

## Comparative genomics of streptococcal species

Joseph J. Ferretti, Dragana Ajdic & W. Michael McShan\*

*Departments of Microbiology & Immunology & \*Pharmaceutical Sciences, University of Oklahoma Health Sciences Center, Oklahoma City, OK, 73190, USA*

Received August 6, 2003

**Microbial genome sequencing has produced an unprecedented amount of new information and insights into an organism's metabolic activities, virulence properties, and evolution. The complete genome sequence has been reported for four different species of streptococci, including *Streptococcus pyogenes*, *S. agalactiae*, *S. pneumoniae* and *S. mutans*. Comparative genome analysis among organisms of the same species not only shows a high degree of similarity in gene content and organization, but also a high degree of sequence heterogeneity as evidenced by the large number of single nucleotide polymorphisms present. Considerable differences were also observed in the number of mobile genetic elements found in each organism, including complete and partial bacteriophage genomes, IS elements, transposons, and plasmids. *S. pyogenes* was the only species to contain complete bacteriophage genomes in its genome, while only *S. pneumoniae* and *S. mutans* contained the full complement of competence genes essential for natural transformation. Comparative genome analysis between the species showed that *S. pyogenes* was more closely related to *S. agalactiae* than with *S. pneumoniae* or *S. mutans*.**

**Key words** Genomes of *Streptococcus pyogenes* - *S. agalactiae* - *S. mutans* - *S. pneumoniae*

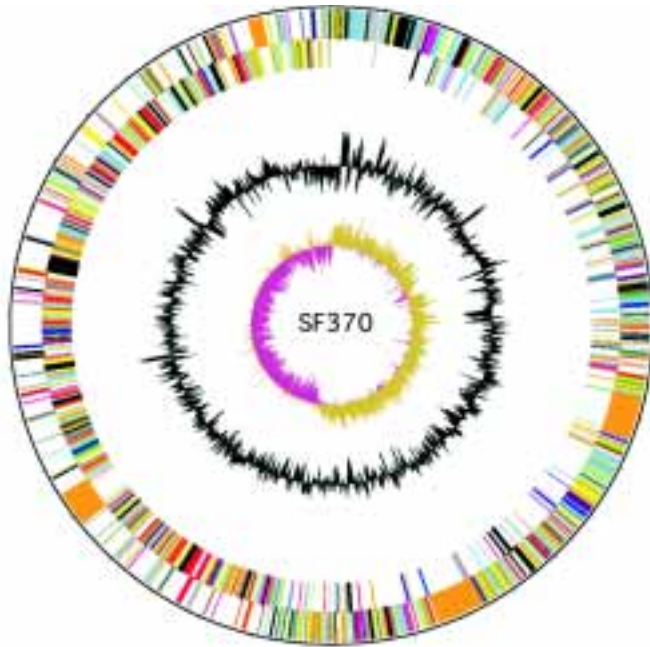
The availability of the complete genome sequence of bacterial genomes in recent years has provided new avenues of investigation for the further understanding of microbial pathogenesis, regulation, diversity, and evolution. Since 1995, over 180 complete genome sequences have been reported, with an approximate 500 genomes in progress (<http://wit.integratedgenomics.com/GOLD/>). The complete genome sequence has been determined for four different species of streptococci, microorganisms known to be pathogens responsible for various human diseases; *e.g.*, *Streptococcus pyogenes* (strep sore throat, flesh-eating disease, scarlet fever, rheumatic fever, acute glomerulonephritis), *S. agalactiae*, (neonatal sepsis and meningitis, adult invasive infections), *S. pneumoniae* (otitis media, sepsis, pneumonia), and *S. mutans* (dental caries, tooth decay). Interestingly, more genomes of bacteria classified as streptococci have been sequenced than any other genus, including several serotypes of each. The eventual

completion of the *S. uberis* and *S. equi* genomes presently in progress (<http://www.sanger.ac.uk/Projects/Microbes/>) will increase this number. In this review, we summarize the major findings of each genome along with a comparative analysis of the same and different species of streptococcal pathogens. Such a comparative analysis can provide further insight into species and strain uniqueness, and importantly, can stimulate new studies and approaches into disease prevention and treatment.

### **Streptococcus genomes**

The complete genome sequences of four different species of streptococci have been reported and a brief description of each is presented below.

(i) *Streptococcus pyogenes* (group A streptococci, GAS): The genomes of six strains of GAS have been or are nearly completed and most of these sequences are



**Fig.** Genome map of the M1 SF370 strain of *S. pyogenes*. The outer two circles show the position of the probable ORFs on the complementary DNA strands. The two inner circles show the %G+C of the sequence and the %G+C deviation by strand, respectively.

or will be publicly available. These include serotype strains of M1<sup>1</sup>, M3<sup>2,3</sup>, M5 (unpublished, see [http://www.sanger.ac.uk/Projects/S\\_pyogenes/](http://www.sanger.ac.uk/Projects/S_pyogenes/)), M18<sup>4</sup>, and M49 (McShan *et al*, unpublished). A typical circular representation of a GAS genome is presented in Figure, showing the location of genes on each strand and illustrating the transcription of genes originating at *oriC* in both directions and terminating at approximately 180°. The overall characteristics of each genome are similar in terms of base composition, gene arrangement, metabolic and physiologic genes, and virulence and pathogenicity genes. However, the recently completed genomes of an M3 strain<sup>2</sup> and the M5 Manfredo strain ([http://www.sanger.ac.uk/Projects/S\\_pyogenes/](http://www.sanger.ac.uk/Projects/S_pyogenes/)) indicate the presence of a large-scale inversion around the origin and terminus as compared to other GAS genomes. These inversions in genomes appear to be typical for such events in maintaining the symmetry of the origin and terminus, and such events may be common features during bacterial genome evolution<sup>5</sup>. However, the significance of these inversions in regard to strain survival or virulence is unknown.

The comparison of the homologous regions of each genome shows that the most common source of

sequence heterogeneity is the large number of single nucleotide polymorphisms found in the coding sequences. The most striking difference between all of the genomes is the number of bacteriophage genomes present, ranging from one to six prophages and accounting for up to 10 per cent of the total base pairs. These phage genomes are similar in functional organization, but show significant differences in size and individual gene content, most likely representing extensive recombination between the phage genomes. Interestingly, these GAS phages are most closely related to those found in the genomes of the non-pathogenic dairy bacterium, *Streptococcus thermophilus*<sup>6</sup> suggesting that horizontal gene flow may be occurring or has occurred in the past between these streptococci with vastly different life styles. However, the phages of *S. pyogenes* differ significantly from the phages of the non-pathogenic streptococci in one important aspect: in most instances, the phages also contain genes specifying virulence factors such as the streptococcal erythrogenic toxins (SPE, pyrogenic exotoxins) that are capable of acting as superantigens that stimulate the immune system to produce shock mediating cytokines. In all of these temperate phages characterized so far, the toxin genes are positioned at the distal end of the integrated phage genome (with respect to the integrase gene), suggesting that these genes may have been acquired in the past through some aberrant excision event where host genomic material was acquired by the phage. However, the ultimate source of the erythrogenic toxin genes and their homologues from other Gram-positive organisms such as *Staphylococcus aureus* remains unknown since no non-phage associated example of these toxins has been found.

With the increased number of GAS genomes sequenced numerous new phage genomes have been identified and consequently the number of SPEs has increased from 3 in the 1960s to over 20 at the present time. Given the high frequency of horizontal gene transfer in the GAS, the numbers of SPEs identified will no doubt increase as additional genomes are sequenced. The integration of these temperate phages generates a duplication of part of the host genome so that host gene functions are typically uninterrupted. However, in the M1 SF370 genome, a defective phage genome (phage SF370.4) is found integrated between two key genes responsible for DNA mismatch repair (MMR), *mutS* and *mutL*<sup>1,6</sup>. This separation is

predicted to interfere with the transcription of *mutL* and thus leave this strain defective for MMR. Since many clinical isolates of pathogenic bacteria are often defective for MMR, it has been proposed that such defects may allow for rapid adaptation in response to environmental stress<sup>7,8</sup>. Therefore, this unusual phage-mediated defect may be important in pathogenesis or survival of some GAS in the human host. Importantly, these overall results indicate the increased pathogenic potential of the GAS and the role of bacteriophages in the horizontal transfer of virulence determinants and evolution of GAS.

Among the complete GAS genomes sequenced to date, all are of the class I type, primarily associated with rheumatic fever and severe invasive infections. The eventual completion of a class II strain associated with acute glomerulonephritis should provide new information about differences between throat and skin strains. Preliminary data in our laboratory so far with an M49 strain (NZ131) indicate that its genome harbours only one bacteriophage genome and approximately sixty unique, non phage-associated genes. A comparison of selected features of GAS genomes is presented in Table I.

(ii) *Streptococcus agalactiae* (group B streptococci, GBS): The complete genome sequences of two of the nine GBS serotypes III<sup>9</sup> and V<sup>10</sup> have been reported. Of particular interest was the finding that most genes apparently unique to specific strains of the same serotype were found clustered in regions (islands). These islands not only contained atypical nucleotide compositions

differing from the 35.7 per cent G+C content of the entire genome, but also contained most of the known or putative GBS virulence factors. Interestingly, all of these islands also contained sequences known to be associated with mobile genetic elements, *e.g.*, insertion sequences, proteins of phages, plasmids, and transposons, suggesting that these islands correspond to horizontal gene transfer events.

A large number of phage and plasmid-related genes were identified in the chromosome of the serotype III strain, however no complete temperate phage genomes were found. Additionally, three copies of an approximate 50 kb sequence were present that had the characteristics of an integrative plasmid. The association of mobile elements and virulence factors in chromosomal islands suggests that they may be pathogenicity islands and thus have an important role in virulence acquisition and genetic diversity.

(iii) *Streptococcus pneumoniae*: Of the more than 90 capsule types, two genomes of *S. pneumoniae* have been sequenced, the original R6 strain (serotype 2) used by Avery *et al*<sup>11</sup> to demonstrate that DNA was the genetic material, and a virulent strain, TIGR4 (serotype 4)<sup>12,13</sup>. Genome sequence analyses confirm the presence of a number of features that make this organism a paradigm for recombination-mediated genetic plasticity<sup>14</sup>. Prime among these features is that *S. pneumoniae* is naturally competent and highly transformable and contains IS, BOX and RUP repetitive elements that account for up to 5 per cent of the genome's sequence. By contrast, the *S. pyogenes* genome shows no evidence of such repeats<sup>15</sup>. These sequences represent hotspots for genetic recombination, accounting for a high degree of heterogeneity among the species. Another important feature is the presence of nearly 400 genes with iterative DNA motifs that can result in phase variation. Interestingly, 25 of these genes appear to be directly associated with the virulence of the organism. There are also a large number of duplicated genes, estimated to be more than 250, as well as multiple gene clusters not present in all strains. Of particular interest with respect to virulence is that the penicillin binding proteins (PBP) contain mosaic structures amenable to recombination, resulting in a lower affinity for penicillin binding and increase in resistance.

**Table I.** Comparison of general features of *S. pyogenes* genomes

	SF370	MGAS315	Manfredo	MGAS8232
<i>emm</i> type	M1	M3	M5	M18
Bp	1,852,442	2,211,488	1,841,271	2,030,921
ORFs	1792	1865	ND	1845
%G+C	38.5	38.6	38.6	38.5
RRNA	6	6	6	6
TRNA	60	60	58	60
TCS	13	20	13+IRR	13+IRR
Phages	4	6	ND	5

ND, not determined; ORF, open reading frames; TCS, two-component regulatory systems

(iv) *Streptococcus mutans*: *S. mutans*, unlike the related pathogens *S. pyogenes* and *S. pneumoniae*, is part of the human oral flora and only incidentally an oral pathogen. As such, it differs from the pathogenic streptococci in several aspects of its basic physiology and in its adaptations to maintain an ecological niche. *S. mutans* is able to metabolize a wide variety of carbohydrates and can synthesize all of its required amino acids<sup>16</sup>. To complement the number of carbohydrates it can use, *S. mutans* devotes a large portion of its coding potential (about 15%) to various transport mechanisms. The number of proteases, peptidases, and other exoenzymes produced by *S. mutans* clearly suggests that it derives resources from host tissues. The analysis of the genome sequence showed that around 16 per cent of the predicted ORFs specified unique genes. In contrast to the closely related low per cent G+C Gram-positive organisms *S. pyogenes*, *L. monocytogenes*, and *L. lactis*<sup>1-4,17,18</sup>, no temperate bacteriophage genomes and no toxin genes are detectable in strain UA159. Finally, *S. mutans* possesses specific virulence factors including adhesins, glucan producing and binding exoenzymes, proteases and cytokine-stimulating molecules which help to protect the bacterium against possible host defenses and maintain its ecological niche in the oral cavity, while contributing to its ability to cause host damage.

**Table II.** Comparison of general features of selected streptococcal genomes

	<i>S. pyogenes</i> <sup>a</sup>	<i>S. agalactiae</i> <sup>b</sup>	<i>S. pneumoniae</i> <sup>c</sup>	<i>S. mutans</i> <sup>d</sup>
bp	1,852,442	2,211,488	2,160,837	2,030,921
ORFs	1792	2118	2236	1963
%G+C	38.5	35.6	39.7	36.8
rRNA operons	6	7	4	5
tRNA species	60	80	58	65
TCS	13	20	13+IRR	13 +1RR
ABC Transporters	36	62	74	65
Phages (complete)	Yes	No	No	No
Transformable	No	No	Yes	Yes

<sup>a</sup> Strain SF370 (type M1) GenBank Accession AE004092

<sup>b</sup> Strain NEM316 (type III) GenBank Accession AL732656

<sup>c</sup> Strain TIGR4 (serotype 4) GenBank Accession AE005672

<sup>d</sup> Strain UA159 (serotype c) GenBank Accession AE014133

ORF, open reading frames; TCS, two-component regulatory systems

## Comparative genomics

The genome sequence of an organism provides information about size of the genome, base composition, complete gene content, physiology and metabolism, content of virulence factors, and lateral gene transfer events. A comparison of general features of selected streptococcal genomes is presented in Table II. Although these organisms have many features in common and evidently have common evolutionary ancestors, the variation of genome size in base pairs and a variance of G+C content of almost 10 per cent indicates that significant differences exist in their genomes. Among the noteworthy differences between the genomes is the content of mobile genetic elements. The GAS are the only species to contain complete bacteriophage genomes, indicating that this mechanism of horizontal gene transfer is an important factor in gene acquisition/loss, strain heterogeneity and in its overall evolution. *S. pneumoniae* and *S. mutans* have highly developed transformation systems, whereas natural transformation is not known to be a common event in GAS or GBS. Although GAS and GBS have many of the genes essential for competence and transformation, the fact that they have lost competence may have occurred because phages have assumed a more important role in population diversity.

An additional noteworthy difference is the number of two component regulatory systems found in GBS, which is higher than in the other organisms. A possible explanation could be that the GBS have greater adaptability than the other streptococci to react and survive in response to fluctuations in the external environment.

**Table III.** Per cent relatedness of streptococcal species based on a comparison of ORFs

	<i>S. pyogenes</i> <sup>a</sup>	<i>S. agalactiae</i> <sup>b</sup>	<i>S. pneumoniae</i> <sup>c</sup>	<i>S. mutans</i> <sup>d</sup>
<i>S. pyogenes</i>	100	77	68	69
<i>S. agalactiae</i>	69	100	67	68
<i>S. pneumoniae</i>	66	72	100	70
<i>S. mutans</i>	65	70	68	100

<sup>a</sup> Strain SF370 (type M1); <sup>b</sup> Strain NEM316 (type III); <sup>c</sup> Strain TIGR4 (serotype 4); <sup>d</sup> Strain UA159 (serotype c)

Genome comparisons can also be made by analyzing open reading frame (ORF) relatedness using a BlastP programme. The results of such a comparison are presented in Table III and the per cent relatedness of the ORFs was calculated as greater than or equal to 50 per cent similarity using the BlastP programme. These data indicate that the GAS are more closely related to GBS and *vice versa* than with either *S. pneumoniae* or *S. mutans*. Such an observation is in agreement with the previously determined phylogenetic relationships among the streptococci based on 16S rRNA sequence analysis<sup>19</sup>.

### Summary & conclusions

Sequence information has provided new information about genes involved in streptococcal virulence and pathogenesis, regulation, metabolism, and physiology. Particularly interesting are the mechanisms of horizontal gene transfer and how these organisms gain new virulence genes to increase their pathogenic potential as well as how horizontal gene transfer mechanisms affect genome plasticity and evolution. Additionally, the sequence data provide new information about the evolution of these organisms and how they change in order to evade human host immune recognition. Future detailed analyses of additional genomes will provide even more information essential for understanding the inter-relationships of these organisms and also the identification of unique targets for drug development and identification of candidate antigens for a universal vaccine. An overarching goal of studies such as these is an improved prevention and treatment of streptococcal diseases.

### Acknowledgment

This work was supported by grant No. AI19304 from the National Institute of Allergy and Infectious Diseases, National Institutes of Health.

### References

1. Ferretti JJ, McShan WM, Ajdic D, Savic DJ, Savic G, Lyon K, *et al*. Complete genome sequence of an M1 Strain of *Streptococcus pyogenes*. *Proc Natl Acad Sci USA* 2001; 98 : 4658-63.
2. Nakagawa I, Kurokawa K, Yamashita A, Nakata M, Tomiyasu Y, Okahashi N, *et al*. Genome sequence of an M3 strain of *Streptococcus pyogenes* reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome Res* 2003; 13 : 1042-55.
3. Beres SB, Sylva GL, Barbian KD, Lei B, Hoff JS, Mammarella ND, *et al*. Genome sequence of a serotype M3 strain of group A streptococcus: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc Natl Acad Sci USA* 2002; 99 : 10078-83.
4. Smoot JC, Barbian KD, Van Gompel JJ, Smoot LM, Chaussee MS, Sylva GL, *et al*. Genome sequence and comparative microarray analysis of serotype M18 group A streptococcus strains associated with acute rheumatic fever outbreaks. *Proc Natl Acad Sci USA* 2002; 99 : 4668-73.
5. Eisen JA, Heidelberg JF, White O, Salzberg SL. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol* 2000; 1 : RESEARCH0011.
6. Canchaya C, Desiere F, McShan WM, Ferretti JJ, Parkhill J, Brussow H. Genome analysis of an inducible prophage and prophage remnants integrated in the *Streptococcus pyogenes* strain SF370. *Virology* 2002; 302 : 245-58.
7. Bayliss CD, Moxon ER. Hypermutation and bacterial adaptation. *ASM News* 2002; 68 : 549-55.
8. Radman M. Enzymes of evolutionary change. *Nature* 1999; 401 : 866-7, 869.
9. Glaser P, Rusniok C, Buchrieser C, Chevalier F, Frangeul L, Msadek T, *et al*. Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. *Mol Microbiol* 2002; 45 : 1499-513.
10. Tettelin H, Massignani V, Cieslewicz MJ, Eisen JA, Peterson S, Wessels MR, *et al*. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. *Proc Natl Acad Sci USA* 2002; 99 : 12391-6.
11. Avery OT, MacLeod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J Exp Med* 1944; 79 : 137-58.
12. Hoskins J, Alborn WE Jr, Arnold J, Blaszcak LC, Burgett S, DeHoff BS, *et al*. Genome of the bacterium *Streptococcus pneumoniae* strain R6. *J Bacteriol* 2001; 183 : 5709-17.
13. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, *et al*. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* 2001; 293 : 498-506.
14. Claverys JP, Prudhomme M, Mortier-Barriere I, Martin B. Adaptation to the environment: *Streptococcus pneumoniae*, a paradigm for recombination-mediated genetic plasticity? *Mol Microbiol* 2000; 35 : 251-9.
15. Mrazek J, Gaynon LH, Karlin S. Frequent oligonucleotide motifs in genomes of three streptococci. *Nucleic Acids Res* 2002; 30 : 4216-21.
16. Ajdic D, McShan WM, McLaughlin RE, Savic G, Chang J, Carson MB, *et al*. Genome sequence of *Streptococcus mutans*

- UA159, a cariogenic dental pathogen. *Proc Natl Acad Sci USA* 2002; 99 : 14434-9.
17. Glaser P, Frangeul L, Buchrieser C, Rusniok C, Amend A, Baquero F, *et al.* Comparative genomics of *Listeria* species. *Science* 2001; 294 : 849-52.
18. Bolotin A, Wincker P, Mauger S, Jaillon O, Malarne K, Weissenbach J, *et al.* The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403. *Genome Res* 2001; 11 : 731-53.
19. Kawamura Y, Hou XG, Sultana F, Miura H, Ezaki T. Determination of 16S rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic relationships among members of the genus *Streptococcus*. *Int J Syst Bacteriol* 1995; 45 : 406-8.

*Reprint requests:* Dr J.J. Ferretti, Department of Microbiology & Immunology, University of Oklahoma Health Sciences Center  
1000 Stanton Young Blvd. Oklahoma City OK 73104, USA  
e-mail: joe-ferretti@ouhsc.edu