**ORAL MICROBIOLOGY
AND IMMUNOLOGY**

Comparative whole-genome analysis of *Streptococcus mutans* isolates within and among individuals of different caries status

**L. Zhang[1], B. Foxman[1], D. R. Drake[2],
U. Srinivasan[1], J. Henderson[1],
B. Olson[2], C. F. Marrs[1], J. J. Warren[3],
M. L. Marazita[4]**

[1]Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI, USA, [2]Department of Endodontics and the Dows Institute for Dental Research, University of Iowa, Iowa City, IA, USA, [3]Department of Preventive and Community Dentistry, University of Iowa, Iowa City, IA, USA, [4]Center for Oral Health Research in Appalachia (COHRA), Center for Craniofacial and Dental Genetics, Departments of Oral Biology, Human Genetics and Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA

Zhang L, Foxman B, Drake DR, Srinivasan U, Henderson J, Olson B, Marrs CF, Warren JJ, Marazita ML. Comparative whole-genome analysis of Streptococcus mutans *isolates within and among individuals of different caries status.*
Oral Microbiol Immunol 2009: 24: 197–203. © 2009 John Wiley & Sons A/S.

Introduction: Genotypic analyses of *Streptococcus mutans* using fingerprinting methods depend on a few genetic loci being different but do not reveal the underlying genome-wide differences between strains.
Methods: We used comparative genomic hybridization (CGH) with 70-mer oligonucleotide microarrays containing open reading frames (ORFs) from *S. mutans* strain UA159 to examine the genetic diversity of 44 isolates from nine children selected from a local study population in Eastern Iowa.
Results: Unique strains (clones) within each child initially identified by arbitrary-priming polymerase chain reaction were confirmed by CGH. There was a wide range of variation in the hybridization patterns of the 1948 ORFs among the test isolates examined. Between 87 and 237 ORFs failed to give a positive signal among individual isolates. A total of 323 of the UA159 ORFs were absent from one or more of the test strains. These 323 variable genes seemed to be distributed across the entire UA159 genome and across all the predicted functional categories.
Conclusion: This set of very close geographically and temporally collected *S. mutans* isolates had a degree of gene content variation as high as a previously examined global set of strains. Comparing the frequency of these variable genes, the majority of which have unknown function, among strains of different origins (i.e. different caries status) could help to determine their relevance in *S. mutans* cariogenicity.

Streptococcus mutans is a bacterial resident of the oral cavity and is considered to be the principal etiological agent of dental caries in humans (13, 23). As *S. mutans* can be isolated from individuals either with or without a history of caries (4, 25), differences in the colonizing *S. mutans* may contribute to the eventual development and progression of caries among different individuals. One difference may be in the bacterial load. High levels of *S. mutans* have been associated with caries (7, 12, 18, 33); however, many people with a high *S. mutans* load because of high-level sucrose consumption do not develop caries (25, 29). Another important factor may be the characteristics of *S. mutans*. Certain *S. mutans* attributes that are believed to be important in caries development, including acidogenicity, aciduricity, biofilm-formation potential and production of glucans, are known to vary among isolates of *S. mutans* and may determine the differences in caries

experience (23, 24, 28, 30, 38). However, correlation of these specific factors with caries in humans has been difficult to establish. Nevertheless, *S. mutans* diversity remains an important area of study because it provides one possible explanation for the contrasting caries status in people colonized with *S. mutans*. The challenge is that the amount of *S. mutans* diversity and its genetic basis are still not well understood.

Molecular fingerprinting techniques such as restriction length fragment polymorphisms, multilocus enzyme electrophoresis, ribotyping and arbitrary-priming polymerase chain reaction (AP-PCR) have been used to demonstrate that *S. mutans* strains are genetically diverse. A large number of clonal types of *S. mutans* have been identified; these vary across both carious and healthy populations (6, 8, 10, 11, 20–22, 26, 34, 35, 39). In a study of *S. mutans* from over 200 individuals, no two individuals exhibited the same clonal type profile unless they were related to each other (5). Genetic fingerprinting methods also demonstrated that transmission of *S. mutans* can occur from parents to children (39), and between children attending the same nursery school class (3). Using restriction length fragment polymorphisms Kulkarni et al. (19) were the first to show that an individual could carry multiple genotypes. Multiple genotypes were isolated from child–parent pairs (27), and from young adults (31). Several studies have found that subjects with active caries had more genotypes than caries-free subjects (2, 9, 15, 31) although others have found the opposite (17). The role of concurrent colonization of multiple genotypes in caries development is unknown. It has been suggested, however, that increased risk of caries in subjects with multiple genotypes could be the result of simultaneous action of strains with different cariogenic potentials (2).

While the above observations were informative, current typing methods provide only limited insight into *S. mutans* genetic diversity. Typing methods survey only a single or a few genetic loci; hence the extent of the genetic differences across the whole genome in *S. mutans* clones is not known. Fingerprinting techniques based on variation in housekeeping genes or on random variations of anonymous loci within the genome are unable to determine whether *S. mutans* clones vary in their overt virulence and concomitant cariogenicity. By contrast, comparative genomic hybridization provides a genome-wide snapshot of differences in *S. mutans* gene content, including those potentially important in virulence. The first comparative genomic hybridization (CGH) study of nine representative *S. mutans* strains revealed extensive heterogeneity in gene content among strains from different geographical locations around the world (42). Considering the genetic differences observed in isolates collected from geographically dispersed populations, we were interested to observe the level of heterogeneity in a single geographic site. CGH provided the opportunity to investigate how many genomic differences exist among multiple genotypes seen within a single individual and the degree of variability and difference among isolates from sites with and without caries. In this study, we used CGH to determine the genetic diversity among multiple *S. mutans* isolates obtained from nine children from a single local study population. Our objectives were to investigate the extent of genome-wide differences among *S. mutans* strains that were geographically and temporally related and to compare AP-PCR and CGH methods for genotyping of *S. mutans* strains.

Materials and methods
Bacterial strains

Clinical *S. mutans* isolates were collected from nine children randomly selected from the 200 participants in a study of caries-inactive and caries-active preschool-age children enrolled in the Head Start program in Eastern Iowa (Table 1). *S. mutans*

Table 1. Streptococcus mutans strains used in this study

Subject ID	DMFS (baseline/ follow up)	Isolate ID	AP-PCR clone type	CGH genotype	No. of deleted ORF by CGH
7	16	AW107-3	AW107CT1	CGH-1	87
		AW107-4	AW107CT1	CGH-1	87
30	3	AW1301-3	AW1301CT1	CGH-2	168
		AW1301-4	AW1301CT1	CGH-2	168
		AW1301-5	AW1301CT1	CGH-2	168
		AW1301-7	AW1301CT1	CGH-2	168
		AW1301-8	AW1301CT1	CGH-2	168
		AW1301-10	AW1301CT1	CGH-2	168
31	1	AW1302-8	AW1302CT1	CGH-3	140
		AW1302-9	AW1302CT1	CGH-3	140
52	0	ARM07-1	ARMCT1	CGH-4	208
		ARM07-2	ARMCT1	CGH-4	208
		ARM07-3	ARMCT1	CGH-4	208
		ARM07-4	ARMCT1	CGH-4	208
		ARM07-5	ARMCT1	CGH-4	208
		ARM07-7	ARMCT1	CGH-4	208
		ARM07-8	ARMCT1	CGH-4	208
		ARM07-9	ARMCT1	CGH-4	208
		ARM07-10	ARMCT1	CGH-4	208
56	6	ARM11-1	ARM11CT1	CGH-5.a	155
		ARM11-3	ARM11CT1	CGH-5.a	155
		ARM11-4	ARM11CT1	CGH-5.b	152
		ARM11-5	ARM11CT1	CGH-5.b	152
		ARM11-6	ARM11CT1	CGH-5.a	155
		ARM11-8	ARM11CT1	CGH-5.a	155
		ARM11-9	ARM11CT1	CGH-5.b	152
		ARM11-10	ARM11CT1	CGH-5.a	155
		ARM11-12	ARM11CT1	CGH-5.a	155
63	56	BRE04-2	BRE04CT1	CGH-6	167
		BRE04-6	BRE04CT1	CGH-6	167
		BRE04-7	BRE04CT1	CGH-6	167
		BRE04-9	BRE04CT2	CGH-7	192
		BRE04-10	BRE04CT1	CGH-6	167
78	16	DUR08-1	DUR08CT1	CGH-8.a	237
		DUR08-2	DUR08CT1	CGH-8.b	235
		DUR08-5	DUR08CT1	CGH-8.a	237
		DUR08-6	DUR08CT2	CGH-9	178
		DUR08-7	DUR08CT2	CGH-9	178
		DUR08-8	DUR08CT1	CGH-8.b	235
		DUR08-10	DUR08CT2	CGH-9	178
112	0	LAM07-3A	LAM07CT1	CGH-10	113
		LAM07-3B	LAM07CT1	CGH-10	113
211	1	SC106-3	SC106CT1	CGH-11	126
		SC106-3	SC106CT1	CGH-11	126

AP-PCR, arbitrary-priming polymerase chain reaction; CGH, comparative genomic hybridization; DMFS, decayed/missing/filled surface; ORF, open-reading frame.

was isolated by spiral-plating original plaque samples in transport media at appropriate dilutions onto Mitis-Salivarius agar + Kanamycin and Bacitracin (i.e. SKMB, a selective medium for *S. mutans*). Initial isolates from this selective/differential medium were then processed for initial identification using a Quad-plate system for the assessment of specific enzyme activities – fermentation of mannitol, sorbitol, raffinose and salicin, and arginine decarboxylase activity. *S. mutans* identification was then finalized by PCR using species-specific sequences described by Oho et al. (32). Up to 10 random colonies from each plaque sample were chosen for this sequential selective scheme. The final number of confirmed *S. mutans* isolates from each child ranged from 2 to 9. A total of 44 isolates from these nine children were included in our analysis. *S. mutans* UA159 is the strain for which the genome has been sequenced (1). We obtained UA159 from the American Type Culture Collection (Manassas, VA) and used it as the reference strain in our microarray experiments.

AP-PCR fingerprinting

The AP-PCR technique was used to obtain DNA fingerprints for all *S. mutans* isolates. AP-PCR was performed as described elsewhere (22) with some modifications. Amplification was conducted in a thermocycler (PTC-200, MJ Research Inc., Waltham, MA) programmed with the following temperature profile: initial 5 min at 94°C, followed by 45 cycles of 30 s at 94°C (denaturation), 30 s at 36°C (annealing), and 1 min at 72°C (elongation). The final elongation cycle was 10 min at 72°C. The amplification mixture (master mix) consisted of: 25.50 μl RNase-free H_2O, 5 μl 10× amplification buffer, 1.0 μl dNTP (10 mM each), 14.0 μl MgCl$_2$ (25 μM), 1 μl 10 μM primers OPA-2 and OPA-13 in separate runs, and 0.50 μl *Taq* polymerase. A 2-μl sample of purified *S. mutans* DNA (25 ng/μl) was added to the master mix. Gel electrophoresis was carried out on the PCR samples. The gel was stained with ethidium bromide and visualized under ultraviolet light. Isolates were considered to belong to the same clonal type if both OPA-2 and OPA-13 fingerprints were identical.

DNA isolation and labeling

Genomic DNA from *S. mutans* was isolated using a GenElute Bacterial Genomic DNA kit (Sigma, St Louis, MO) and sheared into 200- to 1000-base pair fragments in a cup sonicator (Misonix Sonicator 3000 at setting 4 for 20 s). A total of 4 μg sheared DNA per sample was fluorescently labeled with BioPrime Plus Array CGH indirect Genomic Labeling System (Invitrogen, Carlsbad, CA) according to the manufacturer's protocol, using a modified 25× deoxynucleoside triphosphate mix consisting of 12.5 mM each of dAPT, dGTP and cCTP; 2.1 mM dTTP; and 10.4 mM aminoallyl-dUTP. Fluorescence-labeled DNA was purified in Qiagen PCR purification columns. DNA and dye concentrations were determined using a Beckman spectrophotometer.

Oligonucleotide microarrays

S. mutans microarray slides were obtained through the National Institute of Dental and Craniofacial Research (NIDCR) Oral Microbe Microarray Initiative, a collaborative effort between the NIDCR and the National Institute of Allergy and Infectious Disease (NIAID) Pathogen Functional Genomics Resource Center (PFGRC) located at the J. Craig Venter Institute (Rockville, MD). Open reading frames (ORFs) from the *S. mutans* UA159 genome were represented by unique 70-mer oligonucleotides on the genome slides. The UA159 genome currently has 1960 identified ORFs, of which 1948 are represented on the *S. mutans* array.

Microarray hybridization and data process

For each comparative genomic hybridization, equal amounts of DNA (3 μg) from a test strain and the reference strain UA159, labeled with different fluorescence dyes, were combined. Combined DNA was concentrated using a YM-30 Microcon column (Millipore, Billerica, MA) and mixed with hybridization solution to a final volume of 60 μl. Prehybridization, hybridization and posthybridization washes were carried out following the protocol provided by the PFGRC where genome slides were produced (http://pfgrc.jcvi.org/index.php/microarray/protocols.html) with one exception: the ArryIt HybIt hybridization solution from Tele-Chem International was used instead of the suggested formamide-based buffer and hybridization was performed at 62°C. Test strains were hybridized twice with UA159 as the reference on two slides: once with the test strain Alexa Fluor 555 (Cy3 equivalent dye) labeled and UA159 Alexa Fluor 647 (Cy5 equivalent dye) labeled and once with the test strain Alexa Fluor 647 labeled and UA159 Alexa Fluor 555 labeled.

Arrays were scanned with a Chipreader laser scanner (BioRad, Hercules, CA) at 10-μm resolution at 532 and 635 nm. Resulting images were processed using SPOTFINDER software from The Institute for Genomic Research (TIGR) to retrieve probe intensities with a provided annotation file associated with the slide. Normalization (log mean centering) and filtering were performed using TIGR Microarray Data Analysis System (MIDAS). The data quality and the normalization effects were assessed by examining plots of M against A [M = log (reference/test); A = log (reference × test)/2]. The final viewing and clustering of processed data were performed using TIGR MultiExperiment Viewer (MeV). Dye swaps were treated as replicates after normalization. To derive a cut-off point for deletion detection, signal ratios on all ORFs from six UA159 self–self competitive hybridizations were examined. All spots had a log2 intensity ratio (reference/reference) less than 2 or greater than – 2 with the majority concentrated around 0. The range of the signal ratio on reference-test hybridizations was also examined. A set of 50 ORFs representing the range of the signals ratio were selected and their presence in test strains was examined by PCR (see below). Based on these analyses, deleted ORFs were defined as having an average log2 intensity ratio (reference/test) greater than two. Over 200 additional PCR verifications were performed throughout this study and we found that this cut-off point gave low rates of <0.7% false negatives (i.e. ORFs tested positive by PCR but classified by the CGH as absent).

PCR verification

PCR amplifications of selected ORFs in test strains were used to validate microarray hybridization results or to resolve ambiguous hybridization results. Using the UA159 genome sequence, primers internal to the target ORF (or flanking the target ORF if the target ORF was too small) were designed using Lasergenes (DNASTAR, Madison, WI). The PCR amplification was carried out in 25 μl reaction comprising 1× PCR AccuPrime SupreMix II (Invitrogen), 1 μM of each primer, 50 ng DNA template in a PTC-200 DNA Engine thermal cycler (MJ Research). The thermocycling parameters for PCR were as follows: initial denaturation for 2 min at 94°C, 30 cycles of 30 s at 94°C, 30 s at annealing temperature (5°C below the lower melting

temperature (T_m) of the two primers), and 1 min at 72°C, and a final 10 min extension at 72°C. When a PCR failed to generate the expected product, the PCR of the 16S ribosomal RNA gene was performed to confirm the quantity and quality of the DNA templates. Those PCR that were negative for targeted ORF but positive for the 16S RNA gene were considered truly negative for targeted ORFs.

Results

Correlation between genotypes revealed by AP-PCR and CGH

For each of the nine children, multiple *S. mutans* isolates (2–9) were genotyped. DNA fingerprints obtained by AP-PCR using two primers indicated that multiple isolates within seven of the nine children were identical. For the remaining two children, both with high levels of caries, two unique fingerprints were found among five and seven isolates, respectively. Comparing the AP-PCR fingerprints of isolates across children, a total of 11 unique AP-PCR clones could be established and each was assigned a clonal designation (Fig. S1 and Table S1).

AP-PCR genotyping results correlated well with the CGH genotypes based on hybridization patterns of 1948 ORF probes. One CGH clonal type was found in the seven children with a single AP-PCR type, and two CGH clonal types were found in the two children with multiple AP-PCR types. CGH indicated that the two clonal types found in each of the two children were very different, with hybridization pattern differences for 97 and 115 ORFs, respectively. While the extent of differences among clones was difficult to access based on AP-PCR fingerprints generated by gel images, CGH provided a clear genome-wide view of the similarity/difference among *S. mutans* clones from all children. The LAM07CT1 and SC106CT1 were the most closely related clones because they differed by only 25 ORFs at eight genomic loci. By contrast, LAMCT1 differed from DURCT1 by 184 ORFs and these were the most distant clones among all pairwise comparisons.

CGH also showed that minor differences may exist among multiple isolates of the same clone within a child. Among nine isolates from child number 56 designated as ARM11CT1 clone, CGH indicated that there were differences in hybridization patterns for three ORFs: SMU.41, SMU.343, and SMU.1231c. Hybridization differences for two ORFs, SMU.343 and

SMU.176, were also observed among four isolates designated as DUR08CT1. These differences were confirmed by PCR amplification using primer pairs specific to these loci (Fig. S2). Both ARMCT1 and DUR08CT1 were from children with high levels of caries.

Gene content variation among unique test strains

Among the 1948 ORFs within the UA159 genome represented on the microarray, 323 (16.6%) were identified as variable ORFs because they failed to give a positive signal in at least one test genome of the 11 unique strains (Table S1). These variable ORFs were either absent from the test strain or their sequence was sufficiently divergent from that of UA159 to produce positive hybridization signals. The numbers of ORFs absent from individual strains varied significantly, ranging from 87 (4.5%) in clone AW107CT1 to 237 (12.0%) in clone DUR08CT1 (Table 1). The majority of variable ORFs failed to hybridize with more than one strain. Forty-four (13.6%) of the variable ORFs were absent from only one strain and 18 (5.6%) were absent from all 11 strains.

The 323 identified variable ORFs were distributed along the UA159 genome (Fig. 1). However, only 15 (4.6%) of them corresponded to a stand alone deletion without another adjacent variable ORF. Most variable ORFs formed deletion blocks involving variable numbers of ORFs. Occasionally, absent individual variable ORFs or deletion blocks were separated by only a few ORFs. When these variable ORFs separated by no more than two ORFs were considered to belong to the same blocks, the 323 variable ORFs were distributed among 50 genome loci. Loci with large numbers of variable ORFs tended to have a G + C composition atypical for the *S. mutans* genome (Fig. 1). The two largest deletion blocks had 34 ORFs (SMU.191c–SMU.266c) and 38 ORFs (SMU.1329–SMU.1374), respectively. These two blocks corresponded to two genomic islands TnSmu1 and TnSmu2, identified previously (1). These islands were not necessarily present or absent in their entirety in individual strains. Regions corresponding to these genomic islands in individual strains varied in length and complement of ORFs.

Some of the genomic loci potentially associated with important *S. mutans* phenotypes varied among strains. SMU.100–SMU.116, a genomic island associated
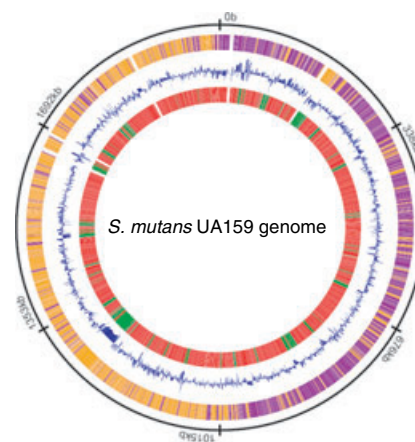

Fig. 1. *Streptococcus mutans* genome features and the distribution of dispensable genes (identified by comparative genomic hybridization) along the circular genome. Going outward: circle 1: variable genes represented by green areas and conserved gene by red; circle 2: percentage GC content; circle 3: predicted genes in positive (violet) and negative (orange) DNA strands; circle 5: genome coordinates. The diagram was created using the MICROBIAL GENOME VIEWER 1.0 (16).

with carbohydrate metabolism, includes genes for sorbose-phosphotransferase and fructose-specific phosphotransferase systems. Four strains lost the genes for fructose and three lost the genes for both uptake systems. One strain also lost the mannose-specific phosphotransferase genes SMU.1878 and SMU.1879. Several genomic loci associated with bacteriocin production also showed heterogeneity. At least one ORF in regions encoding Mutacin 1 (SMU.651–SMU.658) was lost in four strains. Half of the ORFs in region SMU.1803–SMU.1818 associated with *scn* bacteriocin were absent in seven strains. In the region associated with *blp* bacteriocins, only two strains have the same complement of ORFs SMU.1888–SMU.1917 as UA159. The other strains lost 1–15 of the ORFs in this region.

To investigate which functional groups variable ORFs belong to, we classified variable ORFs into clusters of orthologous groups (COGs) (36). Figure 2 shows the number of variable ORFs in each COG category. A large number of variable ORFs could not be classified and are of unknown function. A considerable number of the variable ORFs are involved in DNA replication, recombination and repair, and in carbohydrate transport and metabolism. Relatively few variable ORFs were found to be involved in amino acid nucleotide transport and metabolism, cell wall/membrane biogenesis, translation and coenzyme transport and metabolism.
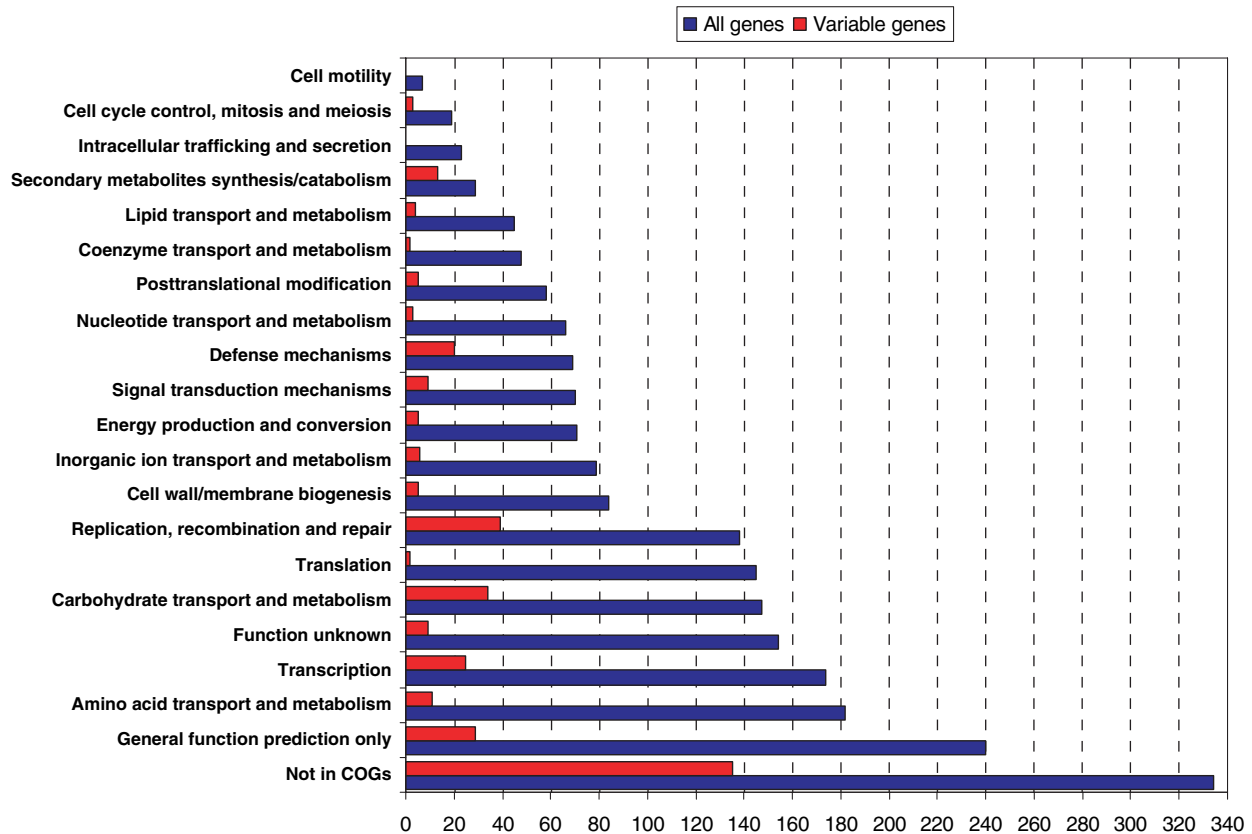
Fig. 2. Number of all (blue) and dispensable (red) genes in *Streptococcus mutans* strain UA159 classified in each of the clusters of orthologous genes (COG) category plus those not classified in the COG database. Dispensable genes are defined as those deleted in at least one test genome. There are 615 genes within the UA159 genome classified as not in COGs and the bar for the corresponding category in the figure is truncated.

Comparison between strains from children with high vs. low caries

Among 11 unique test strains, six were from children with high caries levels and five were from children with no or low-level caries. The number of variable ORFs deleted in high-level caries strains ranged from 87 to 237, and was 113 to 208 for low-level caries strains. There was no correlation between the numbers of variable ORFs and the decayed/missing/filled surface (DMFS) scores of the children. Since a deletion block with multiple variable ORFs is probably the result of a single genetic event at a specific genomic location, strains can also be compared by the number of altered loci instead of individual ORFs. Between 19 and 28 (mean 22) variable genome loci were found in low-level caries strains and 14–36 (mean 26) for the high-level caries strains.

No ORFs were found to be present in all six high-level caries strains but absent in all five low-level caries strains and vice versa. Twenty ORFs were deleted in three to five low-level caries strains but only in

zero to two high-level caries strains. For example, SMU.1896c, a hypothetical protein, was deleted in all five low-level caries strains but only in one high-level caries strain. Similarly, an even larger number of ORFs were deleted in high-level than in low-level caries strains.

Discussion

Genetic fingerprinting of *S. mutans* isolates by AP-PCR and other typing methods indicates that *S. mutans* strains overall are very genetically heterogeneous. While one completely annotated genome sequence of *S. mutans* has been available since 2002, much remains to be discovered in our understanding of the genetic basis of *S. mutans* diversity. We used CGH to determine the genetic variation among local *S. mutans* strains that were geographically and temporally closely related in 44 isolates collected from children enrolled in the Eastern Iowa Head Start program. We found some key and quite interesting findings on *S. mutans* isolates from this Iowa Cohort. Primarily, we found a high degree of gene content variations in this set

of *S. mutans* strains. Second, we found that AP-PCR and CGH were equally good at identifying different clones of *S. mutans* within this set of isolates. Where more than one clone was identified within a single person by AP-PCR, it was confirmed by CGH that these strains were genetically distantly related.

A similar degree of genetic variation was observed in a previous comparative genomic study on nine selected *S. mutans* strains collected between 1966 and 2001 from three continents (41, 42). In this previous CGH study, the number of ORFs missing among nine strains ranged from 113 to 227 (calculated using accompanied online supplemental data) compared to 87 to 237 in our 11 strains. While both our study and published studies found a strain with the same number of deleted ORFs (113), two strains actually had very different sets of 113 absent ORFs (only 67 ORFs were in common) and therefore were different strains. The total number of variable ORFs, defined as absent in at least one *S. mutans* strain, was 323 in our study. Similarly, the total variable ORFs found in nine global strains by Waterhouse et al.

was 385 (42). Three hundred of these variable ORFs were common in both our study and the published studies. Variable ORFs identified in only one study were distributed among variable regions across the genome. The largest block of variable ORFs unique to our study consisted of three ORFs, SMU.1315c to SMU.1317c, encoding a putative ATP-binding protein and two hypothetical proteins. The largest block of variable ORFs unique to the Waterhouse et al. study consisted of eight ORFs (SMU.132 to SMU.143) encoding putative transcriptional regulators and metabolic enzymes. These eight ORFs, however, were consistently present in all our strains. Overall, gene content diversity among *S. mutans* from our local sample was similar to that of the global sample, suggesting that *S. mutans* has a fluid and rapidly evolving genome structure that is likely to be similar to that found in other *Streptococci*. Previous studies on *Streptococcus pneumoniae* and *Streptococcus agalactiae* have shown a great gene content variability. An analysis by Hiller et al. on 17 *S. pneumoniae* genomes indicated that differences per strain pair ranged from 35 to 629 orthologous gene clusters, with each strain's genome containing between 21 and 32% non-core genes (14). Similarly, Tettelin et al. analysed the genomes of eight *S. agalactiae* strains and showed that ∼20% of the genes are not shared among all strains (37).

Our CGH experiments were limited to the detection of genes that are present in the UA159 genome. Other *S. mutans* are likely to carry many additional novel ORFs and genomic islands. Assuming a genome size similar to UA159 and an average size of 1 kilobase for each *S. mutans* gene, we estimate that our test strain, which has 237 absent ORFs by hybridization, could have as many as 237 kilobases of new gene-coding DNA sequences in its genome. This 237-kilobase of DNA sequence represents almost 12% of the 2.03 Megabase UA159 genome; this difference is likely bigger than expected based on natural variations in the genome size. As more *S. mutans* strains are sequenced we will be able to infer more about genome size variations among *S. mutans*. In other *Streptococcus* genomes, sizes ranged 2.13–2.2 for three completely annotated genomes of *Streptococcus agalactiae*, and from 1.84 to 1.94 for 11 completely annotated genomes of *Streptococcus pyogenes* (NCBI). These variations were within 5% of the average genome sizes of these species and well below the 12% we calculated with 11 test

S. mutans. The *S. mutans* strains showing the greatest number of absent ORFs by CGH will be the prime candidates for additional genome sequencing efforts as the probability of discovering novel genes in these strain is high. Using a finite-supragenome model, Hiller et al. predicted that 33 representative *S. pneumoniae* genomes needed to be sequenced to identify 99% of the orthologous gene clusters that are represented in the *S. pneumoniae* population at frequencies of 0.1 (14). Therefore, a substantial number of *S. mutans* genomes need to be sequenced to identify most *S. mutans* genes.

Variable ORFs were distributed across 19 of the 21 COG functional categories. Variable ORFs were over-represented in COG groups involved in defense mechanisms, carbohydrate transport and metabolism, secondary metabolites biosynthesis, and replication, recombination and repair. By contrast, COG categories involved in basic cellular machinery were not represented in variable ORFs to a significant extent. The survival of a pathogen in new environmental niches is increasingly associated with genes that are involved in essential metabolic and catabolic pathways (40). Many of these so called 'life-style genes' have been identified in the variable ORFs of *S. mutans* and may be important in determining pathogenicity of *S. mutans*. These variable genes are often acquired by horizontal transmission of pathogenicity islands that also retained genes such as transposases, integrases and potential ruminants of integrated plasmid. This may explain why a large number of variable genes are associated with replication, recombination and repair.

Comparison of the genetic diversity of *S. mutans* strains obtained from CGH and AP-PCR shows that when performed properly, AP-PCR fingerprinting can reliably distinguish multiple *S. mutans* clones within individual subjects. In our case, it also distinguished clones from different individuals. However, it is difficult to assess the degree of relatedness between any two clones based on fingerprinting patterns in any informative way given that band patterns do not give information about gene content.. CGH provided a genome-wide view of the difference in gene content.

When multiple isolates from two distinct clones from two children with high levels of caries were analysed, CGH indicated minor differences in genetic content within each clone (as determined by differences in hybridization patterns) in a few ORFs. These differences may

represent recent evolutionary changes resulting from active recombination or deletion within the local *S. mutans* population. Further tests are needed to reproduce this finding. If proven to be common phenomena, it would be interesting to examine whether the rate of microevolution differs between caries-active and caries-free populations.

While variation among the 11 unique *S. mutans* strains tested was large, the difference between the six strains from children with high levels of caries and five strains from children with low levels of caries were not significant given the small sample size; however, we made several interesting observations that warrant further investigation. The gene for glucan-binding protein, *gbpA* was present in all 11 strains. Several genes associated with bacteriocin production, carbohydrate metabolism and many of unknown function were differentially deleted in some but not all low- and high-level caries strains. Using a high throughput microarray platform, Library on a Slide (43), we are currently investigating the prevalence of these genes in a population-based sample of *S. mutans* collected from people with and without caries to evaluate whether these genes are associated with *S. mutans* cariogenicity.

Acknowledgments

References

1. Ajdic D, McShan WM, McLaughlin RE et al. Genome sequence of *Streptococcus mutans* UA159, a cariogenic dental pathogen. Proc Natl Acad Sci USA 2002: **99**: 14434–14439.
2. Alaluusua S, Mättö J, Grönroos L et al. Oral colonization by more than one clonal type of mutans streptococcus in children with nursing-bottle dental caries. Arch Oral Biol 1996: **41**: 167–173.
3. Berkowitz RJ. Causes, treatment and prevention of early childhood caries: a microbiologic perspective. J Can Dent Assoc 2003: **69**: 304–307.
4. Carlsson P, Olsson B, Bratthall D. The relationship between the bacterium *Streptococcus mutans* in the saliva and dental caries in children in Mozambique. Arch Oral Biol 1985: **30**: 265–268.
5. Caufield PW. Dental caries – a transmissible and infectious disease revisited: a position paper. Pediatr Dent 1997: **19**: 491–498.

6. de Soet JJ, Nyvad B, Kilian M. Strain-related acid production by oral streptococci. Caries Res 2000: **34**: 486–490.

7. Ellen RP, Banting DW, Fillery ED. *Streptococcus mutans* and *Lactobacillus* detection in the assessment of dental root surface caries risk. J Dent Res 1985: **64**: 1245–1249.

8. Emanuelsson IM. Mutans streptococci – in families and on tooth sites. Studies on the distribution, acquisition and persistence using DNA fingerprinting. Swed Dent J 2001: **148**: 1–66.

9. Emanuelsson IM, Carlsson P, Hamerg K, Bratthall D. Tracing genotypes of mutans streptococci on tooth sites by random amplified polymorphic DNA (RAPD) analysis. Oral Microbiol Immunol 2003: **18**: 24–29.

10. Emanuelsson IR, Li Y, Bratthall D. Genotyping shows different strains of mutans streptococci between father and child and within parental pairs in Swedish families. Oral Microbiol Immunol 1998: **13**: 271–277.

11. Emanuelsson IR, Thornqvist E. Genotypes of mutans streptococci tend to persist in their host for several years. Caries Res 2000: **34**: 133–139.

12. Emilson CG, Klock B, Sanford CB. Microbial flora associated with presence of root surface caries in periodontally treated patients. Scand J Dent Res 1988: **96**: 40–49.

13. Hamada S, Slade HD. Biology, immunology, and cariogenicity of *Streptococcus mutans*. Microbiol Rev 1980: **44**: 331–384.

14. Hiller NL, Janto B, Hogg JS et al. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. J Bacteriol 2007: **189**: 8186–8195.

15. Hirose H, Hirose K, Isogai E, Miura H, Ueda I. Close association between *Streptococcus sobrinus* in the saliva of young children and smooth-surface caries increment. Caries Res 1993: **27**: 292–297.

16. Kerkhoven R, van Enckevort FH, Boekhorst J, Molenaar D, Siezen RJ. Visualization for genomics: the Microbial Genome Viewer. Bioinformatics 2004: **20**: 1812–1814.

17. Kreulen CM, de Soet HJ, Hogeveen R, Veerkamp JS. *Streptococcus mutans* in children using nursing bottles. ASDC J Dent Child 1997: **64**: 107–111.

18. Kristoffersson K, Gröndahl H-G, Bratthall D. The more *Streptococcus mutans*, the more caries on approximal surfaces. J Dent Res 1985: **64**: 58–61.

19. Kulkarni GV, Chan KH, Sandham HJ. An investigation into the use of restriction endonuclease analysis for the study of transmission of mutans streptococci. J Dent Res 1989: **68**: 1155–1161.

20. Lembo FL, Longo PL, Ota-Tsuzuki C, Rodrigues CR, Mayer MP. Genotypic and phenotypic analysis of *Streptococcus mutans* from different oral cavity sites of caries-free and caries-active children. Oral Microbiol Immunol 2007: **22**: 313–319.

21. Li Y, Caufield PW, Emanuelsson IR, Thornqvist E. Differentiation of *Streptococcus mutans* and *Streptococcus sobrinus* via genotypic and phenotypic profiles from three different populations. Oral Microbiol Immunol 2001: **16**: 16–23.

22. Li Y, Caufield PW. Arbitrarily primed polymerase chain reaction fingerprinting for the genotypic identification of mutans streptococci from humans. Oral Microbiol Immunol 1998: **13**: 17–22.

23. Loesche WJ. Role of *Streptococcus mutans* in human dental decay. Microbiol Rev 1986: **50**: 353–380.

24. Longo PL, Mattos-Graner R, Mayer MPA. Determination of mutacin activity and detection of *mutA* genes in *Streptococcus mutans* genotypes from caries-free and caries-active children. Oral Microbiol Immunol 2003: **18**: 144–149.

25. Matee MI, Mikx FH, de Soet JS, Maselle SY, de Graaff J, van Palenstein Helderman WH. Mutans streptococci in caries-active and caries-free infants in Tanzania. Oral Microbiol Immunol 1993: **8**: 322–324.

26. Mattos-Graner RO, Correa MS, Latorre MR, Peres RC, Mayer MP. Mutans streptococci oral colonization in 12–30-month-old Brazilian children over a one-year follow-up period. J Public Health Dent 2001: **61**: 161–167.

27. Mattos-Graner RO, Li Y, Caufield PW, Duncan M, Smith DJ. Genotypic diversity of mutans streptococci in Brazilian nursery children suggests horizontal transmission. J Clin Microbiol 2001: **39**: 2313–2316.

28. Mattos-Graner RO, Napimoga MH, Fukushima K, Duncan MJ, Smith DJ. Comparative analysis of Gtf isozyme production and diversity in isolates of *Streptococcus mutans* with different biofilm growth phenotypes. J Clin Microbiol 2004: **42**: 4586–4592.

29. Mattos-Graner RO, Zelante F, Line RC, Mayer MP. Association between caries prevalence and clinical, microbiological and dietary variables in 1.0 to 2.5-year-old Brazilian children. Caries Res 1998: **32**: 319–323.

30. Mattos-Granner RO, Smith DJ, King WF, Mayer MPA. Water insoluble glucan synthesis by mutans *streptococcal* strains correlates with caries incidence in 12 to 30 month old children. J Dental Res 2000: **79**: 1371–1377.

31. Napimoga M, Kamiya R, Rosa R et al. Genotypic diversity and virulence traits of *Streptococcus mutans* in caries free and caries active individuals. J Med Microbiol 2004: **53**: 697–703.

32. Oho T, Yamashita Y, Shimazaki Y et al. Simple and rapid detection of *Streptococcus mutans* and *Streptococcus sobrinus* in human saliva by polymerase chain reaction. Oral Microbiol Immunol 2000: **15**: 258–262.

33. Ravald N, Birkhed D, Hamp SE. Root caries susceptibility in periodontally treated patients. Results after 12 years. J Clin Periodontol 1993: **20**: 124–129.

34. Redmo Emanuelsson IM, Thornqvist E. Distribution of mutans streptococci in families: a longitudinal study. Acta Odontol Scand 2001: **59**: 93–98.

35. Redmo Emanuelsson IM, Wang XM. Demonstration of identical strains of mutans streptococci within Chinese families by genotyping. Eur J Oral Sci 1998: **106**: 788–794.

36. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 2000: **28**: 33–36.

37. Tettelin H, Masignani V, Cieslewicz MJ et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci USA 2005: **102**: 13950–13955.

38. Tomita Y, Zhu X, Ochiai K et al. Evaluation of three individual glucosyltransferases produced by *Streptococcus mutans* using monoclonal antibodies. FEMS Microbiol Lett 1996: **145**: 427–432.

39. van Loveren C, Buijs JF, ten Cate JM. Similarity of bacteriocin activity profiles of mutans streptococci within the family when the children acquire the strains after the age of 5. Caries Res 2000: **34**: 481–485.

40. Wassenaar TM, Gaastra W. Bacterial virulence: can we draw the line? FEMS Microbiol Lett 2001: **201**: 1–7.

41. Waterhouse JC, Russell RR. Dispensable genes and foreign DNA in *Streptococcus mutans*. Microbiology 2006: **152**: 1777–1788.

42. Waterhouse JC, Swan DC, Russell RR. Comparative genome hybridization of *Streptococcus mutans* strains. Oral Microbiol Immunol 2007: **22**: 103–110.

43. Zhang L, Srinivasan U, Marrs CF, Ghosh D, Gilsdorf JF, Foxman B. Library on a slide for bacterial comparative genomics. BMC Microbiol 2004: **4**: 12.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Figure S1 AP-PCR fingerprints of representative isolates from each clone type.

Figure S2 PCR amplifications of three variable genes.

Table S1 Distribution of 323 variable ORFs among 11 *Streptococcus mutans* clones.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.