Infection, Genetics and Evolution 21 (2014) 269–278

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Comparative genomics of oral isolates of *Streptococcus mutans* by *in silico* genome subtraction does not reveal accessory DNA associated with severe early childhood caries ☆

Argimón Silvia [a,*], Konganti Kranti [b], Chen Hao [b], Alexander V. Alekseyenko [b], Brown Stuart [b], Page W. Caufield [a]

[a] New York University College of Dentistry, Department of Cariology and Comprehensive Care, 345 East 24th St, New York, NY 10010, USA
[b] Center for Health Informatics and Bioinformatics, New York University School of Medicine, 227 East 30th St, New York, NY 10016, USA

ARTICLE INFO

Article history:
Received 17 July 2013
Received in revised form 7 November 2013
Accepted 8 November 2013
Available online 26 November 2013

Keywords:
Comparative genomics
Software
Streptococcus mutans
Dental caries
Virulence
Pathogenesis

ABSTRACT

Comparative genomics is a popular method for the identification of microbial virulence determinants, especially since the sequencing of a large number of whole bacterial genomes from pathogenic and non-pathogenic strains has become relatively inexpensive. The bioinformatics pipelines for comparative genomics usually include gene prediction and annotation and can require significant computer power. To circumvent this, we developed a rapid method for genome-scale *in silico* subtractive hybridization, based on blastn and independent of feature identification and annotation. Whole genome comparisons by *in silico* genome subtraction were performed to identify genetic loci specific to *Streptococcus mutans* strains associated with severe early childhood caries (S-ECC), compared to strains isolated from caries-free (CF) children.

The genome similarity of the 20 *S. mutans* strains included in this study, calculated by Simrank k-mer sharing, ranged from 79.5% to 90.9%, confirming this is a genetically heterogeneous group of strains. We identified strain-specific genetic elements in 19 strains, with sizes ranging from 200 to 39 kb. These elements contained protein-coding regions with functions mostly associated with mobile DNA. We did not, however, identify any genetic loci consistently associated with dental caries, i.e., shared by all the S-ECC strains and absent in the CF strains. Conversely, we did not identify any genetic loci specific with the healthy group. Comparison of previously published genomes from pathogenic and carriage strains of *Neisseria meningitidis* with our *in silico* genome subtraction yielded the same set of genes specific to the pathogenic strains, thus validating our method.

Our results suggest that *S. mutans* strains derived from caries active or caries free dentitions cannot be differentiated based on the presence or absence of specific genetic elements. Our *in silico* genome subtraction method is available as the Microbial Genome Comparison (MGC) tool, with a user-friendly JAVA graphical interface.

1. Introduction

The study of the genetic basis of microbial pathogenesis has been guided by the molecular adaptation of Koch's postulates (Falkow, 1988), i.e., the notion that a disease phenotype should be associated with pathogenic strains of a species that harbor one or more genes associated with a virulence trait. Comparative genomics methods are particularly useful for the identification of genetic elements associated with virulence when pathogenic and non-pathogenic strains of a bacterial species can be isolated. For example, DNA subtractive hybridization has led to the identification of pathogenicity islands and clone-specific markers for several bacterial pathogens (Winstanley, 2002).

With high-throughput sequencing technologies, it is now possible to identify virulence determinants by interrogating *in silico* a large number of whole genomes from bacterial strains isolated from diseased or healthy hosts (Hu et al., 2011). The availability

Abbreviations: S-ECC, severe early childhood caries; CF, caries-free; ORFs, open-reading frames; CDF, chromosomal DNA fingerprinting; PCoA, principal coordinate analysis; UPGMA, unweighted pair group method with arithmetic mean.

* Corresponding author. Tel.: +1 (212) 998 9250.
E-mail address: sa91@nyu.edu (S. Argimón).

1567-1348/$ - see front matter © 2013 The Authors. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2013.11.003

of multiple genome sequences for a bacterial species, has also led to the pan-genome concept, or the sum of the core genes shared among all sequenced strains of a species, and the accessory genes that are present in at least one, but not all strains studied (Tettelin et al., 2005). Popular programs for comparative genomics of whole genomes such as MAUVE (Darling et al., 2004), ACT (Carver et al., 2005), or MUMmer (Laing et al., 2011), can align entire genomes to highlight regions of similarity and synteny, but may not represent the most practical approach for the rapid identification of accessory sequences common to a large group of strains. Instead, a few *in silico* comparative genomics methods have been described that apply the rationale behind DNA subtractive hybridization to whole genome sequences for the identification of group-specific genes. mGenomeSubtractor (Shao et al., 2010) and FindTarget (Chetouani et al., 2001) only take into account protein coding regions. The novel region finder (NRF) module of Panseq (Laing et al., 2011) is available as a standalone version, but it requires knowledge of Perl and Unix-based systems.

We developed an *in silico* genome subtraction method for the rapid identification of genetic elements specific to a group of strains. The accompanying software called Microbial Genome Comparison (MGC) tool is described in detail elsewhere (Chen et al., 2013), and is available as a Java executable from SourceForge. This tool performs *in silico* genomic comparisons independently of feature identification and annotation, an advantage over more comprehensive, yet time-consuming pipelines of comparative genomics. Instead, the MGC tool consists of the *in silico* fragmentation of the genome sequences followed by a series of between groups and within groups blastn queries (Altschul et al., 1997). In this study, we applied the *in silico* genome subtraction method as implemented in the MGC tool to the comparison of 20 genome sequences of *Streptococcus mutans,* commonly referred to as the main etiological agent of dental caries.

Dental caries remains the most common chronic disease of childhood in the United States, with a prevalence of 41% among children 2–11 years of age (Roberts, 2008). Severe early childhood caries (S-ECC) is an extremely destructive form of dental caries affecting the primary dentition of children six years and younger (AAPD, 2004). The association between *S. mutans* and S-ECC has been well documented both by culture-based (Loesche et al., 1975; Marchant et al., 2001; Milnes and Bowden, 1985; Tanner et al., 2011; van Houte et al., 1982) and culture-independent surveys (Becker et al., 2002; Corby et al., 2005; Kanasi et al., 2010) of the dental biofilm. Individuals free of detectable caries can, however, harbor *S. mutans* in their dental biofilm (Ge et al., 2008; Loesche, 1986; Marchant et al., 2001; Tanner et al., 2011), and the presence or total numbers of *S. mutans* are poor predictors of subsequent caries activity (Thenisch et al., 2006).

Whether specific genotypes of *S. mutans* are associated with S-ECC, and are different than genotypes colonizing caries-free (CF) children, has not been determined, but the evidence suggests that strains can differ in virulence (Fitzgerald et al., 1983; Kohler and Krasse, 1990). The remarkable intra-species genetic variability exhibited by *S. mutans* has been extensively documented by restriction enzyme fingerprinting (Caufield and Walker, 1989; Kulkarni et al., 1989), MLST (Do et al., 2010; Nakano et al., 2007), comparative genome hybridization (Zhang et al., 2009), and the comparison of two sequenced genomes (Ajdic et al., 2002; Maruyama et al., 2009). The virulence of *S. mutans* strains may be associated with the presence or absence of regions of accessory DNA, so we queried the genome sequences of 10 S-ECC and 10 CF *S. mutans* strains through *in silico* genome subtraction with the MGC tool to identify differences in their genetic repertoire that correlate with differences in caries experience.

2. Materials and methods

2.1. Bacterial strains

This study included 20 *S. mutans* strains previously isolated, 10 from children diagnosed with severe early childhood caries (S-ECC) and scheduled for extensive caries restorative treatment under general anesthesia at the Bellevue Hospital, New York, NY, and 10 from children diagnosed as being free from detectable caries (caries-free [CF]) (Argimon and Caufield, 2011). Bacterial samples from saliva and pooled plaque of the S-ECC and CF children were collected. Additionally, for S-ECC children plaque samples were obtained from caries lesions. *S. mutans* isolates were selected from mitis salivarius-bacitracin agar medium (MSB) based upon colony morphology (Argimon and Caufield, 2011). The study protocol for human subjects was approved by the Institutional Review Board of New York University School of Medicine and Bellevue Hospital. Our study cohort of S-ECC children conformed to the recent reclassification of S-ECC as hypoplasia-associated severe early childhood caries (HAS-ECC) (Caufield et al., 2012). In most cases, only 1 *S. mutans* genotype was isolated from each subject, though 4 subjects presented 2 genotypes, and 2 subjects presented 3 genotypes. The 20 strains represented 20 different *S. mutans* genotypes by chromosomal DNA fingerprinting (CDF) and arbitrarily-primed PCR (AP-PCR), as previously reported (Argimon and Caufield, 2011). Total genomic DNA was obtained as previously described (Argimon and Caufield, 2011).

2.2. Genome sequencing and assembly

Twenty *S. mutans* genomes, 10 from S-ECC children and 10 from CF children, were multiplexed and library preps were generated using the Illumina TruSeq DNA Sample Prep Kit according to manufacturer's instructions. Libraries were sequenced on the Illumina HiSeq 2000 Genome Analyzer System (coverage of ~100×), and the 50 bp paired-end reads thus generated were assembled *de novo* into contigs using ABySS (Simpson et al., 2009). The contigs for each sample were reordered based on alignment to the reference genome sequence of strain UA159 with Mauve Contig Mover (Rissman et al., 2009).

2.3. Estimation of genome sequence similarity

We used a simple, rapid, computationally efficient and scalable method based on the sharing of short DNA words (k-mers) between genome sequences. The pairwise similarity between genome sequences was estimated with Simrank, which computes the similarity between two sequences as the number of unique k-mers shared, divided by the smallest total unique k-mer count in either sequence (DeSantis et al., 2011). This method requires no annotation of the genomes, treats all portions of the genome equally, and can even be applied to sequence reads that have not been assembled into contigs. The complete genome sequences of *S. mutans* strains UA159, NN20225, GS-5 and LJ23 (Table 1), and *Streptococcus agalactiae* strains (NC_004368.1) and 2603 V/R (NC_004116.1) were included as a reference. A k-mer length of 10 was chosen empirically and validated by comparison of the results to previously published similarity values for *S. agalactiae* genomes (Tettelin et al., 2005). The contigs in each *S. mutans* draft genome sequence were concatenated and separated by a stretch of Ns equal to the length of the k-mer. Dissimilarity (100-similarity) matrices of pairwise comparisons between genome sequences were employed for hierarchical clustering of the genomes based on the unweighted pair group method with arithmetic mean (UPGMA), as implemented in the seqlinkage tool in the Matlab