Article:

Cornejo, OE, Lefebure, T, Pavinski Bitar, PD et al. (11 more authors) (2013) Evolutionary and population genomics of the cavity causing bacteria Streptococcus mutans. Molecular Biology and Evolution, 30 (4). 881 - 893. ISSN 0737-4038

https://doi.org/10.1093/molbev/mss278

Biological Sciences: Microbiology

Bacterial population genomics and the agent of human tooth decay at the dawn of agriculture

Omar E. Cornejo[1], Tristan Lefébure[2,6], Paulina D. Pavinski Bitar[2], Ping Lang[2,7], Vincent P. Richards[2], Kirsten Eilertson[3], Thuy Do[4], David Beighton[4], Lin Zeng[5], Sang-Joon Ahn[5], Robert A. Burne[5], Adam Siepel[3], Carlos D. Bustamante[1], and Michael J. Stanhope[2*]

Affiliations:

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305

[2]Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca, New York 14853, USA

[3]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14850, USA

[4]Department of Microbiology, King's College London Dental Institute and NIHR Biomedical Research Centre at Guy's and St, Thomas' NHS Foundation Trust , Floor 17, Tower Wing, Guy's Hospital, London SE1 9RT, England

[5] Department of Oral Biology, University of Florida, Gainesville, FL 32610, USA

[6]Present address: Université de Lyon, Lyon, F-69003, France; Université Lyon 1, Villeurbanne, F-69622, France; CNRS, UMR5023 Ecologie des Hydrosystèmes Naturels et Anthropisés, Villeurbanne, F-69622, France

[7]Present address: Department of Plant Pathology & Plant-Microbe Biology, Cornell University, Ithaca, NY 14853, USA.

[*]**Corresponding author:**

Michael J. Stanhope. Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca, New York 14853. 607-253-3859; mjs297@cornell.edu

Abstract

Most infectious diseases are believed to have originated after the origin of agriculture. Despite archeological evidence consistent with an increase in the prevalence of cavities after mankind was able to maintain crops, it remains unknown what could have been the etiological agent(s) responsible for this pattern. Here we use population genomic analysis of 57 newly sequenced bacterial genomes, to demonstrate that the human dental caries pathogen *Streptococcus mutans* underwent a historical population expansion about 10,000 years ago (CI-95%: 3,268 – 14,344 ya), placing it at the origin of agriculture. Furthermore, among 73 genes present in all isolates of *S. mutans*, but absent in other species of the mutans taxonomic group, we identify 50 that can be associated with metabolic processes that could have contributed to the successful adaptation of *S. mutans* to its new niche and the dietary changes that accompanied the origin of agriculture. Thus, *S. mutans* is a likely candidate as the etiological agent for the start of human caries and it appears likely that it has played this role in our biology for about the last 10,000 years. This work illustrates the value of comparative population genomic analysis of bacteria species in understanding the origins of human diseases and the basis of adaptive evolution of human pathogens.

Key words: *Streptococcus mutans*, demographic inference, cavities, bacterial evolution, pan and core genome, infectious disease.

Introduction

It has been hypothesized that many infectious diseases could only originate and be maintained after humankind developed agriculture (1-3). The most common explanations for this proposal are: i) epidemics were facilitated by the increase in density of human populations, ii) the increase in transmission of infectious diseases from domesticated (livestock or pets) or commensal (e.g. rats) animals (zoonoses); and iii) the development of diseases associated with food production changes after the origin of agriculture (2-4). An example, in support of this explanation, is the large body of archeological evidence consistent with an increase in the prevalence of dental caries after the development of agriculture (5-7). This pattern has been attributed to changes in diet and the consequent increase in consumption of carbohydrates in human populations after the development of starchy crops, leading to the establishment of infectious agents causing dental caries (5, 7). Despite the archeological evidence, it remains unknown what could have been the etiological agent(s) responsible for the increase in prevalence of cavities after the origin of agriculture.

Numerous studies in physical anthropology have shown an increased prevalence of dental caries in human remains from post-agricultural societies (5% - 50%) when compared with remains of Mesolithic hunter-gatherers (0% - 2%) (5, 7,, 8). Nevertheless, there is no evidence that cariogenic bacteria were associated with humans at this time, or prior to the origin of agriculture; nor is there evidence that any of the current cariogenic bacterial populations are linked with the rise in dental caries found in post-agricultural societies. In order to better understand the increase in dental caries in human populations after the development of agriculture, it is of interest to identify a cariogenic bacterial species with a demographic history that can be traced back to the beginning of agriculture and/or resembles that of the host population after the development of agriculture.

Streptococcs mutans, one the most widely studied cariogenic bacterial species, is known to be clinically associated with the development of human caries (9) and numerous studies have described molecular mechanisms by which this aciduric (resistant to acidic environments) and acidogenic (acid producing) bacterium contributes to the formation of cavities (10). Given the established link between *S. mutans* and human caries, a reasonable prediction is that this organism was associated with the onset of caries in early human history and that it has evolved along with humans for some protracted period. If this were the case, we should be able to detect aspects of *S. mutans* demographic history that could link it to the human disease history. Demographic models inferred from genetic data have an important role in modern population genetic analysis. Because demographic processes affect the accumulation of variation along the entire genome, the analysis of comparative population genome sequence data offers the possibility to address questions about the demographic history of populations. Of particular interest are genome-wide single nucleotide polymorphisms from multiple isolates of the same species representing many thousands of quasi-independent data points. Site frequency spectrum (SFS) methods for the analysis of such data have proven to be a powerful means of assessing demographic history and have recently been applied to questions involving a diversity of organisms (11, 12). Demographic analysis of bacterial species based on population genetic analysis of whole genomes, using the SFS, have yet to be published, although such methods should be entirely applicable if the necessary data were available. We undertook to test the hypothesis that *S. mutans* has been associated with human dental caries from its origins at the beginning of agriculture, by applying SFS population genetic analysis to multiple genome sequences derived from an international collection of *S. mutans*.

Results and Discussion

Next generation technology was used to obtain genome sequences of an international collection of 57 clinical isolates of *S. mutans* (information on isolates and details on sequence coverage and assembly appear in Supplementary Information). *S. mutans* genomes, like those of many other species of *Streptococcus*, are highly dynamic and their overall gene composition differs markedly from one isolate to another, likely due in large part to horizontal gene transfer. As with other bacteria, however, this difference in gene content involves only a portion of the genome, generally referred to as the dispensable component, in contrast to an alternative set of genes common to all strains, known as the core genome. Together these two components comprise the pan-genome of the species (13, 14). The core genome is a clearly identifiable component of *Streptococcus* species, as well as species from other genera, and indeed may represent that set of genes which can best define bacterial species (14-16). In order to conduct population genomic analysis of demographic history in *S. mutans* we needed to identify the core genome components since the necessary genetic information for reconstructing the history of *S. mutans* is contained in those genes that are shared by all isolates of the species. Our comparisons indicate that there are 1490 genes common to all 57 strains (see Fig. S4 in Supplementary Information for estimates of the core and pan-genome of *S. mutans*), out of which 1430 have sufficient information (more than 90% of the gene length for all strains) to perform our population genetic analyses. From the 1430 core genes, we identified 29,805 silent and 21,997 replacement single nucleotide polymorphisms (SNPs). We used principal component analyses (PCA)(17) on the silent sites to inspect the structure of genetic variation in our sample. Consistent with the findings of other studies on *S. mutans* (18), our analysis suggests little genetic differentiation among isolates sampled in different geographic locations (Fig. S7 in Supplementary Information). This facilitates the work of historical demographic reconstruction because single population models can be explored and fit to the data with greater power, since there are fewer numbers of parameters.

To reconstruct the demographic history of *S. mutans*, we employed a maximum likelihood inference method based on the distribution of allele frequencies across silent SNPs, or site frequency spectrum (SFS), and estimated confidence intervals by bootstrapping (see Materials and Methods for details). Four different population models were explored in this framework and the selection of the best-fit model was performed using the Akaike Information Criteria. The large number of singleton (unique) substitutions observed in *S. mutans* SFS is consistent with a recent expansion (Fig. 1a,b). Recently expanded populations leave a signature of mutations found in very low frequency, that have not had chance to disappear, or increase in frequency, by genetic drift. The maximum likelihood analysis shows that the SFS of *S. mutans* is consistent with a demographic scenario in which the population started expanding exponentially around 10,000 years ago (95% CI: 3,268 – 14,344 ya; possible uncertainties in mutation rate and generation time were taken into consideration in the computation of this confidence interval – see Supplementary Information for details; Fig. 1a,b, Table 1) and the absolute fit of the observed and simulated SFS's under this demographic model indicates no significant difference in their distributions (two sided Kolmogorov-Smirnov D = 0.2069, P = 0.564). The fit of the observed data to our simulations suggests that the effective population size of *S. mutans* has increased 4.8 to 5.5 times since the origin of agriculture (Fig. 1c), estimates much larger than those reported for humans (19).

The expected site frequency spectrum of variation is not affected by linkage, but the variance is affected (20, 21). We assessed the prevalence of recombination (gene conversion) among the 58 core genomes analyzed. For this, we used the core genome alignment, similar to the analysis by Leopold et al. (22); and estimated significant gene conversion events among isolates. Our analyses show that there has been extensive gene exchange between lineages represented by the isolates in our sample (Figure 2a), with a wide distribution of gene conversion tract lengths. We performed simulations assuming low recombination rates (four to five orders of magnitude smaller than mutation, between $10^{-12} – 10^{-11}$ subs/generation), and under the same demographic scenario this generates SFS similar to the one observed (Supplementary Information). Given that our actual data has much higher estimated recombination rates, we regard our simulations as highly conservative and therefore strongly supportive of our conclusions of demographic history.

We explored a variety of selection models under a similar maximum likelihood framework to that employed for the demographic fitting, to explain the site frequency spectrum (SFS) of the replacement SNPs (see Materials and Methods). Our analysis suggests that the majority of the changes (70%) that cause amino acid substitutions are under strong negative selection, and the remainder evolve neutrally (Fig. 3). The frequency of rare variants is much higher, and the frequency of common variants much lower, than expected under a neutral model, even after correcting for demographic expansion. This is a pattern consistent with strong purifying selection acting genome-wide (20, 23) and it raises the question of what are the features of molecular adaptation that underlie *S. mutans* successful colonization of, and proliferation in, the human host more than 10,000 years ago.

In order to adapt to the new niche of the "post-agricultural" human mouth, *S. mutans* faced several challenges. Among them, *S. mutans* needed to develop or increase efficiency in the metabolism of new sugars, successfully compete with bacterial species already present in the mouth of humans, develop defenses against increased oxidative stress, and resist the acidic byproducts of its own new efficient carbohydrate metabolism (24). Thus, it is reasonable to expect that even if most of the genome is under strong purifying selection, we should find evidence of adaptive evolution either in the pattern of amino acid changes in proteins involved in these processes, or in the composition of the genes present in the set of *S. mutans* unique core genes that are relevant to conferring an adaptive advantage for the new niche. We explored this question in two ways: i) by performing neutrality tests comparing the odds ratio of replacement to silent divergent vs. polymorphic changes via McDonald-Kreitman (MK) tests, and a Bayesian generalization of the Log-linear model that is the basis for the MK test (SNIPRE, see Materials and Methods); and ii) by identifying the protein domains, as well as the putative metabolic pathways in which these proteins are involved, of the genes present in all isolates of *S. mutans*, but not present in the outgroup *S. ratti* and two other closely related species of the mutans group (namely *Streptococcus macacae* and *Streptococcus criceti*). In particular, we were looking for proteins involved in aciduricity (resistance to acid), sugar metabolism, resistance to oxidative stress, antibiotics, and adherence to human tissue. Strikingly, very few proteins showed signatures of positive selection (more fixed replacement changes than synonymous). MK and SNIPRE tests identified 14 genes that were under positive selection (after Bonferroni correction), all of which are involved in either sugar metabolism or acid tolerance (Table S4 in Supplementary Information). On the other hand, the analysis of proteins present in all isolates of *S. mutans*, but absent in their close relatives (the *S. mutans* unique core genome) suggests that most of these genes are involved in adaptation to the post-agriculture human mouth niche. Of the 1490 genes that conform to the core genome of *S. mutans*, 73 are unique to this species and not found in its putative sister group, *S. ratti* (25, 26), or the mutans streptococci *S. macacae* and *S. criceti* (Fig. 4a). The absence of these putative adaptive genes in other species of the mutans group suggests their acquisition *via* horizontal gene transfer to the *S. mutans* lineage. Consistent with this hypothesis, these proteins tend to be similar to those arising from a wide variety of bacterial species including other oral flora bacteria, as well as taxa which produce lactic acid (Fig. 4b, Table S3, Supplementary Information), and many of them appear to be involved in carbohydrate metabolism (see Supplementary Information for phylogenetic examples highlighting several such cases of putative LGT (lateral gene transfer). An alternative explanation is that these genes arose through vertical descent from one of these close relatives of *S. mutans*, however the genes are not part of the core genome of these other taxa and instead are present in their dispensable genomes, and we simply have not yet sampled them in a single genome sequence. We have identified elsewhere (15) that core genes in one bacterial species can have their origins in the dispensable genome of closely related bacteria. Whatever their precise evolutionary history, these genes are likely key loci in defining the caries-associated phenotype of *S. mutans* and its adaptation to the human mouth environment.

Within this set of *S. mutans* unique core genes, 36 are hypothetical proteins with no similarity to known domains or protein clusters (Fig. 4a). The remaining proteins show similarity with domains of proteins involved in processes of: carbohydrate metabolism, resistance to acidic environments, transcriptional regulation, oxidative stress, metal and peptide translocation, and adhesion to host tissue (Fig. 4a and Tables S3 and S5 in Supplementary Information). In addition, some of these unique core genes contain domains potentially involved in resistance to antimicrobials, suggesting they could be of more recent acquisition (Fig. 4a). Undoubtedly, one of the major challenges that *S. mutans* had to overcome in the environment of the post-agriculture human mouth was surviving at low pH. Although *S. mutans* does not constitute a significant proportion of the oral flora colonizing healthy dentition, it can become numerically significant when there is repeated and sustained acidification of the biofilms associated with excess dietary carbohydrates or impaired salivary function (9). Interestingly, 14 % of the proteins found in the *S. mutans* unique core genome have been shown to be up-regulated in transcriptomic analyses at low pH (27) (binomial test comparison to core genome, P = 0.01). Among these are cation flux pumps that contribute to ionic equilibrium. Although low pH has been considered a primary ecological determinant influencing oral biofilm ecology, oxygen is also a critical factor (28), and it appears to be tolerated much better by commensal streptococci and other members of the normal microbiota than by *S. mutans* (28). In fact, exposure to oxygen strongly inhibits biofilm formation by *S. mutans* and alters the transcriptome and metabolism in a way that renders it less cariogenic (29, 30). Thus, *S. mutans* likely does not compete well in conditions of high redox or oxygen tension. Recently, hydrogen peroxide production by health-associated streptococci, such as *Streptococcus gordonii*, has been demonstrated to strongly inhibit *S. mutans* in mixed culture (31). Thus, while low pH provides strong selective pressure for aciduric species, during fermentable carbohydrate consumption and caries initiation and progression, oxygen may be an equally important environmental factor influencing the composition, biochemistry and pathogenic potential of oral biofilms (32).

S. mutans is also capable of mounting a substantial defense against commensal streptococci. In particular, strains of *S. mutans* produce a variety of lantibiotic and non-lantiobiotic bacteriocins that can kill related organisms (33). Peptide-based quorum-sensing systems, including the ComC competence cascade, multiple two-component systems, density-dependent signaling complexes and global regulatory systems all cooperate to influence the production of bacteriocin-like molecules (34). Interestingly, exposure to air uniformly activates the bacteriocin pathways and endogenous bacteriocin immunity systems, probably as a defense mechanism against competing organisms in immature, comparatively aerobic dental biofilms (29). Therefore, it is significant that the unique core genes of *S. mutans* contain a higher proportion of small peptides and gene products (smaller than 100 amino acids) than the core genome as a whole (approximately 6:1 ratio) that could potentially be involved in signaling and/or gene regulation (binomial test comparison to core genome, P= 1.23e-10; Table S5 in Supplementary Information).

Collectively, these findings indicate that the *S. mutans* unique core genes may represent important pathogen-specific factors that can be targeted with species-specific therapeutics that might decrease the competitive fitness of *S. mutans* without interfering with the propagation of health-associated commensal organisms. This study also suggests that one of the innovations that formed the basis of civilization precipitated a long-term association with an important human pathogen, highlighting the interconnections that exist between our sociocultural and biological evolution.

Materials and Methods

DNA sequencing and alignment. A total of 57 strains of *S. mutans* were selected, representing different sequence types and countries of origin (Supplementary Table S1). Single end sequencing was performed using the Illumina GA2 sequencer, with one lane per strain. This ensured high coverage of the ~2 MB genome of *S. mutans*. Sequence reads were aligned to the *S. mutans* UA159 and *S. mutans* NN2025 complete genomes, respectively, using MAQ (35), with appropriate mapping quality and coverage filters applied to capture the sequence information. De novo assemblies were performed using Velvet (36). Details on the conditions for the selection of the best assemblies are provided in the Supplementary Information. Assembled genomes were annotated using the NCBI PGAAP pipeline. Orthologs were determined by performing an all-versus-all BLASTP search combined with clustering using OrthoMCL2[1], and included all the *S. mutans* de novo assembled genomes and a draft genome sequence for the closely related taxa *S. ratti*. A subsequent OrthoMCL2 comparison was performed using the putative *S. mutans* unique core genome components against two other closely related taxa from the mutans group, *S. criceti* and *S. macacae*. Genome sequence data for 57 strains of *Streptococcus mutans* and single strains each of *Streptococcus ratti* (FA-1), *Streptococcus criceti* (HS-6), and *Streptococcus macacae* (NCTC 11558) have been deposited in GenBank under the following accession numbers: Smu: XXX-XXX (in submission); Sra: XXXX (in submission); Scr: AEUV01000016.1; Sma: AEUW01000012.1.

SNP calling. The 1430 genes constituting the core genome of *S. mutans*, were realigned at protein level to ensure that the alignments were in frame. Synonymous and replacement changes (and potential sites) were estimated following an "in house" pipeline coupled to the dNdS routine implemented in the libsequence suit(37). Because of the deep coverage of our data (>70X) we were confident in the call of rare variants (singletons) and no further sophisticated methods were employed for their identification.

Demographic and selection analysis. Principal Component Analysis (PCA) (38) of synonymous SNPs with frequencies larger than 5%, was performed using the R project for Statistical Computing (http://www.r-project.org/). Rare variants do not contribute to distinguish relatedness among individuals in putative subpopulations. The frequency distribution of variants, or site frequency spectrum (sfs), was calculated for synonymous and replacement changes independently in R. Demographic parameters for different competing models were estimated from the site frequency spectrum of synonymous changes using a diffusion-based approximation implemented in the program δaδI (12) in a maximum likelihood framework. The selection of the best-fit model was done using the Akaike Information Criteria. Changes in population size and time since change in demographics are estimated in 2Neu and 2Ne scaled parameters respectively. To convert these values to actual population sizes (expressed in individuals) and time (in years) we assumed a mutation rate estimated experimentally for bacteria of $5e^{-10}$ subs/site/generation (39), corresponding to $1.87e^{-04}$ subst/silent genome/generation (given there are 374,571 synonymous sites along the genome), and a conservative generation time of 2 divisions per day, as estimated for oral flora *in vivo* (40). Confidence intervals of the parameters were estimated by maximum likelihood fitting of 500 bootstraped data sets (details in Supplementary Information). Recombination was estimated as gene conversion on the core genome alignment of the full data set using Sawyer's algorithm as implemented in GeneConv (41); only significant tracts (after Bonferroni correction) were maintained in the analysis.

Genome wide selection analyses were performed on the replacement site frequency spectrum by a similar diffusion-based approximation as implemented for the demographic analysis and incorporating the action of selection, either as a point mass effects or as a distribution of selective effects, as implemented in PrFreq (23). Again, the best model was selected using the Akaike Information Criteria. We also performed a standard McDonald-Kreitman test (42), and an approach based on a Bayesian Loglinear model, to compare the polymorphism and divergent changes in synonymous and replacement sites on the genes for which an orthologous sequence could be identified in *S. ratti*.

Further details on all these methods can be found in Supplementary Information.

Acknowledgements We thank George Armelagos for input and discussions regarding archeological evidence of dental cavities in pre- and post- agricultural populations. Scott Durkin of JCVI provided some comparative genomic analysis on early drafts of the *S. criceti* and *S. macacae* genome sequences. This work was supported by the National Institute of Allergy and Infectious Disease, US National Institutes of Health, under grant number AI073368-01A2 awarded to M.J.S.

Author contributions PDPB, PL, TD, LZ, and S-JA were involved in various aspects of laboratory technical work; OEC, TL, VPR, and KE conducted data analysis; DB was involved in isolate collections and strain genotyping; OEC, RAB, ACS, CDB, and MJS conceived and designed the study; OEC and MJS wrote the paper.

References

1. Fiennes R (1978) *Zoonoses and the Origins and Ecology of Human Disease* (Academic Press, London).

2. Dobson AP & Carper ER (1996) Infectious diseases and human population history. *Bioscience* 46:115-126.

3. Diamond J (2002) Evolution, consequences and future of plant and animal domestication. *Nature* 418(6898):700-707.

4. Wolfe ND, Dunavan CP, & Diamond J (2007) Origins of major human infectious diseases. (Translated from eng) *Nature* 447(7142):279-283 (in eng).

5. Cohen MN & Armelagos GJ (1984) *Paleopathology at the Origins of Agriculture* (Academic Press, Orlando, FL).

6. Armelagos GJ (1991) Human evolution and the evolution of disease. *Ethn Dis* 1(1):21-25.

7. Lukacs JR (1992) Dental paleopathology and agricultural intensification in south Asia: new evidence from Bronze Age Harappa. (Translated from eng) *Am J Phys Anthropol* 87(2):133-150 (in eng).

8. Formicola V (1987) Neolithic Transition and Dental Changes - the Case of an Italian Site. (Translated from English) *J Hum Evol* 16(2):231-239 (in English).

9. Burne RA (1998) Oral streptococci... products of their environment. (Translated from eng) *J Dent Res* 77(3):445-452 (in eng).

10. van Houte J (1994) Role of micro-organisms in caries etiology. (Translated from eng) *J Dent Res* 73(3):672-681 (in eng).

11. Caicedo AL*, et al.* (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. (Translated from eng) *PLoS Genet* 3(9):1745-1756 (in eng).

12. Gutenkunst RN, Hernandez RD, Williamson SH, & Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. (Translated from eng) *PLoS Genet* 5(10):e1000695 (in eng).

13. Tettelin H*, et al.* (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". (Translated from eng) *Proc Natl Acad Sci U S A* 102(39):13950-13955 (in eng).

14. Tettelin H, Riley D, Cattuto C, & Medini D (2008) Comparative genomics: the bacterial pan-genome. (Translated from eng) *Curr Opin Microbiol* 11(5):472-477 (in eng).

15. Lefebure T, Bitar PD, Suzuki H, & Stanhope MJ (2010) Evolutionary dynamics of complete Campylobacter pan-genomes and the bacterial species concept. (Translated from eng) *Genome Biol Evol* 2:646-655 (in eng).

16. Lapierre P & Gogarten JP (2009) Estimating the size of the bacterial pan-genome. (Translated from eng) *Trends Genet* 25(3):107-110 (in eng).

17. Novembre J & Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. (Translated from eng) *Nat Genet* 40(5):646-649 (in eng).

18. Do T*, et al.* (2010) Generation of diversity in Streptococcus mutans genes demonstrated by MLST. (Translated from eng) *PLoS One* 5(2):e9073 (in eng).

19. Coventry A*, et al.* (2010) Deep resequencing reveals excess rare recent variants consistent with explosive population growth. (Translated from eng) *Nat Commun* 1(8):131 (in eng).

20. Bustamante CD, Wakeley J, Sawyer S, & Hartl DL (2001) Directional selection and the site-frequency spectrum. (Translated from eng) *Genetics* 159(4):1779-1788 (in eng).

21. Zhu L & Bustamante CD (2005) A composite-likelihood approach for detecting directional selection from DNA sequence data. (Translated from eng) *Genetics* 170(3):1411-1421 (in eng).

22. Leopold SR*, et al.* (2009) A precise reconstruction of the emergence and constrained radiations of Escherichia coli O157 portrayed by backbone concatenomic analysis. (Translated from eng) *Proc Natl Acad Sci U S A* 106(21):8713-8718 (in eng).

23. Boyko AR*, et al.* (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. (Translated from eng) *PLoS Genet* 4(5):e1000083 (in eng).

24. Jacobson GR, Lodge J, & Poy F (1989) Carbohydrate uptake in the oral pathogen Streptococcus mutans: mechanisms and regulation by protein phosphorylation. (Translated from eng) *Biochimie* 71(9-10):997-1004 (in eng).

25. Tapp J, Thollesson M, & Herrmann B (2003) Phylogenetic relationships and genotyping of the genus Streptococcus by sequence determination of the RNase P RNA gene, rnpB. (Translated from eng) *Int J Syst Evol Microbiol* 53(Pt 6):1861-1871 (in eng).

26. Hung WC, Tsai JC, Hsueh PR, Chia JS, & Teng LJ (2005) Species identification of mutans streptococci by groESL gene sequence. (Translated from eng) *J Med Microbiol* 54(Pt 9):857-862 (in eng).

27. Gong Y*, et al.* (2009) Global transcriptional analysis of acid-inducible genes in Streptococcus mutans: multiple two-component systems involved in acid adaptation. (Translated from eng) *Microbiology* 155(Pt 10):3322-3332 (in eng).

28. Marquis RE (1995) Oxygen metabolism, oxidative stress and acid-base physiology of dental plaque biofilms. (Translated from eng) *J Ind Microbiol* 15(3):198-207 (in eng).

29. Ahn SJ, Browngardt CM, & Burne RA (2009) Changes in biochemical and phenotypic properties of Streptococcus mutans during growth with aeration. (Translated from eng) *Appl Environ Microbiol* 75(8):2517-2527 (in eng).

30. Ahn SJ & Burne RA (2007) Effects of oxygen on biofilm formation and the AtlA autolysin of Streptococcus mutans. (Translated from eng) *J Bacteriol* 189(17):6293-6302 (in eng).

31. Kreth J, Zhang Y, & Herzberg MC (2008) Streptococcal antagonism in oral biofilms: Streptococcus sanguinis and Streptococcus gordonii interference

with Streptococcus mutans. (Translated from eng) *J Bacteriol* 190(13):4632-4640 (in eng).

32. Abbe K, Carlsson J, Takahashi-Abbe S, & Yamada T (1991) Oxygen and the sugar metabolism in oral streptococci. (Translated from eng) *Proc Finn Dent Soc* 87(4):477-487 (in eng).

33. Balakrishnan M, Simmonds RS, Kilian M, & Tagg JR (2002) Different bacteriocin activities of Streptococcus mutans reflect distinct phylogenetic lineages. (Translated from eng) *J Med Microbiol* 51(11):941-948 (in eng).

34. Martin B, Quentin Y, Fichant G, & Claverys JP (2006) Independent evolution of competence regulatory cascades in streptococci? (Translated from eng) *Trends Microbiol* 14(8):339-345 (in eng).

35. Li L, Stoeckert CJ, Jr., & Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. (Translated from eng) *Genome Res* 13(9):2178-2189 (in eng).

36. Zerbino DR & Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. (Translated from eng) *Genome Res* 18(5):821-829 (in eng).

37. Thornton K (2003) Libsequence: a C++ class library for evolutionary genetic analysis. (Translated from eng) *Bioinformatics* 19(17):2325-2327 (in eng).

38. Patterson N, Price AL, & Reich D (2006) Population structure and eigenanalysis. (Translated from eng) *PLoS Genet* 2(12):e190 (in eng).

39. Ochman H (2003) Neutral mutations and neutral substitutions in bacterial genomes. (Translated from eng) *Mol Biol Evol* 20(12):2091-2096 (in eng).

40. Gibbons RJ (1964) Bacteriology of dental caries. *J Dent Res* 43:SUPPL:1021-1028.

41. Sawyer S (1989) Statistical tests for detecting gene conversion. (Translated from eng) *Mol Biol Evol* 6(5):526-538 (in eng).

42. McDonald JH & Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. (Translated from eng) *Nature* 351(6328):652-654 (in eng).

43. Drake JW (1991) A constant rate of spontaneous mutation in DNA-based microbes. (Translated from eng) *Proc Natl Acad Sci U S A* 88(16):7160-7164 (in eng).

Figure Legends

Figure 1 | Demographic history of *S. mutans*. (a) Schematic representation of *S. mutans* population history. The timeline (in years before present) represents the start of the expansion of cariogenic bacteria after the onset of agriculture, calibrated using an experimentally determined mutation rate for bacteria(43), concomitant with an *in vivo* determined generation time for oral flora bacteria (40) (see Materials and Methods and Supplementary Information for details). (b) The observed distribution of number of synonymous SNPs at a given frequency in the sample of 58 isolates (blue) is shown, as well as the expectation under the parameters that generate the best fit demographic model (dark blue). The difference between the two distributions is not significant. The distribution under a standard neutral model with constant population size is shown in light blue (significant KS, P < 0.0001). (c) The bi-dimensional likelihood profile for combination of parameters ν (ratio of current to ancestral population size) in the x-axis and the time at the beginning of the demographic expansion (scaled in generations / 2Na) in the y-axis. The maximum likelihood value is shown as a white dot and the 95% confidence interval (95%CI) is highlighted as a white dotted line. 95% CI estimated from bootstrapped data can be found in Supplementary Information, Fig. S9.

Figure 2 | Recombination in *S. mutans*. (a) The inferred distribution of recombination tracts (gene conversion) among isolates of *S. mutans*. Gene tracts of the core genome that served as alignment for the estimation of recombination along the genome are represented in blue and red. Tracts of significant gene conversion events detected along the genome are represented in green. (b) The distribution of gene conversion tract lengths, characterized by a wide range of values that follow a geometric distribution.

Figure 3 | Evidence of genome-wide selective constraints in *S. mutans*. The observed distribution of number of replacement SNPs at a given frequency in the sample of 58 isolates is shown in red. The expectation is that replacement changes will have an effect on the fitness of individuals, so it is unlikely that they behave neutrally. Correcting for population expansion inferred from the silent SNPs (Fig. 1), does not account for the excess of singletons observed in the data (light green). On the other hand, a model that allows for selection affecting changes in allele frequency, after correcting for demography, yields a superior fit, suggesting that in the *S. mutans* genome 30% of the replacement changes are neutral and 70% are under strong selection ($\gamma = -17$, where $\gamma =$

$2N_es$, and N_e is the current population size and s is the coefficient of selection).

Figure 4 | Genome map of *S. mutans*. (a) Representation of the forward coding (light blue) and reverse coding (light red) genes comprising the core genome of *S. mutans*. The third inner circle, displays the unique core genes, present in *S. mutans* only, colored by the metabolic functions in which they are involved. The most inner circles present the unique genes shown to be up or down regulated by the impacts coincident with the diet change of humans after the origin of agriculture: starch and sucrose metabolism and low environmental pH. (b) Putative origin of horizontally transferred unique core genes in *S. mutans*.

Table Captions

Table 1 | Selection of demographic models. The logarithm of the maximum likelihood (Ln) for each of the demographic models fit to the data, the number of parameters for each model, and the Akaike Information criteria (AIC = 2*(N free param) – 2*Ln). The models assessed were exponential growth or decay (Exp grow), 2 epoch (constant and instant increase), a bottleneck in the past, combined with exponential growth (Bottle + growth), and 3 epoch (bottleneck, followed by an instantaneous increase). The model with the minimum AIC (Exp grow) was selected as the model that best explains the data.