Evolutionary and Population Genomics of the Cavity Causing Bacteria *Streptococcus mutans*

Omar E. Cornejo,[1] Tristan Lefébure,[†,2] Paulina D. Pavinski Bitar,[2] Ping Lang,[‡,2] 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]

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

[2]Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University

[3]Department of Biological Statistics and Computational Biology, Cornell University

[4]Department of Microbiology, King's College London Dental Institute and NIHR Biomedical Research Centre at Guy's and St. Thomas's NHS Foundation Trust, Guy's Hospital, London, United Kingdom

[5]Department of Oral Biology, University of Florida

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

[‡]Present address: Department of Plant Pathology & Plant-Microbe Biology, Cornell University, Ithaca, NY

*Corresponding author: E-mail: mjs297@cornell.edu.

Associate editor: Ryan Hernandez

Abstract

Streptococcus mutans is widely recognized as one of the key etiological agents of human dental caries. Despite its role in this important disease, our present knowledge of gene content variability across the species and its relationship to adaptation is minimal. Estimates of its demographic history are not available. In this study, we generated genome sequences of 57 *S. mutans* isolates, as well as representative strains of the most closely related species to *S. mutans* (*S. ratti*, *S. macaccae*, and *S. criceti*), to identify the overall structure and potential adaptive features of the dispensable and core components of the genome. We also performed population genetic analyses on the core genome of the species aimed at understanding the demographic history, and impact of selection shaping its genetic variation. The maximum gene content divergence among strains was approximately 23%, with the majority of strains diverging by 5–15%. The core genome consisted of 1,490 genes and the pan-genome approximately 3,296. Maximum likelihood analysis of the synonymous site frequency spectrum (SFS) suggested that the *S. mutans* population started expanding exponentially approximately 10,000 years ago (95% confidence interval [CI]: 3,268–14,344 years ago), coincidental with the onset of human agriculture. Analysis of the replacement SFS indicated that a majority of these substitutions are under strong negative selection, and the remainder evolved neutrally. A set of 14 genes was identified as being under positive selection, most of which were involved in either sugar metabolism or acid tolerance. Analysis of the core genome suggested that among 73 genes present in all isolates of *S. mutans* but absent in other species of the mutans taxonomic group, the majority can be associated with metabolic processes that could have contributed to the successful adaptation of *S. mutans* to its new niche, the human mouth, and with the dietary changes that accompanied the origin of agriculture.

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

Introduction

Streptococcus mutans is known to be one of the most prevalent bacteria in human oral flora and is widely recognized as a key etiological agent of human dental caries (reviewed in Burne 1998). Evidence in support of this latter issue include the following key points: *S. mutans* is frequently isolated from caries lesions; it induces caries formation in animals fed a sucrose-rich diet; the species is highly acidogenic and aciduric (Hamada and Slade 1980; Loesche 1986; van Houte 1994); it can form an effective biofilm by producing surface antigens which promote adhesion to the tooth surface and other bacteria (Hamada and Slade 1980); and it flourishes in cariogenic plaque because it is better able to grow and metabolize carbohydrate in a low pH environment (Bender, et al. 1986; Bender and Marquis 1987; Marquis 1990; Belli and Marquis 1991; Arthur et al. 2011). Despite its recognized importance in this important human disease, there are very few publications using comparative genomics to gain insights on basic biology, evolutionary history and pathogenesis of this organism (Ajdic et al. 2002; Maruyama et al. 2009) and there are but three genome sequences currently on GenBank. A number of studies have demonstrated substantial genetic heterogeneity across clinical isolates of *S. mutans* (Zhang et al. 2009; Arthur et al. 2011; Cheon et al. 2011; Phattarataratip et al. 2011);

however, our present knowledge of gene content variability across the species and its relationship to adaptation and virulence is minimal.

One of the more significant recent discoveries in bacterial genomics is that bacteria species appear to be comprised of both a set of core and dispensable genes, with only the former present in all isolates of that species and with the sum of the two components forming the species pan-genome (or supra-genome). This concept was first introduced for S. agalactiae in 2005 (Tettelin et al. 2005) and is now generally regarded as a principle common to most or all bacteria. Much speculation has centered on the origin, composition, and size of bacteria pan-genomes and whether they are finite or infinite (Tettelin et al. 2008; Lapierre and Gogarten 2009). Recently, we examined the role of the core and dispensable genes in defining two sympatric and closely related species of Campylobacter—C. jejuni and C. coli (Lefebure et al. 2010), and addressed whether their pan-genomes are finite or infinite. We demonstrated, through the analysis of 96 genome sequences, that their pan-genome is indeed finite and that there are unique and cohesive features to each of their genomes defining their genomic identity. The two species have a similar pan-genome size; however, C. coli has acquired a larger core genome and each species has evolved a number of species-specific core genes, possibly reflecting different adaptive strategies. Understanding the pan-genome components of any species of bacteria means that the core genome and the dispensable genome can be identified for the group in question. Comparisons made to the genomes of representative isolates of other closely related species, can then identify the unique core genome of the group—that is, the genes common to all isolates of that group, not present in its closest relatives. This set of unique core genes is of particular interest, because they are amongst the genes likely to define the essence of that group's adaptive specifics.

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 (SNPs) from multiple individuals 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 (Caicedo et al. 2007; Gutenkunst et al. 2009). Demographic analyses 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.

Most of the intraspecific evolutionary analyses of bacterial species have focused on understanding the phylogenetic relationships among bacterial isolates or have concentrated on assessments of population diversity based on seven gene multilocus sequence typing (MLST). Studies based on MLST

suggest that recombination in S. mutans contributes eight times more to the maintenance of genetic diversity than mutation (r/u ~8.3) (Do et al. 2010). Homologous recombination in S. mutans, as in the rest of naturally transformable species, occurs by a process analogous to gene conversion in eukaryotes, by replacement of short fragments of DNA (Smith et al. 1991). Members of the genus Streptococcus have recently been the subject of population genomic analyses (Donati et al. 2010; Croucher et al. 2011; Muzzi and Donati 2011), highlighting the importance of the dispensable gene component, which carries a high proportion of loci associated with different virulence and phenotypic attributes (Suzuki et al. 2011). These studies have also pointed out the large impact of homologous recombination in the maintenance of genetic diversity of the core genome (Donati et al. 2010; Croucher et al. 2011). Given the relatively high rates of homologous recombination in naturally transformable species like S. mutans, we propose employing methods based on the analysis of the SFS to provide insight on the evolutionary history of the species.

In this study, we generated genome sequences of 57 S. mutans isolates, as well as representative strains of the most closely related species to S. mutans, to identify the overall structure and potential adaptive features of the dispensable and core components of the genome, and performed population genetic analyses on the core genome of the species aimed at understanding the demographic history, and impact of selection shaping its genetic variation.

Results

SNPs and the Core Genome

Next generation technology was used to obtain genome sequences of an international collection of 57 clinical isolates of S. mutans. Comparison of gene content among strains indicated that the maximum gene content divergence was approximately 23%, with many strains diverging by 5–15% in number of annotated genes (supplementary fig. S1, Supplementary Material online). On average, each genome contained 1,636 genes. To conduct a population genomic analysis of demographic history in S. mutans, we needed to identify the core genome components because the necessary genetic information for reconstructing the demographic history of S. mutans is contained in those genes that are shared by all isolates of the species. Our comparisons indicated that there were 1,490 genes common to all 57 strains (supplementary fig. S2, Supplementary Material online, for an estimate of the core genome of S. mutans), out of which 1,430 had sufficient information (more than 90% of the gene length for all strains), to perform population genetic analyses; the pan-genome of the species approached 3,296 genes (supplementary fig. S3, Supplementary Material online). From the 1,430 core genes, we identified 29,805 silent and 21,997 replacement SNPs, which summarized across genes, resulted in estimates of sequence polymorphism ranging from 0.5% to 1.9%. We estimated summary statistics and found that the average GC content in S. mutans is approximately 37%, the average Watterson's θ (0.008) is larger than the average
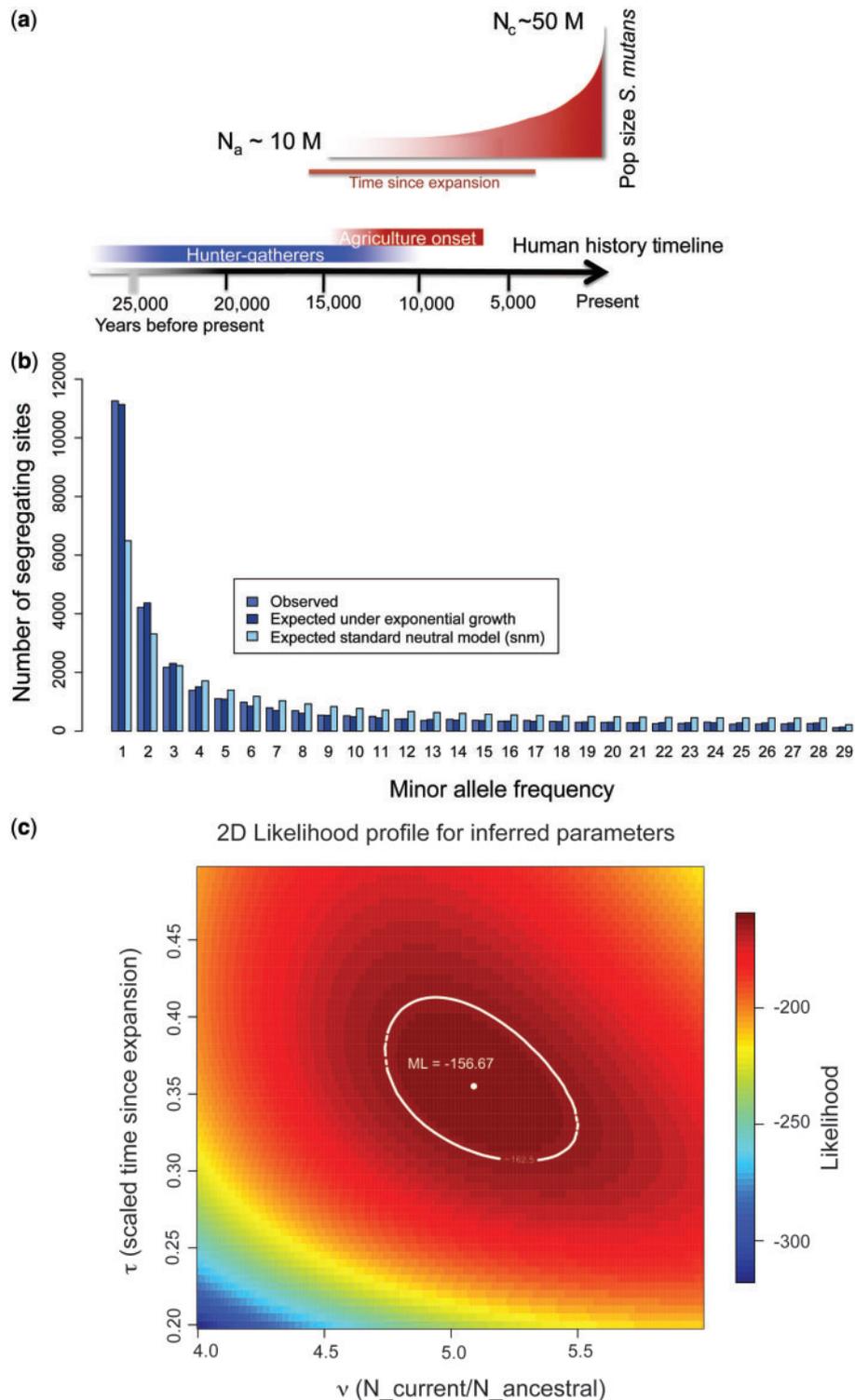
nucleotide pairwise difference (0.005), and this translates into a prevalence of negative Tajima's D across genes, as has been observed in other species of bacteria (Hughes 2005). A graphic representation with the estimates of these summary statistics is included in the Supplementary Material online (supplementary fig. S4, Supplementary Material online). Only 1.16% of the SNPs found across all 1,430 core genes used in the analysis were multi-allelic leaving us with 98.84% bi-allelic SNPs for the demographic and selection analysis. Principal component analysis (PCA) (Novembre and Stephens 2008), employing silent sites from the core genome, was used to inspect the structure of genetic variation. For this, we only considered synonymous polymorphic sites with minor allele frequencies above 5% and $r^2 \leq 0.4$; the correlation between alleles was used as a measure of linkage disequilibrium. The rationale behind the elimination of sites is that alleles in low frequency do not contribute to the clustering of strains and because the information in linked sites is redundant. Nevertheless, PCA performed with the complete data set (all synonymous polymorphic sites from the core genome) produced similar results. Consistent with the findings of other studies on S. mutans (Do et al. 2010), our analysis suggested little genetic differentiation among isolates sampled in different geographic locations (supplementary fig. S5, Supplementary Material online). In addition to the PCA, we performed pairwise Fst analyses with permutation procedures to assess their significance. None of the pairwise Fst between geographic locations was significant. This facilitates the work of historical demographic reconstruction because single population models can be explored and fit to the data with greater power, because there are fewer numbers of parameters. Additional analyses on the dispensable part of the genome suggest that there is no association between the sharing of genes and the geographic location of the samples and like the core genome, there is no indication of population genetic structure evident in the analysis of high frequency dispensable genes.

Demographic History

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 SFS, and estimated confidence intervals (CIs) by bootstrapping. Five different population models were explored (supplementary fig. S6, Supplementary Material online) 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 and 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 (Slatkin and Hudson 1991; Griffiths and Tavare 1994). Estimations performed with $\delta a \delta i$ recover two main parameters under the exponential growth model. The first is ν, the ratio of current to ancestral population size; and the second is τ, the time since the change in demographic, representing in

our case the start of the demographic expansion. τ is equal to time (T) scaled by 2Na, which is the ancestral effective population size. After appropriate rescaling of the inferred parameters, the maximum likelihood analysis suggested that the SFS of S. mutans is consistent with a demographic scenario in which the population started expanding exponentially approximately 10,000 years ago (95% CI$_{classic\_boostrap}$: 3,268–14,344 years ago; possible uncertainties in mutation rate and generation time were taken into consideration in the computation of this CI—see Supplementary Material online for details; fig. 1a and b and table 1) and the absolute fit of the observed and simulated SFS's under this demographic model indicated 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 suggested that the effective population size of S. mutans has increased 4.8 to 5.5 times since the origin of human agriculture (fig. 1c), estimates much larger than those reported for humans (Coventry et al. 2010). The likelihood surface of the parameter search was evaluated to assure that it did not fall into local maxima. Models were searched with different initial parameter values and evaluated if they all converged to the same maxima. The likelihood surface in figure 1c was obtained by finding the likelihood value for each combination of parameters in a grid. We estimated the CIs for the parameters by bootstrapping the synonymous data matrix and fitting the model (supplementary fig. S7, Supplementary Material online).

The expected SFS of variation is not affected by linkage, but that is not true of the variance (Bustamante et al. 2001; Zhu and Bustamante 2005). It is germane therefore to demonstrate that S. mutans, being a naturally transformable bacterial species, undergoes significant homologous gene exchange. We assessed the magnitude of recombination in S. mutans by identifying the number of significant gene conversion events among isolates using Sawyer's algorithm (Sawyer 1989). This algorithm is conservative (Sawyer 1989) and therefore we regard it as an estimate of the minimum set of gene conversion events. Power analyses of the method have concluded that it tends to underestimate the gene conversion events for scenarios in which the rate of gene conversion is high (Mansai and Innan 2010). Our analysis indicates that there has been extensive gene exchange between lineages represented by the isolates in our sample (fig. 2a), with a wide distribution of gene conversion tract lengths. Additionally, our analysis with ClonalFrame suggests that on average, recombination contributes approximately six times more than mutation to the genetic variation in S. mutans ($\rho/\theta \approx 6$: 95% credibility interval ~3.09–19.6; supplementary fig. S8, Supplementary Material online). We performed simulations in ms (Hudson 2002) assuming gene conversion rates similar to those estimated for S. mutans (six times larger than the mutation rate), all under the same demographic scenario estimated from the original data. The mean SFS over 600 simulations is similar to the observed SFS (supplementary fig. S9, Supplementary Material online). We then used the simulations to investigate the demographic inferences performed with $\delta a \delta i$. For each

Fig. 1. Demographic history of *Streptococcus 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 (Drake 1991), concomitant with an in vivo determined generation time for oral flora bacteria (Gibbons 1964) (see Materials and Methods and Supplementary Material online, 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% CI is highlighted as a white dotted line. 95% CI estimated from bootstrapped data can be found in supplementary figure S7, Supplementary Material online.

simulation, we estimated the SFS and inferred maximum likelihood parameters (τ and ν) for a demographic scenario of population expansion in $\delta a \delta i$. The distribution of maximum likelihood estimates of τ and ν present more variability than that obtained with classical bootstrap (τ: 95% CI = 0.2–2.3; ν: 95% CI = 2.5–11.3), and our maximum likelihood estimated parameters ($\tau = 0.355$ and $\nu = 5.09$) fall within the CIs obtained from the parametric bootstrap, suggesting that $\delta a \delta i$ is producing results that are consistent with what would be observed in a 2 Mb chromosome evolving under such a scenario (supplementary fig. S10, Supplementary Material online). Furthermore, in our simulations, the patterns of decay of

Table 1. Selection of Demographic Models.

Model[a]	Ln L[b]	No. of Free[c] Parameters	AIC[d]
Exponential growth	−156.67	2	317.34
2 Epoch	−168.96	2	341.92
Bottleneck + growth	−156.67	3	319.34
3 Epoch	−168.97	4	345.94

[a]The logarithm of the maximum likelihood (Ln) for each of the demographic models fit to the data.
[b]The models assessed were exponential growth or decay (exponential growth), 2 epoch (constant and instantaneous increase), a bottleneck in the past, combined with exponential growth (bottleneck + growth), and 3 epoch (bottleneck, followed by an instantaneous increase).
[c]The number of parameters for each model.
[d]Akaike Information criteria [AIC = 2 × (no. free parameters) − 2 ln]. The model with the minimum AIC (exponential growth) was selected as the model that best explains the data.
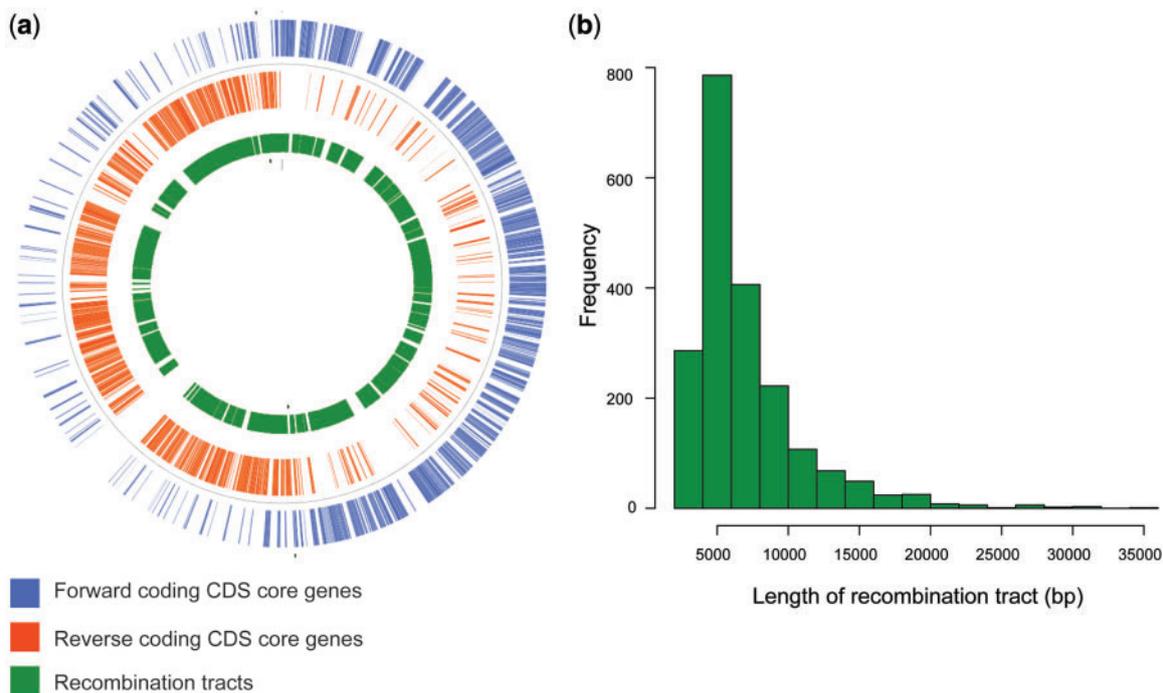
linkage disequilibrium with distance approximately resemble the decay pattern observed in the data.

Adaptation to the Postagriculture Habitat

To understand the role of natural selection in shaping genomic variation in *S. mutans*, we explored a variety of selection models under a similar maximum likelihood framework to that employed for the demographic fitting, used to explain the SFS of the replacement SNPs. Our analysis suggests that the majority of the changes (70%) that cause amino acid substitutions are under strong negative selection, and the remainder evolved 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 (Bustamante et al. 2001; Boyko et al. 2008), and is in agreement with what has been observed in other bacterial species (Hughes 2005).

We explored the molecular adaptation of individual genes of *S. mutans* in two ways: 1) by performing neutrality tests comparing the odds ratio of replacement to silent divergent versus polymorphic changes via McDonald–Kreitman (MK) tests, and a Bayesian generalization of the Log-linear model that is the basis for the MK test (SNIPER); and 2) 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 *S. macacae* and *S. criceti*).

FIG. 2. Recombination in *Streptococcus 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.

Legend:
- Forward coding CDS core genes
- Reverse coding CDS core genes
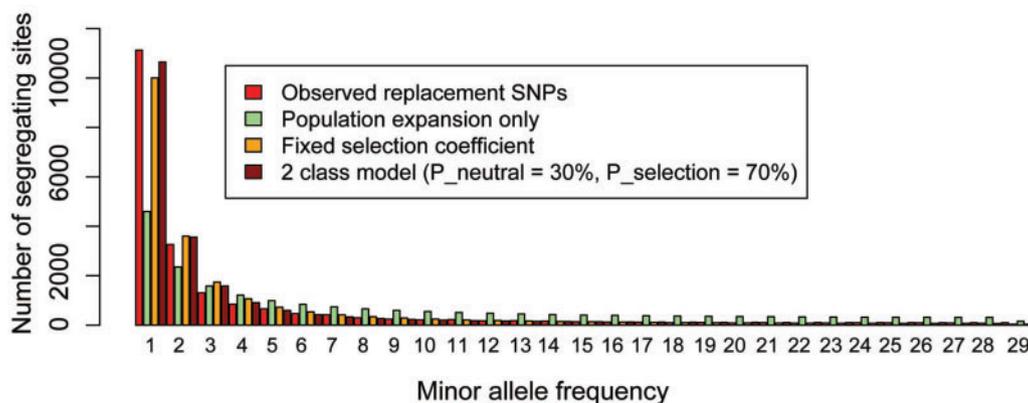- Recombination tracts

FIG. 3. Evidence of genome-wide selective constraints in *Streptococcus 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 $\gamma = 2N_e s$, and N_e is the current population size and s is the coefficient of selection).

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. The annotated genes of *S. ratti* were compared with the 59 *S. mutans* genomes using all versus all BLAST and from the protein clusters we identified the orthologous genes present in both species. A total of 2,033 genes were annotated in *S. ratti* and of these, 1,282 genes were in a 1:1 relationship with *S. mutans*. On average, we estimated the divergence between *S. mutans* and *S. ratti* to be approximately 26%. We estimated very few of these proteins showing signatures of positive selection (more fixed replacement changes than synonymous). The combined estimate of MK and SNIPER tests identified 14 genes that were under selection (after Bonferroni correction), the majority of which are involved in either sugar metabolism or acid tolerance (supplementary table S1, Supplementary Material online). It is possible that the relatively high divergence between *S. mutans* and *S. ratti* reduced the power of these tests, as multiple hits could mask the signature of selection; however, our results are nonetheless consistent with the SFS-based analysis.

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 postagriculture human mouth niche. Of the 1,490 genes that conform to the core genome of *S. mutans*, 148 are present in *S. mutans* and absent in the putative sister group *S. ratti* (Tapp et al. 2003; Hung et al. 2005), whereas 73 loci are unique to *S. mutans* in comparisons involving the three species, *S. ratti*, *S. macacae*, and *S. criceti* (fig. 4a and supplementary table S2, Supplementary Material online). 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 supplementary tables S2 and S3, Supplementary Material online). 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 as the carbohydrate content of the human diet increased 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 (Burne 1998). Interestingly, 14% of the gene products found in the *S. mutans* unique core genome have been shown to be upregulated in transcriptomic analyses at low pH (Gong et al. 2009) (binomial test comparison with core genome, $P = 0.01$). Among these are cation flux pumps that contribute to ionic equilibrium. Analysis of the functional groups (GO terms) of the dispensable genes present in 54, 55, 56, and 57 isolates suggest that there is no enrichment of any particular functional group (P value > 0.05 for all comparisons), when compared with the core genome and none of the genes presented mean ratio of replacement to synonymous changes (k_N/k_S) larger than 1.

Discussion

The genomes of different strains of *S. mutans*, like those of many other species of *Streptococcus*, are highly dynamic and their overall gene composition differs markedly from one isolate to another because of lateral gene transfer (LGT). However, as with other bacteria, 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 (Tettelin et al. 2005, 2008).
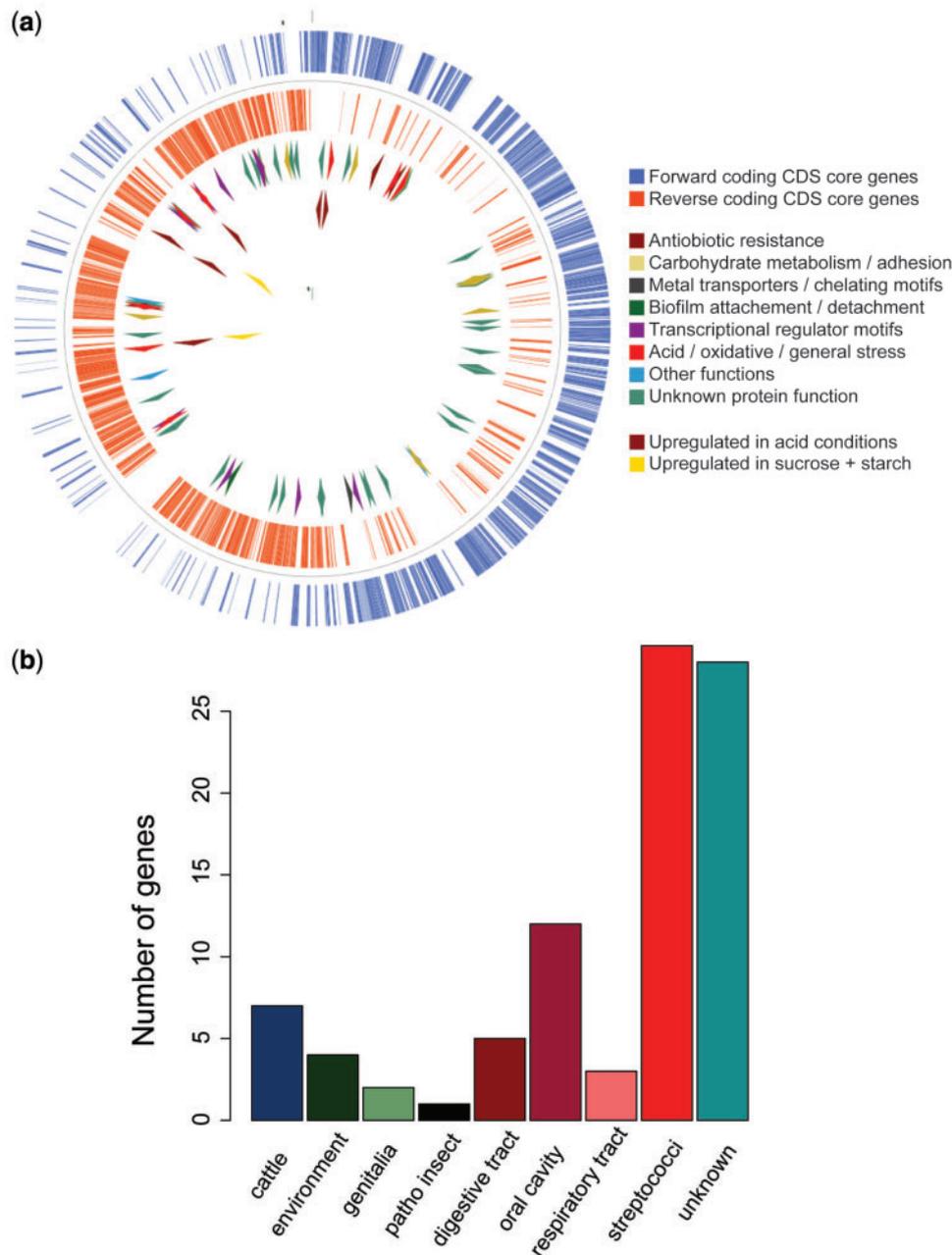
FIG. 4. Genome map of *Streptococcus mutans*. (*a*) Representation of the forward coding (light blue) and reverse coding (light red) genes comprising the core genome of *S. mutans*. The third inner circle, displays the unique core genes, present in *S. mutans* only, colored by the metabolic functions in which they are involved. The most inner circles present the unique genes shown to be upregulated or downregulated 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 laterally transferred unique core genes in *S. mutans*.

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 (Lan and Reeves 2000, 2001; Tettelin et al. 2008; Lapierre and Gogarten 2009; Lefebure et al. 2010). Our analyses suggest that the core genome of the species is readily identifiable and consists of 1,430 genes. The complete gene repertoire of the dispensable component of the genome is much harder to capture. Nevertheless, after examining the accumulation curve of the pan-genome in *S. mutans* (supplementary fig. S3, Supplementary Material online), we suggest that by sampling 58 genomes in the population we are reaching a point at which there is no appreciable increase in the number of new genes as more genomes are included in the analysis.

Demographic History

Our estimates for the timing of the demographic expansion in *S. mutans* coincide with the origin of human agriculture. It has been hypothesized that many infectious diseases could only originate and be maintained after humankind developed agriculture (Fiennes 1978; Dobson and Carper 1996; Diamond 2002). 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%) (Cohen and Armelagos 1984; Formicola 1987; Lukacs 1992). 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 (Cohen and Armelagos 1984; Lukacs 1992). Our timing and expansion estimates are consistent with *S. mutans* as a possible etiological agent of human dental caries; however, it is not possible to rule out other oral bacteria species as also contributing to the development of this disease.

We have attempted to be thorough in our analysis, by including not only rigorous demonstration that *S. mutans* undergoes recombination but also that the methods of analysis are appropriate for the limited conditions of a small genome size, and other constraints imposed by the biology of this organism. Nevertheless, we have not explored all alternative possibilities that could give rise to patterns of variation like those observed in this data set. We have provided a parsimonious explanation for the pattern of variation found in the synonymous substitutions along the core genome; however, the biology of bacterial species is such that there are potential alternative conditions that could give rise to similar patterns. For instance, it is widely appreciated that selection can impact the population dynamics of clonal organisms like bacteria (Dykhuizen 1990a, 1990b, 1993; Guttman and Dykhuizen 1994; Buckee et al. 2008). Although it is possible that a scenario of multiple instances of selection and linkage, because of linkage to replacement sites under selection (Braverman et al. 1995), could explain the excess of singletons in the distribution of synonymous changes, it is difficult to envision a selection scenario, with specific selective pressures, that would explain our observations. Another potential scenario is the existence of different regimes of mutation in time. In laboratory experiments, it has been shown how clones with greater mutation rates (mutators) can facilitate adaptation to new environmental conditions (Sniegowski, et al. 1997; Shaver, et al. 2002; Gerrish et al. 2007). These mutators also have the effect of changing the overall mutation rate in the population, potentially affecting the pattern of observed variation, giving rise to an observed excess of singletons, via a combination of negative selection on replacement changes and variation of the rates at which new synonymous and replacement changes are being generated. Unfortunately, there is no information about the prevalence of mutator strains in natural populations (carrier or clinical) of *S. mutans*, or in experiments performed with laboratory strains. Our analyses of the population genetic structure of the data, consistent with results from previous work (Do et al. 2010), suggest that there is no geographic differentiation. We believe that our simulations under the inferred demographic scenario, which exclude selection, provide a sufficiently parsimonious explanation in agreement with the observed data, to explain the genome wide accumulation of synonymous variation in *S. mutans*. It has been suggested that *S. mutans*

genotypes are maternally transmitted (Alaluusua et al. 1996; Emanuelsson et al. 1998; Emanuelsson and Thornqvist 2000), and this could act to produce a strong pattern of geographic differentiation of the isolates over time. Our results showing a lack of geographic structure are not consistent with such a scenario. Recent work by Klein et al. (2004) suggests a high turnover rate of lineages in the mouth of children after colonization from the mother, which would be consistent with our observations.

Adaptation in the Postagricultural Era

The distribution of replacement variants in the *S. mutans* population is consistent with strong purifying selection acting genome-wide, 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. 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 by-products of its own new efficient carbohydrate metabolism (Jacobson et al. 1989). Or results suggest that only a handful of genes (14 genes) present a pattern that deviates from the neutral expectation. On the other hand, analysis of the function of the unique core genes suggests that a large fraction of these genes are involved in sugar metabolism and pH resistance. The absence of these putative adaptive genes in other species of the Mutans taxonomic group (Tapp et al. 2003) suggests their acquisition via LGT 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, supplementary table S2, Supplementary Material online), and many of them appear to be involved in carbohydrate metabolism. As an example, the gene SMU.438 (supplementary fig. S11, Supplementary Material online) is closely related to a hypothetical protein present in *Lactococcus lactis*, which includes domains corresponding to an activator of 2-hydroxyglutaryl-CoA dehydratase. This protein includes three domains identified in *Lactobacillus*: FGGY family of carbohydrate kinases, N-terminal domain (2), and a CoA-substrate-specific enzyme activase, suggesting that this protein could be involved in carbohydrate metabolism (Klees et al. 1992). The protein SMU.1410 (supplementary fig. S12, Supplementary Material online) is very similar to a putative reductase of *S. gallolyticus* (formerly known as *S. bovis*). *S. gallolyticus* is found in the gastrointestinal tract of ruminants and humans, and has been linked to endocarditis in humans (Hoen et al. 2005). When ruminants consume a diet high in sugars, the fermentable carbohydrates allow for the growth of *S. bovis* (Asanuma and Hino 2002). This organism also exhibits lactic acid fermentation, reducing the pH of the rumen leading to ruminal acidosis (Asanuma and Hino 2002). Reductases are also involved as ion pumps in electron

transport chains. We suggest that this putative reductase is involved in resistance to acidic conditions in *S. mutans* and was obtained via a LGT event from *S. gallolyticus,* or some very closely related taxon. The protein SMU.1561 has been shown to be up-regulated in *S. mutans* in acid conditions (Gong et al. 2009). Our phylogenetic analysis indicates that SMU.1561 (supplementary fig. S13, Supplementary Material online) is closely related to the product of the *trkA* gene from *Ethanoligenens harbinense,* which is another bacteria forming part of the oral flora of humans (Aas et al. 2005). The *trkA* gene encodes a protein involved in the transport of ions across the membrane (Anion permease ArsB/NhaD), which is consistent with the fact that it is upregulated in acidic conditions (Gong et al. 2009). An alternative explanation, to a history of LGT for these unique core genes, is that they arose through vertical descent from one of these close relatives of *S. mutans,* but 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 (Lefebure et al. 2010) 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 may be important loci in defining the caries-associated phenotype of *S. mutans* and its adaptation to the human mouth environment.

Although low pH has been considered a primary ecological determinant influencing oral biofilm ecology, oxygen is also a critical factor (Marquis 1995), and it appears to be much better tolerated by commensal streptococci and other members of the normal microbiota than by *S. mutans* (Marquis 1995). 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 (Ahn and Burne 2007; Ahn et al. 2009). 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 *S. gordonii,* has been demonstrated to strongly inhibit *S. mutans* in mixed culture (Kreth et al. 2008). In addition to genes involved in acid tolerance, the unique core genes of *S. mutans* contain a high proportion of small peptides and gene products that could potentially be involved in oxidative stress tolerance (supplementary table S3, Supplementary Material online). For example, the secreted peptide encoded by SMU.1131c (ciaX) has been shown to modulate signal transduction in the CiaRH two-component system to regulate acid and oxidative stress tolerance and genetic competence (He et al. 2008). A variety of other peptides encoded in the unique core genome are in operons with transporters (e.g., SMU.18, SMU.185, and SMU.390), with transcriptional regulators that control gene expression in response to stressors (e.g., the MarR-type regulators associated with SMU.390, SMU.438, SMU.444, and SMU.631), or with proteins that may influence oxidative stress tolerance (e.g., SMU.545 is genetically linked to the *dpr* gene that encodes a protein necessary for oxidative stress resistance [supplementary table S3, Supplementary Material online]). Of note, a distinct MarR transcriptional

regulator of *S. mutans* (SMU.925) was recently demonstrated to regulate the expression of two ABC-exporters to link (p)ppGpp metabolism, genetic competence and oxidative stress tolerance, possibly through a peptide-based sensing system (Seaton et al. 2011). SMU.1047c is a small peptide genetically linked to the *relQ* operon, which encodes the RelQ (p)ppGpp synthase and three other gene products contributing to oxidative stress tolerance (Kim et al. 2012). Smu.185 is part of the *slo* operon, which regulates manganous ion internalization, oxidative stress tolerance, and expression of a variety of genes including superoxide dismutate. Likewise, SMU.1645 is predicted to mediate tellurite resistance, which is generally associated with thiol-mediated reduction of tellurite, whereas SMU.642 is cotranscribed in a complex operon with a predicted oxidoreductase, competence-associated peptide and peptidase; possibly linking oxidative stress and competence similar to the SMU.925 operon (Seaton et al. 2011). 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 (Abbe et al. 1991); and may be the underlying explanation for the relative prevalence of proteins with roles in oxidative stress carried in the unique core genome of *S. mutans* (supplementary table S3, Supplementary Material online).

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 nonlantiobiotic bacteriocins that can kill related organisms (Balakrishnan et al. 2002). 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 (Martin et al. 2006). The *hdrR* gene (SMU.1854) of the *S. mutans* unique core genome encodes a DNA binding protein that regulates bacteriocin production in a density-dependent fashion through its interactions with the HdrM membrane protein (SMU.1855) to control genetic competence and bacteriocin production (Merritt et al. 2007; Okinaga et al. 2010). 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 (Ahn et al. 2009). 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 with core genome, $P = 1.23e-10$; supplementary table S5, Supplementary Material online).

Conclusion

Taken collectively, our findings suggest that the *S. mutans* unique core genes may represent important pathogen-specific factors that could 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 may have 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 Ortholog Recovery

A total of 57 strains of S. mutans were selected representing different MLST and countries of origin: Brazil ($n = 16$), UK ($n = 29$), Iceland ($n = 5$), Hong Kong ($n = 2$), South Africa ($n = 2$), Turkey ($n = 2$), and USA ($n = 1$) (supplementary table S4, Supplementary Material online). Single end sequencing was performed using the Illumina GA2 sequencer, with one lane per strain. The sequence read lengths averaged 36–46 bp. This ensured high coverage ($\sim70\times$) 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 (Li, Ruan, et al. 2008), with appropriate mapping quality and coverage filters applied to capture the sequence information. Consensus alignments generated from the MAQ mapping were employed to extract, realign, and perform SNP calls in core genome genes; that is, genes for which all strains could be mapped to the reference. De novo assemblies were performed using Velvet (Zerbino and Birney 2008). Several hash lengths (from 21 to 31) and coverage cut-offs (0 to 80) were used, and the best assembly per strain was selected based on a combination of the N50 (which refers to a weighted median with 50% of the entire assembly contained in contigs equal to or greater than this value), the total assembly and the longest contig size. A schematic diagram for the mapping/assembly pipeline is included in supplementary figure S14, Supplementary Material online. Assembled genomes were annotated using the NCBI Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP). This pipeline combines HMM-based predictions with sequence similarity approaches to generate comparisons of the query genes or coding sequences against the gene products available in nonredundant protein databases: Entrez Protein Clusters, the Conserved Domain Database and the COGs (Clusters of Orthologous Groups). The algorithms employed in the pipeline are those implemented in GeneMark and Glimmer (Borodovsky and Mcininch 1993; Lukashin and Borodovsky 1998; Delcher et al. 1999).

To identify the orthologous sequences present in S. ratti, the putative sister group to S. mutans (Tapp et al. 2003), we sequenced the genome of one isolate of S. ratti using 454 GS-FLX with paired end, followed by partial gap closure with PCR. The sequence of this isolate was assembled using Newbler and annotated using PGAAP (supplementary fig. S15, Supplementary Material online). The rationale for sequencing S. ratti and identifying S. ratii/S. mutans orthologs was 2-fold: 1) identify orthologous genes that could be used as outgroups in a selection analysis that involved comparison of

polymorphism and divergence data; and 2) identify core genes present in S. mutans, but not the most closely related species—that is, the putative unique core genome of S. mutans. Orthologs were determined by performing an all-versus-all BLASTP search combined with clustering using OrthoMCL2 (Li et al. 2003), and included all the S. mutans de novo assembled genomes and the genome sequence for the sister group S. ratti. The history of the S. mutans unique genes could involve losses along the S. ratti lineage or gains via LGT along the S. mutans lineage. Core genes present in S. mutans but absent in other closely related mutans group streptococci, in addition to the sister group S. ratti, represent genes most likely to have been acquired by LGT. To identify such genes, we obtained draft genome sequences (454 GS-FLX, paired end) of two other mutans group streptococci, closely related to S. mutans: S. macacae, and S. criceti (Tapp et al. 2003). The sequences of S. criceti and S. macacae were assembled with Newbler and annotated using PGAAP. Once the protein clusters were identified, we performed an all versus all BLAST comparison of the putatively unique S. mutans genes with the annotated genes of these additional two species, using OrthoMCL2 (Li, Stoeckert, et al. 2003). A conservative similarity threshold of 50% was set as a measure of whether a gene in S. mutans had an ortholog in either of the other two more divergent species. The presence or absence of the genes putatively unique to S. mutans, in these other members of the mutans group, helped us distinguish which genes were lost in the S. ratti lineage and which ones were gained by S. mutans. Finally, additional BLAST searches against existing databases (NCBI nr), followed by maximum likelihood phylogenetic analysis of the aligned sequences, were employed to evaluate the evolutionary history and putative donor species of the unique genes identified as gained on the S. mutans lineage. A list of the additional streptococci genomes, included in this analysis (available at NCBI at the time), is provided in supplementary table S5, Supplementary Material online.

Genome sequence data for 57 strains of S. mutans and single strains each of S. ratti (FA-1), S. criceti (HS-6), and S. macacae (NCTC 11558) have been deposited in GenBank under the following accession numbers: S. ratti: AJTZ01000000; S. criceti: AEUV01000016.1; S. macacae: AEUW01000012.1. Accessions for the S. mutans isolates can be found in supplementary table S4, Supplementary Material online.

SNP Calling

The 1,430 genes constituting the core genome of S. mutans, were realigned at the 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 (Thornton 2003). Because of the deep coverage of our data ($>70\times$), we were confident in the call of rare variants (singletons) and no further sophisticated methods were employed for their identification. Each gene alignment was analyzed with the Polydnds program of the Libsequence suite developed by Thornton to estimate the number of

synonymous and nonsynonymous unambiguous changes in the alignment. These data were used in three ways: 1) genome-wide synonymous and nonsynonymous changes were employed to estimate the SFS for the synonymous and nonsynonymous changes respectively, as well as for the PCA to assess structure; 2) the synonymous and nonsynonymous changes per gene, with the inclusion of the outgroup *S. ratti* were used in selection analyses; 3). the number of synonymous and replacement positions for each gene served as denominators in assessing the polymorphism per site, as well as in estimating the mutation parameter θ (2Neu in haploid organisms) for the demographic and selection analyses.

Estimation of Recombination

Recombination, or more formally gene conversion, was estimated on the core genome alignment of the full data set using Sawyer's algorithm as implemented in GeneConv (Sawyer 1989), keeping only those tracts that were judged significant after Bonferroni correction for multiple comparisons, in a manner similar to which it was implemented in Leopold et al. (2009). This algorithm estimates the distribution of the length of identical sequences between any two pairs, considering the distribution of variant sites in the multiple sequence alignment. Using a permutation-based test, any identical fragment between pairs of sequences that is unusually long, as expected from the estimated distribution, is considered to be the result of a gene conversion (recombination) event. In addition to these analyses, we constructed three sets, each comprised of 600 genomic regions of 500-bp long. These sets were used to estimate the relative contribution of recombination to mutation as proposed by Didelot and Falush (2007), and implemented in the program ClonalFrame. The conditions of the run included 100,000 generations of burnin, 100,000 generations for each chain to run, and step sizes for the sampling of 100 generations.

Demographic and Selection Analyses

PCA (Patterson et al. 2006) of synonymous SNPs with frequencies greater than 5%, was performed using the R project for Statistical Computing (http://www.r-project.org/, last accessed December 15, 2012). Rare variants were not included, as these do not contribute to distinguishing relatedness among individuals in putative subpopulations. The frequency distribution of variants, or SFS, was calculated for synonymous and replacement changes independently in R. Demographic parameters for different competing models were estimated from the SFS of synonymous changes using a diffusion-based approximation implemented in the program $\delta a \delta i$ (Gutenkunst et al. 2009) in a maximum likelihood framework. The selection of the best-fit model was done using the Akaike Information Criteria. Changes in effective population size and time since the 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 between

4.08e−10 and 6.93e−10 subs/site/generation (Drake 1991; Ochman 2003), corresponding to 0.112–0.379 subst/genome/year (given that there are 374,571 synonymous sites along the genome), under the conservative assumption of a generation time of two divisions per day, as estimated for oral flora in vivo (Gibbons 1964). This mutation rate is consistent with experimentally determined mutation rates, estimated via fluctuation tests, in wild type clones of another member of the genus, *S. pyogenes*, at approximately $5.3 \times 10e{-}10$ mutations/generation (Scott et al. 2008). CIs of the parameters were estimated by maximum likelihood fitting of 500 bootstrapped data sets (details in Supplementary Material online). We also performed simulations under the estimated demographics and recombination rates, and performed estimations of the parameters in the simulated data sets using $\delta a \delta i$.

Genome wide selection analyses were performed on the replacement SFS by a similar diffusion-based approximation as the one employed for the demographic analysis. The impact of selection on the distribution of replacement mutations genome-wide was evaluated considering models that treat selection as a point mass effect or as a distribution of selective effects. Of these, three main models (one, two, or three categories of sites under selection with constant selection coefficients) were evaluated by maximum likelihood and over bootstrapped data sets, in the program PrFreq (Boyko et al. 2008). Again, the best model was selected using the Akaike Information Criteria. We also performed a standard MK test (McDonald and Kreitman 1991), and an extension of the MK test approach based on a Bayesian Loglinear model, implemented in SnIPRE (Eilertson et al. 2012), to compare the polymorphism and divergent changes in synonymous and replacement sites in the genes for which an orthologous sequence could be identified in *S. ratti* (82% of the core genes in *S. mutans*). The model implemented in SnIPRE is a Bayesian generalization of the log-linear model underlying the MK test that makes use of both genome-wide and gene-specific effects to model the variability among categories of mutations, resulting in increased power to detect the effects of selection on a gene-by-gene basis. It is nonparametric in the sense that it makes no assumptions (and does not require estimation) of parameters such as mutation rate and species divergence time to identify genes under selection.

Supplementary Material

Supplementary material, tables S1–S5, and figures S1–S15 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

References

Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. 2005. Defining the normal bacterial flora of the oral cavity. *J Clin Microbiol.* 43: 5721–5732.

Abbe K, Carlsson J, Takahashi-Abbe S, Yamada T. 1991. Oxygen and the sugar metabolism in oral streptococci. *Proc Finn Dent Soc.* 87: 477–487.

Ahn SJ, Browngardt CM, Burne RA. 2009. Changes in biochemical and phenotypic properties of *Streptococcus mutans* during growth with aeration. *Appl Environ Microbiol.* 75:2517–2527.

Ahn SJ, Burne RA. 2007. Effects of oxygen on biofilm formation and the AtlA autolysin of *Streptococcus mutans*. *J Bacteriol.* 189:6293–6302.

Ajdic D, McShan WM, McLaughlin RE, et al. (19 co-authors). 2002. Genome sequence of *Streptococcus mutans* UA159, a cariogenic dental pathogen. *Proc Natl Acad Sci U S A.* 99:14434–14439.

Alaluusua S, Matto J, Gronroos L, Innila S, Torkko H, Asikainen S, Jousimies-Somer H, Saarela M. 1996. Oral colonization by more than one clonal type of mutans *Streptococcus* in children with nursing-bottle dental caries. *Arch Oral Biol.* 41:167–173.

Arthur RA, Cury AA, Graner RO, Rosalen PL, Vale GC, Paes Leme AF, Cury JA, Tabchoury CP. 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.

Asanuma N, Hino T. 2002. Regulation of fermentation in a ruminal bacterium, *Streptococcus bovis*, with special reference to rumen acidosis. *Animal Sci J.* 73:313–325.

Balakrishnan M, Simmonds RS, Kilian M, Tagg JR. 2002. Different bacteriocin activities of *Streptococcus mutans* reflect distinct phylogenetic lineages. *J Med Microbiol.* 51:941–948.

Belli WA, Marquis RE. 1991. Adaptation of *Streptococcus mutans* and *Enterococcus hirae* to acid stress in continuous culture. *Appl Environ Microbiol.* 57:1134–1138.

Bender GR, Marquis RE. 1987. Membrane ATPases and acid tolerance of *Actinomyces viscosus* and *Lactobacillus casei*. *Appl Environ Microbiol.* 53:2124–2128.

Bender GR, Sutton SV, Marquis RE. 1986. Acid tolerance, proton permeabilities, and membrane ATPases of oral streptococci. *Infect Immun.* 53:331–338.

Borodovsky M, Mcininch J. 1993. Genmark—parallel gene recognition for both DNA strands. *Comput Chem.* 17:123–133.

Boyko AR, Williamson SH, Indap AR, et al. (14 co-authors). 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 4:e1000083.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* 140:783–796.

Buckee CO, Jolley KA, Recker M, Penman B, Kriz P, Gupta S, Maiden MC. 2008. Role of selection in the emergence of lineages and the evolution of virulence in *Neisseria meningitidis*. *Proc Natl Acad Sci U S A.* 105:15082–15087.

Burne R. 1998. Oral *Streptococci*: products of their environment. *J Dental Res.* 77: 445–452.

Bustamante CD, Wakeley J, Sawyer S, Hartl DL. 2001. Directional selection and the site-frequency spectrum. *Genetics* 159:1779–1788.

Caicedo AL, Williamson SH, Hernandez RD, et al. (12 co-authors). 2007. Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet.* 3:1745–1756.

Cheon K, Moser SA, Whiddon J, Osgood RC, Momeni S, Ruby JD, Cutter GR, Allison DB, Childers NK. 2011. Genetic diversity of plaque mutans streptococci with rep-PCR. *J Dent Res.* 90:331–335.

Cohen MN, Armelagos GJ. 1984. Paleopathology at the origins of agriculture. Orlando (FL): Academic Press.

Coventry A, Bull-Otterson LM, Liu X, et al. (29 co-authors). 2010. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun.* 1:131.

Croucher NJ, Harris SR, Fraser C, et al. (23 co-authors). 2011. Rapid pneumococcal evolution in response to clinical interventions. *Science* 331:430–434.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* 27: 4636–4641.

Diamond J. 2002. Evolution, consequences and future of plant and animal domestication. *Nature* 418:700–707.

Didelot X, Falush D. 2007. Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175:1251–1266.

Do T, Gilbert SC, Clark D, et al. (10 co-authors). 2010. Generation of diversity in *Streptococcus mutans* genes demonstrated by MLST. *PLoS One* 5:e9073.

Dobson AP, Carper ER. 1996. Infectious diseases and human population history. *Bioscience* 46:115–126.

Donati C, Hiller NL, Tettelin H, et al. (18 co-authors). 2010. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol.* 11:R107.

Drake JW. 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci U S A.* 88:7160–7164.

Dykhuizen DE. 1990a. Evolution. Mountaineering with microbes. *Nature* 346:15–16.

Dykhuizen DE. 1990b. Experimental studies of natural selection in bacteria. *Annu Rev Ecol System.* 21:373–398.

Dykhuizen DE. 1993. Chemostats used for studying natural selection and adaptive evolution. *Methods Enzymol.* 224:613–631.

Eilertson KE, Booth JG, Bustamante CD. 2012. SnIPRE: Selection inference using a Poisson random effects model. *PLoS Comput Biol.* 8: e1002806.

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

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

Fiennes R. 1978. Zoonoses and the origins and ecology of human disease. London: Academic Press.

Formicola V. 1987. Neolithic transition and dental changes—the case of an Italian site. *J Hum Evol.* 16:231–239.

Gerrish PJ, Colato A, Perelson AS, Sniegowski PD. 2007. Complete genetic linkage can subvert natural selection. *Proc Natl Acad Sci U S A.* 104:6266–6271.

Gibbons RJ. 1964. Bacteriology of dental caries. *J Dent Res.* 43(Suppl): 1021–1028.

Gong Y, Tian XL, Sutherland T, Sisson G, Mai J, Ling J, Li YH. 2009. Global transcriptional analysis of acid-inducible genes in *Streptococcus mutans*: multiple two-component systems involved in acid adaptation. *Microbiology* 155:3322–3332.

Griffiths RC, Tavare S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci.* 344: 403–410.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5:e1000695.

Guttman DS, Dykhuizen DE. 1994. Detecting selective sweeps in naturally occurring *Escherichia coli*. *Genetics* 138:993–1003.

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

He X, Wu C, Yarbrough D, Sim L, Niu G, Merritt J, Shi W, Qi F. 2008. Acid diffusion through extracellular polysaccharides produced by various mutants of *Streptococcus mutans*. *Mol Microbiol.* 70:112–126.

Hoen B, Chirouze C, Cabell CH, et al. (13 co-authors). 2005. Emergence of endocarditis due to group D streptococci: findings derived from the merged database of the International Collaboration on Endocarditis. *Eur J Clin Microbiol Infect Dis.* 24:12–16.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.

Hughes AL. 2005. Evidence for abundant slightly deleterious polymorphisms in bacterial populations. *Genetics* 169:533–538.

Hung WC, Tsai JC, Hsueh PR, Chia JS, Teng LJ. 2005. Species identification of mutans streptococci by groESL gene sequence. *J Med Microbiol.* 54:857–862.

Jacobson GR, Lodge J, Poy F. 1989. Carbohydrate uptake in the oral pathogen *Streptococcus mutans*: mechanisms and regulation by protein phosphorylation. *Biochimie* 71:997–1004.

Kim JN, Ahn SJ, Seaton K, Garrett S, Burne RA. 2012. Transcriptional organization and physiological contributions of the relQ operon of *Streptococcus mutans*. *J Bacteriol*. 194:1968–1978.

Klees AG, Linder D, Buckel W. 1992. 2-Hydroxyglutaryl-CoA dehydratase from *Fusobacterium nucleatum* (subsp. nucleatum): an iron-sulfur flavoprotein. *Arch Microbiol*. 158:294–301.

Klein MI, Florio FM, Pereira AC, Hofling JF, Goncalves RB. 2004. Longitudinal study of transmission, diversity, and stability of *Streptococcus mutans* and *Streptococcus sobrinus* genotypes in Brazilian nursery children. *J Clin Microbiol*. 42:4620–4626.

Kreth J, Zhang Y, Herzberg MC. 2008. Streptococcal antagonism in oral biofilms: *Streptococcus sanguinis* and *Streptococcus gordonii* interference with *Streptococcus mutans*. *J Bacteriol*. 190:4632–4640.

Lan R, Reeves PR. 2000. Intraspecies variation in bacterial genomes: the need for a species genome concept. *Trends Microbiol*. 8:396–401.

Lan R, Reeves PR. 2001. When does a clone deserve a name? A perspective on bacterial species based on population genetics. *Trends Microbiol*. 9:419–424.

Lapierre P, Gogarten JP. 2009. Estimating the size of the bacterial pan-genome. *Trends Genet*. 25:107–110.

Lefebure T, Bitar PD, Suzuki H, Stanhope MJ. 2010. Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. *Genome Biol Evol*. 2:646–655.

Leopold SR, Magrini V, Holt NJ, et al. (15 co-authors). 2009. A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis. *Proc Natl Acad Sci U S A*. 106:8713–8718.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*. 18:1851–1858.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13:2178–2189.

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

Lukacs JR. 1992. Dental paleopathology and agricultural intensification in south Asia: new evidence from Bronze Age Harappa. *Am J Phys Anthropol*. 87:133–150.

Lukashin AV, Borodovsky M. 1998. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*. 26:1107–1115.

Mansai SP, Innan H. 2010. The power of the methods for detecting interlocus gene conversion. *Genetics* 184:517–527.

Marquis RE. 1990. Diminished acid tolerance of plaque bacteria caused by fluoride. *J Dent Res*. 69 (Spec no. 672–675); discussion 682–683.

Marquis RE. 1995. Oxygen metabolism, oxidative stress and acid-base physiology of dental plaque biofilms. *J Ind Microbiol*. 15: 198–207.

Martin B, Quentin Y, Fichant G, Claverys JP. 2006. Independent evolution of competence regulatory cascades in streptococci? *Trends Microbiol*. 14:339–345.

Maruyama F, Kobata M, Kurokawa K, et al. (13 co-authors). 2009. Comparative genomic analyses of *Streptococcus mutans* provide insights into chromosomal shuffling and species-specific content. *BMC Genomics* 10:358.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351:652–654.

Merritt J, Zheng L, Shi W, Qi F. 2007. Genetic characterization of the hdrRM operon: a novel high-cell-density-responsive regulator in *Streptococcus mutans*. *Microbiology* 153:2765–2773.

Muzzi A, Donati C. 2011. Population genetics and evolution of the pan-genome of *Streptococcus pneumoniae*. *Int J Med Microbiol*. 301:619–622.

Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet*. 40:646–649.

Ochman H. 2003. Neutral mutations and neutral substitutions in bacterial genomes. *Mol Biol Evol*. 20:2091–2096.

Okinaga T, Niu G, Xie Z, Qi F, Merritt J. 2010. The hdrRM operon of *Streptococcus mutans* encodes a novel regulatory system for coordinated competence development and bacteriocin production. *J Bacteriol*. 192:1844–1852.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet*. 2:e190.

Phattarataratip E, Olson B, Broffitt B, Qian F, Brogden KA, Drake DR, Levy SM, Banas JA. 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.

Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol*. 6:526–538.

Scott J, Thompson-Mayberry P, Lahmamsi S, King CJ, McShan WM. 2008. Phage-associated mutator phenotype in group A *Streptococcus*. *J Bacteriol*. 190:6290–6301.

Seaton K, Ahn SJ, Sagstetter AM, Burne RA. 2011. A transcriptional regulator and ABC transporters link stress tolerance, (p)ppGpp, and genetic competence in *Streptococcus mutans*. *J Bacteriol*. 193: 862–874.

Shaver AC, Dombrowski PG, Sweeney JY, Treis T, Zappala RM, Sniegowski PD. 2002. Fitness evolution and the rise of mutator alleles in experimental *Escherichia coli* populations. *Genetics* 162:557–566.

Slatkin M, Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129:555–562.

Smith JM, Dowson CG, Spratt BG. 1991. Localized sex in bacteria. *Nature* 349:29–31.

Sniegowski PD, Gerrish PJ, Lenski RE. 1997. Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* 387:703–705.

Suzuki H, Lefebure T, Hubisz MJ, Pavinski Bitar P, Lang P, Siepel A, Stanhope MJ. 2011. Comparative genomic analysis of the *Streptococcus dysgalactiae* species group: gene content, molecular adaptation, and promoter evolution. *Genome Biol Evol*. 3:168–185.

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. *Int J Syst Evol Microbiol*. 53: 1861–1871.

Tettelin H, Masignani V, Cieslewicz MJ, et al. (46 co-authors). 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." *Proc Natl Acad Sci U S A*. 102:13950–13955.

Tettelin H, Riley D, Cattuto C, Medini D. 2008. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol*. 11:472–477.

Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* 19:2325–2327.

van Houte J. 1994. Role of micro-organisms in caries etiology. *J Dent Res*. 73:672–681.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 18:821–829.

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

Zhu L, Bustamante CD. 2005. A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics* 170: 1411–1421.