Infection, Genetics and Evolution xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Review

Molecular epidemiology and genomics of group A Streptococcus

Debra E. Bessen [a,*], W. Michael McShan [b], Scott V. Nguyen [b], Amol Shetty [c], Sonia Agrawal [c], Hervé Tettelin [c]

[a] Department of Microbiology & Immunology, New York Medical College, Valhalla, NY 10595, USA
[b] University of Oklahoma Health Sciences Center, Department of Pharmaceutical Sciences, College of Pharmacy, Oklahoma City, OK 73117, USA
[c] Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA

ARTICLE INFO

Article history:
Received 9 July 2014
Received in revised form 11 October 2014
Accepted 13 October 2014
Available online xxxx

Keywords:
Group A Streptococcus
Streptococcus pyogenes
Genomics
Population biology
Epidemiology
Evolution

ABSTRACT

Streptococcus pyogenes (group A Streptococcus; GAS) is a strict human pathogen with a very high prevalence worldwide. This review highlights the genetic organization of the species and the important ecological considerations that impact its evolution. Recent advances are presented on the topics of molecular epidemiology, population biology, molecular basis for genetic change, genome structure and genetic flux, phylogenomics and closely related streptococcal species, and the long- and short-term evolution of GAS. The application of whole genome sequence data to addressing key biological questions is discussed.

Contents

1. Taxonomy, habitats and disease . . . 00
2. Molecular epidemiology . . . 00
3. Population genetics . . . 00
4. Molecular mechanisms underlying genetic change . . . 00
5. Whole genome sequences . . . 00
6. Mobile genetic elements . . . 00
7. Phylogenomics of GAS and closely related species . . . 00
8. Long-term evolution: population genomics and tissue tropisms . . . 00
9. Short-term evolution: Comparative genomics in epidemics and outbreaks . . . 00
10. The shaping forces of natural selection . . . 00
Addendum . . . 00
Acknowledgements . . . 00
References . . . 00

Abbreviations: AGR, accessory gene region; C4BP, C4b-binding protein; Cas, CRISPR-associated proteins; CC, clonal complex; CGH, comparative genomic hybridization; CRISPR, clustered, regularly interspaced short palindromic repeat; *D*, genetic diversity; DNase, deoxyribonuclease; *emm*, encodes M protein; FCT region, encodes pili and T-antigen; FnBP, fibronectin-binding protein; GAS, group A streptococci; GBS, group B streptococci; GCS, group C streptococci; GGS, group G streptococci; H, haplotype; HGT, horizontal gene transfer; ICE, integrative and conjugative element; LCB, locally collinear block; MGE, mobile genetic element; MLST, multilocus sequence typing; MMR, mismatch repair; MSCRAMMs, microbial surface components recognizing adhesive matrix molecules; R_o, basic reproductive rate; RD, region of difference; Sda, streptococcal DNase (also, Sdn and Spd); SDD, *Streptococcus dysgalactiae* subspecies *dysgalactiae*; SDE, *Streptococcus dysgalactiae* subspecies *equisimilis*; Se, *Streptococcus equi* subspecies *equi*; SF, subfamily; SLV, single locus variant; SOF, serum opacity factor; Spe, streptococcal pyrogenic exotoxin; Spn/Nga, NAD+-glycohydrolase; SpyCI, streptococcal phage-like chromosomal island; sRNA, small regulatory RNA; ST, sequence type; Sz, *Streptococcus equi* subspecies *zooepidemicus*; URT, upper respiratory tract.

* Corresponding author. Tel.: +1 914 594 4193.
E-mail addresses: debra_bessen@nymc.edu (D.E. Bessen), William-McShan@ouhsc.edu (W.M. McShan), Scott-Van-Nguyen@ouhsc.edu (S.V. Nguyen), AShetty@som.umaryland.edu (A. Shetty), SAgrawal@som.umaryland.edu (S. Agrawal), tettelin@som.umaryland.edu (H. Tettelin).

http://dx.doi.org/10.1016/j.meegid.2014.10.011
1567-1348/

1. Taxonomy, habitats and disease

The importance of *Streptococcus pyogenes* as a human pathogen led to development of well-used clinical microbiology tools for its identification. Most notably, *S. pyogenes* forms large colonies and produces β-hemolysis following growth on blood agar, and is serologically distinguished from many other streptococcal species by its group carbohydrate that is covalently linked to the peptidoglycan cell wall. The term group A streptococci (or GAS) is commonly used as an alternative to *S. pyogenes*. Taxonomy based on 16S rRNA places *S. pyogenes* in what was historically referred to as the pyogenic (pus-forming) division of streptococci. The closest genetic relative of GAS is *Streptococcus canis* (group G streptococci; GGS) (Facklam, 2002). Other close genetic relatives, all of which lie within the pyogenic division and display β-hemolysis following growth on blood agar, include *Streptococcus dysgalactiae* subspecies *equisimilis* (SDE; largely group C streptococci [GCS] and GGS), *S. dysgalactiae* subspecies *dysgalactiae* (SDD), *Streptococcus equi* subspecies *zooepidemicus* (Sz) and *equi* (Se) (both are GCS), and *Streptococcus agalactiae* (group B streptococci [GBS]). Organisms that are (mostly) restricted to humans are GAS and SDE, whereas *S. canis*, SDD and *S. equi* subspecies Sz and Se primarily cause disease in other mammalian hosts. Environmental reservoirs appear to be non-existent or highly limited for the streptococcal species that cause human disease.

GAS is a human-specific pathogen that is highly prevalent worldwide, causing ~750 million infections per year (Carapetis, 2007; Carapetis et al., 2005), mostly at the throat (pharyngitis, tonsillitis) and skin (non-bullous impetigo). The epithelium of the throat and skin are the primary ecological niches of GAS and importantly, the tissue sites for most new GAS acquisitions and transmissions. Invasive disease is a relatively rare outcome of GAS infection, whereby the organism gains access to normally sterile tissue via the upper respiratory tract (URT) or breaks in the skin; from the bloodstream, GAS can disseminate to numerous deep tissues within the host. Although invasive GAS disease exacts a heavy toll in terms of morbidity and mortality, it is often an evolutionary dead end because an organism infecting deep tissue usually lacks an efficient means for transmission to a new host.

Symptomless throat infections are also known to occur, whereby the patient lacks overt clinical symptoms yet mounts a specific immune response to GAS antigens (i.e., clinically inapparent). Asymptomatic carriage of GAS at the URT – which is presumed to elicit little or no immune response due to the semi-quiescent state of the organism – can often achieve rates of >20% among school-aged populations (Kaplan, 1980). Although the organism is believed to be in a semi-quiescent, slow-growing state during carriage at the URT, transmission to new hosts can occur. Skin carriage appears to occur under endemic conditions (Anthony et al., 1976), but overall, it is less well documented. Thus, at least for GAS strains colonizing the oropharynx, a commensal-like state appears to be the predominant lifestyle.

The relative prevalence of pharyngitis and impetigo due to GAS varies in accordance with geographical location and season. Pharyngitis prevails in temperate regions and peaks in winter months, during which people spend extended time indoors and transmission occurs via a respiratory route. Impetigo is associated with warm, humid climates and is mostly observed in tropical and sub-tropical regions, or during summer months. Children are the primary targets of superficial GAS infections, and impetigo tends to afflict a slightly younger age cohort. Overall, there are spatial (geography) and temporal distances (winter versus summer seasons) that act to physically separate many organisms having a predilection for causing pharyngitis from those strains having a high tendency to cause impetigo. It is of great interest as to whether the spatial–temporal distances between GAS strains causing pharyngitis versus impetigo reduces the number of opportunities for horizontal gene transfer (HGT) which in turn, can shape the population genetic structure of a bacterial species (addressed in Section 3).

2. Molecular epidemiology

Pioneering work by Dr. Rebecca Lancefield aimed to understand the basis for protective immunity to GAS infection, and led to the development of a serological typing scheme based on the anti-phagocytic M protein surface fibrils (Lancefield, 1962). More than 80 distinct M types were identified, whereby protective immunity to GAS is M type-specific. The M type-specific determinants map to the fibril tips, encoded by the 5′ end of *emm* genes. More recently, a sequence-based *emm* typing scheme was implemented, based on extensive nt sequence differences at the 5′ end of the *emm* gene, whereby a unique *emm* type is defined as having <92% sequence identity over the nt sequence corresponding to the first 30 codons of the mature M protein (Beall et al., 1996). Among the 234 *emm* types recognized to date are >1200 distinct allelic forms of the *emm* type-specific regions of *emm* genes, known as *emm* subtypes (Beall, 2014). Virtually all contemporary epidemiological studies define GAS isolates according to their *emm* type and therefore, *emm* type provides the primary framework for understanding the population biology and genetic structure of this species.

All GAS isolates harbor an *emm* gene. In addition, many GAS strains have paralogous *emm*-like genes lying immediately upstream and downstream of *emm*, and a few strains have only the downstream *emm*-like locus. Thus, a given GAS strain can have one or two *emm*-like genes, in addition to *emm*; the upstream *emm*-like gene is often referred to as *mrp*, and the downstream *emm*-like gene is often referred to as *enn*. The paralogous *mrp* and *enn* genes lack *emm* type-specific determinants, and are also distinct from the *emm* locus within their 3′ end regions which encode a peptidoglycan-spanning domain. Based on the structure of the cell wall-spanning domain, there are four major forms or subfamilies (SF) of *emm* and *emm*-like genes, whereby *emm* is either SF-1 or SF-2, *mrp* is always SF-4, and *enn* is SF-1 or SF-3 (Hollingshead et al., 1994). Furthermore, there are five distinct chromosomal arrangements (Haanes et al., 1992; Hollingshead et al., 1993; Podbielski, 1993) of *emm* and *emm*-like genes and their SF forms, designated *emm* patterns A through E; *emm* patterns B and C are rare and grouped together with pattern A strains (referred to as *emm* pattern A–C), due to their structural similarities (i.e., all have an SF-1 *emm* gene and lack *mrp*). Pattern D and E strains have *mrp* and the SF-3 form of *enn*, but are distinct in that their *emm* genes are of the SF-1 and SF-2 forms, respectively. The SF-1 (58 aa residues) and SF2 (39 aa residues) forms of the peptidoglycan-spanning domain are markedly different in length, the functional significance of which has yet to be determined.

The *emm* pattern genotype has been determined for 170 of the 234 currently recognized *emm* types (McGregor et al., 2004; McMillan et al., 2013), including the most common ones. Thirty-six (21%) *emm* types are pattern A–C, 64 (38%) are *emm* pattern D, and 66 (39%) are *emm* pattern E; two *emm* types have a rearranged *emm* region (no pattern group is assigned), and only two (fairly rare) *emm* types are found in association with >1 *emm* pattern group. Importantly, the *emm* pattern group displays a statistically significant association with tissue site of infection, whereby GAS strains of the *emm* pattern A–C genotype show a statistically significant association with pharyngitis (throat specialists) and *emm* pattern D isolates show a statistically significant association with impetigo (skin specialists) (Bessen, 2009; Bessen et al., 2000b, 1996; Bessen and Lizano, 2010).

Please cite this article in press as: Bessen, D.E., et al. Molecular epidemiology and genomics of group A Streptococcus. Infect. Genet. Evol. (2014), http://dx.doi.org/10.1016/j.meegid.2014.10.011