Curr Microbiol
DOI 10.1007/s00284-017-1305-z

Exploring the Genomic Diversity and Cariogenic Differences of *Streptococcus mutans* Strains Through Pan-Genome and Comparative Genome Analysis

Peiqi Meng[1,2] · Chang Lu[3] · Qian Zhang[1] · Jiuxiang Lin[2] · Feng Chen[1]

Received: 2 April 2017 / Accepted: 11 July 2017

Abstract Pan-genome refers to the sum of genes that can be found in a given bacterial species, including the core-genome and the dispensable genome. In this study, the genomes from 183 *Streptococcus mutans* (*S. mutans*) isolates were analyzed from the pan-genome perspective. This analysis revealed that *S. mutans* has an "open" pan-genome, implying that there are plenty of new genes to be found as more genomes are sequenced. Additionally, *S. mutans* has a limited core-genome, which is composed of genes related to vital activities within the bacterium, such as metabolism and hereditary information storage or processing, occupying 35.6 and 26.6% of the core genes, respectively. We estimate the theoretical core-genome size to be about 1083 genes, which are fewer than other *Streptococcus* species. In addition, core genes suffer larger selection pressures in comparison to those that are less widely distributed. Not surprisingly, the distribution of putative virulence genes in *S. mutans* strains does not correlate with caries status, indicating that other factors are also responsible for cariogenesis. These results contribute to a more understanding of the evolutionary characteristics and dynamic changes within the genome components of the species. This also helps to form a new theoretical foundation for preventing dental caries. Furthermore, this study sets an example for analyzing large genomic datasets of pathogens from the pan-genome perspective.

Introduction

Streptococcus mutans (*S. mutans*) is one of the most prevalent bacteria in human oral flora and is widely recognized as a key etiological agent in human dental caries [11]. *Streptococcus mutans* is both highly acidogenic and aciduric [26, 40, 65], while also forming an effective biofilm on tooth enamel by producing surface antigens which promote adhesion to tooth surfaces and other bacteria [26]. Furthermore, this bacterium occasionally causes bacteremia, abscesses, and infective endocarditis [51, 52]. *Streptococcus mutans* isolates vary among a range of phenotypic properties, including polysaccharide serotypes, sugar utilization, macromolecule binding, acid production, acid tolerance [67] and biofilm formation [46]. These phenotypic variations may underlie individual variations in cariogenicity. Until now, a number of studies have demonstrated a substantial genetic heterogeneity across clinical isolates of *S. mutans* [6, 13, 54, 71].

Comparative genomics is a popular method for identifying determinants of microbial virulence. This method compares genomes from different strains and identifies similarities and differences, revealing common molecular mechanisms that can be related to phenotypic features. It can also be used to explore microbial virulence factors and tracking pathogenic mechanisms [33], which may lay the

Electronic supplementary material The online version of this article (doi:10.1007/s00284-017-1305-z) contains supplementary material, which is available to authorized users.

✉ Feng Chen
chenfeng2011@hsc.pku.edu.cn

[1] Central Laboratory, Peking University School and Hospital of Stomatology, No. 22, Zhongguancun South Avenue, Haidian District, Beijing 100081, People's Republic of China

[2] Department of Orthodontics, Peking University School and Hospital of Stomatology, Beijing 100081, People's Republic of China

[3] Department of Periodontics, Peking University School and Hospital of Stomatology, Beijing 100081, People's Republic of China

Published online: 17 July 2017

foundation for developing novel vaccines [21]. Comparative genomics is primarily used to investigate intraspecies variation [30, 62] and a large number of studies have indicated that the genetic variability within bacterial species is much larger than the variation found in the other domains of life. The gene content between pairs of isolates can diverge by as much as 30% within a species, such as *Streptococcus pneumoniae* [68]. This intraspecies genome variation can be resolved by several explanations, including genome reduction, genome rearrangements, gene duplication, and acquisition of new genes through lateral gene transfer (LGT) [23, 37]. Thus, it is important to note that the genomic sequence from one strain is not always representative of the entire species, suggesting that information from multiple isolates is needed to understand the dynamic nature of the species's genome. This led to Tettelin et al. developing the concept of a "pan-genome" for microorganisms [62], which represents the sum of genes that can be found within a given bacterial species [50, 66]. The pan-genome is composed of a "core-genome" and a "dispensable genome." The former consists of genes that are shared by all strains within the species, which are primarily related to basic biological functions and typical phenotypic properties of the organism. The dispensable genome is comprised of genes that only exist in some strains or specific genes of a certain strain, most of which are often associated with environmental adaptations, such as the production of antibiotics, drug tolerance, and virulence. Overall, this genome reflects the diversity of the strains. Since the original proposal of the pan-genome in 2005, this concept has been applied to describe a variety of microorganisms and in evolutionary analyses for *Streptococcus agalactiae* [62] and *Streptococcus pneumoniae* [37].

The development of high-throughput sequencing and bioinformatics, coupled with the first genomic sequence of a *S. mutans* isolate (*S. mutans* UA159) published in 2002 [2], has allowed for more complete/draft genomes of *S. mutans* isolates to be sequenced. This provides new resources for analyzing evolution of the species and genetic variation within *S. mutans* based on the pan genome. Argimón et al. performed whole genome comparisons of *S. mutans* strains associated with severe early childhood caries, as well as strains isolated from caries-free children in order to identify genetic loci specific to caries formation [5]. Liao et al. identified genotypic changes in a fluoride-resistant *S. mutans* strain and examined potential functions of the identified mutations by comparing gene expression between the fluoride-sensitive and fluoride-resistant strains [39]. However, these comparative genomic studies of *S. mutans* only included a relatively small number of species genomes, reducing the confidence in the robustness of the estimated pan-genome and core genome sizes. In addition, multiple studies have concluded that the distribution of putative virulence genes among *S. mutans* strains does not correlate with caries development [4]. Nevertheless, these results need to be verified or falsified by using a larger number of isolates with more advanced methods, including pan-genomics. Thus, we assembled the draft genomes of 11 new *S. mutans* isolates after sequencing. These genomes then were combined with complete/draft genomes from 172 *S. mutans* isolates in order to analyze pan-genomic features and the variation distribution of cariogenic genes among different isolates through comparative genomics. This analysis may contribute to dental caries prevention at the genomic level in the future.

Materials and Methods

Materials

This study used 172 *S. mutans* genomes published by Dec, 31th, 2016. The genomic data were obtained from the genomes released to the public database NCBI (ftp://ncbi.nlm.nih.gov/genomes/refseq/bacteria/Streptococcus_mutans/all_assembly_versions/), including 6 complete genomes and 166 draft genomes for which complete genomes were not available. Characteristics of all the *S. mutans* strains were acquired from NCBI (https://www.ncbi.nlm.nih.gov/genome/) and related genome publications [1, 2, 5, 8, 45].

Methods

Genome Sequencing, Assembly, and Annotation

Whole-genome sequencing of 11 *S. mutans* strains was performed using a shotgun high-throughput sequencing approach on an Illumina HiSe 2000 platform by generating 100 bp paired-end read libraries. An average of 470 Mb of high-quality data were generated for each strain, corresponding to a sequencing depth of 114–366 fold. The paired-end reads were de novo assembled using SOAPdenovo2 [42]. The individual genome assemblies of 11 strains have been deposited at DDBJ/ENA/GenBank under the project number PRJNA377637 with individual accession numbers listed in Online Resource 1. Raw reads for 11 strains have been deposited in the sequence read archive under the sample accession IDs listed in Online Resource 1.

The coding sequences of genes were predicted for each sequenced genome using Glimmer v3.02 [16]. In order to perform functional annotation, the amino-acid sequences of predicted coding sequences were blasted against the non-redundant protein sequence database with a criterion of

e-value $<1e^{-5}$. In addition, rRNAs and tRNAs of the 11 assembled genomes were predicted using RNAmmer [36] and tRNAscan [41], respectively.

Identification and Functional Classification of Homologous Clusters

In addition to the 11 *S. mutans* draft genomes assembled in this study, 172 previously published *S. mutans* genomes were also included in the following analysis. All extracted protein sequences from 183 genomes were adjusted to a prescribed format and were grouped into homologous clusters using OrthoMCL [22] based on sequence similarity. The BLAST reciprocal best hit algorithm [48] was applied with the criterion of e-value $<1e^{-5}$, identity >40%, and length coverage of a gene >50%, and Markov Cluster Algorithms [20] were employed with an inflation index of 1.5 to complete cluster analysis. A conservation value (CV) for each homologous cluster was assigned based on the number of strains covered by the cluster. For example, a CV value of 180 would be assigned to a cluster that contained genes from 180 strains. The functional category of each homologous cluster was determined by performing BLAST against the Cluster of Orthologous Groups database (http://www.ncbi.nih.gov/COG/) with a criterion of e-value $<1e^{-5}$ and identity >40%.

Pan-Genome and Core-Genome Analysis

Identified homologous clusters were parsed using Perl scripts to estimate the size of the pan- and core-genomes for each additional genome sequenced. In order to take into consideration of core genes that are possibly missed during genome sequencing and assembly, a correction step was introduced for the calculation of core-genome size, in which any one gene that is only absent in one of the 177 draft genomes was still regarded as core gene. The number of total genes/core genes provided by each added new genome depends on the selection of previously added genomes. For an N number of given strains, the number of all possible combinations is C(183, N). If the value of C(183, N) is greater than 8000, only 8000 random combinations were used. If fewer, all possible combinations were used. The final size of the pan- or core-genomes was the average value of all used combinations.

Virulence Factors Distribution

In order to explore the distribution of species group-specific virulence factors, *S. mutans* virulence factors were collected from previous publications [4, 14, 56, 57]. Information regarding the existence of relevant genes of these virulence factors was extracted from annotation reports for each strain in the NCBI database (ftp://ncbi.nlm.nih.gov/genomes/refseq/bacteria/Streptococcus_mutans/all_assembly_versions/). The virulence factors distribution and abundance in 183 *S. mutans* genomes were displayed as a heatmap using HemI software (http://hemi.biocuckoo.org/) [17].

Results and Discussion

Distribution and Identification of Homologous Clusters

The detailed information of the 183 *S. mutans* strains are shown in Online Resource 2. The genome size varies from 1.74 Mb (*S. mutans* TCI-298) to 2.73 Mb (*S. mutans* str. B16 P Sm1) with an average value of 1.95 Mb (Online Resource 3a), and the genomic GC content ranges from 35.7 (*S. mutans* PKUSS-9) to 37.1% (*S. mutans* LJ23, *S. mutans* B23Sm1 and *S. mutans* B112SM-A) (Online Resource 3b).

As defined previously, CV was calculated for each homologous cluster based on the number of strains for each cluster covered. Different CVs represent the distribution of gene variations from different strains in each cluster. A larger CV indicates a wider spread for the cluster in the species and a greater conservation for the gene. In this study, a total 333,069 protein-coding genes from 183 *S. mutans* strain genomes were grouped into 3978 homologous clusters.

Among the 3978 clusters, the number of specific genes whose CV was 1 is 906, and there were 5 specific genes in each strain on average, which is much less than observed in other species. Previous genomic analyses have revealed that the average number of specific genes in each strain is 39 and 34, respectively, for *Haemophilus influenza* [31] species and *Streptococcus pneumoniae* species [30]. Specific genes are usually the result of genes newly generated during evolution process, whereas, for prokaryotic organisms, LGT is the major driving force to obtain new genes [24]. From the perspective of the pan-genome, as the number of sequenced genomes increases, more homology for previously identified specific genes is obtained from the newly sequenced genomes. Thus, these specific genes might be redefined as "dispensable genes" as more genomes are sequenced. This may explain why the number of specific genes decreases as the number of sequenced genomes increases. As this study included more genomes in the analysis than in previous work, the average number of specific genes identified was consequently lower.

Besides, there were 905 homologous clusters with CVs that were 183. This is smaller than the number of core genes because the core genome defined in our study

includes a part of clusters with CVs of 182 other than those of 183, which will be elaborated in the next section.

Compared with clusters whose CVs are 1 or 183, the number of clusters with CVs ranging from 2 to 182 was relatively small (Fig. 1a). In order to avoid bias generated in the analysis due to few clusters with medium CVs, we divided these clusters into three groups based on CV; clusters with CVs of 2–7, 8–139, and 140–182 were grouped together. In addition, clusters with CVs of 1 or 183 were grouped separately. In this way, the 3978 clusters were separated into five groups.

The five cluster groups were assigned to Cluster of Orthologous groups classifications of clusters and analyzed for the functional distribution of clusters with different degree of conservation (Fig. 1b). Based on the increase in conservation for the clusters, the proportion of genes related to metabolism, as well as hereditary information storage and processing, were increased, while genes with functions that are poorly characterized were decreased. Metabolism and hereditary information storage/processing composed a majority of the genes whose CVs were 183, making up 35.6 and 26.6% of the clusters, respectively, whereas cellular processes and signaling made up 20.3% of the clusters. Meanwhile, the majority of specific genes is poorly characterized and may be involved in specific adaptations that help the *S. mutans* strains survive in novel environments.

In addition, the nucleotide sequences from each cluster were aligned and the d*n*/d*s* ratio was computed. The d*n*/d*s* ratio is commonly regarded as one of the most popular and reliable measures of evolutionary pressures on protein-coding sequences [35]. Initially, $dn/ds < 1$ was interpreted as negative selection, $dn/ds = 1$ as neutral, and $dn/ds > 1$ as positive selection [44]. However, recent studies have indicated that this kind of interpretation is too rough, as it has been found that $dn/ds < 1$ can occur under both negative and positive selection [35]. Even so, it has been universally acknowledged that the smaller the d*n*/d*s* ratio the larger the selection pressure. After eliminating clusters with CVs of 1 or where the ds value is 0, 2634 clusters remained for evolutionary pressures analysis. The relationship between the d*n*/d*s* distributions of clusters and their conservations were analyzed (Fig. 1c); it was observed that the larger the CV, the more conservative the cluster, and the smaller the d*n*/d*s* ratio. It can be inferred that core genes are subject to larger selection pressures in the evolutionary process when compared to dispensable genes, and that their sequences are more conservative.

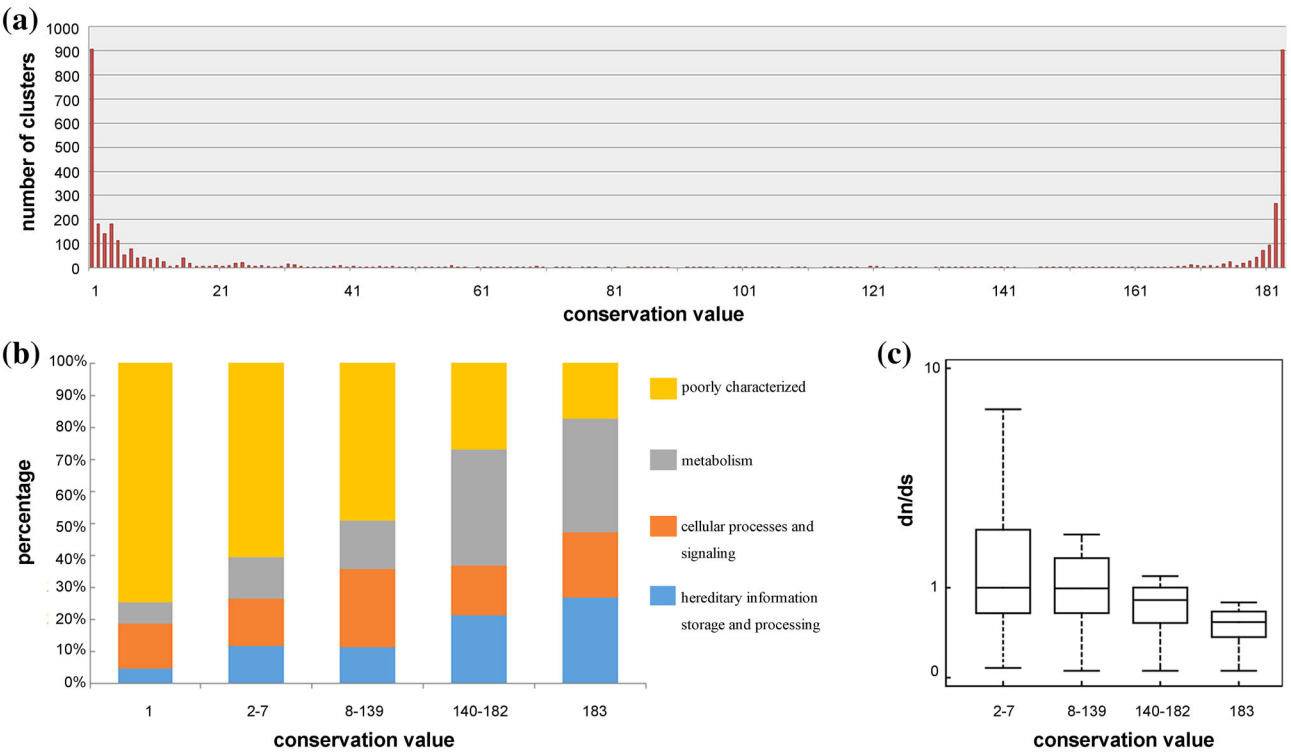

Fig. 1 Distribution characteristics of homologous clusters grouped by conservation values. **a** The number of homologous clusters for each different conservation value. **b** COG function classifications of genes with different degrees of conservation. **c** The box plot showing the d*n*/d*s* value distributions of homologous clusters with different degrees of conservation

Pan-Genome and Core-Genome Analysis

Pan-Genome

A pan-genome analysis of 183 *S. mutans* strains was performed, including the eleven *S. mutans* strains sequenced in this study and 172 *S. mutans* strains with genomes available in the NCBI database on Dec 31th, 2016. Figure 2 depicts gradual expansion of the pan-genome and contraction of the core-genome of *S. mutans* as genomes have been added to the dataset. The pan-genome size increased steadily without reaching a plateau. Applying a power-law regression model ($y = a + bx^c$) based on Heap's Law [29] resulted in a model that fit all of the data points at a very high confidence ($R^2 = 0.99996$) (detailed information of all core and pan-genome modeling are given in Online Resource 4). Based on this model, *S. mutans* has an "open" pan-genome and the expected average new gene number for each additional new genome was estimated to be 5. In contrast, Cornejo et al. proposed a finite pan-genome for *S. mutans* by using a special "pseudogene cluster" identification process to exclude ~30% of the rare genes that were considered to be pseudogenes [15]. However, Cornejo et al. did not provide the detailed parameters they obtained from fitting and also observed a high rate of LGT in *S. mutans*, where many genes were acquired from related streptococci and bacterial strains. Such high rates of LGT might also lead to a continuously growing pan-genome. Moreover, *S. mutans* is found to colonize diverse habitats including oral cavity, blood and the intestinal and urinogenital tracts [67]. It has been reported that an open pan-genome is likely to be typical in species that colonize multiple environments and have multiple ways of exchanging genetic material, indicating that it is reasonable that *S. mutans* may have an open pan-genome.

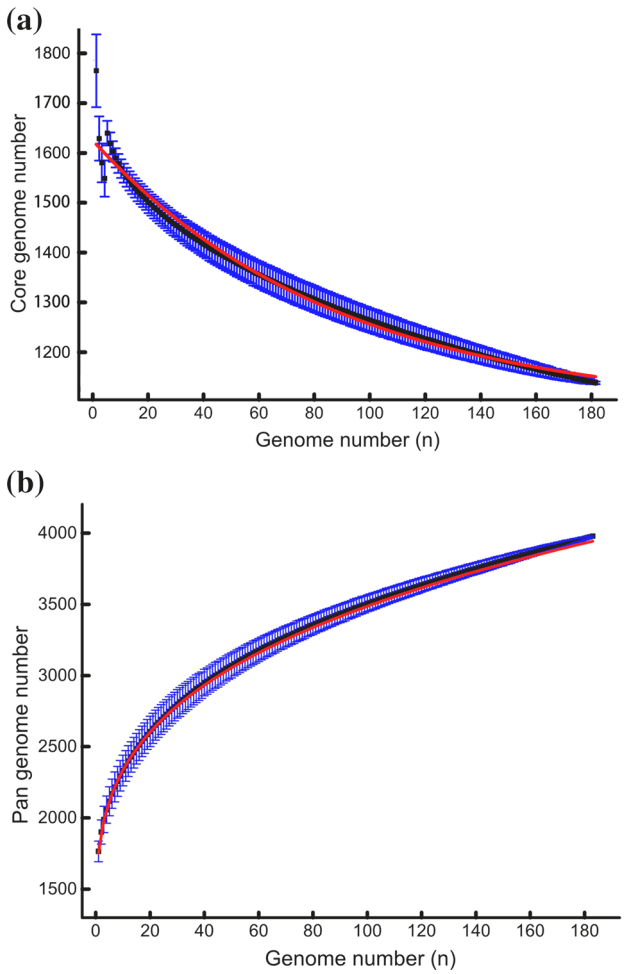

Fig. 2 Core and pan-genome model of 183 *S. mutans* genomes. **a** Core-genome model of *S. mutans*. **b** Pan-genome model of *S. mutans*. In the two figures, *black squares* are the medians of the core-genome sizes calculated by randomly sampling 8000 different genome combinations of *n* genomes out of 183 genomes. *Blue bars* are the standard deviation of the medians. The *red curve* is the fitting result

Previous pan-genome studies in other organisms have revealed that both *Bacillus cereus* and *Streptococcus pneumoniae* also have open pan-genomes, while *Bacillus anthracis* and *Ureaplasma urealyticum* have finite pan-genomes [63]. This may represent that different pan-genome characteristics reflect habitat diversity of different species, while this may also just indicate the various features by which different organisms acquire or lose genes. Together, this infers that *S. mutans* has a strong ability to acquire genes from its surroundings.

Core-Genome

In contrast to the pan-genome, estimation of the *S. mutans* core-genome indicates that genes shared in all strains decrease with each genome addition, until it finally reaches a plateau. In order to estimate the theoretical core-genome size that could be achieved using an infinite number of *S. mutans* genomes, an exponential regression core-genome model was applied [$F_c(n) = \kappa_c \exp(-n/\tau_c) + \Omega$) as proposed previously by Tettelin et al. [62] to fit the median data points of the core-genome sizes. The detailed information of fitting parameters is given in Online Resource 4. Using these fitting results to describe the core-genome of *S. mutans*, the theoretical core-genome size (Ω) was estimated to be ~1083 genes, which is fewer than other *Streptococcus* species. Previous studies have estimated that the core-genome size of *S. pyogenes* [67], *S. pneumoniae* [50], and *S. agalactiae* [62] to be 1400, 1647, and 1800 genes, respectively. Besides, Cornejo et al. [15] and Song et al. [57] determined that the core-genome size of *S. mutans* was 1490 and 1370 using 57 and 67 *S. mutans* genomes, respectively. This is clearly different from the core-genome size estimated in this study, although all of the *S. mutans*

genomes used by above Cornejo et al. and Song et al. were included in the calculations. These differences could be caused by a variety of reasons, such as different methods and parameter settings used for determining orthologs. Obviously, we have used a more stringent process to determine orthologs, leading to a smaller core genome size of *S. mutans* estimated.

Moreover, it is worth noting that the core gene number in each genome varies slightly due to the involvement of duplicated genes and paralogs between shared clusters, and it is not unusual to find variability in the number of core genes among strains. This once again implies genomic plasticity among *S. mutans* strains in different niches [43]. Detailed information of all the core genes in *S. mutans* UA159 is provided in Online Resource 5.

Furthermore, phylogenetic trees of the 183 *S. mutans* strains were constructed using various algorithms with the pan-genome profile (Online Resource 6b) and single SNP information of the core-genome (Online Resource 6a and Online Resource 7).

Distribution of Virulence Factors

The virulence factors in *S. mutans* play an important role in helping the organism to conquer various niches and inhibit competing species. Herein, all the virulence factors possibly related to the development of dental caries were collected (Online Resource 8–11). Then the distribution of these factors was analyzed among the 183 *S. mutans* strains involved in this study (Fig. 3).

First, *S. mutans* must adhere to the tooth surface and accumulate in sufficient numbers to cause dental caries. Therefore, its ability to colonize is considered to be central to the pathogenic capability of this organism [26]. Several studies have suggested that the adherence of *S. mutans* to the tooth surface involves both a sucrose-dependent (glucan-mediated accumulation) and independent (initial attachment) process [59]. Sucrose-independent adhesion is primarily mediated by cell surface antigen SpaP [9], which facilitates binding of the bacteria to the enamel pellicle, as well as by cell wall-associated protein WapA [55] and surface adhesion AdcA [18], a Zn-binding ABC transporter-type lipoprotein. All the genes which code for these three proteins are dispensable genes, especially *spaP*, whose CV is 74 and can be found in the genomes of *S. mutans* strains isolated from oral cavities with or without caries. Sucrose-dependent adhesion is mediated by synthesized glucan, which binds to the glucan receptor at the *S. mutans* membrane surface. The glucan receptors include glucan-binding proteins and glucosyltransferase (GTF) [70]. There are four types of glucan-binding proteins that are encoded by *gbpA*, *gbpB*, *gbpC*, and *gbpD*. *gbpD* has been suggested to encode an exoenzyme and was a core

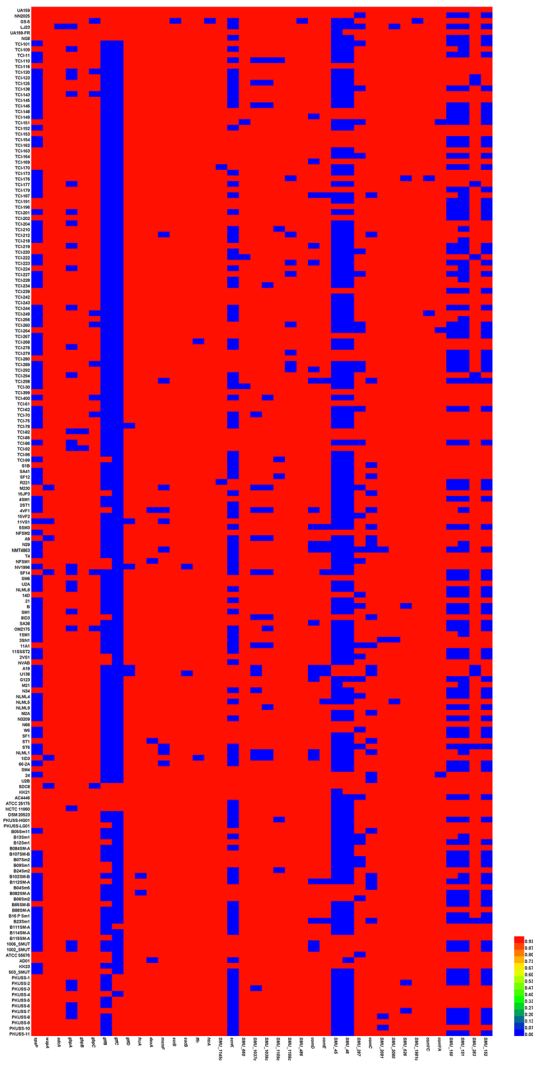

Fig. 3 Heatmap of the distribution of 40 putative cariogenic virulent genes which are not core genes across 183 *S. mutans* strains. These genes correspond to virulence factors listed in Online Resource 8–11 that are not in *bold*. Gene copy number of each virulence-related gene is indicated by the color key ranging from *blue* (absent) to *red*

gene of 183 *S. mutans* genomes based on the results of this study. In addition, the protein products of *gbpA* and *gbpC* enable cell aggregation in the presence of dextran, while g*bpB* encodes a peptidoglycan hydrolase which is also important for cell wall integrity and glucan binding. These three genes are dispensable genes and can be found in both caries-active and caries-free individuals. Meanwhile, GTF, another glucan receptor, catalyzes the synthesis of adhesive glucans from sucrose, and *S. mutans* has at least three different types of GTFs: GTF-I, GTF-SI, and GTF-S, which are encoded by the *gtfB*, *gtfC,* and *gtfD* genes, respectively. The genes *gtfB* and *gtfC* seem to play an important role for *S. mutans* virulence [61]. However, both of them are dispensable genes and are much less widely distributed than the *gtfD* gene. This may be related to the

use of a large number of draft genomes rather than the complete genomes involved in this, as the two genes could have been left out during sequencing and assembly. In addition, as the tandemly associated *gtfB* and *gtfC* genes could give rise to a new gene *gtfBC* following homologous recombination [69], the recombinant gene *gtfBC* might be omitted during annotation. Although the three kinds of GTFs could interact with each other and the ratio of different GTFs does affect the nature of the final glucan product [53], exactly how *S. mutans* uses the different GTFs remains unclear.

Once established on the tooth surface, *S. mutans* must produce acids in order to drive enamel demineralization. Therefore, acidogenicity is the second cariogenic feature of the organism. It is unsurprising that *S. mutans* is capable of fermenting a wide variety of carbohydrates and appears to metabolize sucrose to acid more rapidly than most any other Gram-positive organism [47]. All the genes which code for ten fermentation enzymes proved to be core genes in this study, implying that each of the *S. mutans* strains possesses a complete glycolytic pathway. However, no *S. mutans* strain is capable of aerobic respiration and the genes required for the aerobic electron transport chain are not present. In the glycolytic pathway, the intermediate product pyruvate is reduced to various fermentation products, especially lactic acid, which is main component that leads to decreased pH in dental plaques and is an initial factor in caries. Lactic acid production is catalyzed by lactate dehydrogenase, which is commonly regarded as a major virulence factors in *S. mutans*. Lactate dehydrogenase is encoded by the gene *ldh*, which is also one of the core genes, indicating its importance to the organism.

Like many bacteria, *S. mutans* has the capacity to accumulate storage polysaccharides when provided with excess carbohydrates. The accumulation of intracellular polysaccharides, a glycogen–like material, and of sucrose-derived extracellular polymers has been shown to lead directly to caries formation [12, 58]. Genes involved in storing intra/extracellular polysaccharides include *glgA*, *glgB*, *glgC*, and *glgD* and these four genes belong to the core-genome based on this study. The protein products are concerned with synthesizing glycogen, which may indicate that each *S. mutans* strain is capable of storing glycogen. Furthermore, the ability of accumulating glycogen may be necessary for the survival of the organisms during nutrient limitation in order to adapt to diverse habitats. Simultaneously, this property helps the species to continue fermentation in the absence of exogenous food supplies and may contribute to virulence.

As dental plaque becomes acidic during carbohydrate fermentation, the acid-tolerance of *S. mutans* also serves as a virulence factor. Comparisons of the relative aciduricity between oral bacteria have demonstrated that strains of *S. mutans* are more acid-tolerant than all other bacteria examined, with the exception of lactobacilli [28] and bifidobacteria [64]. This property appears to be based primarily on the presence of a membrane-bound, acid-stable, proton-translocating F0F1 ATPase that can maintain an intracellular pH of 7.5 [7]. The genes involved in this process include *atpA*, *atpB*, *atpC*, *atpD*, *atpE*, *atpF*, *atpG*, and *atpH*, which encode components of the ATPase and are all present as core-genes among *S. mutans* strains. In addition to this mechanism, acidic environmental changes can trigger a variety of adaptive responses in the organism. Genes previously demonstrated to be involved in the acid stress response include *dltC* [10], *ffh* [25], *dnaK* [38], and *sloR* [19]. Most of these are core genes, while others are widely distributed. For instance, the gene *ffh*, whose CV is 181 is only absent in the strains *S. mutans* TCI-268 and *S. mutans* 1ID3.

In addition, bacterial transduction two-component systems (TCSs) also play an important role in acid tolerance. A typical two-component regulatory system consists of a trans-membrane sensor histidine kinase (HK), which senses pH changes in the environment, and a cytoplasmic response regulator (RR), which enables the cell to respond via regulation of gene expression [60]. Although stand-alone genes ("orphans") coding for HKs or RRs have also been reported [3]. The *S. mutans* genome encodes 14 TCSs as well as an orphan RR [56], and detailed information of them is given in Online Resource 9. Most of the genes involved in TCSs are widely distributed. Genes coding for TCS-1, TCS-2, TCS-4, TCS-9, TCS-10, TCS-12, and Orphan RR1, as well as for RRs of TCS-3, TCS-5, TCS-6, and TCS-11 belong to the core-genome. All of these TCSs are related to acid tolerance as well as sucrose-dependent biofilm formation except for TCS-12, whose function is still unknown. Furthermore, TCS-2, TCS-5, TCS-6 and TCS-11 also play central roles in mutacin production, which will be expounded in the following section.

As is stated above, LGT is quite universal in *S. mutans* and contributes to to the open pan-genome of the species. It has been reported that LGT is greatly facilitated by competence development, a complex process involving multiple protein components and sophisticated regulatory networks that trigger the capacity of bacterial cells to take up exogenous DNA from the environment [49]. This phenomenon is frequently encountered in *S. mutans*, which resides in multispecies biofilm that harbors copious amounts of DNA released via cell lysis. Competence development-related systems found in *S. mutans* are listed in Online Resource 10. Among these proteins, ComX, which drives the transcription of the so-called "late-competence genes" required for genetic transformation plays a central role. In our study, both ComX and the "late-competence genes" regulated by ComX are highly conserved, which is consistent with previous researches [57]. On the

other hand, the distribution of genes involved in the upstream signal pathways regulating the activity of ComX varies. Some of them are presented as core-genes, such as SepM, which is an extracellular protease that is involved in the processing of 21-competence-stimulating peptide (CSP) to the mature 18-CSP [32]. Others are distributed less widely in *S. mutans*. For instance, ComB and ComC are missing from dozens of strains in our study, both from caries-active and caries-free individuals, indicating the complexity of the competence regulatory networks.

The last but not the least, bacteriocins produced by *S. mutans*, namely mutacins also contribute to the virulence of the species. Since *S.mutans* resides in dental plaque, a multispecies biofilm community that harbors approximately 700 different types of microorganisms [34]. To inhibit the growth of competing bacteria, *S. mutans* secretes a wide variety of mutacins. Distribution of identified mutacins and mutacin-immunity proteins is summarized in Online Resource 11. An interesting new result is that all of the genes coding for mutacin-V-related mutacins or mutacin-immunity proteins belong to the core-genome based on our study. It has been reported that mutacin-V is a single peptide and has broad antimicrobial activities, ranging from *Streptococcus mitis* to micrococcus [27]. Unlike mutacin-V, mutacin IV has been demonstrated to contain two peptides, NlmA and NlmB [6]. Both of them are required for optimum activity of mutacin IV. However, in our study genes encoding these two peptides are not so widely distributed, with a CV of 102 and 88, respectively.

All in all, similar to previous studies, the distribution of putative virulence genes in *S. mutans* strains does not correlate with caries status from the perspective of comparative genomics, which may suggest that there are several other factors contributing to cariogenesis. Nevertheless, we set an example for analyzing large datasets of pathogen genomes from the perspective of the pan-genome. This method could be also used for other organisms, such as lactobacillus to further explore the pathogenic microorganisms of dental caries.

Acknowledgements This study was supported by funding from Peking University School of Stomatology (PKUSS20130210).

Compliance with Ethical Standards

Conflicts of interest None.

References

1. Aikawa C, Furukawa N, Watanabe T et al (2012) Complete genome sequence of the serotype k *Streptococcus mutans* strain LJ23. J Bacteriol 194:2754–2755. doi:10.1128/JB.00350-12
2. Ajdić D, McShan WM, McLaughlin RE et al (2002) Genome sequence of *Streptococcus mutans* UA159, a cariogenic dental pathogen. Proc Natl Acad Sci USA 99:14434–14439. doi:10.1073/pnas.172501299
3. Alm E, Huang K, Arkin A (2006) The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. PLoS Comput Biol 2:e143. doi:10.1371/journal.pcbi.0020143
4. Argimón S, Caufield PW (2011) Distribution of putative virulence genes in *Streptococcus mutans* strains does not correlate with caries experience. J Clin Microbiol 49:984–992. doi:10.1128/JCM.01993-10
5. Argimón S, Konganti K, Chen H et al (2014) Comparative genomics of oral isolates of Streptococcus mutans by in silico genome subtraction does not reveal accessory DNA associated with severe early childhood caries. Infect Genet Evol 21:269–278. doi:10.1016/j.meegid.2013.11.003
6. Arthur RA, Cury AADB, Graner ROM et al (2011) Genotypic and phenotypic analysis of *S. mutans* isolated from dental biofilms formed in vivo under high cariogenic conditions. Braz Dent J 22:267–274
7. Bender GR, Sutton SV, Marquis RE (1986) Acid tolerance, proton permeabilities, and membrane ATPases of oral streptococci. Infect Immun 53:331–338
8. Biswas S, Biswas I (2012) Complete genome sequence of *Streptococcus mutans* GS-5, a serotype c strain. J Bacteriol 194:4787–4788. doi:10.1128/JB.01106-12
9. Bowen WH, Schilling K, Giertsen E et al (1991) Role of a cell surface-associated protein in adherence and dental caries. Infect Immun 59:4606–4609
10. Boyd DA, Cvitkovitch DG, Bleiweis AS et al (2000) Defects in D-alanyl-lipoteichoic acid synthesis in *Streptococcus mutans* results in acid sensitivity. J Bacteriol 182:6055–6065
11. Burne RA (1998) Oral streptococci. products of their environment. J Dent Res 77:445. doi:10.1177/00220345980770030301
12. Burne RA, Chen YY, Wexler DL et al (1996) Cariogenicity of *Streptococcus mutans* strains with defects in fructan metabolism assessed in a program-fed specific-pathogen-free rat model. J Dent Res 75:1572–1577. doi:10.1177/00220345960750080801
13. Cheon K, Moser SA, Whiddon J et al (2011) Genetic diversity of plaque mutans streptococci with rep-PCR. J Dent Res 90:331–335. doi:10.1177/0022034510386375
14. Conrads G, de Soet JJ, Song L et al (2014) Comparing the cariogenic species *Streptococcus sobrinus* and *S. mutans* on whole genome level. J Oral Microbiol 6:26189. doi:10.3402/jom.v6.26189
15. Cornejo OE, Lefébure T, Bitar PDP et al (2013) Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutans*. Mol Biol Evol 30:881–893. doi:10.1093/molbev/mss278
16. Delcher AL, Bratke KA, Powers EC et al (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23:673–679. doi:10.1093/bioinformatics/btm009
17. Deng W, Wang Y, Liu Z et al (2014) HemI: a toolkit for illustrating heatmaps. PLoS ONE 9:e111988. doi:10.1371/journal.pone.0111988
18. Dintilhac A, Claverys JP (1997) The adc locus, which affects competence for genetic transformation in Streptococcus pneumoniae, encodes an ABC transporter with a putative lipoprotein homologous to a family of streptococcal adhesins. Res Microbiol 148:119–131. doi:10.1016/S0923-2508(97)87643-7
19. Dunning DW, McCall LW, Powell WF et al (2008) SloR modulation of the Streptococcus mutans acid tolerance response involves the GcrR response regulator as an essential intermediary. Microbiology 154:1132–1143. doi:10.1099/mic.0.2007/012492-0

20. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–1584
21. Fan MW, Bian Z, Peng ZX et al (2002) A DNA vaccine encoding a cell-surface protein antigen of Streptococcus mutans protects gnotobiotic rats from caries. J Dent Res 81:784. doi:10.1177/0810784
22. Fischer S, Brunk BP, Chen F et al (2011) Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new Ortholog groups. Curr Protoc Bioinform 35:6.12.1–6.12.19. doi:10.1002/0471250953.bi0612s35
23. Fraser-Liggett CM (2005) Insights on biology and evolution from microbial genome sequencing. Genome Res 15:1603–1610. doi:10.1101/gr.3724205
24. Gogarten JP, Townsend JP (2005) Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol 3:679–687. doi:10.1038/nrmicro1204
25. Gutierrez JA, Crowley PJ, Cvitkovitch DG et al (1999) Streptococcus mutans ffh, a gene encoding a homologue of the 54 kDa subunit of the signal recognition particle, is involved in resistance to acid stress. Microbiology 145:357–366. doi:10.1099/13500872-145-2-357
26. Hamada S, Slade HD (1980) Biology, immunology, and cariogenicity of streptococcus mutans. Microbiol Rev 44:331
27. Hale JD, Ting YT, Jack RW et al (2005) Bacteriocin (mutacin) production by Streptococcus mutans genome sequence reference strain UA159: elucidation of the antimicrobial repertoire by genetic dissection. Appl Environ Microbiol 71:7613. doi:10.1128/AEM.71.11.7613-7617.2005
28. Harper DS, Loesche WJ (1984) Growth and acid tolerance of human dental plaque bacteria. Arch Oral Biol 29:843–848
29. Heaps HS (1978) Information retrieval: computational and theoretical aspects. Academic Press, New York
30. Hiller NL, Janto B, Hogg JS et al (2007) Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome. J Bacteriol 189:8186–8195. doi:10.1128/JB.00690-07
31. Hogg JS, Hu FZ, Janto B et al (2007) Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. Genome Biol 8:R103. doi:10.1186/gb-2007-8-6-r103
32. Hossain MS, Biswas I (2012) An extracelluar protease, SepM, generates functional competence-stimulating peptide in Streptococcus mutans UA159. J Bacteriol 194:5886–5896. doi:10.1128/JB.01381-12
33. Kreikemeyer B, McIver KS, Podbielski A (2003) Virulence factor regulation and regulatory networks in Streptococcus pyogenes and their impact on pathogen–host interactions. Trends Microbiol 11:224–232
34. Krishnan K, Chen T, Paster BJ (2016) A practical guide to the oral microbiome and its relation to health and disease. Oral Dis 23:276. doi:10.1111/odi.12509
35. Kryazhimskiy S, Plotkin JB (2008) The population genetics of dN/dS. PLoS Genet 4:e1000304. doi:10.1371/journal.pgen.1000304
36. Lagesen K, Hallin P, Rødland EA et al (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100–3108. doi:10.1093/nar/gkm160
37. Lefébure T, Stanhope MJ (2007) Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. Genome Biol 8:R71. doi:10.1186/gb-2007-8-5-r71
38. Lemos JA, Luzardo Y, Burne RA (2007) Physiologic effects of forced down-regulation of dnaK and groEL expression in *Streptococcus mutans*. J Bacteriol 189:1582–1588. doi:10.1128/JB.01655-06
39. Liao Y, Chen J, Brandt BW et al (2015) Identification and functional analysis of genome mutations in a fluoride-resistant *Streptococcus mutans* strain. PLoS ONE 10:e0122630. doi:10.1371/journal.pone.0122630
40. Loesche WJ (1986) Role of *Streptococcus mutans* in human dental decay. Microbiol Rev 50:353–380
41. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964
42. Luo R, Liu B, Xie Y et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18. doi:10.1186/2047-217X-1-18
43. Marri PR, Hao W, Golding GB (2006) Gene gain and gene loss in streptococcus: is it driven by habitat? Mol Biol Evol 23:2379–2391. doi:10.1093/molbev/msl115
44. Martinez-Medina M, Aldeguer X, Lopez-Siles M et al (2009) Molecular diversity of *Escherichia coli* in the human gut: new ecological evidence supporting the role of adherent-invasive *E. coli* (AIEC) in Crohn's disease. Inflamm Bowel Dis 15:872–882. doi:10.1002/ibd.20860
45. Maruyama F, Kobata M, Kurokawa K et al (2009) Comparative genomic analyses of Streptococcus mutans provide insights into chromosomal shuffling and species-specific content. BMC Genom 10:358. doi:10.1186/1471-2164-10-358
46. Mattos-Graner RO, Napimoga MH, Fukushima K et al (2004) Comparative analysis of Gtf isozyme production and diversity in isolates of *Streptococcus mutans* with different biofilm growth phenotypes. J Clin Microbiol 42:4586–4592. doi:10.1128/JCM.42.10.4586-4592.2004
47. Minah GE, Loesche WJ (1977) Sucrose metabolism by prominent members of the flora isolated from cariogenic and non-cariogenic dental plaques. Infect Immun 17:55–61
48. Moreno-Hagelsieb G, Latimer K (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics 24:319–324. doi:10.1093/bioinformatics/btm585
49. Morrison DA (1997) Streptococcal competence for genetic transformation: regulation by peptide pheromones. Microb Drug Resist 3:27. doi:10.1089/mdr.1997.3.27
50. Muzzi A, Donati C (2011) Population genetics and evolution of the pan-genome of Streptococcus pneumoniae. Int J Med Microbiol 301:619–622. doi:10.1016/j.ijmm.2011.09.008
51. Nakano K, Nomura R, Matsumoto M et al (2010) Roles of oral bacteria in cardiovascular diseases—from molecular mechanisms to clinical cases: cell-surface structures of novel serotype k *Streptococcus mutans* strains and their correlation to virulence. J Pharmacol Sci 113:120–125
52. Nomura R, Nakano K, Taniguchi N et al (2009) Molecular and clinical analyses of the gene encoding the collagen-binding adhesin of *Streptococcus mutans*. J Med Microbiol 58:469–475. doi:10.1099/jmm.0.007559-0
53. Ooshima T, Matsumura M, Hoshino T et al (2001) Contributions of three glucosyltransferases to sucrose-dependent adherence of *Streptococcus mutans*. J Dent Res 80:1672–1677. doi:10.1177/00220345010800071401
54. Phattarataratip E, Olson B, Broffitt B et al (2011) *Streptococcus mutans* strains recovered from caries-active or caries-free individuals differ in sensitivity to host antimicrobial peptides. Mol Oral Microbiol 26:187–199. doi:10.1111/j.2041-1014.2011.00607.x
55. Russell MW, Harrington DJ, Russell RR (1995) Identity of Streptococcus mutans surface protein antigen III and wall-associated protein antigen A. Infect Immun 63:733–735
56. Song L, Sudhakar P, Wei W et al (2012) A genome-wide study of two-component signal transduction systems in eight newly sequenced mutans streptococci strains. BMC Genom 13:128. doi:10.1186/1471-2164-13-128

57. Song L, Wang W, Conrads G et al (2013) Genetic variability of mutans streptococci revealed by wide whole-genome sequencing. BMC Genom 14:430. doi:10.1186/1471-2164-14-430
58. Spatafora G, Rohrer K, Barnard D et al (1995) A Streptococcus mutans mutant that synthesizes elevated levels of intracellular polysaccharide is hypercariogenic in vivo. Infect Immun 63:2556–2563
59. Staat RH, Langley SD, Doyle RJ (1980) Streptococcus mutans adherence: presumptive evidence for protein-mediated attachment followed by glucan-dependent cellular accumulation. Infect Immun 27:675–681
60. Stock AM, Robinson VL, Goudreau PN (2000) Two-component signal transduction. Annu Rev Biochem 69:183–215. doi:10.1146/annurev.biochem.69.1.183
61. Tamesada M, Kawabata S, Fujiwara T et al (2004) Synergistic effects of streptococcal glucosyltransferases on adhesive biofilm formation. J Dent Res 83:874–879. doi:10.1177/154405910408301110
62. Tettelin H, Masignani V, Cieslewicz MJ et al (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci USA 102:13950–13955. doi:10.1073/pnas.0506758102
63. Tettelin H, Riley D, Cattuto C et al (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11:472–477
64. Valdez RM, Dos Santos VR, Caiaffa KS et al (2016) Comparative in vitro investigation of the cariogenic potential of bifidobacteria. Arch Oral Biol 71:97. doi:10.1016/j.archoralbio.2016.07.005
65. Van HJ (1994) Role of micro-organisms in caries etiology. J Dent Res 73:672. doi:10.1177/00220345940730031301
66. Waterhouse JC, Russell RR (2006) Dispensable genes and foreign DNA in *Streptococcus mutans*. Microbiology 152:1777–1788. doi:10.1099/mic.0.28647-0
67. Waterhouse JC, Swan DC, Russell RRB (2007) Comparative genome hybridization of *Streptococcus mutans* strains. Oral Microbiol Immunol 22:103–110. doi:10.1111/j.1399-302X.2007.00330.x
68. Wu C, Cichewicz R, Li Y et al (2010) Genomic island TnSmu2 of *Streptococcus mutans* harbors a nonribosomal peptide synthetase-polyketide synthase gene cluster responsible for the biosynthesis of pigments involved in oxygen and H_2O_2 tolerance. Appl Environ Microbiol 76:5815–5826. doi:10.1128/AEM.03079-09
69. Yamashita Y, Bowen WH, Kuramitsu HK (1992) Molecular analysis of a *Streptococcus mutans* strain exhibiting polymorphism in the tandem gtfB and gtfC genes. Infect Immun 60:1618–1624
70. Yamashita Y, Bowen WH, Burne RA et al (1993) Role of the *Streptococcus mutans* gtf genes in caries induction in the specific-pathogen-free rat model. Infect Immun 61:3811–3817
71. Zhang L, Foxman B, Drake DR et al (2009) Comparative whole-genome analysis of Streptococcus mutans isolates within and among individuals of different caries status. Oral Microbiol Immunol 24:197–203. doi:10.1111/j.1399-302X.2008.00495.x