Molecular Epidemiology, Ecology, and Evolution of Group A Streptococci

Debra E. Bessen  
*New York Medical College*

Pierre R. Smeesters

Bernard W. Beall

Follow this and additional works at: https://touroscholar.touro.edu/nymc_fac_pubs

Part of the Medicine and Health Sciences Commons

**Recommended Citation**


This Article is brought to you for free and open access by the Faculty at Touro Scholar. It has been accepted for inclusion in NYMC Faculty Publications by an authorized administrator of Touro Scholar. For more information, please contact touro.scholar@touro.edu.
Molecular Epidemiology, Ecology, and Evolution of Group A Streptococci

DEBRA E. BESSEN,1 PIERRE R. SMEESTERS,2 and BERNARD W. BEALL3

1Department of Microbiology and Immunology, New York Medical College, Valhalla, NY 10595; 2Department of Pediatrics, Queen Fabiola Children’s University Hospital, and Molecular Bacteriology Laboratory, Université Libre de Bruxelles, Brussels, 1020, Belgium; 3National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30333

ABSTRACT The clinico-epidemiological features of diseases caused by group A streptococci (GAS) is presented through the lens of the ecology, population genetics, and evolution of the organism. The serological targets of three typing schemes (M, T, SOF) are themselves GAS cell surface proteins that have a myriad of virulence functions and a diverse array of structural forms. Horizontal gene transfer expands the GAS antigenic cell surface repertoire by generating numerous combinations of M, T, and SOF antigens. However, horizontal gene transfer of the serotype determinant genes is not unconstrained, and therein lies a genetic organization that may signify adaptations to a narrow ecological niche, such as the primary tissue reservoirs of the human host. Adaptations may be further shaped by selection pressures such as herd immunity. Understanding the molecular evolution of GAS on multiple levels—short, intermediate, and long term—sheds insight on mechanisms of host-pathogen interactions, the emergence and spread of new clones, rational vaccine design, and public health interventions.

ECOLOGY: HOST TISSUES AND DISEASES

The group A streptococcus (GAS; Streptococcus pyogenes) is a free-living organism. Its ecological niche appears to be quite narrow, and its only known biological host of import is humans. There may be occasional or very rare natural infections in non-human primates and other mammals (1, 2).

The primary sites for GAS colonization of the human host involve two tissues: the oropharyngeal mucosal epithelium of the upper respiratory tract (URT) and the superficial layers of the epidermis. GAS have no known environmental reservoir of significance, and transmission is almost exclusively person to person, via either respiratory droplets or direct contact. Thus, the throat and skin of the human host are the primary habitats for GAS. It is at these two tissue sites that GAS undergo successful reproductive growth and transmission of progeny to new hosts (Fig. 1).

Streptococcal pharyngitis and impetigo are superficial, self-limiting infections that usually cause a mild illness that resolves within 2 weeks, even if left untreated. Both are primarily diseases of childhood. Once colonization is established at the URT, the organism can cause a symptomatic pharyngitis, with or without tonsillitis. Symptomatic infection caused by GAS is typically characterized by a copious purulent exudate. Oropharyngeal

Received: 1 April 2018, Accepted: 13 April 2018, Published: 6 September 2018
Editors: Vincent A. Fischetti, The Rockefeller University, New York, NY; Richard P. Novick, Skirball Institute for Molecular Medicine, NYU Medical Center, New York, NY; Joseph J. Ferretti, Department of Microbiology & Immunology, University of Oklahoma Health Science Center, Oklahoma City, OK; Daniel A. Portnoy, Department of Molecular and Cellular Microbiology, University of California, Berkeley, Berkeley, CA; Miriam Braunstein, Department of Microbiology and Immunology, University of North Carolina-Chapel Hill, Chapel Hill, NC, and Julian I. Rood, Infection and Immunity Program, Monash Biomedicine Discovery Institute, Monash University, Melbourne, Australia
Correspondence: Debra E. Bessen, debra_bessen@nymc.edu © 2018 American Society for Microbiology. All rights reserved.
infection can be cleared via a protective antibody that is produced by the host in response to infection. Alternatively, a “clinically inapparent” infection can arise wherein the infected individual lacks obvious clinical symptoms of illness, yet a specific immune response is mounted and directed to antigens of the infecting organism (3).

Another type of GAS-human interaction that can occur at the URT is asymptomatic carriage, where a vigorous host immune response to many/most GAS antigens is lacking (4). The carrier state can persist for weeks or months, and the organism (presumably) remains capable of transmission to a new host. However, relative to acute pharyngitis, the number of organisms present during carriage may be vastly reduced, and GAS are often in an altered dormant-like state signified by a relative lack of M protein expressed on the cell surface. Intracellular invasion of epithelial cells by GAS may contribute to persistent infection (5). Microcolony and biofilm formation, and the bacterial communities that arise, may also be a key part of the ecology of the GAS carrier state (6).

At the skin, GAS survive and replicate below the dry cornified layer, provoking a strong inflammatory response that gives rise to the purulent lesions of non-bullous impetigo, a form of pyoderma. Recovery of GAS from normal skin can precede impetigo by ~10 days (7, 8). However, it is not entirely clear whether there exists stable GAS colonization and/or a true long-term carrier state at the skin; uncertainty stems from the duration of GAS association with normal human skin in the absence of inflammation (i.e., disease), which may be transient (9). Breaks in the stratum corneum, which can be quite minor and barely noticeable, are a requirement for GAS to establish a firm foothold at this tissue site and cause disease. The scabies mite infests the epidermal layer just below the stratum corneum; scabies can cause intense

FIGURE 1 Transmission routes for GAS. From the throat, GAS can transmit via the respiratory route to a new host, where it either causes pharyngitis or persists in a quiescent carrier state. In a carrier state, the organism is presumed to be only weakly transmissible. Transmission from the throat to skin is relatively rare. From an impetiginous skin lesion, the organism can be transmitted by direct contact to the slightly damaged skin of a new host or to other damaged skin sites on the same host, causing multiple skin lesions. GAS can also be transmitted from a skin lesion to the throat of the same host, but it is widely assumed to enter a carrier state and is only weakly transmissible. From either the throat or skin, the organism can invade normally sterile deeper tissue, but this is rare relative to superficial infections.
itching and is a key risk factor in the development of GAS impetigo in many tropical communities (10–12). Also, it is not uncommon for *Staphylococcus aureus* to be a copathogen along with GAS in impetiginous skin lesions (13, 14). At least some strains of GAS that cause impetigo can subsequently spread from skin lesions to colonize the URT, peaking at 2 to 3 weeks post-skin infection (7, 8) (Fig. 1).

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 cold winter months, whereas impetigo thrives during warm, humid weather and is far less prevalent. In many tropical regions, such as the Northern Territory of Australia, GAS impetigo is far more common than pharyngeal infection among the indigenous population, and there are no discrete seasonal peaks in the incidence of disease (10). Climate and the level of hygiene strongly influence the prevalence of impetigo, whereas indoor crowding is a major risk factor for URT infection. These risk factors largely reflect the primary modes of transmission.

Worldwide there are an estimated >616 million cases of GAS pharyngitis per year (13), and >162 million cases of impetigo at any given time (16). In a study conducted ∼40 years ago in the United States, for children between the ages of 5 and 7 years, it was estimated that an average of half experience one GAS infection annually (17). During the peak season, asymptomatic carriage rates at the URT range from (very) roughly 15 to 20% or higher; this value depends on the exact cohort (18). Bacteria tend to attain high numbers and densities during purulent infection. Large bacterial counts yield large infective doses, which in turn can lead to the exposure of several susceptible hosts with the minimum infective dose, thereby augmenting transmission and spread.

A primary reason why GAS is a major public health problem is because of its ability to cause invasive (iGAS) disease and trigger nonsuppurative (i.e., nonpustule-forming) sequelae. The latter diseases are of an autoimmune (acute rheumatic fever [ARF]) or immune complex (acute glomerulonephritis) nature. Collectively, the severe GAS diseases are responsible for >500,000 deaths each year (15).

In iGAS disease, bacteria gain access to deep tissue that is normally sterile. Consistent with pharyngitis providing the major GAS reservoir in temperate regions, iGAS disease is also more prevalent in these regions during the winter months (19). Intracellular invasion or paracellular transport of GAS at the URT may facilitate its access to the bloodstream. iGAS infection can be particularly severe in cases of necrotizing fasciitis and toxic shock syndrome. Although iGAS disease is associated with elevated rates of morbidity and mortality, its frequency of occurrence is relatively low. Population-based surveys of iGAS disease over the past 2 decades estimate the annual incidence at ∼3 to 4 per 100,000 people per year in the United States and Europe, although more recently, the incidence of GAS disease in the United States has increased to ∼5 per 100,000 people per year (20); worldwide, there are an estimated 663,000 new iGAS cases per year (15). Thus, the overall incidence of iGAS disease is ∼1,000-fold lower than the incidence of new pharyngitis and impetigo infections. Invasion of normally sterile tissue is often a result of “bacteremia without focus.” When a lack of evidence for a superficial infection is coupled to GAS growth being largely restricted to deep tissue, and the patient is sequestered and nonambulatory, iGAS disease becomes a transmission dead end for the infecting bacterium.

From the ecological perspective of long-term evolution, GAS fitness is largely shaped by adaptations to its primary tissue reservoirs; acquisition of traits that enhance fitness in iGAS disease is probably coincidental. In North America, the composition of the iGAS strain population largely overlaps with the population of GAS strains recovered from superficial (pharyngitis) infections (21), although a few strains appear to have invasive indices at the more extreme high or low ends (22).

Cutaneous wound infections can be considered a mild form of iGAS disease, wherein the histopathology is distinct from nonbullous impetigo. Depending on the study, wound infections are occasionally classified under the broader umbrellas of “pyoderma” or “skin lesions.” Ecthyma is a severe form of impetigo with extensive dermis involvement. Impetigo largely afflicts younger children and is mostly prevalent in communities with a warm humid climate (tropics, summer), whereas in contrast, GAS wound infections occur in all age groups and in all geographic regions, including those where URT disease predominates.

Transmission of GAS leading to cutaneous wound infections can occur from one tissue to another site within an individual (e.g., URT → broken skin), through shared skin-penetrating devices (e.g., needles), and possibly by direct person-to-person contact via open wounds. Skin contact may be an important mode of transmission in a recent GAS outbreak among adults that include intravenous drug users (IVDUs) and homeless people in a temperate zone (23, 24).
ECOLOGY: TRANSMISSION
The superficial nature of pharyngitis and impetigo provides GAS with an easy exit from infected hosts and an easy entrance into new hosts. Transmission is strongly influenced by the GAS agent and human host, coupled with both microenvironmental and macroenvironmental factors. The dynamics of both the GAS and human host populations largely dictate the scope of GAS disease outbreaks and epidemics.

Macroenvironment Risk Factors
As discussed above (“Ecology: Host Tissues and Diseases”), the primary environmental risk factors for GAS pharyngitis and impetigo are cold winters and tropical conditions, respectively. Copathogens that may elevate the risk for GAS infection include the scabies mite and S. aureus for impetigo and, possibly, undefined infectious agents for pharyngitis.

GAS-Microenvironment Factors
The microenvironments at the throat and skin differ from one another in numerous ways. Also, each tissue undergoes fluctuations in local conditions as a GAS infection evolves. To initiate infection, the GAS organism needs to be reasonably well adapted to the local tissue microenvironment. Distinctions between classic throat and skin strains of GAS (see “Epidemiology: Global Trends” below) are probably rooted, at least in part, in host tissue-specific adaptations. GAS cell surface components implicated in transmission and/or mediating colonization in a new host include the hyaluronic acid capsule and pili, among numerous other extracellular proteins (25–29). Yet the recent emergence and successful spread of emm89 iGAS strains is coupled to a loss in capsule (30, 31) (see “Epidemiology: Building Evolutionary Models” below). Conceivably, GAS transmission factors are strain specific.

At the URT, GAS initially encounter saliva and mucus secretions that contain mucins, secretory IgA, digestive enzymes, and antimicrobial peptides. At the broken surface just below the cornified layer of the skin, GAS face the early stages of wound healing: an inflammatory response in the epidermis is characterized by infiltration of low numbers of polymorphonuclear leukocytes and small amounts of plasma containing Iggs and complement, plus antimicrobial peptides. As the infection evolves, the bacteria deploy virulence factors that are able to thwart host defenses, allowing GAS to undergo a net increase in growth, with a concomitant increase in infiltration by inflammatory cells leading to the accumulation of a purulent exudate (32).

Competition with normal flora is a key determinant of GAS colonization at the URT. In general terms, competition can be stiffest between highly related organisms. The closest genetic relative of GAS/S. pyogenes that is also a common human URT colonizer is Streptococcus dysgalactiae subspecies equisimilis (SDSE). SDSE usually have cell wall-associated group C or G group carbohydrate and are largely restricted to humans, with >60% genome sequence homology to S. pyogenes (33, 34). A few uncommon lineages of SDSE have group A (CHO-A) or group L carbohydrate (thus, the term “GAS” is largely equivalent to “S. pyogenes,” and in this report, the two terms are completely equivalent unless otherwise noted). Recently, a widely disseminated and longstanding iGAS lineage of SDSE was described that appears to have arisen from a recombination event involving CHO-A biosynthesis genes from S. pyogenes (35).

SDSE bacteria are largely commensals of the oropharynx but are capable of causing many of the diseases attributed to S. pyogenes (36, 37). Like its close cousin, SDSE expresses capsules composed of hyaluronic acid, cell surface proteins that include M protein and T-antigens, and exotoxins such as streptolysin O; however, SDSE lacks some critical S. pyogenes virulence factors, such as the secreted cysteine proteinase SpeB. Besides the potential sharing of human host receptors for adherence to the epithelium, relatively little is known about how these two species—GAS and SDSE—compete for nutrients and other critical resources (38). Signaling between GAS, SDSE, and other streptococcal species has been demonstrated (39), and GAS are susceptible to killing by bacteriocins produced by other streptococcal species (40). Much remains to be uncovered regarding the role of the microbiome in shaping the success of GAS in colonizing and/or causing infection at the URT.

GAS physiology is heavily influenced by the nature and concentration of carbon sources (e.g., glucose), oxygen, and its toxic byproducts, among numerous other factors present in different tissues under shifting conditions (41, 42). Changes in the local microenvironment, coupled with increased bacterial density and nutrient depletion as the organisms reproduce and yield new progeny, impact the transcriptional program of the GAS cell, which in turn directs activation or repression of numerous genes associated with pathogenesis, and presumably transmissibility as well (43).

Acquisition of GAS at the URT does not necessarily lead to an infection, and in the asymptomatic carrier state, the local microenvironment is substantially different from the purulent environs. In the carrier state,
virulence factor gene expression and the requirements for adaptation appear to be quite different (e.g., 44, 45).

Human (Host) Factors
In general terms, successful transmission of a pathogen depends on the availability of susceptible hosts, as well as the ratio of susceptible to resistant hosts in the context of the number of person contacts and potential exposures (46). Host genetic susceptibility (or resistance) to superficial GAS infections has not been characterized, and evidence for its existence is largely lacking.

Highly efficacious immunological protection against GAS infection is strain specific (i.e., M type-specific) (47), although a host immune response directed toward numerous other antigenic targets can confer at least partial protection (48, 49). Herd immunity is probably a critical host determinant for circumscribing the degree to which a given GAS strain can successfully transmit and spread within a human population. Thus, the possibility of future immunization with an M type-based vaccine is likely to reshape the composition of the GAS strain population (see “Epidemiology: Strategies for M Protein-Based Vaccines” below).

GAS (Agent) Factors
A key open question is whether GAS engage in a form of evolutionary bet-hedging via transmitting a bolus of organisms with a diverse array of transcriptional profiles, to increase the odds that at least one physiological form will colonize and thrive within an individual host. Alternatively, perhaps there exists a singular transcriptional profile that is deemed optimal for transmission within the wider human host population. Or perhaps a mixture of both strategies is used. Mutations in transcriptional regulatory genes and/or promoter regions may enhance the transmissibility of GAS (see “Evolution: Short Term” below).

GAS exhibit very high diversity in their cell surface proteins, in terms of both the presence or absence of a given protein, and in the sequence heterogeneity among products of the same locus. Several of the heterologous cell surface proteins may impart tissue-specific adaptations (reviewed in reference 50, 51). They also serve as key epidemiological markers (see next section). Genetic changes that lead to alterations in the antigenic composition of the GAS cell surface may aid the organism in escaping herd immunity.

Epidemiology: Molecular Markers
The use of serological typing to categorize the pathogens of the Streptococcus genus (i.e., cell wall group carbohydrate), as well as the strains within a pathogenic streptococcal species, has a long and deep history. Unlike pneumococci, GAS lack a structurally diverse and immunogenic polysaccharide capsule. Instead, GAS express several cell surface proteins which collectively impart a broad repertoire of antigenic diversity. Underlying the serological phenotype is an extensive history of horizontal gene transfer (HGT) and recombination among the cell surface protein genes, plus accumulated point mutations, that has been shaped (at least in part) by the strong diversifying selection pressures of the host immune response. The antigenic structure of GAS cell surface proteins, and the underlying genotype of both the cell surface protein and core housekeeping genes, is highly complex.

Numerous typing schemes have been used to aid in understanding the epidemiology and underlying organization of the S. pyogenes species. A summary of typing (and related) schemes, and their applications and limitations, is provided in Table 1. Each epidemiological molecular marker is discussed in detail below.

M Protein
During the 1920s, work began that aimed at understanding the basis for protective immunity to GAS infection. Antibodies raised to extractable surface antigens, called M proteins, led to opsonophagocytosis of the strain from which the M protein was derived (52). However, antibody directed to the M protein of one organism often failed to confer protection against many other GAS isolates tested. A serological typing scheme arose through the development of antibodies directed to M proteins of different isolates. More than 80 distinct M types were identified, and protective immunity to GAS infection is, in large part, M type specific.

Copurifying with the M type-specific material was M-associated protein (MAP), which has two major antigenic forms (53, 54). The MAP I antigen was largely associated with ARF isolates, and MAP II was highly prevalent among serum opacity factor (SOF)-producing organisms. The MAP antigens were subsequently identified as part of the conserved C repeat region (CRR) of M protein (55), which occupies the COOH-terminal half of the surface-exposed portion of M protein. The CRR has two major antigenic forms known as class I and class II (56). The structure of M protein is depicted in Fig. 2A.

emn types
A nucleotide sequence-based emm typing scheme that closely parallels the M serotype scheme was developed ~2 decades ago (57). emm typing is based on the extensive
### TABLE 1 Epidemiological typing methods for group A streptococci

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Description of method</th>
<th>Applications</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>M serotyping</td>
<td>Immunoprecipitation, using hyperimmune rabbit antiserum</td>
<td>No longer in wide use</td>
<td></td>
</tr>
<tr>
<td><em>emm</em> typing</td>
<td>Sequencing, yielding &gt;92% nt identity over the first 90 bases encoding the processed (mature) M protein of an <em>emm</em> type reference strain</td>
<td>Most commonly used typing method; highly reproducible; low cost; large database at CDC; can be used to infer <em>emm</em> clusters and <em>emm</em> pattern groups with reasonable accuracy</td>
<td>Based on oligonucleotide primers (via PCR); needs better integration with whole-genome sequencing projects</td>
</tr>
<tr>
<td><em>emm</em> subtyping</td>
<td>Sequencing of 180 nt corresponding to the 10 COOH-terminal codons of the signal sequence and first 50 codons of the mature M protein</td>
<td>Further stratifies <em>emm</em> types into unique alleles; performed in parallel with <em>emm</em> typing; large database and BLAST server are maintained at the CDC</td>
<td></td>
</tr>
<tr>
<td><em>emm</em> pattern grouping</td>
<td>PCR-based mapping, based on 3’ ends of <em>emm</em> and <em>emm</em>-like genes</td>
<td>Assignment of organisms of known <em>emm</em> types into groupings that correspond reasonably well to tissue site preferences for superficial infection at the throat and/or skin</td>
<td>Inferences based on <em>emm</em> type may occasionally be incorrect; ∼30% of <em>emm</em> types await <em>emm</em> pattern assignment; PCR-based mapping is laborious and limited by design of “universal” oligonucleotides</td>
</tr>
<tr>
<td><em>emm</em> cluster typing</td>
<td>Based on phylogeny and four bioinformatic criteria, related to surface-exposed portion of M proteins; corresponds to binding by several human serum proteins</td>
<td>Used to extend information on <em>emm</em> typing; can be easily inferred from <em>emm</em> type; binding motifs can be used to infer binding properties of M proteins</td>
<td>Inferences based on <em>emm</em> type may occasionally be incorrect; ∼29% of <em>emm</em> types await <em>emm</em> cluster assignment</td>
</tr>
<tr>
<td>SOF typing</td>
<td>Inhibition of the serum opacity reaction by hyperimmune rabbit serum</td>
<td>No longer in wide use</td>
<td></td>
</tr>
<tr>
<td><em>sof</em> typing</td>
<td>Sequencing of ~700 bp region near 5’ end of <em>sof</em></td>
<td>Data can be generated though sequencing of PCR products or whole-genome sequencing</td>
<td><em>sof</em> type assignment needs to undergo standardization; evidence for extensive recombination within the <em>sof</em> type region</td>
</tr>
<tr>
<td>T serotype</td>
<td>Bacterial cell agglutination, using hyperimmune rabbit antiserum</td>
<td>In limited use by specialized laboratories, to supplement <em>emm</em> typing</td>
<td>Inconsistent results; labor-intensive; hyperimmune serum is limited and difficult to standardize</td>
</tr>
<tr>
<td>FCT region typing</td>
<td>PCR-based mapping; whole genome sequence</td>
<td>Can be used as an initial step in further characterization of pilin genes</td>
<td>Assignment of the FCT region needs to undergo further definition and standardization</td>
</tr>
<tr>
<td>FCT gene typing</td>
<td>Sequenced-based</td>
<td>Under development for all pilin subunit and adhesin genes</td>
<td></td>
</tr>
<tr>
<td>MLST</td>
<td>Nucleotide sequence of seven (partial) housekeeping genes, from PCR products or whole-genome sequences</td>
<td>Used to define “clones” (i.e., STs) and clonal complexes (i.e., CCs); often coupled with <em>emm</em> type; alleles are assigned via simple web-based user queries and curation; database of alleles, STs, and isolates is maintained on PubMLST server (Oxford)</td>
<td>Provides only a small snapshot of the core genome; isolate database is skewed toward unique genotypes (i.e., not population-based)</td>
</tr>
<tr>
<td>Whole-genome sequencing</td>
<td>Complete or partial genome assembly</td>
<td>Phylogenetic relationships of closely related organisms via the core genome sequence; investigations of local outbreaks and widespread epidemics; routine surveillance; pipelines for analysis of typing and antibiotic resistance genes</td>
<td>Core genome definition lacks standardization for the diverse range of strains in the species; assembly problems in regions of extensive sequence diversity and/or recombination; bioinformatic analysis can be time-consuming</td>
</tr>
</tbody>
</table>

*nt, nucleotide.*
nucleotide sequence differences at the 5′ end of the emm gene, whereby a unique emm type is defined as having <92% nucleotide identity to any other emm type, over the sequence corresponding to the first 30 codons of the mature M protein (Fig. 2A). Virtually all contemporary epidemiological studies define GAS isolates according to their emm type. Currently, there are ~243 recognized emm types (58).

Identification of the emm gene is not always obvious, and the problem can be due to paralogous emm-like genes mapping immediately upstream (mrp) or downstream (enn) of emm in numerous GAS strains (see next section; Fig. 2B). Confusion about emm type assignments has arisen wherein enn is directly fused to the 5′ end region of emm as the result of DNA excision arising from recombination between the emm and enn genes.

FIGURE 2. M protein structure and emm pattern gene arrangements. (A) Key features of M protein are shown, including the type-specific determinants, cell wall-spanning domain, and C repeat region. (B) Chromosomal arrangement of emm and the flanking emm-like genes (mrp, enn) gives rise to five emm patterns, which form three main groupings (A-C, D, E). Transcription of emm and emm-like genes is positively regulated by Mga, which is encoded by mga, which lies immediately upstream of the emm region.
Among the ~243 emm types recognized to date, there are ~1,632 distinct allelic forms of the emm type-specific region, known as emm subtypes (58). The emm subtype alleles are based on sequence differences relative to the typing reference, within a 180-bp segment corresponding to the last 10 codons of the predicted signal peptide coding region and the first 50 codons corresponding to the NH2-terminus of the mature M protein molecule (59).

emm patterns
Many strains of GAS have emm-like genes (mrp, emm) positioned immediately upstream and/or downstream of emm on the chromosome (60–63). Near the 3’ end of each emm and emm-like gene lies a region that encodes one of four divergent cell wall-spanning domains that can be distinguished by oligonucleotides (60, 64). Using a PCR-based mapping approach, the content of emm and emm-like genes and their arrangement on the chromosome was determined for GAS isolates of many different emm types. The resulting emm region maps are referred to as “emm patterns”; three major emm pattern groupings are recognized (Fig. 2B). M proteins of emm pattern E tend to be class II, whereas the emm patterns A-C and D groupings largely overlap with the class I CRR (60).

To date, the emm pattern genotype has been ascertained for GAS isolates of 170 of the ~243 recognized emm types (Table 2). With only two exceptions uncovered thus far (emm54, emm218), multiple isolates sharing the same emm type also share the same emm pattern grouping. The majority of emm types are represented by emm pattern groups D and E, accounting for >75% of the emm types analyzed.

emm clades and clusters
An emm cluster typing system was recently developed, based on the entire surface-exposed portion of M protein and its capacity to bind six human serum proteins (65). Phylogenetic analysis of >1,000 emm genes corresponding to 174 emm types reveals 2 main clades (X and Y), and 16 well-supported clusters, accounting for 82% of the emm types analyzed. The relationships between emm clades, clusters, and patterns for 174 emm types are illustrated in Table 3. Nearly all pattern E emm types (98%) belong to clade X, whereas 92% of pattern A-C emm types are clade Y.

In contrast to patterns A-C and E emm types, the pattern D emm types are present in both clades X and Y (Table 3) (65). The highly specialized plasminogen-binding emm cluster D4 of clade Y represents the largest emm pattern D grouping (n = 30 emm types). Within clade X, emm clusters E5 and E6 constitute a second group of pattern D emm types (n = 14); these two clusters include emm types of other pattern groups as well. Most of the remaining pattern D emm types are interspersed throughout clade Y (n = 16). Overall, the emm cluster scheme, based on the phylogeny of the entire surface-exposed portion of M protein, is highly consistent with the emm pattern scheme that is based on the COOH-terminal peptidoglycan-spanning domain of emm and emm-like flanking genes.

The emm cluster typing system is easy to implement since the emm cluster can be inferred from the emm type (66–69). Importantly, the emm type → emm cluster scheme is also highly informative of the functional capacity of the M protein molecule, based on consensus binding motifs for numerous human host plasma proteins (65).

M protein functional domains
As summarized in recent reviews (63, 70), the surface-exposed non-type-specific portion of M proteins (i.e., central region) contains functional domains that bind human host proteins such as IgG (various subclasses), IgA, plasminogen, and fibrinogen. The M type-specific determinants at the NH2-terminus bind two complement regulators, C4b-binding protein (C4BP) and factor H-like 1.

Binding capacities strongly correlate with emm cluster (65). For example, binding of human IgG is mostly restricted to M proteins of emm clusters E1 through E6. Similarly, binding of C4BP is limited to M proteins of emm clusters E1, E3, E4, and E6. IgA binding is evident only in emm clusters E1, E4, and E6 products, whereas fibrinogen binding is restricted to M proteins of clade Y, and plasminogen binding is restricted emm cluster D4 (51, 65).

Although much less is known about the M-like proteins Mrp and Enn (Fig. 2B), binding properties shared with M protein have been described (reviewed in reference 63), and in some instances, they may use similar
binding motifs. Additional investigation of the structure and functional binding properties of the M-like proteins, and structure of the \textit{mrp} and \textit{enn} genes, may eventually lead to a higher-resolution typing scheme that includes the entirety of the \textit{emm} chromosomal region.

**Serum Opacity Factor**

SOF promotes the opacification of serum, wherein the serum opacity reaction can be neutralized by hyperimmune antiserum produced in animals (71). Not all GAS strains produce SOF, but SOF typing was often used as an adjunct for strains that posed challenges for M typing. Of note, SOF producers have a strong association with strains that harbor the MAP II antigen (72). In addition, SOF-producing strains tend to exhibit diminished beta-hemolysis (73).

More recent studies show SOF to be a proteinaceous virulence factor that is either bound to the GAS cell surface or secreted. SOF has an enzymatic domain that interacts with high-density lipoprotein, leading to the opacification of serum, plus a fibronectin-binding domain (74–77). The genes encoding M protein (\textit{emm}) and SOF (\textit{sof}) map \(~\sim\~\)15 kb apart on the GAS chromosome; as with \textit{emm} and \textit{emm}-like genes, transcription of \textit{sof} is positively regulated by the Mga regulon (78, 79). \textit{sof} types have been defined based on \(~\sim\~\)600 to 750 bp positioned near the 5’ end of the \textit{sof} gene (80). However, \textit{sof} typing is far less developed than \textit{emm} typing, and attempts to generate a well-supported phylogeny for the \textit{sof} type-specific determinants have largely failed due to extensive intragenic recombination (81).

The relationships between the presence or absence of SOF activity, or the \textit{sof} gene, and the \textit{emm} pattern and \textit{emm} clade as inferred from the \textit{emm} type, are summarized in Table 4. Based on the strong associations between the MAP II material, class II M proteins, and \textit{emm} pattern E (60), it is not surprising that there are strong correlations between \textit{emm} pattern E and the presence of the \textit{sof} gene and/or a positive serum opacification reaction (50, 80, 82, 83). Exceptions include several \textit{sof}-positive strains within \textit{emm} cluster E6 that are assigned to the \textit{emm} pattern D grouping via PCR-based mapping (\textit{emm59}, \textit{emm81}, \textit{emm85}; \textit{emm12} isolates (pattern A–C) have an inactive \textit{sof} gene. Despite the close genome map positions of \textit{sof} and \textit{emm}, a single \textit{emm} type can be found in association with more than one \textit{sof} type and vice versa (80), indicative of HGT of \textit{emm} or \textit{sof} to new genetic backgrounds and intraspecies recombination within the greater \textit{emm} region.

**T-Antigens**

T-typing is based on agglutination of trypsin-treated GAS cells with hyperimmune rabbit serum (84, 85). The T-typing antisera is produced by immunization with...
TABLE 4 Relationships between SOF/sof and emm patterns and clades*

<table>
<thead>
<tr>
<th>emm type-based grouping</th>
<th>No. of emm types</th>
<th>% SOF activity positive</th>
<th>% sof gene positive</th>
<th>SOF/sof positive emm types</th>
<th>SOF/sof negative emm types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern A–C</td>
<td>22</td>
<td>0</td>
<td>5.6</td>
<td>12b</td>
<td>All others</td>
</tr>
<tr>
<td>Pattern D</td>
<td>42</td>
<td>7.3</td>
<td>7.1</td>
<td>59, 81, 85</td>
<td>All others</td>
</tr>
<tr>
<td>Pattern E</td>
<td>48</td>
<td>97.9</td>
<td>97.9</td>
<td>All others</td>
<td>15</td>
</tr>
<tr>
<td>Clade X</td>
<td>57</td>
<td>87.7</td>
<td>87.7</td>
<td>All others</td>
<td>15, 34, 42, 51, 65, 67, 99</td>
</tr>
<tr>
<td>Clade Y</td>
<td>56</td>
<td>0</td>
<td>1.8</td>
<td>12</td>
<td>All others</td>
</tr>
</tbody>
</table>

*Based on analysis of 113 emm types, as reported in reference 83.

The SOF product of sof12 is typically inactive.

trypsin-treated GAS cells, followed by adsorption of antisera to undigested GAS cells of different T types; ~22 T types have been serologically defined. The T6 antigen gene (lee6) was the first one cloned (86) and was subsequently mapped to a region of the GAS genome that had been previously characterized as having genes encoding microbial surface components recognizing adhesive matrix molecules that bind fibronectin and collagen (87–89). The acronym FCT (fibronectin-binding, collagen-binding, T-antigen) is widely used to describe this genome region (87). SDSE isolates also express similar T-antigens (90), as do group B streptococci (91, 92).

T-antigens are equivalent to pili. Within the FCT region lie several genes that contribute to the production of pili whose assembly is mediated by sortases (48) (reviewed in references 93–95). Seven FCT region forms are depicted in Fig. 3 (82, 92); the FCT-3 and FCT-4 regions are associated with the vast majority of emm types (82). For the most prevalent FCT region forms (i.e., FCT-1, 3, 4, and 5) found among GAS strains of different emm types, the pilin backbone (BP) proteins and many of the ancillary pilin (AP) proteins that act as adhesins display extensive sequence heterogeneity (87, 92, 96).

The relationships between T-typing sera and recombinant pilin proteins has been examined (48, 92, 97). In perhaps the most thorough study to date, using 15 T-typing sera, agglutination of trypsin-treated GAS cells corresponds most closely to reactivity with both the BP and AP1 (presumptive adhesin) subunits by immunoblot (92). For a T3/13/B3264 strain (Alab49; FCT-3), the recombinant AP1 subunit (i.e., Cpa) is strongly reactive by immunoblot with each of the anti-T3, anti-T13, and anti-B3264 sera, whereas the BP (i.e., FctA) is bound by anti-T3 sera only (97). However, there is very little amino acid sequence homology between Cpa and FctA, as is also the case for the other adhesin–backbone pilin subunit pairs (82, 98).

That some T-typing sera display immunoreactivity to multiple recombinant pilin proteins seems likely due to incomplete adsorption of the antisera and/or the presence of antigenic epitopes shared among pilin subunits belonging to the variety of GAS strains used for immunization. Possible unmasking of immunological determinants via the trypsin pretreatment used in the cell agglutination assay may further complicate efforts to draw direct connections between T-serotypes and specific pilin protein variant forms. Adding even further complexity to the T-typing scheme are the numerous GAS isolates that agglutinate with multiple T-typing sera, such as T3/13/B3264 or T8/25/Imp19; many other T-serotype combinations (>60 in total) are also observed (83). Overall, the T-typing sera displaying the least cross-reactivity with multiple pilin forms are T2, T4, and T6 (92); these appear to be best represented by the FCT-6, FCT-5, and FCT-1 regions, respectively (Fig. 3), although within these FCT regions lie at least some extensive pilin gene diversity (see, e.g., reference 99 for FCT-1).

T-serotype and M-serotype and/or emm-type data for GAS isolates (n > 20,000) from two large collections (the CDC and the University of Minnesota reference laboratory), recovered in the United States from about the 1980s to the mid-2000s (83), was collated and is summarized in Table 5. There are no strict sampling criteria for this assemblage; it seems reasonable to assume that it largely consists of a combination of isolates causing iGAS disease and strains with high prevalence. In this skewed, non-population-based sample, the dominant T types among emm pattern A-C strains are T1, T11 and/or T12, and T3, T13, and/or B3264, reflecting the highly prevalent M/emm types M1, M12, and M3, respectively.

T-non-typeable (TNT) GAS probably represent isolates that fail to produce pili; they account for ~9% of total isolates but >20% of emm pattern D isolates (Table 5). The relatively large numbers of TNT isolates among pattern D strains may reflect the disproportional presence of nra (82); Nra functions as a repressor or activator of pilus gene expression in a strain-dependent manner (88, 100). Nearly 60% of emm pattern D isolates are T3, T13, and/or B3264 (Table 5), which
appears to be largely represented by the FCT-3 region form (82, 87, 92, 97). The most frequent M/emm types among pattern E isolates of this collection are M4 and M28 (83), which largely reflects the dominance of the T4 and/or T28 serotype grouping (Table 5). Several correlations between individual loci within the FCT region and emm pattern and/or emm type have been noted (50, 82, 92, 101).

Within an emm cluster group, there can be a very wide array of T types (based on Table 5; detailed data not shown) (83). For example, the emm cluster E3 isolates (13 emm types) are found in association with T5, T27, and/or T44 (28%); TNT (16%); T8, T25, and/or Imp19 (11%); T14 (11%); T9 (9%); T4 and/or T28 (9%); T3, T13, and/or B3264 (6%); and T11 and/or T12 (4%), among others. In contrast, cluster D4 isolates (21 emm types, all encoding plasminogen-binding M proteins) are far more homogeneous, with 72% associated with T3, T13, and/or B3264 (another 16% are TNT). Although emm genes within a cluster are more closely related to each another than to emm of other clusters, the extent to which surface pili contribute to the overall antigenic cell surface repertoire appears to vary widely by cluster.

For GAS sharing the same emm type, individual isolates often differ in their T-serotype (80, 83). For example, while a substantial majority of M/emm type 4 isolates from the Table 5 study are T4, there are other M/emm4 isolates that have T types T4/28, T8/25/Imp19, T3/13, and TNT (83) (detailed data not shown). M/emm4 isolates typically have sof4, but there are exceptions here as well (80). The unique combinations of M/emm, SOF/sof, and T types is indicative of HGT of the epidemiological marker genes among GAS.
Unlike M type-specific antibodies, there is no known human host protective effect mediated via T type-specific serum; this may be a consequence of immunoreactivity directed toward trypsin-treated target antigens that are not encountered during natural infection. However, antiserum raised to recombinant pilin subunits or intact pili can be protective in experimental assays (48, 102).

Shortcomings of the T-typing cell agglutination assay include a general lack of good resolution, the time-consuming process, and the requirement for standardized rabbit serum. Because T-typing sera display high levels of immunological cross-reactivity to different recombinant pilin polypeptides, and many GAS isolates agglutinate with multiple T-typing sera, it can be difficult to predict the FCT region genotype form based solely on the T-agglutination pattern.

At present, there is no single widely recognized genotyping scheme for the FCT region form and pilin genes (35, 82, 92, 96), but the need for one seems quite compelling given the importance of the FCT region in strain diversity as well as virulence.

### Multilocus Sequence Typing and Core Housekeeping Genes

The first study to examine the population structure of GAS utilized multilocus enzyme electrophoresis (103). Isolates recovered from cases of severe iGAS disease correspond to several distinct lineages as defined by electrophoretic type. The multilocus enzyme electrophoresis approach was replaced by the more standardized multilocus sequence typing (MLST). For GAS, MLST is based on the nucleotide sequences of seven (partial) core housekeeping genes (104). MLST data are maintained at a user-interactive website (www.pubmlst.org/spyogenes) and currently lists 966 distinct sequence types (STs) (105); numerous investigators from throughout the world have generously contributed to this rich data set. Using goeBURST (106), 133 clonal complexes (CC) as defined by single locus variants (SLVs) have been detected. Figure 4 provides a chromosomal map of the relative positions of the core housekeeping genes and the emm and FCT regions.

#### TABLE 5 Relationships between emm pattern and T types for >20,000 GAS isolates

<table>
<thead>
<tr>
<th>T type</th>
<th>% Total</th>
<th>% Pattern A–C</th>
<th>% Pattern D</th>
<th>% Pattern E</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>18.73</td>
<td>34.64</td>
<td>0.93</td>
<td>0.14</td>
</tr>
<tr>
<td>T2</td>
<td>3.30</td>
<td>0.00</td>
<td>0.00</td>
<td>7.85</td>
</tr>
<tr>
<td>T6</td>
<td>6.19</td>
<td>11.09</td>
<td>1.97</td>
<td>0.34</td>
</tr>
<tr>
<td>T9</td>
<td>0.82</td>
<td>0.00</td>
<td>1.39</td>
<td>1.81</td>
</tr>
<tr>
<td>T14</td>
<td>0.88</td>
<td>0.11</td>
<td>4.29</td>
<td>1.55</td>
</tr>
<tr>
<td>T15</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>T18</td>
<td>0.59</td>
<td>1.10</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>T22</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>T23</td>
<td>0.07</td>
<td>0.12</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>T3, 13, and/or B3264</td>
<td>16.10</td>
<td>17.14</td>
<td>58.63</td>
<td>10.75</td>
</tr>
<tr>
<td>T4 and/or 28</td>
<td>15.88</td>
<td>0.15</td>
<td>3.59</td>
<td>37.24</td>
</tr>
<tr>
<td>T5, 27, and/or 44</td>
<td>4.36</td>
<td>4.90</td>
<td>0.70</td>
<td>4.03</td>
</tr>
<tr>
<td>T8, 25, and/or Imp19</td>
<td>6.60</td>
<td>1.13</td>
<td>2.90</td>
<td>13.82</td>
</tr>
<tr>
<td>T11 and/or 12</td>
<td>18.17</td>
<td>20.80</td>
<td>4.52</td>
<td>16.17</td>
</tr>
<tr>
<td>Nontypeable</td>
<td>8.29</td>
<td>8.83</td>
<td>21.09</td>
<td>6.25</td>
</tr>
</tbody>
</table>

- Based on analysis of T-serotyping data (83) according to emm pattern groups or clades inferred from M or emm type (65, 126, 127), for 21,231 isolates, whereby 271 isolates had >1 grouping and were counted more than once. M or emm types ranged from types 1 through 124.
A given pattern A-C emm type has a high tendency to be associated with a single ST or CC (104, 107, 108). However, the relationships between emm type and ST are not strict, and emm types of pattern groups D and E are often found in association with distant STs (see "Lateral Movement of emm" below). These data parallel findings on the complex relationships between M/emm types, sof types, and T types (80, 83) and, taken together, provide supporting evidence for the HGT of emm from one genetic background to another.

Whole-Genome Sequences

The first completely assembled GAS genome was that of an emm1 isolate (109). In the years that followed, numerous complete GAS genomes were reported for many of the most prevalent emm types found in resource-rich nations (updated listings are presented in recent reviews [50, 110]). Via next-generation sequencing technologies, many thousands of genomes of GAS isolates have undergone nucleotide sequence determination (e.g., 24, 30, 111–115). In nearly all studies, core genome sequences are compared for isolates sharing the same emm type, and numerous key insights on genetic changes that underlie enhanced transmission and/or virulence have been uncovered. The core genome data can be used to build evolutionary models of outbreaks and epidemics (see "Epidemiology: Building Evolutionary Models" and "Evolution: Short-Term" below).

A core genome MLST has not yet been developed for GAS. Although it is a reasonable concept for some bacterial species, its utility for GAS would probably be rather limited due to the ultrafine resolution this approach would provide for GAS, coupled with the extensive genetic recombination that characterizes this species. Core genome-based phylogenies of GAS of different emm types and/or CCs tend to yield lineages with deep branches, indicative of large genetic distances (50, 110, 111, 116).

The use of genome sequence data for surveillance of GAS has begun (35, 117–120). The CDC recently implemented a whole-genome sequence-based typing scheme for population-based iGAS strain surveillance conducted through the Active Bacterial Core program (35). For large-scale strain surveillance, the ability to obtain detailed information on typing and resistance features within a single output, from a single raw data input source, has largely precluded the need for multiple typing protocols and antibiotic susceptibility testing. The GAS typing pipeline automatically restricts an emm type to the sequence linked to emm consensus primer 1 (121) to improve accurate identification of emm and assigns an emm subtype (i.e., allele) as well. Partial sequences of BP genes and sof are also captured. In addition to emm subtype, the emm-like genes and MLST alleles and STs are identified, as is gacI, which is specific to CHO-A biosynthesis. Among numerous other genes, antimicrobial resistance determinants are also assessed.

A key feature of genome sequence-based active iGAS surveillance is the capacity to detect temporally and geographically related GAS clusters. For GAS isolates that are nearly identical at the single nucleotide polymorphism (SNP) level, the circumstances of where and when the isolates were recovered can be examined to help establish whether there exist disease clusters that may be well suited for an intervention (e.g., health care facilities, vulnerable patient populations such as the homeless, and IVDUs). Thus, next-generation sequencing coupled with an appropriate bioinformatics pipeline can be used as an autodetection tool to help guide public health interventions.

EPIDEMIOLOGY: GLOBAL TRENDS

Infection at the Epithelium of the Throat and Skin

Past studies show that GAS strains bearing certain M types have a strong tendency to cause pharyngitis but not impetigo. Similarly, there are M types often recovered from impetigo skin lesions but rarely from cases of tonsillopharyngitis (8, 122, 123). This observation gave rise to the important concept of classical throat and skin strains that are largely distinct from one another in their epidemiological and disease associations.

There are strong correlations between emm pattern groups A-C and D and recovery from cases of pharyngitis versus impetigo, respectively (50, 124, 125). Since the emm type is highly predictive of the emm pattern group for nearly all emm types examined (126, 127), reasonable inferences can be made about emm pattern grouping based on knowledge of emm types. Such inferences were made on an expanded scale for >5,000 GAS isolates derived from cases of pharyngitis or impetigo in 29 population-based surveillance studies from throughout the world (Table 6). In 23 studies on pharyngitis, emm pattern A-C and D isolates constituted a mean average of 41% and 6%, respectively. In six studies of impetigo, the mean average for emm pattern A-C and D isolates was reversed, at 9% and 47%. For pharyngitis versus impetigo groups, the difference between the fraction of A-C and D isolates in each surveillance study is highly significant (Table 6). emm pattern E isolates account for almost equal fractions...
of throat and skin infections; the mean average percentage of \textit{emm} pattern E isolates shows no significant difference from pattern A-C isolates at the throat or pattern D isolates at the skin.

Based on the strong clinico-epidemiologic associations, the \textit{emm} pattern A-C strains are designated “throat specialists,” whereas the pattern D strains are “skin specialists.” As a group, \textit{emm} pattern E strains are considered “generalists.” Importantly, the \textit{emm} pattern serves as a useful genotypic marker for the preferred tissue site of infection by GAS (32, 50, 124, 125, 128).

The \textit{emm} cluster typing scheme can also be applied to the 29 population-based surveillance studies on GAS pharyngitis and impetigo (Table 7). Both pharyngitis and impetigo isolates are almost evenly split between clades X and Y, reflected by fractional ratios of percent pharyngitis to percent impetigo approaching 1. However, within clade Y, the distribution of pharyngitis and impetigo isolates are highly skewed among the clusters, with clusters A-C1-5 disproportionately represented by pharyngitis isolates, and clusters D1-5 disproportionately represented by impetigo isolates.

### TABLE 6

\textit{emm} pattern assignments for GAS recovered in 29 population-based surveys for pharyngitis or impetigo

<table>
<thead>
<tr>
<th>Disease association</th>
<th>No. of studies</th>
<th>Mean average % pattern A–C</th>
<th>Mean average % pattern D</th>
<th>Mean average % pattern E</th>
<th>Paired t-test(^b) (% pattern A–C versus D)</th>
<th>Paired t-test(^b) (% pattern A–C versus E)</th>
<th>Paired t-test(^c) (% pattern D versus E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pharyngitis</td>
<td>23</td>
<td>40.5 ± 19.8</td>
<td>5.8 ± 12.0</td>
<td>53.7 ± 16.0</td>
<td>&lt;0.001</td>
<td>N.S.</td>
<td>n/a</td>
</tr>
<tr>
<td>Impetigo</td>
<td>6</td>
<td>9.3 ± 6.7</td>
<td>47.4 ± 8.1</td>
<td>43.4 ± 7.0</td>
<td>&lt;0.001</td>
<td>n/a</td>
<td>N.S.</td>
</tr>
</tbody>
</table>

\(^b\)Paired t-tests are two-tailed.

\(^c\)N.S., nonsignificant; n/a, not applicable.

Several predicted binding motifs within M protein strongly correlate with \textit{emm} clusters (65). Based on the 29 population-based surveys, the fractional ratio of pharyngitis to impetigo isolates is highly skewed for the fibrinogen-binding clusters of clade Y (high ratio value, with pharyngitis dominant) and the plasminogen-binding cluster D4 (low ratio value, with impetigo dominant) (51). The differential binding properties for human host proteins, by M proteins in accordance with disease, are potential determinants of tissue tropism. How modulation of the coagulation-fibrinolytic pathways (129) may translate into GAS tissue tropisms is an active line of inquiry.

### Invasive Disease

As detailed above (“Ecology: Host Tissues and Diseases”), iGAS disease is extremely rare relative to GAS infection at the two primary ecological niches. When iGAS disease does occur, it is largely a transmission dead end, with the possible exception of direct contact transmission via open cutaneous wounds among underserved adult populations. Thus, iGAS infections have minimal

### TABLE 7

\textit{emm} cluster and clade assignments for GAS recovered in 29 population-based surveys for pharyngitis or impetigo

<table>
<thead>
<tr>
<th>\textit{emm} clade</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{emm} cluster</td>
</tr>
<tr>
<td>No. of isolates</td>
</tr>
<tr>
<td>No. of pharyngitis isolates</td>
</tr>
<tr>
<td>No. of impetigo isolates</td>
</tr>
<tr>
<td>% of total pharyngitis</td>
</tr>
<tr>
<td>% of total impetigo</td>
</tr>
<tr>
<td>Fractional ratio of pharyngitis to impetigo</td>
</tr>
<tr>
<td>Fractional ratio of impetigo to pharyngitis</td>
</tr>
</tbody>
</table>

\(^a\)Based on references 50, 51, and 135 and Table 6.

\(^b\)Calculated as the ratio of the percentage of total pharyngitis isolates to the percentage of the total impetigo isolates.

\(^c\)n.a., not applicable; n.d., not determined.

Key ratios are highlighted (bold)
impact on shaping the long-term evolution of GAS. The occurrence of iGAS disease is strongly affected by host risk factors and to a lesser extent by the genotype of the infecting GAS strain.

In the United States and Europe, most iGAS disease depends on transmission of GAS via a respiratory route (Fig. 1). Whole-genome sequencing of hundreds of GAS isolates has provided important details on genetic changes that enhance transmission and/or invasiveness (see “Evolution: Short-Term” below). The invasive index value provides insight on whether a GAS strain has an inherent ability to better gain access to deep tissue relative to its overall prevalence in the host population (22, 130). The emm types associated with iGAS disease have been intensively studied and are well described (e.g., 131–133).

Nonbulbous impetigo caused by GAS usually arises in young children in tropical or subtropical climates. GAS can also cause slightly deeper cutaneous infections that are considered “pyoderma,” and while they are not considered part of the iGAS disease spectrum, these cutaneous infections are distinct from typical impetigo in terms of patient demographics (i.e., adult, temperate regions) and host risk factors (e.g., IVDU, homeless, military) (23, 24, 134–136). Of interest is the strong association between certain pattern D emm types (specifically, emm59 and emm81) and recent infections and outbreaks of deeper cutaneous infections. The emm59 and emm81 types are notable for their grouping in cluster E6 and tend to harbor sof as well; conceivably, these are “super-specialist” strains containing a hybrid mix of virulence factors that shift the modes of transmission and GAS disease course. Most other GAS-causing skin infection outbreaks among IVDUs are sof-positive emm pattern E strains (137, 138).

Acute Rheumatic Fever
Acute rheumatic fever (ARF) follows an inadequately treated GAS infection of the URT in a genetically predisposed host by ∼3 to 6 weeks and can lead to rheumatic heart disease (RHD) via autoimmune attack of the heart valves (139–143). With >500,000 annual deaths, half attributed to RHD, GAS ranks high among all infectious causes of morbidity and mortality. Recent estimates of the prevalence of RHD are ≥15 to 30 million cases worldwide (13, 144–146); RHD is the leading cause of mortality from acquired heart disease in individuals <50 years old (147) and of all cardiovascular disease in children in resource-poor nations (148). Today, ARF and RHD are largely diseases of poverty, with an occasional resurgence in disease elsewhere (149).

Early epidemiological studies of ARF outbreaks in institutional settings showed that only a subset of GAS strains trigger ARF (11, 55, 140, 150–156). Historically, “rheumatogenic” GAS strains were referred to by their M protein serotypes, with M3, M5, M6, and M18 being highly prevalent. The concept of “nonrheumatogenic” strains is exemplified by the RHD patients of Irvington House during the preantibiotic era, who experienced a GAS outbreak (M type 4) in the absence of recurrent ARF attacks (155). Of ten M types often associated with ARF (M types 1, 3, 5, 6, 14, 18, 19, 24, 27, and 29 [157]), nine are emm pattern A-C (one is pattern E); of five relatively common M types that are rarely associated with ARF (M types 2, 4, 12, 22, and 28 [157]), four are emm pattern E (one is emm pattern A-C) (126, 127).

ARF-associated isolates recovered in Hawaii differ from the classic rheumatogenic emm types (158). Atypical emm types are also found in tropical Australia, where the prevalence of RHD is among the highest in the world (159). Over time, it has become much less clear which emm types are best assigned the “rheumatogenic” label, even though it is widely accepted that not all GAS strains trigger ARF in a genetically susceptible host (150, 155).

The potential role of GAS skin infection in triggering ARF and RHD in certain settings (e.g., noninstitutional, tropical) has been an ongoing debate since the 1960s (160). For example, the indigenous population of Australia suffers from a high incidence of ARF and RHD, yet GAS is far more frequently recovered from impetigo lesions than from the throat. Moreover, the strains recovered from the URT in this host population often belong to the emm pattern D skin specialist group or pattern E generalists. The epidemiological cooccurrence of GAS skin infections and RHD supports the idea that impetigo may contribute to this chronic disease. More extensive studies are needed to better understand skin infections and their potential role in ARF and RHD. The possibility that the closely related species SDSE might also trigger episodes of ARF or worsen RHD deserves deeper consideration as well (161, 162).

Epidemiology: Diversity within Host Communities
Within a human host community over a narrow time frame, numerous GAS strains are in circulation. For the 29 population-based collections of GAS recovered from cases of pharyngitis and impetigo, the number of emm types and the Simpsons diversity index was calculated.
The number of emm types for the impetigo collections averaged nearly twice that of the pharyngitis collections ($t < 0.01$, unpaired $t$-test, 2-tailed). The Simpson’s diversity index ($D$), which incorporates the number of recovered organisms into the calculation, also shows highly significant differences between the impetigo and pharyngitis collections, wherein the impetigo group has a higher mean average $D$ value (a $D$ value of 1 reflects maximal diversity). An analysis that includes all GAS, irrespective of clinical association, also shows fewer numbers of emm types accounting for a large proportion of isolates in resource-rich countries, where URT infections tend to predominate (131). The large numbers of circulating emm types is likely the consequence of expansive GAS strain migration, even into geographically remote communities (108). The biological basis for the difference in diversity among GAS in host populations experiencing pharyngitis versus impetigo may be tied to the mechanisms for transmission: respiratory droplets versus direct contact. It may be that the respiratory route is more highly efficient, such that a single infected individual can infect many more new hosts (i.e., higher basic reproductive rate), and this in turn leads to dominance by fewer strains, resulting in a lower $D$ value.

### EPIDEMIOLOGY: EVOLUTION OF MARKERS (emm)

Although emm type diversity within a host community is relatively low for the emm pattern A-C throat specialists compared to the pattern D skin specialists (Table 8), there is relatively high diversity among emm pattern A-C organisms in terms of the number of emm subtypes (i.e., alleles based on the 5' end of the emm gene) that evolved worldwide (Table 9). The average mean number of emm subtypes identified for pattern A-C strains is 22.3 subtypes per emm type, compared to only 4.0 and 6.9 for pattern D and E strains, respectively, based on current data available in the CDC online database (58), which receives numerous deposits from investigators throughout the world.

### Table 8 emm type diversity within communities for the 29 population-based surveys for pharyngitis and impetigo

<table>
<thead>
<tr>
<th>Disease association</th>
<th>No. of studies</th>
<th>Mean average no. of emm types</th>
<th>Mean average $D$ index$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pharyngitis</td>
<td>23</td>
<td>17.4 ± 7.8</td>
<td>0.875 ± 0.065</td>
</tr>
<tr>
<td>Impetigo</td>
<td>6</td>
<td>33.8 ± 8.1</td>
<td>0.959 ± 0.012</td>
</tr>
</tbody>
</table>

Unpaired $t$-test, 2-tailed $t = 0.00499 < 0.001$

$^a$Based on references 50, 51, and 135 and Tables 6 and 7.

$^b$Simpson’s diversity index ($D$) (248) is based on emm types.

The >5-fold difference in the number of emm subtypes for throat specialists versus skin specialists may be partly due to sampling bias that skews in favor of GAS invasive and pharyngitis isolates in resource-rich nations, such as the highly prevalent emm1 type (pattern A-C). In addition, the tandem sequence repeats that overlap the emm subtype-determining region of emm5 and emm6 genes (pattern A-C) are a mutational hotspot due to DNA slipped-strand mispairing (163). However, biological factors such as host immune selection promoting immune escape (164) may also play a role.

Nucleotide sequence alignment of emm subtype alleles for each emm type-specific region, derived from >500 global GAS isolates ($n = 105$ emm types), reveals average ratios of nonsynonymous substitutions per nonsynonymous site (Ka) and synonymous substitutions per synonymous site (Ks) of 4.9, 1.5, and 1.3 for emm types of the emm pattern A-C, D, and E groups, respectively (107). Thus, diversifying (positive) selection acting on the emm type-specific sequence appears to be strongest for emm pattern A-C strains.

A likely source of diversifying selection acting on the emm type region is the protective host immune response, which can drive immune escape. Conceivably, the immunological pressures at the oropharynx (pattern A-C throat specialists) are more intense than those at the skin (pattern D skin specialists), although an underlying molecular mechanism is not presently obvious. Perhaps herd immunity builds up more quickly with a respiratory mode of transmission compared to direct skin contact; as herd immunity builds, the number of susceptible hosts dwindles, leading to more frequent occurrences of immune escape. In further support of a role for herd immunity in shaping the emm type content of GAS populations, newly acquired URT infections in older children are skewed toward less common emm types, consistent with exposure and protective immunity to the more prevalent emm types in early childhood (165).

The collection of M protein types assigned to a single emm cluster have high sequence similarity relative to M proteins of other clusters (65). Indeed, an M type-specific region sharing the same cluster as other M types
often elicits cross-protective antibodies (see “Epidemiology: Strategies for M Protein-Based Vaccines” below). Ultimately, this cross-protective effect may aid in the design of highly efficacious vaccines that have broad coverage but contain only a limited selection of type-specific epitopes. Whereas 90% of pattern D and E emm types belong to 1 of the 16 emm clusters, nearly half of pattern A-C emm types are standalone and do not cluster with any other emm type, highlighting once again a distinct dynamic for the evolution of many pattern A-C emm types.

Another possible factor at play in shaping the evolution of emm type-specific sequences is functional constraints, such as those imposed by binding C4BP via the type-specific region of numerous M proteins (166–168). Whole GAS organisms that bind C4BP tend to harbor M types associated with emm pattern groups D and E (167). Thus, there may be limits to diversification within M proteins that bind C4BP for the GAS strains that utilize this virulence strategy to enhance survival. Also, potentially related to selection pressures acting on GAS is the finding that engineered mutants with emm gene deletions have a (near) complete loss in the antiphagocytic effect for pattern A-C strains, whereas the loss is only partial for pattern D and E strains (169).

Overall, the throat specialists are characterized by fewer distinct emm types within a human host community but high diversification worldwide within the type-specific region (i.e., more emm subtypes). The skin specialists have about twice the number of emm types within a community, but diversification within the type-specific region is ∼5-fold lower based on the number of emm subtypes or Ka/Ks ratios. The emm pattern E generalists are most similar to the skin specialists in these values.

The “emm type” and “emm cluster” definitions are categorical, and the noted differences between throat and skin specialists may partly reflect the placement of the categorical “cutoff” value along a continuum of sequence divergence. Factors that shape the GAS popula-

<table>
<thead>
<tr>
<th>emm grouping category</th>
<th>Name of group</th>
<th>No. of emm types represented</th>
<th>No. of emm subtypes represented</th>
<th>Average no. of emm subtypes per emm type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern</td>
<td>A-C</td>
<td>35</td>
<td>781</td>
<td>22.31</td>
</tr>
<tr>
<td>Pattern</td>
<td>D</td>
<td>64</td>
<td>253</td>
<td>3.95</td>
</tr>
<tr>
<td>Pattern</td>
<td>E</td>
<td>67</td>
<td>460</td>
<td>6.85</td>
</tr>
<tr>
<td>Clade</td>
<td>X</td>
<td>85</td>
<td>518</td>
<td>6.09</td>
</tr>
<tr>
<td>Clade</td>
<td>Y</td>
<td>89</td>
<td>1,014</td>
<td>11.39</td>
</tr>
<tr>
<td>Cluster</td>
<td>E1</td>
<td>5</td>
<td>40</td>
<td>8.0</td>
</tr>
<tr>
<td>Cluster</td>
<td>E2</td>
<td>15</td>
<td>84</td>
<td>5.6</td>
</tr>
<tr>
<td>Cluster</td>
<td>E3</td>
<td>19</td>
<td>119</td>
<td>6.3</td>
</tr>
<tr>
<td>cluster</td>
<td>E4</td>
<td>17</td>
<td>150</td>
<td>8.3</td>
</tr>
<tr>
<td>Cluster</td>
<td>E5</td>
<td>7</td>
<td>15</td>
<td>2.1</td>
</tr>
<tr>
<td>Cluster</td>
<td>E6</td>
<td>18</td>
<td>98</td>
<td>5.4</td>
</tr>
<tr>
<td>Cluster</td>
<td>Clade X singletons</td>
<td>4</td>
<td>12</td>
<td>3.0</td>
</tr>
<tr>
<td>Cluster</td>
<td>A-C1</td>
<td>2</td>
<td>7</td>
<td>3.5</td>
</tr>
<tr>
<td>Cluster</td>
<td>A-C2</td>
<td>2</td>
<td>26</td>
<td>13.0</td>
</tr>
<tr>
<td>Cluster</td>
<td>A-C3</td>
<td>5</td>
<td>118</td>
<td>23.6</td>
</tr>
<tr>
<td>Cluster</td>
<td>A-C4</td>
<td>5</td>
<td>111</td>
<td>22.2</td>
</tr>
<tr>
<td>Cluster</td>
<td>A-C5</td>
<td>3</td>
<td>140</td>
<td>46.7</td>
</tr>
<tr>
<td>Cluster</td>
<td>D1</td>
<td>3</td>
<td>13</td>
<td>4.3</td>
</tr>
<tr>
<td>Cluster</td>
<td>D2</td>
<td>5</td>
<td>22</td>
<td>4.4</td>
</tr>
<tr>
<td>Cluster</td>
<td>D3</td>
<td>2</td>
<td>4</td>
<td>2.0</td>
</tr>
<tr>
<td>Cluster</td>
<td>D4</td>
<td>31</td>
<td>144</td>
<td>4.6</td>
</tr>
<tr>
<td>Cluster</td>
<td>D5</td>
<td>3</td>
<td>6</td>
<td>2.0</td>
</tr>
<tr>
<td>Cluster</td>
<td>Clade Y singletons</td>
<td>28</td>
<td>423</td>
<td>15.1</td>
</tr>
<tr>
<td>Clade Y singletons</td>
<td>Pattern A-C</td>
<td>16</td>
<td>385</td>
<td>24.1</td>
</tr>
<tr>
<td>Clade Y singletons</td>
<td>Pattern D</td>
<td>8</td>
<td>16</td>
<td>2.0</td>
</tr>
<tr>
<td>Clade X singletons</td>
<td>Pattern E</td>
<td>2</td>
<td>6</td>
<td>3.0</td>
</tr>
</tbody>
</table>

*Based on the CDC online database (58).
tion structure, as defined by \textit{emn} type, likely include the route of transmission, environmental risk factors, the host immune response, tissue-specific host factors, binding functions of M protein, and intrinsic GAS factors that influence mechanisms for genetic change.

**EPIDEMIOLOGY: STRATEGIES FOR M PROTEIN-BASED VACCINES**

Numerous vaccine strategies for protection against GAS infections were recently reviewed (170–172). Briefly, vaccine strategies can be divided into M protein and non-M protein target antigens, with the latter including streptolysin O, C5a peptidase, CHO-A, pili, and the enzyme arginine deiminase (102, 173–178). However, only the vaccine candidates that incorporate M protein as the target antigen have attained human clinical trial status (179).

M protein vaccines based on the conserved antigens of the CRR (Fig. 2A) have undergone extensive investigation (49, 180–185). Since all M proteins have a CRR, a CRR-based vaccine has the theoretical advantage of providing broad coverage against all GAS strains; several formulations are currently, or will soon be, in clinical trials. Potential drawbacks of a CRR-based vaccine include immune escape. The CRR region of M protein, which includes MAP I/II and the class I/II antigenic epitopes, is not 100% conserved. For example, among 175 \textit{emn} types from a worldwide strain collection, there are 22 alleles of the CRR-based J8 vaccine candidate (65). Importantly, some of the amino acid residues within the CRR of the M protein display signatures of diversifying selection. Immune escape is a possibility for any vaccine target, and current knowledge indicates that it deserves careful consideration for a CRR-based vaccine.

The second category of M protein-based vaccine candidates targets the NH$_2$-terminal type-specific determinants (179, 186). Pioneering work in the 1940s and 1950s showed that type-specific serum is responsible for protective immunity against the homologous \textit{emn} type (187–190); in contrast, minimal to no protective effects were observed for infection by heterologous GAS types. A 26-valent M type-specific recombinant polypeptide vaccine is the only GAS vaccine to have reached phase II clinical trials, whereupon it was found to be safe and highly immunogenic (179). The “type-specific protection” paradigm of GAS immunity has led to concerns about the potential for low coverage by the multivalent vaccine, especially in tropical settings, where the number of GAS \textit{emn} types can be very high (131, 191–195).

Recent data indicate that cross-protective immunity between closely related \textit{emn} types can induce broader protection via multivalent vaccines than type-specific immunity might otherwise predict. An initial hypothesis on cross-protective immunity was raised following genetic analysis of GAS strains collected in Brazil (196, 197). Cross-protective opsonization by hyperimmune rabbit antiserum generated via immunization with M type-based vaccines was subsequently demonstrated using a 30-valent vaccine (186) and an experimental 4-valent vaccine designed to protect against the 17 \textit{emn} types belonging to \textit{emn} cluster E4 (198).

Superficial skin infection can stimulate a robust antibody response. A recent study of humoral immunity acquired after impetigo infection in Fijian schoolchildren demonstrates that 38% of subjects had a ≥4-fold increase in anti-M-type-specific IgG titers that match the M type of the infecting strain (199). Importantly, cross-protective \textit{emn} cluster-specific immunity was evident for clusters E4, E6, and D4, although the levels of cross-protection were variable depending on both the patient and \textit{emn} cluster. That the cross-reactive immune responses frequently align with \textit{emn} cluster groups holds promise for the design of M type-based vaccines having broad coverage. More data are needed to further solidify the clinical and experimental findings obtained to date; identifying immunoassays to help define correlates of protection (CoPs) is a high-priority goal (200). Serotype replacement may potentially be an issue for GAS following widespread immunization with M type-based vaccines, as is the case for \textit{Streptococcus pneumoniae} and serotype-based capsule vaccines (201, 202).

**EPIDEMIOLOGY: BUILDING EVOLUTIONARY MODELS**

MLST data can be used to build models of evolutionary descent via eBURST or goeBURST, which are clustering algorithms based on SLVs (and/or double locus variants) among the seven core housekeeping genes (106, 203). The single mismatched allele that defines an SLV pair can arise by either point mutation or homologous recombination, whereby the latter may introduce >1 SNP. For GAS, the combined total length of the seven housekeeping (partial) alleles is 3,134 bp. However, 3,134 bp represents only a fraction of the GAS core genome; for example, \textit{emn}3 type GAS have a core genome of ∼1,645 kb (113), which is ∼524-fold longer in sequence than the MLST alleles combined. Based on these values, it can be estimated that an MLST-defined SLV pair of GAS would differ by an average minimum
of >500 SNPs within its core genome if genetic change were strictly due to point mutation or intraspecies recombination between very closely related organisms. However, if there was extensive HGT and recombination between more distally related GAS, the number of SNPs that differentiate the SLV pair is expected to be even higher. Given that the core genome of \textit{emm3} type iGAS strains has an estimated 29 to 133 SNPs per isolate (i.e., short-term evolution) (115), the models for evolutionary descent based on the SLVs of MLST using only seven core housekeeping genes may be a better capture of intermediate stages of evolution, which can be quite informative.

MLST data in combination with epidemiological marker genotypes can be compared to whole-genome sequence data. A good example for comparison is the \textit{emm89} strains causing iGAS disease in Europe and the United States, which underwent a marked increase in prevalence during the 2010s (30, 31, 111, 204). Using the ST and GAS isolate data deposited at www.pubmlst.org/spyogenes (105), a goeBURST plot that includes the recent epidemic \textit{emm89} strains is shown in Fig. 5A. The newly emerged epidemic clade of \textit{emm89} corresponds to ST101 (and its descendants). Based on demographic information (i.e., dates of recovery), it seems reasonable to assume that ST407 is ancestral to ST101. Notably, ST407 is recovered in association with two distinct \textit{emm} types: at least one other ST407 GAS isolate in the PubMLST database has \textit{emm78} in place of \textit{emm89}, as do several other SLVs and double locus variants of ST407. Alignment of the 180-bp \textit{emm} subtype regions of \textit{emm78.0} (ST253) and \textit{emm89.0} (ST407, ST101) reveals <60% identity for both nucleotide and amino acid sequences, indicative of substantial divergence (ClustalW; data not shown) and supporting a model wherein \textit{emm78} and/or \textit{emm89} underwent HGT and recombination onto an ST407-related genetic background. Both \textit{emm78} and \textit{emm89} are \textit{emm} pattern E but belong to different \textit{emm} clusters (E1 and E4, respectively).

Examination of the FCT regions of ST407 and ST101 \textit{emm89} type GAS isolates reveals highly divergent Cpa and FctA proteins, corresponding to the AP1 and BP subunits of pili, respectively (Fig. 5B). The distinct Cpa and FctA forms appear to coincide with T serotypes T11 (ST407) and T13 (ST101) (35). The housekeeping gene locus that distinguishes ST407 and ST101 is \textit{yqiL}, which lies adjacent to the FCT region (Fig. 4); \textit{yqiL}01 (ST401) differs from \textit{yqiL}03 (ST101) by three nucleotide substitutions, highly suggestive of diversification via recombination. In contrast to the eBURST analysis of MLST data, whole-genome sequencing and analysis of the core genomes of \textit{emm89} isolates provides a much higher level of resolution (30, 35, 111, 112). Importantly, several SNPs and indels that distinguish the \textit{emm89} clade 2 → clade 3 transition lead to alteration of virulence factor expression: loss of the hyaluronic acid capsule (\textit{AhasA}) and transcriptional upregulation of toxin genes encoding Nga and streptolysin O.

The genetic relationships among GAS strains harboring \textit{emm89} and ~20 additional \textit{emm} types were inferred based on ~75,000 SNPs of the core genome (111). The phylogeny shows that the \textit{emm89} isolates of the three clades detailed above, and corresponding to CC407 (ST101, ST407), cluster within the same major lineage; ST142 isolates (double locus variants of ST407; Fig. 5A) also cluster on that same deep branch. However, GAS of other \textit{emm} types, plus \textit{emm89}-ST380 isolates belonging to another CC (204), each cluster along distinct lineages with deep branches (111). ST380 differs from both ST101 and ST407 at all seven housekeeping alleles, indicative of HGT of \textit{emm89} to a distant genetic background.

The MLST data for \textit{emm89} strains captures two different modalities of HGT involving \textit{emm}: replacement of a nascent \textit{emm} by a new \textit{emm} type on the original genetic background (i.e., recipient-new clone pairs, with two \textit{emm} types and one ST) and lateral movement of an \textit{emm} type to a new genetic background (i.e., donor-new clone pairs, with one \textit{emm} type and two STs) (107). Within CC407 (recipient) there is evidence for the former process, involving replacement of \textit{emm78} with \textit{emm89} or vice versa (Fig. 5A). A possible donor of \textit{emm89} to CC407 is \textit{emm89}-ST380 (latter scenario); alternatively, ST380 may have been the recipient of \textit{emm89} from a CC407 donor. Comparisons of completely assembled genomes of possible donor-recipient-new clone pairings may aid in establishing the directionality of horizontal \textit{emm} gene movements.

Based on the connections within CC407, the seven housekeeping genes of MLST appear to recapitulate the more expansive core genome sequence data reasonably well. However, MLST fails to capture SNPs arising in core virulence-related genes that may alter the trajectory of GAS disease in a host population. What MLST may be best tuned to, when combined with typing data on \textit{emm}, pilin genes, and/or \textit{sof}, are HGT events wherein a typing gene is present on distant STs/CCs or multiple typing genes are recovered in association with the same ST/CC (such as divergent Cpa-FctA in \textit{emm89}-CC407; Fig. 5B). For GAS, MLST relationships can provide an intermediary link between short- and long-term evolution and may help delineate the genetic roots of the emer-
FIGURE 5 Model for evolutionary descent (CC407). (A) goeBURST diagram of CC407, which includes single representatives of 41 STs and is based on data from www.pubmlst.org/spyogenes (105). Unless noted otherwise (orange shaded boxes), all isolates (n = 114) are emm89 and were recovered from (i) the 1990s through 2010s, (ii) Europe or North America, and (iii) URT infections or iGAS disease. For evolutionary analysis of emm89 (formerly PT4245 (250)) isolates based on the core genome sequence (31, 111), clade 1 strain MGAS11027 is ST407, whereas clade 2 and epidemic clade 3 strains (MGAS23530 and MGAS27061, respectively) are ST101; the epidemic emm89 strain H293 from the United Kingdom is also ST101 (30). Note that ST407 corresponds to GAS with either emm89 or emm78 (105). (B) Amino acid sequence identities among predicted FCT region gene products of ST407 (MGAS11027) and ST101 (H293, MGAS23530) strains (all emm89) via ClustalW alignments.
gence of new strains that have unique repertoires of cell surface antigens and their linked virulence functions.

**EVOLUTION: MOLECULAR MECHANISMS OF GENETIC CHANGE**

Core genes of GAS contain a historical record of numerous homologous recombination events (111, 205–208). Intraspecific gene exchange between GAS strains most likely occurs during cocolonization of the URT or coinfection of impetigo lesions with multiple strains (209). GAS can grow as biofilms or form microcolonies during infection within both tissue niches (210, 211), and the high bacterial density that is attained likely facilitates HGT. There is also ample evolutionary evidence for gene exchange between GAS and other species, such as SDSE (212–218).

A major unresolved puzzle of GAS biology is the molecular mechanisms that govern HGT, leading to the high rates of homologous recombination that is inferred from its evolutionary history. In principle, nonspecific packaging of chromosomal fragments into bacteriophage capsids could provide a vehicle for HGT of bacterial DNA (i.e., generalized transduction). In addition to bacteriophage, GAS are also rich in integrative and conjugative elements (91). The degree to which random DNA fragments may slip through during the conjugal process remains unclear. A conserved RNA uracil methyltransferase (rnm) gene provides an integration hotspot for both bacteriophage and integrative and conjugative elements in GAS, as well as other related species (219). Natural transformation via high-frequency uptake of naked DNA has not been demonstrated for GAS, although competence gene homologues are present in the GAS genome and (very) low efficiency transformation is detectable in biofilms (220–222). However, GAS secrete numerous DNases that degrade extracellular DNA, which may render the transforming substance unavailable for uptake by a competent cell.

A mutator phenotype that results in an increased mutation rate has been characterized for GAS. This defect in DNA mismatch repair is controlled by excision and reintegration of a streptococcal phage-like chromosomal island (SpyCl) at a site between the mutS and mutL loci (223) and is present in the genomes of numerous GAS strains (50). SpyCl yields an ∼100-fold increase in the rate of spontaneous mutation, to as high as ∼10⁻⁷ to 10⁻⁸ mutations per generation. A possible relationship between the presence of SpyCl and CCs with a relatively low ratio of recombinational to mutational events remains to be explored.

**EVOLUTION: LONG TERM**

Core Housekeeping Genes

As previously stated, core genome-based phylogenies of GAS of different *emm* types and/or CCs tend to yield lineages with deep branches, indicative of large genetic distances (50, 110, 111, 116). Whole-core genome phylogenies depicted as SplitsTree graphs (neighbor-net analysis) show recombination between GAS of different *emm* types located deep within the network. Furthermore, there is no obvious clustering of the major lineages in accordance with *emm* pattern grouping or by any other recognized parameter.

Numerous methods have been used to assess the relative amount of recombination versus mutation that contributes to genetic change in GAS, based on the seven core housekeeping genes used in MLST (205–208). Together, the findings support the notion that GAS exhibit relatively high levels of recombination which, at least by some measures, is roughly comparable to the levels observed for the pneumococcus.

Using MLST data, the relative amount of recombinative versus mutation among the *emm* pattern-defined subgroups of GAS was examined (107, 224). Findings based on several methodologies all point to the same general trend, wherein the pattern A-C throat specialists have the highest tendency to diversify by mutation, and pattern D skin specialists and pattern E generalists have a greater tendency to diversify by recombination. The data on a larger role for mutation, relative to recombination, in genetic diversification among the classical throat strains is consistent with the molecular evolutionary analysis of *emm* subtypes (see “Epidemiology: Evolution of Markers (*emm*)” above; Table 9) (107). Differences among the *emm* pattern-defined groups of GAS in terms of their mechanisms for genetic change underscore different dynamics in shaping their long-term population structures.

GAS strains may encounter ecological separation due to their tissue site preferences for infection (throat versus skin), geographic partitioning (temperate versus tropical) and temporal distances (winter versus summer). Ecological barriers can give rise to allopatric speciation, and therefore, signatures of early stages of speciation within the GAS population may be evident. However, analyses of housekeeping alleles provide strong support for highly fluid HGT and homologous recombination of housekeeping genes among all three *emm* pattern groupings (101, 128, 224). These findings extend to isolates known to be recovered from the URT versus impetigo lesions, regardless of *emm* pattern group (224).
In summary, despite wide spatial-temporal distances between many GAS strains, there is no clear evidence for diminished opportunities for HGT of core housekeeping genes between organisms of the different \(emm\) pattern groups. Numerous statistical measures all point to a population genetic structure for the \(S.\ pyogenes\) species that is highly nonclonal (205–208). Against a background of highly random genetic change in core housekeeping genes, loci that exhibit strong linkage with \(emm\) pattern genotypes are reasonably good candidates for playing a direct role in the adaptations leading to throat versus skin infection (e.g., 50, 101).

**Lateral Movement of \(emm\)**

For the same global set of >500 GAS isolates used to calculate Ka/Ks ratios within the \(emm\) type-specific region (see “Epidemiology: Evolution of Markers (\(emm\))” above), the number of HGT events as a function of \(emm\) type was measured (107). Movement of \(emm\) to a distal ST is quantified as recovery of an \(emm\) type in association with >1 ST differing by \(\geq 5\) of the 7 housekeeping alleles; data show an average of 0.17, 0.60, and 0.75 lateral transfer events per \(emm\) type for the pattern A-C, D, and E groups, respectively. This finding supports the notion that for the classical throat strains, but less so for skin strains or generalists, \(emm\) type is often a reasonably accurate marker for ST or CC (104, 107, 108). In another measure of HGT and recombination, the number of \(emm\) types associated with \(emm\) type-variable STs (i.e., STs recovered in association with >1 \(emm\) type) was calculated; the number of \(emm\) type replacements per ST was 0.12, 0.15, and 0.02 for pattern A-C, D, and E isolates, respectively (107). Overall, there were three times as many events captured for HGT of \(emm\) to a distant ST \((n = 63)\) than for replacement of \(emm\) on the same ST \((n = 21)\).

Taken together, these data shed light on GAS strain competition surrounding the narrow time-space wherein the donor, recipient, and newly emerged clone all coexist. Strain competition can be mediated through host/herd immunity (225) and perhaps via other host-pathogen interactions as well. For the throat specialist strains, competition appears to be most intense between strains sharing the same \(emm\) type, because organisms of the original ST that served as the (presumptive) \(emm\) donor are rarely recovered. In contrast, for generalists, competition at the level of \(emm\) type appears to be considerably weaker, as evidenced by the frequent recovery of the same \(emm\) type in association with multiple distant ST backgrounds. How other cell surface antigens, such as SOF and/or pili, might contribute to the competitive sphere will require study of lateral movements of their genes and alleles within the GAS population.

**Full Complement of Epidemiological Markers Considered**

How do the host’s specific immune defenses perceive the incoming antigenic repertoire present on the GAS cell surface? Except during carriage, the Mga regulon that upregulates the expression of \(emm\) and \(emm\)-like genes and \(sof\) is typically active. The pilus T-antigens appear to be expressed by >90% of the GAS population. The expression/nonexpression of the hyaluronic acid capsule can potentially mask/unmask some or all the GAS cell surface proteins. Setting aside the variance in the transcriptional program, which can have a profound effect on transmission and virulence (see “Evolution: Short Term” below), the collection of epidemiological markers provides a useful construct for defining GAS “strains.”

Figure 6 summarizes the relative diversity or presence/absence of each major epidemiological marker in accordance with the \(emm\) pattern-defined groupings of throat specialists, skin specialists, and generalists. Importantly, \(emm\) type is rarely recovered in association with multiple \(emm\) pattern groupings (Table 2), thereby allowing reasonably valid inferences to be drawn. The pattern-and cluster-defined \(emm\) subtype alleles reveal stark differences between the throat specialist, skin specialist, and generalist groupings regarding diversification (Table 9). Major distinctions also lie in the relative distributions of the \(sof\) locus (Table 4), \(emm\)-like genes (see Fig. 2B), and FCT region forms (Fig. 3) (82).

Do highly diverse \(emm\) subtypes compensate for the dearth of \(emm\)-like genes and \(sof\) among the throat specialists—and vice versa for the generalists? The diversity among \(mrp\) and \(emm\) alleles is modest, but they too have undergone intraspecies HGT (226); selection may be tied to functional binding attributes of the Mrp or \(Emm\) proteins (e.g., IgA, IgG) (63, 227–230). The \(sof\) gene also undergoes rather extensive intraspecies HGT (80); however, \(sof\) gene movement is nonrandom in that it is largely restricted to the \(emm\) pattern E generalists (Table 4). Yet the relationships between individual \(emm\) types and \(sof\) types are highly varied, with many unique combinations arising via HGT to distant genetic backgrounds (35).

Among the throat specialist group is a wide variety of FCT region forms (Fig. 6) (82). The FCT-1 region is the most prevalent form among representative pattern A-C organisms of unique \(emm\) types; FCT-1 is characterized by an AP1 adhesin (FctX) and major BP (Tee) that are markedly divergent from the highly variable
Cpa and FctA pilins of FCT-3 and FCT-4 (82, 92). Not shown in Fig. 6 is FCT-2, which is (almost) exclusive to the highly prevalent emm1 organisms; the AP1 and BP pilins of FCT-2 are akin to a distant lineage of Cpa and FctA, respectively (82). Thus, the unusually high success (i.e., fitness) of emm1 strains may be due, in part, to its antigenically unique pilus structure. Overall, the absence of sof and emm-like genes among the pattern A-C strain grouping may be compensated for by the antigenic and structural diversity among its pili.

The skin specialist group appears to be the most homogeneous regarding epidemiological marker loci content, largely lacking sof but having both mrp and emm, and an FCT-3 region skewed heavily toward T3/13/ B3264 (Table 5); this genotype is coupled to the lowest diversity among emm subtypes (Table 9). Also, the skin specialist group is dominated by emm cluster D4, encoding the plasminogen-binding M protein (51, 65). However, past lateral movement of emm type to distant genetic backgrounds is rather abundant among the skin specialists and is ~2-fold greater than that of emm type replacement on the same ST (107) (see “Lateral Movement of emm” above); there is also ample HGT of housekeeping alleles. Conceivably, the emm pattern D skin specialists represent a more recent subpopulation of GAS (i.e., younger) that has less diversity within any one gene, but the genes themselves are highly recombinatorial. Alternatively, the cell surface protein repertoire of skin specialists may be subject to less intense pressures for diversifying selection, perhaps due to its microenvironment and/or the host immune response at the skin.

The relative lack of pattern D and E emm types recovered in association with FCT-1 or FCT-2 regions may be indicative of functional constraints that would otherwise lower fitness, such as inadvertent adherence to a suboptimal tissue via specific host receptors. As a group, the emm pattern E generalists display all the emm region epidemiological markers and tend to have either the large FCT-4 region or FCT-5 (Fig. 6), both of which harbor the prtF1 gene that encodes a fibronectin-binding protein (Fig. 3). Pattern E strains are also characterized by relatively high levels of recombination, involving both emm and the seven core housekeeping genes of MLST (50, 231, 232). Conceivably, within the generalist grouping are discrete throat or skin specialist strains whose phenotype is masked by the organizational framework (i.e., emm pattern group) that is imposed.

Historically GAS “strains” have been defined (primarily) by M serotype and secondarily by T and/or Sof serotype in combination with M type. Over the past 2 decades, emm typing has largely replaced GAS serotyping. The multitude of combinations of emm type and ST underscores the HGT of emm type, which is also consistent with the diverse array of M, T, and Sof serotype combinations observed. Even within a CC, STs are found in association with multiple emm types and/or divergent pilin genes (e.g., Fig. 5). Therein lies a conundrum for studying the population biology of GAS: How does one best define “strain?” This may or should be one of the next new challenges to be addressed by the GAS field.

![FIGURE 6 Trends in the distribution of epidemiological markers. An overview of the diversity and/or occurrence of epidemiological marker genes among three clinico-epidemiologic phenotype groupings of GAS. The relative distribution of FCT region forms is in accordance with one representative isolate for 94 unique emm types, as reported in reference 82; however, given the extensive HGT of emm and/or FCT genes to distant genetic backgrounds, the present analysis of the FCT region distribution may be far from complete.](image-url)
EVOLUTION: SHORT-TERM
The monophyletic grouping of GAS strains in accordance with emm type and/or MLST-defined CC readily allows for inference of ancestral-descendent pairs within a branch, especially when coupled to epidemiological collections of strains with known dates of recovery. Whole-genome sequences of hundreds or thousands of GAS isolates sharing the same emm type and collected over a narrow spatial-temporal distance have captured epidemic waves of clones that are distinct from presumptive ancestral clones (24, 30, 111, 113–115, 233). An important finding is that within an emm type (specifically, emm1 and emm3), iGAS isolates arise from all major lineages occupied by URT isolates and are genetically similar to the circulating pharyngitis strains from which they evolved (114, 234).

Comparative genomics of GAS recovered within a narrow spatial-temporal distance, and in association with a defined bacteria-host ecology (i.e., pharyngitis, asymptomatic carriage, and/or iGAS disease), have revealed high levels of diversifying selection at certain loci, particularly those involved in regulating the phenotypic expression of virulence factors (24, 30, 44, 111, 113–115, 233–240). Affected loci include bhaABC (hyaluronic acid capsule biosynthesis), corRS (two-component system that acts as a global regulator), rgg/ropB (standalone response regulator), mga (response regulator and activator of emm transcription), ial (sensor kinase), and nga-slo (exotoxins). There are some recurrent themes, wherein downregulation of emm enhances URT carriage, downregulation of speB enhances iGAS disease, and upregulation of nga-slo enhances URT transmission. The contribution of capsule to GAS ecology seems more complex, and perhaps it is GAS strain dependent. In fact, in general terms, all mutations that alter fitness are context dependent and, accordingly, may be strain dependent. In addition to SNPs in critical virulence loci, GAS genomes are profoundly shaped by mobile genetic elements (which are often not included as part of the core genome); bacteriophage and integrative and conjugative elements can harbor critical virulence genes as well (e.g., 114).

EVOLUTION (OR NOT): RESURGENCE OF SCARLET FEVER
Scarlet fever (SF) was a dreaded GAS disease through the early 1900s; it manifests as a URT infection coupled to a pathognomonic scarlatina rash triggered by the action of erythrogenic toxin, which was subsequently characterized as a T-lymphocyte superantigen. The occurrence of SF in the United States and Europe declined mid-century and was largely sporadic in recent decades. However, in 2011 there was an alarming increase in the number of reported cases of SF in Hong Kong and in China soon thereafter (241–243). This upsurge was largely attributed to emm12 GAS organisms that were resistant to macrolides and tetracycline. Comparative genome sequence analysis with other emm12 strains revealed a novel prophage harboring the superantigen genes spec and ssa (113, 244). However, the Hong Kong SF outbreak was multiclonal in origin.

A resurgence of SF cases also occurred in the United Kingdom, beginning in ~2013 (245). Comparative genomics failed to reveal a single clone or a single gene (e.g., a superantigen gene on a mobile genetic element) associated with those disease cases. As in Asia, the increase in SF incidence was associated with multiple lineages of GAS (246, 247).

Taken together, the recent upsurge in SF appears to be multifactorial and likely includes host-related factors. Numerous GAS lineages are responsible for SF, although resistance to antibiotics may have fueled the spread of some clones. Possible shifts in the content of superantigen genes may collectively have had an effect. The two cohorts—children in Asia and Europe and largely nonrural—possibly had a history of past exposure to a similar set of prevalent emm types and strains (131), and their herd immunity may be similar despite the geographic distance. Conceivably, an unknown copathogen may affect host susceptibility to SF and development of the scarlatina rash. Heightened clinician awareness and better reporting and surveillance, and/or changes in clinical practice, can also contribute to a higher disease incidence. The mysterious underpinnings of the recent upsurge in SF cases in Asia and Europe underscore the contributory role of factors other than genetic changes in GAS in shaping the epidemiology of GAS disease.

EVOLUTION AND POPULATION DYNAMICS
A central problem concerning GAS is the difficulty in connecting the evolutionary histories of different emm type-defined strains. There are large genetic gaps between extant organisms—but also rampant gene flow. So, what became of the evolutionary intermediates? Was there an extended series of population bottlenecks whereby new clones outcompeted their progenitors? And perhaps each successful new clone bore a new emm subtype plus changes in other genes, until a very unique and highly fit organism with a far more divergent emm
type emerged and gained a strong foothold in the human host population. Perhaps the genetic intermediates went extinct due to low fitness in the context of shifting host-environmental conditions, or perhaps they remain extant, lurking in low numbers in a semiquiescent state and have yet to be sampled.

When did GAS emerge as a human pathogen? Was the GAS progenitor a human commensal? Or was the progenitor an animal pathogen that has yet to be recovered via sampling? Did the new GAS pathogen acquire key genes from a trading partner as a pivotal step in its emergence? What are the genotypes that confer adaptation of GAS to the primary ecological niches of the throat and skin? Are GAS commensals at heart, and should one view their ecology primarily through that lens?

What was the structure of the human population at the time GAS emerged as a human pathogen—highly migratory with intermingling or isolated tribes? What is the minimum host population size necessary to sustain GAS transmission? What becomes of a GAS strain if/when it runs out of susceptible hosts? Does the organism persist in an altered carrier state? Why do there exist ~243 emm types? And why are there even far larger numbers of unique combinations of M, SOF, and T antigens? Is the number of unique combinations of cell surface proteins related to characteristics of the human host population? If so, how?

What are the selective pressures shaping the GAS population of today? Is the ability to cause a clinically inapparent infection a more recent adaptation, driven by social behavior in the form of seeking antibiotics when ill, thereby giving the stealth invader a survival advantage? Why did SF recently re-emerge in Asia and the United Kingdom? Why did acute rheumatic fever undergo a renaissance in the United States during the 1980s? How might a future vaccine alter its evolutionary trajectory? A better understanding of the ecology and evolution of GAS, and its population genetic structure, may ultimately lead to interventions that limit its impact on human health.

ACKNOWLEDGMENTS
Funding support for this work is from the NIH (grant AI-117088 to D.E.B.) and FNRS research grants (PDR T.0255.16 and CDR J.0019.17 to P.R.S.).

The authors dedicate this work to Susan K. Hollingshead (1953–2017), a coauthor of earlier editions of this chapter and a pioneer in the field of group A streptococcal genetics and evolution.

REFERENCES


Group A Streptococci


Group A Streptococci


Streptococcus dysgalactiae strains were recovered from isolates collected in the United States. Bessen et al.


