and L2) in the interserotype network. The MGD of 14,247 SNPs between the L1 and L2 genomes is greater than the MGD of 11,548 SNPs among the serotype M5, M6, M18, and M23 genomes. Of note, the emm89 L1 to L2 MGD is greater than the emm89 L1 to M53 genome MGD of 14,194 SNPs.
section 1, in the supplemental material). Although genes \((n = 182)\) located within the eight distinct regions of HGT differentiating clade 1 and clade 2 comprise only 11% of the gene content, they accounted for 24% of the differentially expressed genes at exponential growth, a highly nonrandom occurrence \((P < 0.0001)\). Importantly, genes encoding many key virulence factors had significantly different transcript levels, including the fibronectin/collagen/T-antigen (FCT) region pilin genes, nga-ifs-slo, speG, ideS, ska, sclA, fba, enn, emm, mrp, and mga \((49)\). Collectively, these findings demonstrate that the genome segments that had been horizontally acquired and retained on the evolutionary pathway leading from clade 1 to clade 2 strains have contributed disproportionately to remodeling the global transcriptome, including many virulence genes, and argue that they are likely not selectively neutral.

The genomic changes accruing in the molecular evolution of clade 2 to clade 3 are of considerable interest because they are associated with the emergence, dissemination, and recent rapid increase in the frequency of \(emm89\) invasive infections recorded in many countries \((46, 61–65)\). In contrast to the 11% of the gene content reshaped by HGT in the transition from clade 1 to clade 2, a more modest 1% was reshaped in the transition from clade 2 to clade 3. Despite this modest 1% change we found that in comparing the transcriptomes of clade 2 strain MGAS23530 with clade 3 strain MGAS26844, 4% and 11% of the genes were differentially expressed at the exponential and stationary growth phases, respectively (Fig. 7A; see also Table S3, section 2, in the supplemental material). Genes located within regions of HGT were significantly overrepresented among the differentially expressed genes in exponential growth \((P < 0.0001)\). Included among the 28 genes with significantly increased expression in exponential growth were the critical virulence genes nga-ifs-slo (Fig. 7B). To confirm that increased expression of \(nga\) and \(slo\) is a trait broadly common to clade 3 strains, we assessed the expression of these genes by quantitative PCR (qPCR) in 11 strains selected to represent the range of genetic and geographic diversity present in the \(emm89\) major population (Fig. 7E and F). The subclades represented by these 11 isolates encompass 1,120 (94%) of the 1,193 strains of the major \(emm89\) population. All 5 of the clade 3 strains had significantly greater \(nga\) and \(slo\) expression than all 6 of the clade 1 and 2 strains \((P < 0.001)\). This is consistent with previous findings for Nga NADase activity assessed for 27 strains of the cohort \((50)\). Importantly, significantly increased transcription of \(nga-ifs-slo\) was asso-

**TABLE 1** HGT recombination blocks separating GAS \(emm89\)/M89 clades

| Block | Clades | Start\(a\) | Stop\(a\) | Length (kb) | SNPs | Genes | M-like | % ID
|-------|--------|-----------|-----------|-------------|------|-------|--------|------|
| RB1\(^b\) | C1-C2 | 92,389 | 164,162 | 71,774 | 411 | 72 | M2 | 88.74\(^b\)
| RB2 | C1-C2 | 295,481 | 297,574 | 2,094 | 9 | 2 | M2 | 100.00
| RB3\(^c\) | C1-C2 | 773,487 | 780,634 | 7,148 | 55 | 8 | M2 | 99.55
| RB4\(^c\) | C1-C2 | 794,417 | 800,659 | 6,243 | 28 | 8 | M2 | 99.55
| RB5 | C1-C2 | 921,261 | 960,297 | 39,037 | 100 | 41 | M2 | 99.68
| RB6 | C1-C2 | 1,022,619 | 1,030,407 | 7,789 | 20 | 6 | M2 | 99.97
| RB7 | C1-C2 | 1,543,651 | 1,561,165 | 17,515 | 103 | 13 | M2 | 97.17
| RB8 | C1-C2 | 1,577,916 | 1,597,447 | 19,532 | 138 | 21 | M2 | 99.58
| RB9 | C2-C3 | 86,603 | 88,366 | 1,764 | 12 | 2 | M5/M23 | 99.38
| RB10 | C2-C3 | 145,163 | 155,569 | 10,407 | 59 | 11 | M1/M12 | 98.52
| RB11 | C2-C3 | 244,407 | 244,758 | 352 | 5 | 1 | M5 | 100.00
| RB12 | C2-C3 | 1,472,262 | 1,473,025 | 764 | 9 | 1 | M12 | 100.00
| RB13 | C2-C3 | 1,558,898 | 1,559,698 | 801 | 7 | 2 | M49 | 99.75
| RB14 | C2-C3 | 1,693,613 | 1,694,805 | 1,193 | 6 | 2 | M5/M6 | 100.00
| RB15 | C3-SC3D | 341,762 | 359,579 | 17,818 | 106 | 21 | M1 | 99.67

\(a\) The start and stop positions provided are relative to the MGAS23530 genome.

\(b\) The first 18.6 kb and last 39.6 kb are M2-like (>99% identity [ID]); however, the central 13.5 kb FCT pilus-encoding region is unlike that of any other sequenced GAS \(emm\)-type.

\(c\) RB3 and RB4 likely represent a single HGT event that encompasses the intervening streptin lanthiobiotic synthesis genes, thus resulting in a larger single recombination of 26,697 bp.
associated with the emergence and epidemic increase in S. pyogenes emm1 invasive infections (32, 34, 35).

Additional genetic changes that differentiate epidemic clade 3 strains from the most recent predecessor clade 2 strains are acquisition of phage 27061.1 encoding speC and spd1 and loss of the hasABC capsule synthesis genes. To explore the role these genetic changes have potentially played in contributing to the emergence of the epidemic clade 3 strains, we inspected transcript data for the speC and spd1 genes and hasABC virulence factor genes between the preepidemic (clade 1 and clade 2) and epidemic (clade 3) emm89 representative strains. Transcript levels of speC and spd1 were significantly greater for the preepidemic clade 1 MGAS11027 strain than for the epidemic clade 3 MGAS26844 strain at both phases of growth assessed (Fig. 7C). The finding of significantly lower levels of speC and spd1 transcripts in the genetically representative epidemic clade 3 strain further argues that presence of these virulence factors in the clade 3 lineage is unlikely to have conferred a fitness advantage relative to clade 1 strains and therefore is an unlikely mechanism for the emergence of the epidemic clone and displacement of the predecessor clade 1 and clade 2 strains (46). Similarly, although the epidemic clade 3 strains are incapable of producing the antiphagocytic hyaluronic acid (HA) capsule due to HGT-mediated loss of the hasABC genes, the transcript data indicate that this gene loss was likely not responsible for a significant decrease in capsule production between the clade 2 and 3 strains. We found that transcription of hasABC was very...
weak in clade 2 strain MGAS23530 under both growth conditions assessed (Fig. 7D), arguing that capsule production was already negligible before the HGT-mediated loss of the hasABC genes by the clade 3 lineage. Capsule production was strong only for clade 1 strain MGAS11027 at the exponential growth phase.

We next investigated the molecular basis for the differences in capsule production using all strains of clades 1 and 2. Sequence variation in the hasABC promoter has been reported to alter transcription and capsule production (66). Inspection of the genome sequence data, coupled with Sanger sequencing of the hasABC promoter for all clade 1 and 2 strains, identified two major variants (see Fig. S6A in the supplemental material). These promoter variants corresponded to strong clade 1 strain MGAS11027 and weak clade 2 strain MGAS23530 hasABC transcription. Whereas the two promoter variants are equally represented among clade 1 strains, the vast majority (88.5%) of clade 2 strains had the weak transcription variant (see Fig. S6B and C in the supplemental material). Expression of hasA correlated perfectly with the promoter variant (Fig. 7G), which is consistent with results of HA production assays previously reported for 27 strains of the cohort (50). Importantly, hasA transcript levels for strains with the weak promoter variant were not significantly different from those of the clade 3 strains that lack the hasABC genes. Thus, the evolution of clade 3 from a clade 2 progenitor strain likely involved a transition from very little capsule production to no capsule production. This again argues that loss of the hasABC genes is not an unlikely mechanism for the epidemic emergence and displacement of the predecessor lineages. Whereas some S. pyogenes outbreaks have been associated with strains having a hyperencapsulation phenotype (33, 67) we are unaware of a body of epidemiological data associating GAS epidemic outbreaks with a loss of capsule phenotype. To summarize, the global transcriptome data comparing the preepidemic and epidemic strains show that neither production of phage-encoded virulence factors SpeC and Spd1 nor lack of production of the antiphagocytic HA.
capsule is a characteristic unique to the emergent clade 3 strains relative to the predecessor clade 1 and 2 strains and therefore does not correspond to the epidemic increase in invasive infections.

The very recent emergence of SC-3D strains is temporally associated with a single HGT event in which SC-3D strains acquired an 18 kb sequence that includes 21 genes, including genes encoding the secreted virulence proteins SpyA, a C3-like ADP-ribosyltransferase, and SpeJ, a pyrogenic exotoxin superantigen (49, 51–53). On the basis of the nearly identical sequences, this 18 kb region likely was acquired from an epidemic emm1 clone donor. Differentially expressed genes accounted for 2% and 11% of the genome at the exponential and stationary growth phases, respectively, in comparisons of the transcriptomes of strain MGAS26844 (clade 3) and MGAS27520 (SC-3D) (Fig. 7A; see also Table S3, section 3, in the supplemental material). This was the lowest number of differentially expressed genes among the four genetically representative strains studied, consistent with SC-3D strains being a recently emerged closely genetically related subset of the epidemic clade 3 strains.

Further transcriptome remodeling and epidemic perpetuation. Discovery of significant alteration of transcriptomes caused by HGT events, and the role in emergence and dissemination of clade 3 organisms, led us to investigate the hypothesis that additional transcriptome remodeling contributed to perpetuating the emm89 epidemic. We tested this hypothesis by focusing on SC-3D strains, because these organisms disproportionately increased in frequency in Finland starting from 2013 (Fig. 2A; see also Fig. S4 and Table S1 in the supplemental material). Given the relatively modest number of genes differentially expressed between MGAS26844 (clade 3) and MGAS27520 (subclade 3D), we interrogated the genome data for candidate polymorphisms that may further alter the transcriptome and potentially influence pathogen behavior. Analysis of the genome sequences of the 33 SC-3D strains found unique single amino acid replacements in gene regulators CovR (S130N) and LiaS (K214R). These polymorphisms were prevalent among the SC-3D strains; 11 strains had the CovR (S130N) change, and 6 strains had the LiaS (K214R) change (see Fig. S4). In contrast, none of the other 1,185 emm89 or 3,615 emm1 strains (32) studied had these polymorphisms. The branching of the strains with these mutations in the inferred phylogeny and their absence in other S. pyogenes strains identify identity by descent rather than identity by independent mutation (i.e., commonality by evolutionary convergence).

Repeated recovery of clonal progeny with either the CovR (S130N) or LiaS (K214R) polymorphisms from invasive episodes has not been reported previously and thus was unexpected. Because relatively little is known about liaS in S. pyogenes, we elected to study the LiaS (K214R) polymorphism in more detail. Consistent with our altered-transcriptome hypothesis, RNAseq analysis showed that the transcriptome of strain MGAS27710 LiaS (K214R) differed from that of SC-3D LiaS wild-type strain MGAS27520, including significant changes in expression of several virulence genes (data not presented). However, as these two strains are not isogenic, the extent to which the altered transcriptome was due to the LiaS (K214R) polymorphism could not be assessed. To address this issue, we constructed a LiaS (K214R) isogenic mutant from parental strain MGAS27556 and conducted RNAseq analysis. We found that, compared to the wild-type parental strain, the LiaS (K214R) isogenic mutant had 127 and 70 differentially expressed genes in exponential-phase growth and stationary-phase growth, respectively (see Table S3, section 4, in the supplemental material). Virulence genes significantly increased in expression by the LiaS (K214R) isogenic mutant included all 9 genes of the streptolysin S biosynthesis operon (sagaABCDEFGHI) in exponential phase and speG encoding streptococcal pyrogenic exotoxin G in stationary phase.

The capacity of the CovR (S130N) and LiaS (K214R) naturally occurring mutant strains to repeatedly cause serious infections means that they can effectively spread between hosts and implies that they are not attenuated in the ability to survive in the upper respiratory tract, the more common S. pyogenes niche. Consistent with this idea, we found that the naturally occurring mutant strains had an enhanced ability to survive in human saliva in vivo relative to SC-3D wild-type strain MGAS27520 (Fig. 8H). These results contrast with data showing that strains with other covR or covS (covR/S) mutations have reduced survival in human saliva relative to wild-type strains (68).

Comparative strain virulence. The epidemiological, comparative genomic, and transcriptome data demonstrate that clade 1, 2, and 3 organisms are genotypically and phenotypically distinct and strongly suggest differences in virulence. To test this hypothesis, the three genetically distinct reference strains for each clade were compared in mouse and nonhuman primate models of necrotizing fasciitis (NF) (69–71). Epidemic clade 3 reference strain MGAS26844 was significantly more lethal and caused significantly greater tissue damage in the mouse NF infection model than the two preepidemic reference strains (Fig. 8A and B). Moreover, relative to clade 1 reference strain MGAS11027, epidemic clade 3 strain MGAS26844 caused significantly larger lesions with greater tissue damage in a nonhuman primate model of NF (Fig. 8C to E).

Concluding comment. We have used S. pyogenes as a model pathogen for studying the evolutionary genomics of epidemic disease and the molecular basis of bacterial pathogenesis. The organism is a strict human pathogen, causes abundant infections worldwide, and has a relatively small genome (~1.8 Mb). In addition to its propensity to cause epidemic waves, the availability of comprehensive, population-based strain collections from many countries, coupled with the fact that humans are its only natural host, means that the history of underlying events that generate genomic diversity is not obscured by molecular processes occurring in nonhuman hosts or environmental reservoirs. These factors afford considerable advantages in the use of S. pyogenes as a model system compared to many other pathogenic bacteria such as E. coli, S. enterica, and S. aureus.

The primary goal of our study was to determine if genomic changes linked with the origin and perpetuation of human epidemic disease have remodeled global gene expression and altered virulence in the model pathogen S. pyogenes. We were especially interested in determining the effect, if any, of horizontally acquired genome segments on global gene expression and virulence of the progeny strains. Despite the importance of bacterial pathogens in human and veterinary health, remarkably few studies have addressed how transcriptome remodeling contributes to the origin and perpetuation of epidemics. Zhou et al. (26) studied diversity in 149 genomes of S. enterica serovar Paratyphi A and used the resulting data to speculate that most recent increases in frequencies of bacterial diseases are due to environmental changes rather than to the novel evolution of pathogenic bacteria. In essence, it was suggested that many epidemics and pandemics of bacterial disease in humans did not involve recent evolution of particularly
virulent organisms but instead reflected chance environmental events. A similar conclusion was reached in studies of other pathogens, for example, *Yersinia pestis*, *S. enterica* serovar Agona, *Mycobacterium tuberculosis*, *Mycobacterium leprae*, and *Shigella sonnei* (17, 25). Although this may be the case for some pathogens, on the basis of the full-genome data from 4,815 strains, human patient information (33), analysis of isogenic mutant strains, RNAseq studies, and experimental animal infection, we arrive at a fundamentally different conclusion for *emm*98 and *emm1* S. *pyogenes*, organisms that have caused epidemics involving tens of millions of human infections in the last 30 years. In particular, our results unambiguously show that newly emerged clones causing epidemic disease are more virulent than previously circulating wild-type strains MGAS27520. No growth, <10 CFU/ml for a 1:10 dilution. IM, intramuscularly; NHP, nonhuman primate.

![FIG 8 Virulence assays. (A) Kaplan-Meier survival curve for mice (n = 25/strain) inoculated intramuscularly in the right hind limb with 2.5 × 10⁸ CFU. The genetically representative epidemic strain (MGAS26844) was significantly more lethal than the preepidemic strains throughout the period of observation. The index of the strains compared in panel A applies to panels A to G. (B) Histopathology scores for muscle tissue sections as determined by pathologists blind to the infecting strain. Data represent means (n = 5 assessments/strain) ± standard errors of the means (SEM). (C) Cynomolgus macaques were inoculated intramuscularly in the anterior thigh with 1.0 × 10⁷ CFU/kg of body mass. Shown at the same magnification are micrographs of muscle tissue sections from the site of inoculation. (D and E) Epidemic strain MGAS26844 caused significantly larger lesions (D) with greater tissue destruction (E) than preepidemic strain MGAS11027. (F and G) Although the bacterial burdens were similar at the site of inoculation (F), they were significantly greater for the epidemic strain than for the preepidemic strain at the distal margin (G) showing greater dissemination. *P* values for panels B, D, E, F, and G were determined with the Mann-Whitney test. (H) Viability of naturally occurring variant strains MGAS28980 CovR (S130N) and MGAS27552 LiaS (K214R) in human saliva persisted for 2 and 4 weeks longer, respectively, than that of wild-type strain MGAS27520. No growth, <10 CFU/milliliter. IM, intramuscularly; NHP, nonhuman primate.\]

![Image 325x626 to 437x710]

![Image 440x626 to 552x710]

![Image 325x626 to 437x710]
strains \((n = 1,178)\) were collected as part of comprehensive population-based public health surveillance of GAS invasive infections conducted in the United States, Finland, and Iceland between 1995 and 2014. The remaining \(emm89\) strains were recovered from invasive disease cases in Ontario, Canada, and from a pharyngitis case in Italy. A subset of this population has been previously studied, and preliminary genetic findings have been presented (35, 48).

**Genome sequencing.** Isolation of chromosomal DNA, generation of paired-end libraries, and multiplexed sequencing were accomplished as described previously (32, 35) using Illumina (San Diego, CA) instruments (HiSeq2500, MiSeq, and NextSeq). Whole-genome sequencing data for the 1,200 isolates studied were deposited in the NCBI Sequence Read Archive.

**Reference genome assembly, annotation, and polymorphism discovery.** The bioinformatics tools used for assembling and annotating the reference genomes and for identifying and analyzing polymorphisms in the population studied are described in Text S1 in the supplemental material. Complete genome sequences for genetically representative strains MGAS11027, MGAS23530, and MGAS27061 were deposited in the NCBI GenBank database. MGAS11027, MGAS23530, and MGAS27061 were deposited in the BEIR strain repository.

**Phylogenetic inference and population structure.** The bioinformatics tools used for sequence alignments, detection, and filtering of HGT polymorphisms, for clustering and phylogenetic inference, and for analysis of the population structure are described in Text S1 in the supplemental material.

**Gene content and mobile genetic element analysis.** The known GAS pangenome core and accessory gene content was determined based on 30 complete genomes of 18 different \(emm\)-types (see Table S2 in the supplemental material) as described in Text S1 in the supplemental material. Among the 53,336 coding sequences (CDSs) of the 30 genomes, PanOCT identified 3,338 orthologous clusters, culled by BLAST reciprocal-best-hit analysis to 2,835 on the basis of the criterion of no two clusters sharing >95% amino acid identity. A GAS pseudo-pangenome sequence of ~3 Mbp was generated by concatenating onto the \(emm89\) MGAS23530 reference genome all accessory gene content not already present in the genome, starting with \(emm89\) strains MGAS11027 and MGAS27061, and then the remaining 27 genomes by \(emm\)-type (i.e., \(emm1\), \(emm2\), \(emm3\), etc.). Based on mapping of the \(emm89\) reference genome sequencing reads to the GAS-30 pangenome, an RPKM (reads per kilobase of transcript per million reads mapped) value of >50 corresponded to gene presence. A phage was called present if a minimum of 80% of its gene content represented in the GAS-30 pangenome was found to be present. Reads not mapping to the GAS-30 pangenome were assembled de novo using SPAdes. Resultant contigs with greater than 100 nucleotides were queried against the NCBI nonredundant database using BLAST to determine their nature.

**Construction of isogenic mutant strains.** The construction of the \(liaS\) isogenic mutant strain was accomplished by allelic exchange as previously described (35). Briefly, MGAS27556 LiaS (K214R) was generated by introducing the \(liaS\) A641G SNP into wild-type strain MGAS27556, using DNA amplified from strain MGAS27552, a clinical isolate with a naturally occurring \(liaS\) A641G SNP (i.e., LiaS K214R substitution) as the template. Successful introduction of the desired SNP and the absence of spontaneous spurious mutations were confirmed in candidate isogenic mutants by whole-genome sequencing. Primers, plasmids, and restriction enzymes used in the construction are listed in Text S1 in the supplemental material.

**Transcriptome sequencing and expression analysis.** Whole-genome transcriptional analysis was conducted for strains genetically representative of the clades and subclades studied using RNAseq as previously described (35, 82) with minor modifications. Briefly, RNA was isolated from triplicate cultures grown in Todd-Hewitt broth supplemented with yeast extract (THY). Multiplexed libraries were subjected to single-end sequence analysis (50 bp) to a high depth value (~10 million reads/sample) with an Illumina HiSeq2500 instrument. RNAseq reads were mapped to the genome of the most closely related \(emm89\) reference strain (for example, clade 3 strains were mapped to the genome of reference strain MGAS27061). Use of multiple reference sequences was critical, as the use of a single common reference did not permit accurate quantitative read mapping to the divergent sequences in the regions of HGT. RNAseq data were normalized, and genes statistically differently expressed following Benjamini–Hochberg correction at a minimum 1.5-fold change in mean transcript level were identified using the bioinformatics tools provided in Text S1 in the supplemental material. RNAseq transcriptome data were deposited in the NCBI Gene Expression Omnibus database. Expression levels of the key virulence genes \(nrg\), \(slo\), and \(hasA\) were assessed by quantitative real-time PCR in triplicate for 11 strains genetically representative of the most abundant subclusters of the major population using primers, probes, and protocols previously described (50). The significance of strain-to-strain differences in expression was assessed by one-way analysis of variance (ANOVA).

**Experimental animal infections.** The virulence of serotype \(emm89\) reference strains MGAS11027, MGAS23530, and MGAS26844 was assessed using mouse and nonhuman primate models of necrotizing fasciitis (32, 69–71). These strains have a wild-type (i.e., the most commonly occurring) allele for all major transcription regulators, including \(covR/S\), \(ropB\), and \(nrg\). All animal experiments were approved by the Institutional Animal Care and Use Committee of Houston Methodist Research Institute.

**Accession numbers.** Whole-genome sequencing data for the 1,200 isolates studied were deposited in the NCBI Sequence Read Archive under accession number SRP059971. Complete genome sequences for genetically representative strains MGAS11027, MGAS23530, and MGAS27061 were deposited in the NCBI GenBank database under accession numbers CP013838, CP013839, and CP013840, respectively. MGAS11027, MGAS23530, and MGAS27061 were deposited in the BEIR strain repository under accession numbers NR-33707, NR-33706, and NR-50285, respectively. RNAseq transcriptome data were deposited in the NCBI Gene Expression Omnibus database under accession number GSE76816.

**SUPPLEMENTAL MATERIAL**

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppldoi:10.1128/mBio.00403-16/-/DCSupplemental.

- Text S1, DOCX file, 0.1 MB.
- Figure S1, TIF file, 3 MB.
- Figure S2, PDF file, 0.2 MB.
- Figure S3, PDF file, 1 MB.
- Figure S4, PDF file, 0.2 MB.
- Figure S5, PDF file, 0.5 MB.
- Figure S6, PDF file, 0.4 MB.
- Table S1, DOCX file, 0.4 MB.
- Table S2, DOCX file, 0.1 MB.
- Table S3, DOCX file, 0.4 MB.

**ACKNOWLEDGMENTS**

We thank FiRe (the Finnish Study Group on Antimicrobial Resistance) Chris A. Van Beneden, Bernard Beall, and the Active Bacterial Core surveillance (ABCs) of the Emerging Infections Programs network of the Centers for Disease Control and Prevention (CDC); Kathryn Stockbauer and Helen Chifotides for critical comment and editorial assistance; and Hanne-Leena Hyyrylainen, Kai Puhakainen, and Francesca Latronico for microbiological and epidemiological assistance.

This project was supported in part by the Fondren Foundation, Houston Methodist Hospital; the Academy of Finland (grant 255636); and the European Society of Clinical Microbiology and Infectious Diseases Training Fellowship 2011 and the Federation of European Societies of Microbiology Research Fellowship 2011-1 (awarded to M.C.D.L.).
REFERENCES


Beres et al.