Infection, Genetics and Evolution 9 (2009) 179–188

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Gene content differences across strains of *Streptococcus uberis* identified using oligonucleotide microarray comparative genomic hybridization

Ping Lang [a], Tristan Lefébure [a], Wei Wang [b], Ruth N. Zadoks [c,1], Ynte Schukken [c], Michael J. Stanhope [a,*]

[a] Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA
[b] Life Sciences Core Laboratories Center, Cornell University, Ithaca, NY 14853, USA
[c] Quality Milk Production Services, 22 Thornwood Drive, Cornell University, Ithaca, NY 14850-1263, USA

ARTICLE INFO

Article history:
Received 29 May 2008
Received in revised form 16 October 2008
Accepted 26 October 2008
Available online 14 November 2008

Keywords:
Bovine mastitis
Streptococcus uberis
Microarray
Host-adaptation

ABSTRACT

Streptococcus uberis is one of the principal causative agents of bovine mastitis. The organism is typically considered an environmental pathogen. In this study, two multilocus sequence typing (MLST) schemes and whole genome DNA microarrays were used to evaluate the degree and nature of genome flexibility between *S. uberis* strains. The 21 isolates examined in this study arise from a collection of 232 international isolates for which previous epidemiological and preliminary genotyping data existed. The microarray analysis resulted in an estimate of the core genome for *S. uberis*, consisting of 1530 ORFs, among 1855 tested, representing 82.5% of the *S. uberis* 0140J genome. The remaining ORFs were variable in gene content across the 21 tested strains. A total of 26 regions of difference (RDs), consisting of three or more contiguous ORFs, were identified among the variable genes. Core genes mainly encoded housekeeping functions, while the variable genes primarily fell within categories such as protection responses, degradation of small molecules, laterally acquired elements, and two component systems. Recombination detection procedures involving the MLST loci suggested *S. uberis* is a highly recombinant species, precluding accurate phylogenetic reconstructions involving these data. On the other hand, the microarray data did provide limited support for an association of gene content with strains found in multiple cows and/or multiple herds, suggesting the possibility of genes related to bovine transmissibility or host-adaptation.

1. Introduction

Streptococcus uberis is one of the main causes of bovine mastitis, which remains the most economically important infectious disease of dairy cattle throughout the world (Leigh, 1999). Because of its widespread occurrence in the environment, *S. uberis* is usually considered an environmental pathogen. Strain typing data show that a large variety of strains can cause mastitis, supporting the idea that *S. uberis* is an opportunistic pathogen of environmental origin (Leigh, 1999; Douglas et al., 2000; Wieliczko et al., 2002; McDougall et al., 2004). However, recent epidemiological studies suggest that predominant strains and clonal complexes occur in many herds (Phuektes et al., 2001; Zadoks et al., 2001, 2003; Coffey et al., 2006; Tomita et al., 2008). The occurrence of predominant strains suggests cow-to-cow transmission of the pathogen, and/or the possible existence of a host-adapted ecotype that differs from environmental strains of *S. uberis* in chronicity of infection, as well as perhaps ability to cause mastitis (Zadoks et al., 2003; Zadoks, 2007; Tomita et al., 2008).

S. uberis has been characterized using a variety of DNA-based methods, including restriction endonuclease fingerprinting (Hill and Leigh, 1989), random-amplified polymorphic DNA (RAPD) typing (Gillespie et al., 1998; Wieliczko et al., 2002; Zadoks et al., 2003), repetitive extragenic palindromic (REP)-PCR (Wieliczko et al., 2002), ribotyping (Zadoks et al., 2004, 2005b) and pulse-field gel electrophoresis (Douglas et al., 2000; Phuektes et al., 2001; Wieliczko et al., 2002). Recently, two multilocus sequence typing (MLST) schemes were developed to enhance our understanding of the population biology of this important pathogen (Zadoks et al., 2005a; Coffey et al., 2006). Zadoks et al. (2005a) designed an MLST scheme based on six loci involving both housekeeping and virulence genes and subsequently Coffey et al. (2006) established an MLST scheme which makes use of seven housekeeping genes. Based on the Coffey et al. scheme, an MLST profile database

* Corresponding author. Tel.: +1 607 253 3859; fax: +1 607 253 3083.
E-mail address: mjs297@cornell.edu (M.J. Stanhope).
[1] Current address: Moredun Research Institute, Pentland Science Park, Bush Loan, EH26 0PZ Scotland, United Kingdom.

1567-1348/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2008.10.015

containing the sequences of all alleles and the allelic profiles (STs) is maintained at the University of Oxford (http://pubmlst.org/suberis).

These two MLST schemes have great advantages over the previous molecular typing techniques as tools to investigate *S. uberis* epidemiology and genetic diversity on a global scale. However, these analyses of variation are based on a limited number of genetic loci and tell us nothing about the variability and evolution of genome composition of the species. Currently, a single genome sequence for *S. uberis* has become publicly available (Ward et al., in press). Recent studies of comparative genomics of *Streptococcus*, as well as other genera, indicate that many bacteria species are typified by both a core and pan genome, where the core genome consists of the genes common to all strains for that taxon, and the pan genome consists of the core genome plus a dispensable portion, which consists of genes present in one or more strains, but not common to all (Tettelin et al., 2005; Kettler et al., 2007; Lefebure and Stanhope, 2007). Thus, a single genome sequence does not provide the ability to assess the core genome composition of a species, which is arguably that which defines the essence of that taxon (see for e.g. Lan and Reeves, 2001). Microarray hybridizations are an effective means of assessing gene presence and absence across strains of a species and have been used for this purpose in several streptococci and other pathogens (e.g. Salama et al., 2000; Dorrell et al., 2001; Hakenbeck et al., 2001; Fukiya et al., 2004; Zhou et al., 2004; Gressmann et al., 2005; Brochet et al., 2006). They are a particularly good way to determine core genome composition, but because of the use of a single, or at least limited number of reference genomes, not an effective way to evaluate pan genome composition. The purpose of this study is to make use of the recently available *S. uberis* genome sequence concomitant with MLST evaluations of genotype and microarray hybridizations of carefully chosen *S. uberis* isolates to (1) determine the core genome composition of the species *S. uberis* and (2) evaluate the degree and nature of genome flexibility between *S. uberis* strains.

2. Materials and methods

2.1. Bacterial strains

Bacterial isolates were obtained from samples collected in California and New York (USA), The Netherlands, France, Italy, and New Zealand. Of 21 isolates, 1 originated from bulk tank milk from cows, 19 from cow or quarter milk samples, and 1 from a buffalo milk sample. Species identity of all isolates was confirmed by means of PCR (Hassan et al., 2001). Isolates were selected from a collection of 232 isolates that had been previously characterized by means of MLST using the method of Zadoks et al. (2005a), representing multiple continents, countries, and herds. Isolates were further characterized, based on epidemiological, and these preliminary genotyping data, as belonging to sequence types (STs) that were unique in the collection of 232 isolates, or conversely, were found in multiple animals in a herd, or in multiple herds. This latter category included a ST that caused an outbreak of *S. uberis* mastitis in a dairy herd in the Netherlands (Zadoks et al., 2003), and also found to be predominant in a herd from France and one from New York State. For several STs that were isolated from multiple herds, two isolates representing different herds were included in the analysis. These earlier data then, served as a means of choosing isolates for the present comparative analysis, encompassing a range of genotypic diversity, geographic diversity, as well as putative differences in epidemiology. *S. uberis* strain 0140J was used as reference strain for microarray analysis (see below). Preliminary epidemiology and genotyping data of ours indicate that 0140J has the same six gene MLST profile as a set of isolates from a *S. uberis* outbreak in California. Multiple animals in that herd were infected with the same strain, and 0140J originates in the UK, so this also means this ST is found in more than one country. We recognize that a tentative classification of isolates as either unique ST, or recurring ST found in multiple animals/multiple herds, is not without its shortcomings, at least partly because a larger sample set may ultimately reveal that some isolates currently recognized as unique ST, may appear in multiple cows, multiple herds, and different countries, and also because isolates with the same ST can sometimes display different PFGE patterns, and therefore not necessarily be genetically closely related. Nonetheless, it is true that the designation of unique ST does arise from an international collection of 232 isolates, and this does suggest something different about the adaptive attributes of these compared to STs more commonly associated with mastitis. An overview of isolates is provided in Table 1. Isolates were also characterized using the MLST method of Coffey et al. (2006). Isolate and source data can be found in Cornell's publicly accessible strain database PathogenTracker 2.0 at www.pathogentracker.net.

2.2. Multilocus sequence typing and recombination detection

MLST was performed by sequencing the seven housekeeping genes (*arc*C, *ddl*, *gki*, *rec*P, *tdk*, *tpi* and *yqi*L) comprising the Coffey et al. (2006) scheme, as well as the six genes of the Zadoks et al. scheme (*cpn*60, *gap*C, *opp*F, *pau*A, *sod*A and *tuf*). In the case of the Coffey et al. loci, allele numbers were defined by comparing our sequences to the *S. uberis* MLST database http://pubmlst.org/suberis. Sequence types were defined by analyzing both the allelic profile of the 7 MLST genes from the Coffey et al. scheme and all 13 MLST genes; eBURST (Feil et al., 2004) was used to assess possible relationships between STs. Putative recombinational histories involving these loci were investigated using MaxChi (Smith, 1992; Smith et al., 1993), implemented in the START program (Jolley et al., 2001), Phi (Bruen et al., 2006) implemented in Splitstree (Huson, 1998; Huson and Bryant, 2006), and a phylogenetic approach involving heatmap plots of the approximately unbiased (AU) test (Shimodaira, 2002). This latter approach was originally proposed by Susko et al. (2006) with subsequent refinements by us, detailed in Lefebure and Stanhope (2007). Briefly, it involves testing for the rejection of a set of topologies by a set of orthologous genes using the approximately unbiased test (Shimodaira, 2002). When a gene rejects a tree that is supported by the majority of the other genes, this gene is considered to have been laterally transferred. Tested topologies include the individual gene trees obtained by phyML ($GTR + G_4 + I$ model of evolution with a BIONJ starting tree (Guindon and Gascuel, 2003)), with the addition of the tree obtained with PAUP ($GTR + G_4 + I$ model of evolution, NJ starting tree, and a TBR branch-swapping algorithm) reconstructed from the concatenation of all genes. The AU test is then applied using CONSEL (Shimodaira and Hasegawa, 2001). AU test results (p-values for the rejection of each tree) are plotted using heatmaps obtain with *R*. In order to evaluate whether a tree is not rejected because it is not significantly different or because it is simply unresolved we complement this initial AU/heatmap analysis with a second set of analyses, intended to quantify the amount of supported and incongruent phylogenetic signal between two gene trees. This approach relies on the discovery of well-supported bipartitions (i.e. branches), that are not observed on other trees, as measured by non-parametric bootstrap analysis (Felsenstein, 1985), thus revealing incongruence between gene histories.