FEDERAL UNIVERSITY OF MINAS GERAIS

BIOLOGIC SCIENCES INSTITUTE

BIOINFORMATIC POST-GRADUATE INTERUNIT PROGRAM

Master dissertation

GENOME SEQUENCING AND COMPARATIVE GENOME ANALYSIS OF *Streptococcus dysgalactiae* subsp. *dysgalactiae,* AN EMERGING PATHOGEN OF NILE TILAPIA

By: Alexandra Antonieta Urrutia Zegarra

Advisor: Prof. Henrique César Pereira Figueiredo, DMV, Ph.D.

Belo Horizonte, MG, August of 2017

Alexandra Antonieta Urrutia Zegarra

GENOME SEQUENCING AND COMPARATIVE GENOME ANALYSIS OF THE EMERGING FISH PATHOGEN *Streptococcus dysgalactiae* subsp. *dysgalactiae*

Dissertation presented to the Bioinformatic Post-Graduate Interunit Program of the Federal University of Minas Gerais to obtain the title of Master in Bioinformatics.

Concentration area: Genomic Bioinformatics

Advisor: Prof. Henrique César Pereira Figueiredo, DMV, Ph.D.

Belo Horizonte, MG, August of 2017

"*Joq quimico tatichiy kawsay*"

Paul Ehrlich

"Taytayta, mamayta ñañaytawan.

Tioyta Jose Antonio kunan astawan qaylla ch'askakunamanta kashan."

ACKNOWLEDGMENTS

Thanks to the Federal University of Minas Gerais for the training offered;

To the CNPq, FAPEMIG, CAPES and the Ministry of Fishery and Aquaculture for the funding of this project;

To the coordination, professors and colleagues of the Post-Graduation course in Bioinformatics on the UFMG;

To the members of the dissertation defense committee for accepting the invitation to evaluate this work;

To my advisor Prof. Dr. Henrique César Pereira Figueiredo for the opportunity, for sharing his experience in order to improve the quality of this project and specially for letting me be part of his amazing team in AQUACEN;

To the AQUACEN team for letting me feel as part of the family, to Felipe for all the teaching and all the knowledge shared;

To the AQUAGIRLS for their awesome friendship, millions of laughs and incredible support;

To Julio Cesar H. S. for the biggest support all over these couple of years in Brazil;

To my family, you are my everything, and this achievement is as mine as yours, I love you so much;

To all of you, thank you very much.

SUMMARY

ABBREVIATION LIST

GCS: Group C streptococci

GGS: Group G streptococci

GAS: Group A streptococci

SDD: *Streptococcus dysglactiae* subsp. *dysgalactiae*

SDE: *Streptococcus dysglactiae* subsp. *equisimilis*

PFGE: Pulsed-Field Gel Electrophoresis

NGS: Next generation sequencing

PEL: Paired-end library

MPL: Mate-pair library

WGC: Whole genome coverage

NGS: Next Generation Sequencing

MLST: Multilocus Sequence Typing

MLSA: Multilocus Sequence Analysis

VFDB: Virulence Factor Database

GEI: Genomic Island

PAI: Pathogenicity Island

RI: Resistance Island

CDS: Coding DNA Sequence

ABSTRACT

Streptococcus dysgalatiae subsp. *dysgalactiae* (SDD) is a Gram-positive cocci, that autoaggregates in saline solution, it is catalase negative and forms long chains in growth medium. On fish, the disease is characterized with clinical signs of septicaemia and a typical form of necrosis in the caudal peduncle with a high mortality rate. In 2002, it caused the first outbreak in southern Japanese farms and during the subsequent years fish farms all over the country suffered losses. On Brazil, outbreaks of streptococcosis are common in the freshwater fish species Nile tilapia, *Oreochromis niloticus* (L.) and in 2007, the first disease outbreak caused by SDD was spotted on the state of Ceará. Nowadays it is considered as an emergent pathogen therefore, considering the importance of a complete genome to characterize this pathogen; a next-generation sequence genome initiative was managed. Three strains, SD64, SD92 and SD192, were sequenced and assembled in order to perform genomic comparative analysis within other SD strains. Results confirm robust and coherent cluster within S. *dysgalactiae* subsp. *equisimilis* (SDE) and SDD strains. MLST analysis also showed additional host dependent clustering within SDD strains, this presumably shows that the SDD strains maybe host-adapted. Plus, higher similarity within SDE strains than between SDD strains reveals that even within the same subespecies the strains have different features among them. Final results propose SDD adaptation to changing environments and new hosts presumably involved with the acquisition of virulence factor and other features from other species.

DOCUMENT STRUCTURE

The dissertation was divided into three chapters:

- the first chapter consists of a bibliographical revision of the previous topics needed for the study and the objectives;
- the second one includes the results obtained in the analyzes performed, in this chapter an article in the form of scientific paper is presented. The article, is entitled "Comparative genomics of three *Streptococcus dysgalactiae* subsp. *dysgalactiae* strains isolated from fish" and will be submitted to the Frontier in Microbiology or Aquaculture Journal; and
- Finally, in the last chapter, concluding remarks together with future perspectives are presented.
- The scientific content generated during the course of this work is attached at the end of the current document as annexes:
    - Scientific paper entitled "Use of MALDI-TOF Mass Spectrometry for the Fast Identification of Gram-Positive Fish Pathogens " published on *Frontiers in Microbiology* journal. (https://doi.org/10.3389/fmicb.2017.01492)
    - Abstract and banner presented during the X-Meeting 2016 - 12th International Conference of the Brazilian Association of Bioinformatics and Computational Biology, entitled "The first complete genome sequence of *Streptococcus dysgalatiae* subsp. *dysgalactiae* an emerging fish pathogen"

CHAPTER I

Introduction

Streptococcus dysgalactiae

Streptococcus dysgalactiae is a Gram-positive bacterium usually found in animals, it can be isolated from the udders of cows with mild mastitis and from blood and tissues of lambs with polyarthritis (Gaviria & Bisno, 2000). This species was initially considered as non-pathogenic for humans; however, it is now recognized as an increasingly important human pathogen that may cause several diseases (Hughes, Wilson, Brandt, & Spellerberg, 2009). It has been characterized in veterinary medicine as the cause of bovine mastitis (Whist, Østerås, & Sølverød, 2007) and recently as an important fish pathogen (Netto, Leal, & Figueiredo, 2011). While there are phylogenetic analyzes based on rRNA sequences which suggested that *S. dysgalactiae* is closely related to *Streptococcus pyogenes*due to the similar clinical situations, sharing niches for colonization and the evolutionary relationship indicates lateral gene transfer interspecies (Davies, McMillan, Van Domselaar, Jones, & Sriprakash, 2007).

The name was first used in 1932, by Dierhofer who described a streptococci of veterinary origin (Diernhofer, 1932). In addition, Frost reported the discovery of a similar human pathogen, which he named *Streptococcus equisimilis* (Frost, 1940). In parallel, Rebecca Lancefield incorporated a method of classification of streptococci based on their carbohydrate-antigens and successfully described both of the previously named *Streptococcus* as belonging to group C and group G (Lancefield, 1933) respectively. Years later, the *S. dysgalactiae* isolated from bovine was reported to be identical to *S. equisimilis*, except for the absence of beta-hemolysis (Breed, Murray, & Hitchens, 1948). Lancefield's grouping method was the favorite within laboratories at the time, which resulted in the disuse of the previously coined names *S.*

dysgalactiae and *S. equisimilis*, this led to both species losing standing in nomenclature when they were not included on the Approved Lists of Bacterial Names (Skerman & Sneath, 1980).

Three years later, the name *S. dysgalactiae* was revived, but only as reference of the alpha-hemolytic, group C strains of bovine origin (Garvie, Farrow, & Collins, 1983). Following DNA hybridization studies, revealed extensive similarities between *S. dysgalactiae, S. equisimilis*, and streptococci belonging to serogroups G and L which exhibited high levels of DNA-DNA binding and therefore belonged to a single species: *S. dysgalactiae* (Farrow & Collins, 1984). However, subsequent molecular investigations indicated heterogeneity within this new species, and led to a subdivision in 1996. Vandamme divided *S. dysgalactiae* into two subspecies: *S. dysgalactiae* subspecies *equisimilis* and *S. dysgalactiae* subspecies *dysgalactiae* (Vandamme, Pot, Falsen, Kersters, & Devriese, 1996). And as this classification is an ongoing debate the subspecies are now characterized as *S. dysgalactiae* subsp. *equisimilis*, a large human colony formed by group C and G streptococci, and *S. dysgalactiae* subsp. *dysgalactiae*, group C streptococci (GCS) with an animal origin (Rantala, 2014).

In the year 2004, in Japan, a GCS was isolated from cultures of *Seriola dumerili* and *S. garvieae.* The bacteria isolated were Gram-positive cocci, self-aggregated in saline solution, forming large chains in culture medium, catalase negative and alfa-hemolytic in blood agar. An almost complete genetic sequence of 16S rDNA from two isolated strains was determined and compared to the available strains in the databases (Nomoto, *et al.*, 2004). *S. dysgalactiae* was identified based on the results of the 16S rDNA sequence and the serological properties of the Lancefield groups. The severe necrotic lesions observed in the experiments were the same as those found in fish naturally infected; this was the first report of a fish infection of *S. dysgalactiae.* This infection was then characterized as severe necrotic lesions of the caudal peduncle associated with high mortality (Nomoto, *et al.*, 2006).

Since then, *S. dysgalactiae* has been isolated from different origins and geographic locations as showed on Table 1.

Host	**Country**	**Year of Isolation**	**Reference**
Seriola garviae	Japan	2004	(Nomoto, *et al.*, 2004)
Seriola. lalandi	Japan	2007	(Abdelsalam, Eissa, & Chen, 2015a)
Seriola quinqueradiata	Japan	2007	(Abdelsalam, Eissa, & Chen, 2015a)
Seriola. dumerili	Japan	2006	(Abdelsalam, Eissa, & Chen, 2015a)
Rachycentron canadum	Taiwan	2008	(Abdelsalam, Chen, & Yoshid, 2010)
Liza alata	Taiwan	2007	(Abdelsalam, Chen, & Yoshid, 2010)
Mugil cephalus	Taiwan	2005	(Abdelsalam, Eissa, & Chen, 2015a)
Trachinotus ovatus	China	2007	(Zhou, Li, Ma, & Liu, 2007)
Acipenser schrenckii	China	2009	(Yang & Li, 2009)
Acipenser baerii	China	2009	(Pan, *et al.*, 2009)
Ctenopharyngodon idella	China	ND	(Abdelsalam, Eissa, & Chen, 2015a)
Carassius carassius	China	ND	(Abdelsalam, Eissa, & Chen, 2015a)
Liza haematocheila	China	ND	(Abdelsalam, Eissa, & Chen, 2015a)
Trachinotus blochii	China	2008	(Abdelsalam, Chen, & Yoshid, 2010)
Oreochromis sp	Indonesia	2004	(Abdelsalam, Eissa, & Chen, 2015a)
Lutjanus stellatus	Malaysia	2004	(Abdelsalam, Eissa, & Chen, 2015a)
Trachinotus blochii	Malaysia	2005	(Abdelsalam, Eissa, & Chen, 2015a)
Oreochromis niloticus	Brazil	2007	(Netto, Leal, & Figueiredo, 2011)
Oncorhynchus mykiss	Iran	2008	(Pourgholam, *et al.*, 2011)
Oreochromis spp.	Egypt	2015	(Abdelsalam, Elgendy, Shaalan, Moustafa, & Fujino, 2017)

ND: Not determined.

Table 1. - Geographic distribution of *Streptococcus dysgalactiae* isolated from fish.

However, in contrast to other fish pathogenic, streptococci diseases related to *S. dysgalactiae* had been restricted mostly to the Asian continent, until the year 2009, when a report of an isolation and description of an infection of *S. dysgalactiae* on Nile tilapia (*Oreochromis niloticus* L.*)* from Brazil was described (Netto, Leal, & Figueiredo, 2011).

On 2014, Costa and collaborators performed studies of the genotyping of SD strains isolated from infected fish (Costa, Leal, Leite, & Figueiredo, 2014), 21 strains among four farms in different Brazilian states were isolated and characterized using pulsed-field gel electrophoresis (PFGE), ERIC-PCR, REC-PCR and *sodA* gene sequencing. Identical sequences of the *sodA* gene were obtained from all the isolates, ERIC-PCR and REP-PCR were unable to discriminate within isolates. However, the study probed PFGE as the best genotyping method for this pathogen and establish three different genetic patterns, based on a similarity threshold of 80%, all of them showing a relationship with its state of origin. The three strains selected for this work represent each one of those patterns.

Next generation sequencing

Ongoing revolution in sequencing technology has led to the production of sequencing machines with dramatically lower costs and higher throughput than the technology of just few years ago (Mardis, 2008). Since it was described by Sanger in 1977, sequencing has undergone major changes, from long sessions in the laboratory to the generation of large amounts of data, in a short time emulating a mass production of biological data. Next generation sequencing (NGS) impact on genomics is in turn causing a revolution in genetics that, because of a variety of factors, will fundamentally change the nature of genetic experimentation (Mardis, 2008).

Over the years, many sequencing platforms have been developed, from these, Roche 454 pyrosequencing, Ion Torrent Personal Genome Machine (PGM), and Illumina HiSeq with their bench-top versions (454 Jr, PGM, and MiSeq, respectively) have been extensively applied to bacterial genome sequencing (Loman, *et. al.*, 2012).

454 technology is based on pyrosequencing, a non-electrophoretic, bioluminescence method that measures the realese of inorganic pyrophosphate by proportionally converting it into visible light using a serie of enzymatic reactions (Metzker, 2010). The light emitted is directly proportional to the amount of a particular nucleotide incorporated (up to the level of detector saturation). Hence, for runs of multiple nucleotides (homopolymers), the linearity of response can exceed the detector sensitivity, at which indel errors can occur in those reads (Mardis, 2008)

IonTorrent is based on the detection of hydrogen ions that are released during the polymerization of DNA (Rothberg, *et al.,* 2011). IonTorrent has within its major benefits rapid sequencing speed and low operating costs which has been possible by the avoidance of modified nucleotides and optical measurements (Perkel, 2011). This platform is a suitable option for microbiology studies, provided that researchers are consistent in DNA extraction methods, PCR protocols, and bioinformatics pipeline (Indugu, *et. al.,* 2016).

In other hand, the Illumina Genome Analyzer was first introduced in 2006 and it is based on the concept of 'sequencing by synthesis' (SBS), after the fragments amplification each cycle will incorporate a base followed by an imaging step to identify the added nucleotide (Mardis, 2008). Interests on studies using Illumina have increased mainly due to lower cost per sequence than other platforms, enabling high-throughput microbial ecology at the greatest coverage yet possible (Caporaso, *et. al.,*2012). Previous studies from our group (Pereira, *et. al.,* 2016) show that, sequences generated by different technologies are closer one by other, turning the comparative genomic analysis into a more confident task.

The ongoing revolution of the NGS era led to many impacts on the genomic research, one of the biggest impacts was Comparative Genomics, which allowed that sequenced genomes in different benchtop or labs worldwide may be compared on structure and functional features (Metzker, 2010; Edwards & Holt, 2013) According to Touchman, comparative genomics is a field of biological research in which the genome sequences of different species are compared (Touchman, 2010). One of the first comparison by sequence method proposed (Woese, Winker, & Gutell, 1990) was based on a the classification of the small-subunit 16S rRNA gene sequences, since then other technologies have emerged; microarrays are a collection of DNA probes arrayed on a solid support and are used to assay, through hybridization, the presence of complementary DNA (Becquet, *et. al.,* 2002; Willenbrock, *et. al.,* 2007; Gresham, Dunham, & Botstein, 2008). Multilocus sequence typing (MLST), is a technique that examines the genome at multiple ‘housekeeping’ gene loci (Maiden, *et. al.*, 1998), by whole-genome alignment approach and searching for highly conserved sequences across multiple species, it allowed scientists to identified critical functional elements (Bejerano *et al.*, 2004; Fleischmann, 2002). These data lead researchers to obtain a global survey of all genetic differences, as well as information on genome structure with respect to rearrangements (Hu, Xie, Lo, Starkenburg, & Chain, 2011).

As NGS techniques appeared and advanced allowed the researchers a better comparison of whole genome sequences provides a highly detailed view of how organisms are related to each other at the genetic level (Touchman, 2010). Multiple draft genome sequences at once, introduced the pan-genome studies (Tettelin, *et al.*, 2005; Rasko, *et. al.*, 2008; Bentley 2009). For the insights of this work, comparison analysis between the genomes from this study along with 27 sequences of *Streptococcus dysgalactiae* available on the NCBI were performed as showed on the following paper chapter.

Objectives

The aim of this work was to sequence and assembly three bacterial genomes of *Streptococcus dysgalactiae* subsp. *dysgalactiae* strains, SD64, SD42 and SD142, isolated from different outbreaks and states of Brazil. Also, a comparative analysis with these isolates together with the SDD ATCC-27957 strain, isolated from mastitis bovine infection, was performed to compare the genomes at species-level.

CHAPTER II

Paper

Comparative genome analyses of three *Streptococcus dysgalactiae* subsp. *dysgalactiae* strains isolated from fish

Alexandra U Zegarra[1], Felipe L Pereira[1], Alex F Carvalho[1], Fernanda A Dorella[1], Carlos A G Leal[1], Henrique C P Figueiredo[1*]

[1] National Reference Laboratory for Aquatic Animal Diseases (AQUACEN) of Ministry of Agriculture, Livestock and Food Supply, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

Correspondence: Henrique Cesar Pereira Figueiredo, E-mail: figueiredoh@yahoo.com, AQUACEN, Federal University of Minas Gerais, School of Veterinary, Department of Preventive Veterinary Medicine. Address: Av. Antônio Carlos 6627, Pampulha 30161-970, Belo Horizonte, Minas Gerais, Brazil. Phone/Fax number: +55 31 3409 2126.

Authors' e-mails:

AUZ: alexandra.auz@gmail.com
FLP: felipe@flpsw.com.br
AFC: alexficar@gmail.com
FAD: fernandadorella@gmail.com
CAGL: carlos.leal@vet.ufmg.br
HCPF: figueiredoh@yahoo.com

Abstract

Streptococcus dysgalatiae subsp. *dysgalactiae* (SDD) is an important emergent fish pathogen, associated with high mortality rate. The infection is characterized by septicemia and several abscesses in the muscle of the caudal peduncle. In 2007 on Brazil the first outbreak caused by SDD was reported in the state of Ceará. With the aim to understand the genomic structure of SDD and the main traits potentially associated with its virulence and evolution, the genomes of three strains (SD64, SD92 and SD142) isolated from diseased fish, which, belong to different farms and pulse-field gel electrophoresis profiles, were sequenced, assembled and compared. An expected coverage of ~178-fold, ~39% GC content and 133 contigs were obtained on average, genomic comparison between these strains and the isolated from bovine mastitis strain, SDD ATCC 27957, showed similarity of 69%-71%. *In-silico* PCR using characterized primers of genes involved on virulence and antibiotic resistance amplified *sagA, slo* and *tn1207*, virulence genes and *gyrB* and *parC* resistance genes on all the strains, the *emm* and the *isp.1* virulence gene and the *gyrA* and *parE* resistance genes were only found on the bovine strain. Pan-genome analysis showed 1,563 core protein code sequences shared within all the species, only one phage was found and scored as "intact" within the fish isolated strains, further studies of pathogenicity within the strains showed that although the genomes are similar, not all the genomic islands are shared between all the SDD strains. Our findings provide an insight of the differences between SDD strains which may be a basis for a more specific identification and suggest a possible specific host adaptation together with the acquisition of new features by horizontal gene transfer.

Introduction

Streptococcus dysgalactiae is a Gram-positive cocci that auto-aggregates in saline solution, forms long chains in growth medium, it is catalase negative and α/β-hemolytic on blood agar (Vieira, *et. al.,* 1998; Jensen & Kilian, 2012). It is usually found in animals and can be isolated from udders of cows with mild mastitis and from blood and tissues of lambs with polyarthritis (Gaviria & Bisno, 2000). This species was initially considered as non-pathogenic for humans, however, it is now recognized as an increasingly important human pathogen causing several diseases (Hughes, Wilson, Brandt, & Spellerberg, 2009). It has been characterized in veterinary medicine as the cause of bovine mastitis (Whist, Østerås, & Sølverød, 2007) and recently as an important fish pathogen (Netto, Leal, & Figueiredo, 2011)

The first streptococci fish outbreak was reported on cultured rainbow trout in Japan (Hoshina, Sano, & Morimoto, 1958), since then it has spread worldwide, both in wild and cultured fish. There are a few different species of streptococci that are considered as potential fish pathogens: *Lactococcus garvieae, Lactococcus piscium, Streptococcus iniae, Streptococcus agalactiae, Streptococcus parauberis, Vagococcus salmoninarum* (Toranzo, Magariños, & Romalde, 2005), *Streptococcus ictaluri* (Shewmaker *et al.*, 2007) and *Streptococcus phocae* (Romalde *et al.*, 2008).

In the year of 2002, an infection similar to the one caused by *L. garvieae* began to affect both vaccinated and non-vaccinated yellowtail (*Seriola quinqueradiata*) and amberjack (*Seriola dumerili)* on Japan. This was the first outbreak of streptococci caused by *Streptococcus dysgalactiae* subsp. *dysgalactiae* (SDD) reported (Nomoto *et al*., 2004). Since then this pathogen has been isolated from kingfish (*S. lalandi*), yellowtail (*S. quinqueradiata*) and amberjack (*S. dumerili*) in Japan, cobia (*Rachycentron canadum*), basket mullet (*Liza alata*) and gray mullet (*Mugil cephalus*) in Taiwan, golden pomfret (*Trachinotus ovatus*), amur sturgeon (*Acipenser schrenckii*), Siberian sturgeon (*Acipenser baerii*), grass carp (*Ctenopharyngodon idella*), crucian carp (*Carassius carassius*), Soiny mullet (*L. haematocheila*) and pompano (*Trachinotus blochii*) in China, hybrid red tilapia (*Oreochromis* sp.) in Indonesia, white spotted snapper (*Lutjanus stellatus*) and pompano (*T. blochii*) in Malaysia (Abdelsalam, M., Eissa, A., & Chen S. C., 2015a). Meanwhile, in Brazil outbreaks of streptococcoci are common in the freshwater fish Nile tilapia (*Oreochromis niloticus* L.), and on the year 2007, the first disease outbreak caused by SDD was described in the state of Ceará (Netto, Leal, & Figueiredo, 2011).

The infection on fish causes a disease characterized by systemic multifocal inflammatory reaction, microabscess, severe septicemia, and high mortality rates with pathognomonic necrotic ulcers at the caudal peduncle region (Nomoto, 2004, 2006; Netto, Leal, & Figueiredo, 2011; Abdelsalam, Asheg, & Eissa, 2013).

Next Generation Sequencing represents a remarkable tool for the analysis and development of results that will allow to clarify and further differentiate these definitions. However, highly repetitive genomes due to the presence of regions that code for phage sequences, transposons, plasmid, or ribosomal RNA (rRNA) (Bashir A, 2012) still represent a huge challenge in the genome assembly matters (Fricke & Rasko, 2014; Mariano, *et. al.*, 2015). Even though, several strategies are being used to perform the scaffold based assembly process, for example: (i) scaffolding by reference, (ii) scaffolding by mate-pair libraries, or (iii) scaffolding by optical maps (Mariano, *et. al.*, 2016), so far, the problem still persists as a bioinformatics dare.

At the moment, there is no information about whole genome shotgun sequences of SDD isolated from fish, therefore the aim of this study was to generate data that allowed to perform genomic comparisons analysis between SDD strains from different hosts. Thus, the genomic characterization will improve the understanding of this important emerging pathogen.

Material and Methods

Bacterial strains

Strains SD64, SD92 and SD142 were selected from the culture collection of the National Reference Laboratory for Aquatic Animal Diseases (AQUACEN). The strains thar were selected belonged to each one of the three genotypes previously identified among 21 isolates of diseased Nile tilapia during the 2007 and 2011 outbreaks on four different Brazilian farms located in Ceará and Alagoas states (Costa, *et. al.*, 2014).

Identification and early evaluation were performed, in previous studies of our group (Costa, *et. al.*, 2014; Assis, *et. al.*, 2017). The isolates were thawed, streaked onto Todd Hewitt agar (BD) and incubated at 28 ° C for 24 h for DNA extraction.

DNA extraction

The scraping of half a plate of good growth bacterial culture was resuspended in 400 µl TE buffer with 10 mg/ml of lysozyme added. The suspension was incubated for 16 h – 18 h in a 37°C dry bath. After this time 20 ul of a 20 mg/ml proteinase K solution (Qiagen, USA) was added and incubated at 56 °C for 30 min. DNA was extracted using the Maxwell 16 Tissue DNA Purification Kit (Promega), then the solution was transferred to the self-extracting cartridge of Maxwell 16 Research Instrument (Promega, USA), according to the manufacturer's instructions.

Next-generation Sequencing

The sequencing of the three strains was performed using Ion Torrent Personal Genome Machine (PGM). Different libraries were constructed for the strains: A library of 400 bp for the SD64 strain and a library of 200 bp for the SD92 and SD142 strain. The libraries were constructed as follow: 0,1 µg of genomic DNA was used. Sequencing process began with the fragmentation of genomic DNA using the Ion Shear TM Plus Reagents Kit (Life Technologies, USA), barcoding was performed using the Ion Xpress Fragment Library kit and Ion Xpress™ Barcode Adapters (Life Technologies). Size selection, both for 200 bp and 400 bp fragments, was performed with 2% E-Gel® SizeSelect™ Agarose Gels (Invitrogen, USA). Quantification for the library of 400 bp was performed using Ion Library Quantitation Kit (Life Technologies). Later, the libraries were amplified with the OneTouch Template 200 kit and with the OneTouch Template 400 kit (both from Life Technologies), respectively, on the Ion One Touch™ 2 (Life Technologies) and enriched on the Ion OneTouch™ ES (Life Technologies). After annealing the sequencing primer, binding the Ion Torrent PGM Sequencing Polymerase and loading the Ion 318 v2 Chip (Life Technologies) according to the manufacturer's protocols. The enriched libraries were sequenced using correspondingly the Ion Torrent PGM 400 bp and Ion Torrent PGM 200 bp Sequencing Kits (Life Technologies), on the PGM. Finally, the sequencing and signal processing was performed using Torrent Suite 4.2.1. All of the kits were used according to the manufacturer’s recommendations.

Data trimming, Assembly and Gap Filling

The quality of the raw data was analyzed using the Quick Read Quality Control version 1.30.0 package on the Program R (Buffalo V., 2012). Quality trimming, adaptors and barcode removal were performed using an *in-house* script (https://github.com/aquacen/fast_sample). Only reads with a Phred Quality score >= 20 were considered in the assembly (-q 20 parameter of *in-house* script). Assemblies were performed using SPAdes version 3.8.0 (Bankevich, 2012) using "--iontorrent" parameter, and were compared using QUAST version 3.2 (Gurevich, Saveliev, Vyahhi, & Tesler, 2013) using default parameters.

SD64 Gap Filling

DNA from the SD64 was extracted, isolated and sent to OpGen Inc. (Gaithersburg, Maryland, USA) in order to obtain the optical map for the SpeI restriction enzyme. The map composed of 272 fragments was used to map the assembled using MapSolver software version 3.2.0 (OpGen, USA) in sequence placement tool. Parameters were set to Maximum Allowed Places = 4 and Minimum Score for Local = 2. Additionally, the contigs were used to construct scaffolds with the CONTIGuator 2.0 software (Galardini, Biondi, Bazzicalupo, & Mengoni, 2011) with parameters set by default, using as the genome reference the complete sequence of *Streptococcus dysgalactiae* subsp. *equisimilis* (SDE) AC-2713 (GenBank accession number: HE858529). The scaffolds were constructed by the concatenation of overlapped contigs on the Optical Map and CONTIGuator alignment contigs. If gaps existed, they were closed using CLC Genome Workbench 7 (Qiagen) by filling with recursive mapping of reads the contig flanking regions until an overlapping region was found. The generated super contigs were then used as "--trusted-contigs" parameter and an assembly was re-executed with SPAdes, like described above. The new contigs were then mapped into the optical map and the procedure was repeated until the whole-genome coverage (WGC) (e.g., Optical Map alignment with assembled contigs) could no longer be improved.

Bioinformatics analysis

The genomes included in these analyzes were the sequenced strains of SDD of this work (SD64, SD92 and SD142) and the ATCC 27957 strain available as a draft genome on the GenBank database of National Center for Biotechnology Information (NCBI) (Accession number: NZ_CM001076.1, isolated from bovine*),* hereafter called "SDD group"; and a second group, hereafter called "SD available group", composed by the SDE group along with *S. dysgalactiae* strains with sequenced genomes available at GenBank (Accession numbers in Supplementary Table 1).

The SDD group was characterized using the MLST schema available for the *Streptococcus dysgalactiae* subsp. *equisimilis* on the PUBmlst webserver (Jolley & Maiden, 2010), which uses the DNA sequence of seven housekeeping genes *(gki, gtr, murI, mutS, recP, xpt* and *atoB*). The sequence of each of the seven housekeeping genes was extracted using the sequences of the PUBmlst database as template and the BLAST webserver (http://www.ncbi.nlm.nih.gov/blast) for the alleles search. To establish the links between all the Sequence Types (STs) the software geoBURST version 1.2.1 (Francisco, Bugalho, M., & Carriço, 2009) was used.

A Multilocus Sequence Analysis (MLSA) was also performed as described previously (Jensen & Kilian, 2012), the sequences of seven housekeeping (*map, pfl, ppaC, pyk, rpoB, sodA* and *tuf*) genes were concatenated and compared. Comparisons for this analysis were made within the SDD group, the SD available group and the sequences depositated by Jensen and collaborators (Jensen & Kilian, 2012). The extraction and trimmingof gene sequences were performed like described above for MLST analysis. The sequence were subjected to phylogenetic analysis using the minimun evolution algorithm with the Kimura two-parameter substitution model and a bootstrap of 1000 repetitions.

In order to identify the conserved genomic regions within the SDD group, as well as the possible rearrangement between them, the software Mauve version 2.3.1 (Darling A. C., 2004) was used. Before the alignment, the contigs were reordered, using the ATCC-27957 strain as reference. Then the alignment was performed using the progressiveMauve (Darling A. E., 2010) method. Both steps were run with all the parameters by default. Furthermore, comparison of the genomes was also performed using Gegenees version 2.2.1 (Agren, Sundström, Håfström, & Segerman, 2012). Parameters were set as "Accurate" (i.e., Fragment size = 200 bp and Step size = 100 bp) and a threshold in heatmap analysis was set equals to 0%. The result was then exported as a nexus file and a phylogenomic tree was created using SplitsTree (Huson, 1998) v. 4.14.2. This analysis included the SD available group for a comprehensive comparative analysis of the entire species.

Virulence and resistance genes where searched using a set of F/R primers (*n* = 93) described in previous works (Rato, M., Nerlich, A., & Bergmann, R., 2011; Pinho, *et. al.,* 2010; Yan, *et. al.,* 2000; Maeda, *et. al.,* 2011; Ding, *et. al.,* 2016; Abdelsalam, *et. al.,* 2015b; Nishiki, 2011), and available at Supplementary File 1, by *in silico* PCR using the software FastPCR version 6.6.01 (Kalendar, Lee, & Schulman, 2009), the search was performed for the SDD group. "Circular sequence", "Restrict analysis to F/R primers pairs" and "Probe search" options were set as true, leaving the other parameters set as default. Further comparisons were made in order to identify virulence gence following the work of Suzuki and collaborators (Suzuki, *et al.*, 2011) and comparing the sequence of 129 *Streptococcus* virulence genes retrived from the Core Dataset (Genes associated with experimentally verified VFs only) of the Virulence Factor Database (VFDB) (Chen, 2005); comparison were only considered when the Blast best hit had a e-value <= 1e-10 (Pearson, 2013). Addittionally, a broader comparison including not only the *Streptococcus* genes but the full dataset of genes related to known and predicted VFs on the whole database of the VFDB was performed, for this analysis only Blast best hits with a percentage query coverage and an identity percentage larger than 90% were considered.

Moreover, orthoMCL software version 0.9 (Li, Stoeckert, & Roos, 2003), with all the parameters set by default, was used in order to identify orthologous groups within the SDD group. GIPSy software (Soares, *et al.*, 2016) was used to predict genomic islands with default parameters; the strain set as the non-pathogenic subject in all of the analysis was the

Streptococcus thermophilus CNRZ1066 (Accession number NC_006449). The results were plotted using BRIG software (Alikhan, Petty, Zakour, & Beatson, 2011), version 0.95 with default parameters. Predictions of the genomic islands were performed on SDD ATCC-27957 strain, thus the interpretation of the results should be done by evaluating the percentage of similarity between these parts of the genome. Prophinder (Lima-Mendez, Van Helden, Toussaint, & Leplae, 2008) and Phaster (Arndt D. , Grant, Sajed, Liang, & Wishart, 2016) were used in order to find phages within the strains. Both of them were used with all the parameters set by default.

Results

De novo Assembly and SD64 Gap Filling

The Ion Torrent libraries resulted in a total of reads of 1384323, 1602890, and 1427894, with the mean fragment size of 264 bp, 253 bp, and 246 bp, with an expected coverage of ~174-, ~193-, and 167-fold, and with a GC content of ~39 %, for SD64, SD92 and SD142, respectively. Figure 1 shows the quality score per base position (Figure 1A), the GC content variation (Figure 1B) and distribution of the read's length (Figure 1C). Using an *in-house* script, 5.5 %, 12 %, and 2 % of reads, for SD64, SD92, SD142, respectively, were discarded due to poor quality and short length of the reads. In summary, the assemblies results in 131, 138, and 132 contigs, with N50 of 32945 bp, 32312 bp, 32948 bp, and largest contig of 120861 bp, 120572 bp, 120573 bp, for SD64, SD92, and SD142, respectively. Quast software report is showed on Table 1.

In order to perform the SD64 gap filling, the assembled contigs were aligned with the optical map (Figure 2A), with an initial WGC of 51.95%, and generated scaffolds with *S. dysgalactiae* subsp. *equisimilis* ATCC-2713 (Figure 2B), resulting in 87 contigs oriented. Five initial scaffolds were constructed using alignment on optical map and scaffolds information, which lead to a WGC increase of 60.89% (Figure 3A). In subsequent assemblies (see SD64 Gap Filling on Material and Methods section) contigs were aligned to the optical map, increasing the WGC first to 70.51%, and finally to 83.93% (Figure 3B, Figure 3C).

MLST and MLSA analyses

Figure 4A shows all the STs from SDE available on PubMLST together with the STs from the SDD group (pointed with red arrows). Two new STs were formed, one for the strains isolated from fish and another for the strain with bovine origin.

Results also showed that the ST-246 profile, available at PubMLST database, has only a Single Locus Variation (SLV) with the fish ST profile, which represents a close relation pattern (Figure 4B). Metadata of the ST-246 reports it like a SDE isolated from fish. On the other hand, the bovine strain was found as part of a clonal complex (CC308). (Figure 4C). Figure 5 shows the result tree of the MLSA analysis. SDD group formed a specific-clade, in 100 % of bootstrap repetitions, that is in accordance with previously showed by Jensen and collaborator (Jensen & Kilian, 2012).

Genome similarity

The progressiveMauve algorithm showed (Figure 6) a high number of locally collinear blocks. This feature represents that even considering draft genomes, a high number of rearrangements (i.e., deletions, duplications, inversions, and translocations) of genetic material crossing over the strains chromosome. Also, a high similarity all over the genomes of this work is showed on the Gegenees heatmap plot, varying from 96.4% to 98.92% (Figure 7). This score lows to 69.51 % when the strains within the SDD group were compared. The percentage of identity between the SDE genomes was between 79.46% and 99.65%. Additionally this result is showed as a phylogenomic tree on Figure 8.

Virulence and Antibiotic Resistance analyses

Regarding the virulence genes on the *in-silico* PCR results of the SDD SD64, SD92 and SD142 strains, the following genes were amplified: *sagA*, *slo, tn1207/f10394.4 lj, NAPlr, eno,* and *sof-FD*. Whereas, on the SDD ATCC-27957 strain, the amplified genes were *sagA*, *slo, tn1207/f10394.4 lj, NAPlr, eno, isp.1* and *emm* genes. In this case, *slo* gene amplified with an unexpected size (Table 2).

Furthermore, the results of the *in-silico* PCR for the genes related to antibiotic resistance (*gyrB* and *parC)* and the composite transposon *tn1207.3/f10394.4 lj* were amplified (Table 3). On the SDD ATCC-27957 strain, besides these genes, the *gyrA* and *parE* genes also associated to resistance were amplified. Both, the *parE* and *tn1207.3/f10394.4 lj* genes amplified with an unexpected size.

Finally, the Table 4 shows the values of the identity percentage obtained within the SDD group strains and VFDB Core Dataset for *Streptococcus*. SDD SD64, SD92 and SD142 strains matched with *hasC*, *fbp54*, *gbs0630*, *gbs0631*, and *gbs0632* genes. The SDD ATCC-27957 strain matched with *mf/spd*, *hasC*, *fbp54*, *sda*, *gbs0630*, *gbs0631*, and *gbs0632* genes. Else, for the comparison using the VFDB Full Database 94, 87, 77 and 56 hits where found for the SDD SD64, SD92, SD142 and ATCC-27957 strains respectively.

Phage analysis

Prophinder predicted no prophages for the strains of this study. On the other hand , Phaster found and scored one phage as “intact” and five other phages as “incomplete” both for the SD64 and the SD92 strain, whereas for the SD142 one “questionable” and six “incomplete” phages were found. Finally for ATCC-27957 strain, five phages were found, of these two of them were scored as “intact” and the others scored as “incomplete” (Table 5). The list of the products found on each phage predicted, together with the Phaster results, are available on the Supplementary Table 2.

Island analysis

Gipsy predicted 34 genomic islands (GEIs) for the SDD ATCC-27957 strain, of them 11 were unclassified genomic islands (Figure 9), 13 were pathogenicity islands (PAIs) and 10 were resistance islands (RIs). The PAI2, PAI7 and PAI13 had its prediction score described as “Strong”, PAI1 and PAI12 scored “Weak” and the other were classified as “Normal” PAIs. As for RIs, the RI3, RI9 and RI10 were catalogued as “Strong”, RI8 as “Weak” and the other of the RIs as “Normal”. Finally, the 11 GEIs remained unscored due to its low concentration of specific factors. Apparently, all of the predicted islands are at least partially preserved, showing PAI10, PAI12 and PAI13, R10, GEI7 and GEI11 as the less preserved within the strains, all the rest of GEIs are shared by all the sequenced genomes from the SDD group. Corresponding genes and some of products that may be interesting in matters of virulence factors for each island are shown on Supplementary Table 3.

Pan-genome from fish and bovine SDD isolates

OrthoMCL analysis gave as result of 1,563 protein coding sequence (CDS) shared within the entire SDD group (Figure 10): the SDD core genome. There are 117, 74 and 68 exclusive CDS to the strains, isolated from infected fish, SD64, SD92 and SD142, respectively. These strains also share 515 CDS only within them (the accessory genome of SDD isolated from fish) and 39 CDS more are shared between at least one strain of this study and the SDD ATCC-27957 strain. Meanwhile, there are 384 exclusive CDS to the SDD strain ATCC-27957 isolated from bovine mastitis. Exclusive CDS and their products are listed on the Supplementary Table 4.

Discussion

Streptococcus dysgalactiae subsp. *dysgalactie* is an important emerging pathogen, usually characterized in veterinary medicine as the cause of bovine mastitis (Wyder, *et. al.*, 2011), however, at the moment, the increasing number of fish infection reports suggest its critical expansion as a pathogen of importance (Abdelsalam, M., Asheg, A., & Eissa, A. E., 2013). In fish, the infection is characterized by septicemia, severe necrotic ulcers on the caudal peduncle with a high mortality rate (Nomoto, *et. al.*, 2004; 2006; Netto, L. N., Leal, C. A., & Figueiredo, H. C., 2011). Faced with this problem, NGS technologies offer solutions on genomic studies that not only allow to characterize the nature and biological aspects of the organism, but also help the understanding of its pathology and treatment. However a complete genome sequence of SDD is still expected.

Throughout the history of *S. dysgalactiae* the difficulty of good typing among subspecies has been common (Garvie & Collins, 1983; Farrow, J. A., & Collins, M. D., 1984), there are even studies that indicate that strains of *S. dysgalactiae* subsp. *equisimilis* (SDE) of animal origin are genetically diverse from the ones of human origin and future reclassifications are suggested (Jensen, A., & Kilian, M., 2012; Pinho, *et. al.*, 2016). It is interesting to note that although the MLST analysis does not allow a proper separation between SD subspecies, the relationship between STs respect to the host type might suggest that, like pointed out in the work of Pinho and collaborators (2016) within SD horse strains, it is likely that the fish isolates may represent a recent strain adapted to fish hosts. There are three isolates from fish deposited for the ST-246, with the following IDs: 1242, 1243 and 1314, all of them with Asiatic origin (Japan and Singapore), this may suggest that, both, Asian and Brazilian SDD are related. MLSA analysis could also reaffirm this notion due to the separation within the SDD group with a 100% of bootstrap repetitions. Also, both genome similarity and phylogenomic analysis showed the

SDD and the SDE strains separated. A segregation between the SDD group reaffirms the previous results that suggest a host adaptation on this subspecies. This may be explained with the study synteny analysis that show rearrangements that may allow the SDD group strains to have different traits within them. Previous studies of these groups of streptococci indicate they may undergo into significant genome rearrangement due to horizontal transfer, and other recombination related such as insertion or deletions (Towers, *et. al.,* 2004; Sachse, *et. al.,* 2002; Richards, *et. al.,*2011).

In-silico PCR showed virulence genes *sagA* and *slo* that encode for Streptolysin S and Streptolysin O, respectively, are present in SDD strains. Both streptococcal hemolytic exotoxins that are suspected as a zoonotic character of rheumatic fever (Kłos & Wójkowska-Mach, 2017). Also, the *NAPlr* and *eno* genes, that also were found in the analysis, they have been already described as important agents during fish infection causing adhesion to host epithelial cells and the presence of wall-associated plasminogen binding proteins (Abdelsalam, Fujino, Eissa, Chen, & Warda, 2015b). The *sof-FD* gene responsible of the serum opacification, activity previously described as an important virulence factor on fish infections (Nishiki, 2011), also was found. By the other side, the *emm* gene, an important virulence factor gene, even used to *S. dysgalactiae* pre-genomics typing, was only found in the bovine strain, as previous studies confirmed (Suzuki, *et. al.,* 2011; Abdelsalam, M., Eissa, A., & Chen, S.-C., 2015a). Furthermore, antibiotic microbial resistance genes *gyrB* and *parC* were amplified on all of the strains, DNA gyrase subunit B and DNA topoisomerase IV subunit A, respectively, have been associated as quinolone resistance regions (Maeda, *et. al.,* 2011). And, as for *tn1207.3/f10394.4 lj,* which was also found, it has been established as a mobile element containing genetic sections associated with the resistance to erythromycin (D'ercole, 2005)

Moreover, the results within the comparison against the VFDB (Table 4), the SDD strains had hit with the following virulence factors: *hasC*, *fbp54* and *sda*. The *hasC* gene is related on the production of a hyaluronic capsule as a mechanism to avoid phagocytosis on bacteria (Schrager H. M., *et. al.,* 1998; Bisno, A. L., Brito, M. O., & Collins, C. M., 2003); the *fbp54* gene is related to the fibronectin binding proteins which are known to participate and mediate cellular invasion (Kreikemeyer, B., Talay, S., & Chhatwal, G., 1995; Rocha, C., & Fischetti, A., 1999). While the *sda*, described as a dnase, which due to its digestive activity, may contribute to the bacteria mobility within the host (Podbielski A, *et. al.,* 1996; Bisno, A. L., Brito, M. O., & Collins, C. M., 2003). The *mk/spd* gene was only found SDD ATCC-27957 strain and corresponds to a dnase with the previously described function. On the other hand the virulence factors *gbs0630, gbs0631* and *gbs0632* were found as exclusive for the fish isolates strains but they were characterized as virulence factors only by association *gb0630* and *gb0631* are putative class C sortases and *gbs0632* is a putative tip adhesin protein with an unknow function (Glaser, , *et. al.,* 2002) .

In-silico PCR results also showed that based on the information from previous studies (Rato, *et. al.,* 2010; 2011) on Group A, Group C and Group G Streptococci (GAS/GCS/GGS) none bacteriophage-associated virulence genes (*speC, speJ, speI, speH, ssa, mf4, slaA, speA3 speK, speL, speM, spd1* and *sdn)* was found on our strains. However, the SDD ATCC-27957 strain *in-silico* PCR amplified the *speM* and *slaA* genes.

Else, contrary of previous studies (Suzuki, *et. al.*, 2011) no homology prophage was found within SDD fish isolated strains and the “M3 GAS phage 315.3”. The intact prophages found on SD64 strain, corresponds to the phage “Streptococcus phage phiNJ2”, reported on a strain of *Streptococcus suis* (Tang, *et. al.*, 2013). While, the phage “Streptococcus phage A25” in SD92 strain that was predicted as intact, have been reported on a strain of *Streptococcus pyogenes* (Accession number: NC_028697.1). Both pathogens: *S. suis* and *S. pyogenes* had not been yet reported on fish infectios at the moment, however this prophages may confer additional features for the adaptation of the SDD strains.

In other way of horizontal-gene transfer, 13 PAIs and 10 IRs were predicted. Most of these islands are composed of genes without a proven virulence factor for this subspecies, however although *S. agalactiae* is consider an usual pathogen on fish (Mian G. F., *et al.*, 2009) and previous studies have been carried out demonstrating mechanisms that help environmental adaptation and acquisition of potential virulence factors, between this two species (Richards, *et. al.*, 2011), our studies did not find any horizontal genetic transfer between them. What can be conjectured is that the presence of genomic islands predicted in fish and bovine strains may be due to horizontal gene transfer, which shows that even though there are certain differences between strains there is a continuous flow of genetic information between them.

In conclusion, the present work showed the first comparative genomic analyzes within the SDD from different hosts, identifying the virulence factors, due to its origins, presumably as a result of horizontal transfer. Delimitations between subspecies of SDD were found, however a study including a more comprehensive collection of isolates, both from fish and from mammals, may draw a better delimitation of the host-pathogen interaction.

Conflict of interests

The authors declare no conflict of interest.

Author contribution

AUZ, FLP and HCPF wrote the manuscript. AUZ performed bioinformatics analyses. FAD and AFC: performed the experiments at bench. HCPF conceived and designed the experiments and coordinated all analyses of the project. All authors read and approved the final manuscript.

Acknowledgements

We thank the Coordination for the Improvement of Higher Education Personnel (CAPES), the National Council of Technological and Scientific Development (CNPq) and Ministry of Agriculture, Livestock and Food Supply (MAPA) for financing this study.

References

Abdelsalam, M., Asheg, A., & Eissa, A. E. (2013). Streptococcus dysgalactiae: an emerging pathogen of fishes and mammals. *International Journal of Veterinary Science and Medicine*, 1(1), 1-6.

Abdelsalam, M., C. S., & & Yoshida, T. (2009). Surface properties of Streptococcus dysgalactiae strains isolated from marine fish. *Bull. Eur. Assoc. Fish Pathol.*, 29, 15-23.

Abdelsalam, M., Chen, S. C., & Yoshida, T. (2010). Dissemination of streptococcal pyrogenic exotoxin G (spegg) with an IS-like element in fish isolates of Streptococcus dysgalactiae. *FEMS microbiology letters*, 309(1), 105-113.

Abdelsalam, M., Eissa, A., & Chen, S.C. (2015a). Genetic diversity of geographically distinct Streptococcus dysgalactiae isolates from fish. *Journal of Advanced Research, 6*(2), 233-238.

Abdelsalam, M., Fujino, M., Eissa, A. E., Chen, S. C., & Warda, M. (2015b). Expression, genetic localization and phylogenic analysis of NAPlr in piscine Streptococcus dysgalactiae subspecies dysgalactiae isolates and their patterns of adherence. *Journal of advanced research*, 6(5), 747-755.

Abdelsalam, M., Nakanishi, K., Yonemura, K., Itami, T., Chen, S. C., & Yoshida, T. (2009). Application of Congo red agar for detection of Streptococcus dysgalactiae isolated from diseased fish. *Journal of Applied Ichthyology*, 25(4), 442-446.

Agren, J., Sundström, A., Håfström, T., & Segerman, B. (2012). Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. *PloS one,* 7(6), e39107.

Alikhan, N. (2011). BRIG 0.95 Manual. 1-53.

Alikhan, N. F., Petty, N. K., Zakour, N. L., & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC genomics*, 12(1), 402.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. 175-176.

Arndt, D., Grant, J., Marcu, A., Sajed, T., Pon, A., Liang, Y., & Wishart, D. (8 de 7 de 2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic acids research, 44*(W1), W16-21.

Assis, G. B., Pereira, F. L., Zegarra, A. U., Tavares, G. C., Leal, C. A., & Figueiredo, H. C. (2017). Use of MALDI-TOF Mass Spectrometry for the Fast Identification of Gram-Positive Fish Pathogens. *Frontiers in Microbiology*, 8, 1492.

Bankevich, A. N. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5), 455-477.

Bashir A, K. A.-S. (2012). A hybrid approach for the automated finishing of bacterial genomes. *Nat Biotechnol*, 30:701–7.

Bisno, A. L., Brito, M. O., & Collins, C. M. (2003). Molecular basis of group A streptococcal virulence. *The Lancet infectious diseases, 3*(4), 191-200.

Boisvert, S. L. (2010). Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of computational biology*, 17(11), 1519-1533.

Bolger, A. M. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, btu170.

Buffalo V., &. C. (2012). *qrqc: Quick Read Quality Control. R package version 1.30.0,* . Fonte: Bioconductor: http://github.com/vsbuffalo/qrqc

Cameron, M., Saab, M., Heider, L., McClure, J., Rodriguez-Lecompte, J., & Sanchez, J. (2016). Antimicrobial Susceptibility Patterns of Environmental Streptococci Recovered from Bovine Milk Samples in the Maritime Provinces of Canada. *Frontiers in veterinary science, 3*, 79.

Chen, L. Y. (2005). VFDB: a reference database for bacterial virulence factors. *Nucleic acids research*, 33(suppl_1), D325-D328.

Christin, C., Hoefsloot, H., Smilde, A., Hoekman, B., Suits, F., Bischoff, R., & Horvatovich, P. (1 de 1 de 2013). A Critical Assessment of Feature Selection Methods for Biomarker Discovery in Clinical Proteomics. *Molecular & Cellular Proteomics, 12*(1), 263-276.

Costa, F. A. (2014). Genotyping of Streptococcus dysgalactiae strains isolated from Nile tilapia, Oreochromis niloticus (L.). *Journal of Fish Diseases*, 37(5), 463–469.

Darling, A. C. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, 14(7), 1394-1403.

Darling, A. E. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS one*, 5(6), e11147.

David Metzgar1, *. (2017). The M protein of group A Streptococcus is a key virulence factor and a clinically relevant strain identification marker.

Davies, M., McMillan, D., Beiko, R., Barroso, V., Geffers, R., Sriprakash, K., & Chhatwal, G. (1 de 6 de 2007). Virulence Profiling of Streptococcus dysgalactiae Subspecies equisimilis Isolated from Infected Humans Reveals 2 Distinct Genetic Lineages That Do Not Segregate with Their Phenotypes or Propensity to Cause Diseases. *Clinical Infectious Diseases*, *44*(11), 1442-1454.

D'ercole, S. P. (2005). Distribution of mef (A)-containing genetic elements in erythromycin-resistant isolates of Streptococcus pyogenes from Italy. *Clinical microbiology and infection*, 11(11), 927-930.

Ding, Y., Zhao, J., He, X., Li, M., Guan, H., Zhang, Z., & Li, P. (2016). Antimicrobial resistance and virulence-related genes of Streptococcus obtained from dairy cows with mastitis in Inner Mongolia, China. *Pharmaceutical Biology, 54*(1), 162-167.

Ercole, S., Petrelli, D., Prenna, M., Zampaloni, C., Catania, M., Ripa, S., & Vitali, L. (2005). Distribution of mef(A)-containing genetic elements in erythromycin-resistant isolates of Streptococcus pyogenes from Italy. *Clinical Microbiology and Infection, 11*, 927-930.

Farrow, J. A., & Collins, M. D. (1984). Taxonomic studies on streptococci of serological groups C, G and L and possibly related taxa. *Syst. App. Microbiol.*, 483-493.

Feil, E. J., C., L. B., Aanensen, D. M., Hanage, W. P., & Spratt, B. G. (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *Journal of bacteriology*, 186(5), 1518-15.

Francisco, A. P., Bugalho, M., R. M., & Carriço, J. A. (2009). Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC bioinformatics*, 10(1), 152.

Fricke, W. F., & Rasko, D. a. (2014). Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nature Reviews. Genetics*.

Galardini, M., Biondi, E. G., Bazzicalupo, M., & Mengoni, A. (2011). CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source code for biology and medicine*, 6(1), 11.

Garvie, E., & Collins, M. (1983). Streptococcus dysgalactiae (Diernhofer) nom. Rev. *Int J Syst Bacteriol.*, 33: 404–405.

Gaviria, J. M., & Bisno, A. L. (2000). Group C and G streptococci. Streptococcal infections: clinical aspects, microbiology and molecular pathogenesis. *Oxford University Press*, 238-254.

Glaser, P., Rusniok, C., Buchrieser, C., Chevalier, F., Frangeul, L., Msadek, T., & Trieu-Cuot, P. (2002). Genome sequence of Streptococcus agalactiae, a pathogen causing invasive neonatal disease. *Molecular microbiology*, 45(6), 1499-1513.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), 1072-1075.

Hoshina, T., Sano, T., & Morimoto, Y. (1958). A Streptococcus pathogenic to fish. *J. Tokyo Univ. Fish*, 44(5).

Hughes, J. M., Wilson, M. E., Brandt, C. M., & Spellerberg, B. (2009). Human infections due to Streptococcus dysgalactiae subspecies equisimilis. *Clinical Infectious Diseases*, 49(5), 766-772.

Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics (Oxford, England),*, 14(1), 68-73.

Jensen, A., & Kilian, M. (2012). Delineation of Streptococcus dysgalactiae, its subspecies, and its clinical and phylogenetic relationship to Streptococcus pyogenes. *Journal of Clinical Microbiology*, 50(1), 113–126.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic acids research*, 36(suppl 2), W5-W9.

Jolley, K. A., & Maiden, M. C. (2010). BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC bioinformatics*, 11(1), 595.

Kalendar, R., Lee, D., & Schulman, A. H. (2009). FastPCR software for PCR primer and probe design and repeat search. *Genes, Genomes and Genomics*, 3(1), 1-14.

Kawamura, Y., Fujiwara, H., Mishima, N., Tanaka, Y., Tanimoto, A., Ikawa, S., . . . Ezaki, T. (2003). First Streptococcus agalactiae Isolates Highly Resistant to Quinolones, with Point Mutations in gyrA and parC. *ANTIMICROBIAL AGENTS AND CHEMOTHERAPY, 47*(11), 3605-3609.

Kłos, M., & Wójkowska-Mach, J. (2017). Pathogenicity of Virulent Species of Group C Streptococci in Human. *Canadian Journal of Infectious Diseases and Medical Microbiology*.

Kreikemeyer, B., Talay, S., & Chhatwal, G. (1995). Characterization of a novel fibronectin-binding surface protein in group A streptococci. *Mol. Microbiol.*, 17(1):137-145.

Lehri, B., Seddon, A. M., & Karlyshev, A. V. (2017). The hidden perils of read mapping as a quality assessment tool in genome sequencing. *Scientific Reports*, 7.

Li, L., Stoeckert, C., & Roos, D. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research, 13*(9), 2178-89.

Lima-Mendez, G., Van Helden, J., Toussaint, A., & Leplae, R. (2008). Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics, 24*(6), 863-865.

Maeda, Y. M., Goldsmith, C. E., Coulter, W. A., M. C., Millar, B. C., & Elborn, J. S. (2010). Molecular characterization and phylogenetic analysis of quinolone resistance-determining regions (QRDRs) of gyrA, gyrB, parC and parE gene loci in viridans group streptococci isolated from adult patients with cystic fibrosis. *Journal of antimicrobial chemotherapy*, 66(3), 476-486.

Mariano, D. C., Pereira, F. L., Ghosh, P., Barh, D., Figueiredo, H. C., Silva, A., & Azevedo, V. A. (2015). MapRepeat: an approach for effective assembly of repetitive regions in prokaryotic genomes. *Bioinformation*, 11(6), 276–9.

Mariano, D. C., Sousa, T. d., Pereira, F. L., Aburjaile, F., Barh, D., . . . Azevedo, V. A. (2016). Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of Corynebacterium pseudotuberculosis strain 1002. *BMC Genomics*, (17): 315.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1), pp-10.

Maye, S., Flynn, J., Stanton, C., Fitzgerald, G., & Kelly, P. (19 de 6 de 2017). Bovine intra-mammary challenge with Streptococcus dysgalactiae spp. Dysgalactiae to explore the effect on the response of Complement activity. *Journal of Dairy Research*, 1-7.

Metzgar, D., & Zampolli, A. (2011). The M protein of group A Streptococcus is a key virulence factor and a clinically relevant strain identification marker. *Virulence*, 2(5), 402-412.

Mian, G. F., Godoy, D. T., Leal, C. A. G., Yuhara, T. Y., Costa, G. M., & Figueiredo, H. C. P. (2009). Aspects of the natural history and virulence of S. agalactiae infection in Nile tilapia. *Veterinary microbiology*, 136(1), 180-183

Netto, L. N., Leal, C. A., & Figueiredo, H. C. (2011). Streptococcus dysgalactiae as an agent of septicaemia in Nile tilapia, Oreochromis niloticus (L.). *Journal of fish diseases*, 34(3), 251-254.

Nishiki, I. H. (2011). Cloning and expression of serum opacity factor in fish pathogenic Streptococcus dysgalactiae and its application to discriminate between fish and mammalian isolates. *FEMS microbiology letters*, 323(1), 68-74.

Nomoto, R., Kagawa, H., & Yoshida, T. (2008). Partial sequencing of sodA gene and its application to identification of Streptococcus dysgalactiae subsp. dysgalactiae isolated from farmed fish. *Letters in applied*.

Nomoto, R., Munasinghe, L., & Jin, D. (2004). Lancefield group C Streptococcus dysgalactiae infection responsible for fish mortalities in Japan. *Journal of fish*.

Nomoto, R., Unose, N., Shimahara, Y., Nakamura, A., Hirae, T., Maebuchi, K., . . . Yoshida, T. (2006). Characterization of Lancefield group C Streptococcus dysgalactiae isolated from farmed fish. *Journal of Fish Diseases, 29*(11), 673-682.

Onmus-Leone, F., Hang, J., Clifford, R. J., Yang, Y., Riley, M. C., Kuschner, R. A., & Lesho, E. P. (2013). Enhanced De Novo Assembly of High Throughput Pyrosequencing Data Using Whole Genome Mapping. *PLoS ONE*, 8(4), 2–10.

Pearson, W. R. (2013). An Introduction to Sequence Similarity (“Homology”) Searching. *Current Protocols in Bioinformatics / Editoral Board*.

Pinho, M. D., Erol, E., Ribeiro-Gonçalves, B., Mendes, C. I., Carriço, J. A., Matos, S. C., & Ramirez, M. (2016). Beta-hemolytic Streptococcus dysgalactiae strains isolated from horses are a genetically distinct population within the Streptococcus dysdysgalactiae taxon. *Scientific reports,* 6.

Pinho, M. D., Melo-Cristino, J., Ramirez, M., & Infections, P. G. (2010). Fluoroquinolone resistance in Streptococcus dysgalactiae subsp. equisimilis and evidence for a shared global gene pool with Streptococcus pyogenes. *Antimicrobial agents and chemotherapy,* 54(5), 1769-1777.

Podbielski, A., Zarges, I., Flosdorff, A., & Weber-Heynemann, J. (1996). Molecular characterization of a major serotype M49 group A streptococcal DNase gene (sdaD). *Infection and immunity, 64*(12), 5349-5356..

Poyart, C., Quesne, G., Boumaila, C., Trieu-Cuot, P., & Liu, J. (1 de 12 de 2001). Rapid and Accurate Species-Level Identification of Coagulase-Negative Staphylococci by Using the sodA Gene as a Target. *Journal of Clinical Microbiology, 39*(12), 4296-4301.

Putman, M., Van Veen, H., Degener, J., & Konings, W. (2017). The lactococcal secondary multidrug transporter LmrP confers resistance to lincosamides, macrolides, streptogramins and tetracyclines. *Microbiology, 0119*(147), 49-2873.

Rato, M., Bexiga, R., Nunes, S., Vilela, C., & Santos-Sanches, I. (2010). Human group A streptococci virulence genes in bovine group C streptococci. *Emerging infectious diseases, 16*(1), 116-9.

Rato, M., Nerlich, A., & Bergmann, R. (2011). Virulence gene pool detected in bovine group C Streptococcus dysgalactiae subsp. dysgalactiae isolates by use of a group A S. pyogenes virulence microarray. *Journal of Clinical Microbiology*, 49(7), 2470-2479.

Richards, V. P., Lang, P., Bitar, P. D., Lefébure, T., Schukken, Y. H., Zadoks, . . . Stanhope, M. J. (2011). Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted Streptococcus agalactiae. *Infection, Genetics and Evolution*, 11(6), 1263-1275.

Rocha, C., & Fischetti, A. (1999). Identification and characterization of a novel fibronectin-binding protein on the surface of group A streptococci. *Infect. Immun.*, 67(6):2720-2728.

Rognes, T., T., F., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584.

Romalde, J. L., Ravelo, C., Valdés, I., Magariños, B., De La Fuente, E., San Martín, C., ... & Toranzo, A. E. (2008). Streptococcus phocae, an emerging pathogen for salmonid culture. *Veterinary microbiology*, *130*(1), 198-207.

Sachse, S., Seidel, P., Gerlach, D., Günther, E., Rödel, J., Straube, E., & Schmidt, K. H. (2002). Superantigen-like gene (s) in human pathogenic Streptococcus dysgalactiae, subsp. equisimilis: genomic localisation of the gene encoding streptococcal pyrogrogenic exotoxin G (speGdys). *FEMS Immunology & Medical Microbiology*, 34(2), 159-167.

Schrager, H. M., Albertí, S., Cywes, C., Dougherty, G. J., & Wessels, M. R. (1998). Hyaluronic acid capsule modulates M protein-mediated adherence and acts as a ligand for attachment of group A Streptococcus to CD44 on human keratinocytes. *Journal of Clinical Investigation*, *101*(8), 1708.

Shewmaker, P. L., Camus, A. C., Bailiff, T., Steigerwalt, A. G., Morey, R. E., & Maria da Glória, S. C. (2007). Streptococcus ictaluri sp. nov., isolated from channel catfish Ictalurus punctatus broodstock. *International journal of systematic and evolutionary microbiology*, *57*(7), 1603-1606.

Soares, S. C., Geyik, H., Ramos, R. T., G., P. H., Barbosa, E. G., Baumbach, J., & Azevedo, V. (2016). GIPSy: Genomic island prediction software. *Journal of Biotechnology,*, 232, 2–11.

Sutcliffe, J., Tait-Kamradt, A., & Wondrack, L. (1996). Streptococcus pneumoniae and Streptococcus pyogenes resistant to macrolides but sensitive to clindamycin: a common resistance pattern mediated by an efflux system. *Antimicrobial Agents and Chemotherapy*, 40(8), 1817-1824.

Suzuki, H., Lefébure, T., Hubisz, M. J., Pavinski Bitar, P., Lang, P., Siepel, A., & Stanhope, M. J. (2011). Comparative genomic analysis of the Streptococcus dysgalactiae species group: gene content, molecular adaptation, and promoter evolution. *Genome biology and evolution*, 168-185.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & & Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution*, 30(12), 2725-2729.

Tang, F., Bossers, A., H., Lu, C., & Smith, H. (2013). Complete Genome Sequence of the Streptococcus suis Temperate Bacteriophage varphiNJ2. *Genome Announc*.

Toranzo, A., Magariños, B., & Romalde, J. (5 de 2005). A review of the main bacterial fish diseases in mariculture systems. *Aquaculture, 246*(1-4), 37-61.

Towers, R. J., Gal, D., McMillan, D., Sriprakash, K. S., Currie, B. J., Walker, M. J., & ... & Fagan, P. K. (2004). Fibronectin-binding protein gene recombination and horizontal transfer between group A and G streptococci. *Journal of clinical microbiology*.

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics*, 13(1), 36.

Vélez, J., Cameron, M., Rodríguez-Lecompte, J., Xia, F., Heider, L., Saab, M., . . . Sánchez, J. (2017). Whole-Genome Sequence Analysis of Antimicrobial Resistance Genes in Streptococcus uberis and Streptococcus dysgalactiae Isolates from Canadian Dairy Herds. *Frontiers in veterinary science, 4*, 63.

Vieira, v., Teixeira, L., Zahner, V., Momen, H., Facklam, R., Steigerwalt, A., . . . Castro, A.. Genetic relationships among the different phenotypes of Streptococcus dysgalactiae strains. *International Journal of Systematic Bacteriology, 48*(4), 1231-1243.

Whist, A. C., Østerås, O., & Sølverød, L. (2007). Streptococcus dysgalactiae isolates at calving and lactation performance within the same lactation. *Journal of dairy science*, 90(2), 766-778.

Wyder, A. B., Boss, R., Naskova, J., Kaufmann, T., Steiner, A., & Graber, H. U. (2011). Streptococcus spp. and related bacteria: their identification and their pathogenic potential for chronic mastitis–a molecular approach. *Research in veterinary science*, 91: 349 - 57.

Yan, S. S., Fox, M. L., M., H. S., Stock, F., Gill, V. J., & Fedorko, D. P. (2000). Resistance to multiple fluoroquinolones in a clinical isolate of Streptococcus pyogenes: identification of gyrA and parC and specification of point mutations associated with resistance. *Antimicrobial agents and chemotherapy*, 44(11), 3196-3198.

Zhang, W., Rong, C., Chen, C., & Gao, G. F. (2012). Type-IVC secretion system: a novel subclass of type IV secretion system (T4SS) common existing in gram-positive genus Streptococcus. *PLoS One*, 7(10), e46390.

Zhou, Y., Liang, Y., Lynch, K., Dennis, J., & Wishart, D. (2011). PHAST: A Fast Phage Search Tool. *Nucleic Acids Research*.

Tables

Table 1. Results from Quast summarizing the assemblies features for the assemblies of each one of the strains of this study.

	SDD64	**SDD92**	**SDD142**
Number of contigs > 0 bp	286	210	208
Number of contigs > 500 bp	131	138	132
Number of contigs > 1000 bp	120	120	116
Number of contigs > 5000 bp	79	78	81
N50	32945	32312	32948
Largest Contig	120861	120572	120573
Total length	2129995	2082786	2126294

Table 2. Results of the *in-silico* PCR regarding the virulence genes amplified.

Strain	Gene	Primers	Expected Size (bp)	Product in-silico PCR (bp)	Reference primer source
SD64 **SD92** **SD142**	*sagA*	5'-gatgataccccgataaggataa 5'-tacttcaaatattttagctact	487	487	(Rato *et. al.,* 2011)
	slo	5'-acggcagctcttatcatt 5'-gacctcaaccgttgctttgt	487	487	(Rato *et. al.,* 2011)
	NAPlr	5′-gttaaagttggtattaacggt 5´-ttgagcagtgtaagacatttc	1157	1157	(Abdelsalam, *et. al.,* 2015b)
	eno	5′-atgtcaattattactgatgt 5′-ctattttttaagttataga	1307	1306	(Abdelsalam, *et. al.,* 2015b)
	sof-FD	5′-ggmgtwgatttacarggwgc 5′-ctgcmgctccaataaywgtta	3329	3329	(Nishiki, 2011)
ATCC-27957	*emm*	5'-tattcgcttagaaaattaa 5'-gcaagttcttcagcttgttt	Variable	5117	(Rato *et. al.*,2011)
	sagA	5'-gatgataccccgataaggataa 5'-tacttcaaatattttagctact	487	480	(Rato *et. al.*,2011)
	slo	5'-acggcagctcttatcatt 5'-gacctcaaccgttgctttgt	487	8291	(Rato *et. al.*,2011)
	isp.1	5'-ggttgaagtcaaaggcaccataa 5'-caactgaaaaaaccccagagcc	429	416	(Rato *et. al.*,2011)
	NAPlr	5′-gttaaagttggtattaacggt 5´-ttgagcagtgtaagacatttc	1157	1157	(Abdelsalam, *et. al.,* 2015b)
	eno	5′-atgtcaattattactgatgt 5′-ctattttttaagttataga	1307	1306	(Abdelsalam, *et. al,* 2015b)

Table 3. Results of the *in-silico* PCR regarding the resistance genes amplified

Strain	Gene	Primers	Expected Size (bp)	Product in-silico PCR (bp)	Reference primer source
SD64 **SD92** **SD142**	*gyrB*	5'-acatcdgcatcrgtcat 5'-gaagtdgtiaaratyacbaaycg	470	470	(Maeda et al., 2011)
	parC	5'-caaaacatgtcccttgagga 5'-ctagctttgggatgatcaatcat	520	587	(Yan et al., 2000)
	tn1207.3/f10394.4 lj	5'- tcttcgccgcataaaccctatc 5'-cctttgaccaatgaagtgaccttt	453	452	(Rato et al., 2010)
ATCC-27957	*gyrB*	5'-acatcdgcatcrgtcat 5'-gaagtdgtiaaratyacbaaycg	470	470	(Maeda *et al*., 2011)
	parC	5'-caaaacatgtcccttgagga 5'-ctagctttgggatgatcaatcat	520	515	(Yan *et al*., 2000)
	gyrA	5'-agtttyatygaytaygcbatgag 5'-ccrggnandacttccat	614	584	(Maeda *et al*., 2011)
	parE	5'-tcyarwcygcyatyacyaagg 5'-gcdccdatngtrtaratcat	390	8852	(Maeda *et al*., 2011)
	tn1207.3/f10394.4 lj	5'-cctttgaccaatgaagtgaccttt 5'-cctttgaccaatgaagtgaccttt	453	8359	(Rato *et al*., 2010)

Table 4. Hit table showing the percentage of identity of the SDD strains of this study against the 129 Streptococcus virulence factors from VFDB

Strain/Gene	***mf/spd***	***hasC***	***fbp54***	***sda***	***gbs0630***	***gbs0631***	***gbs0632***
ATCC 27957	80%	91%	87%	81%	0%	0%	0%
SD64	0%	98%	88%	0%	90%	97%	95%
SD92	0%	98%	88%	0%	90%	97%	95%
SD142	0%	98%	88%	0%	90%	97%	95%

Table 5. Table showing the results of Phaster, a software for phage prediction.

Strain	**Phage**	**Score**	**CPP(%)**	**Accession Number**
SD64	Streptococcus phage phiNJ2	Intact	50.81	NC_019418.1
	Streptococcus phage T12	Incomplete	15.38	NC_028700
	Streptococcus prophage 315.2	Incomplete	46.15	NC_004585
	Streptococcus phage phiARI0923	Incomplete	72.72	NC_030946
	Streptococcus prophage 315.1	Incomplete	31.81	NC_004584
	Lactococcus phage 28201	Incomplete	11.62	NC_031013
SD92	Streptococcus phage A25	Intact	76	NC_028697
	Streptococcus phage T12	Incomplete	15.38	NC_028700
	Streptococcus prophage 315.2	Incomplete	44.23	NC_004585
	Prochlorococcus phage P-SSM2	Incomplete	33.33	NC_006883
	Streptococcus phage phiARI0131-2	Incomplete	53.33	NC_031941
	Streptococcus phage phiNJ2	Incomplete	52.42	NC_019418.1
SD142	Streptococcus prophage 315.1	Incomplete	33.3	NC_004584
	Streptococcus prophage 315.2	Incomplete	54.76	NC_004585
	Streptococcus phage T12	Incomplete	15.38	NC_028700
	Lactococcus phage 28201	Incomplete	13.04	NC_031013
	Streptococcus phage phiNJ2	Incomplete	50	NC_019418
	Streptococcus phage A25	Questionable	78.26	NC_028697
	Streptococcus phage phiARI0746	Incomplete	35.71	NC_031907

CPP: Common Phage Percentage

Figure legends

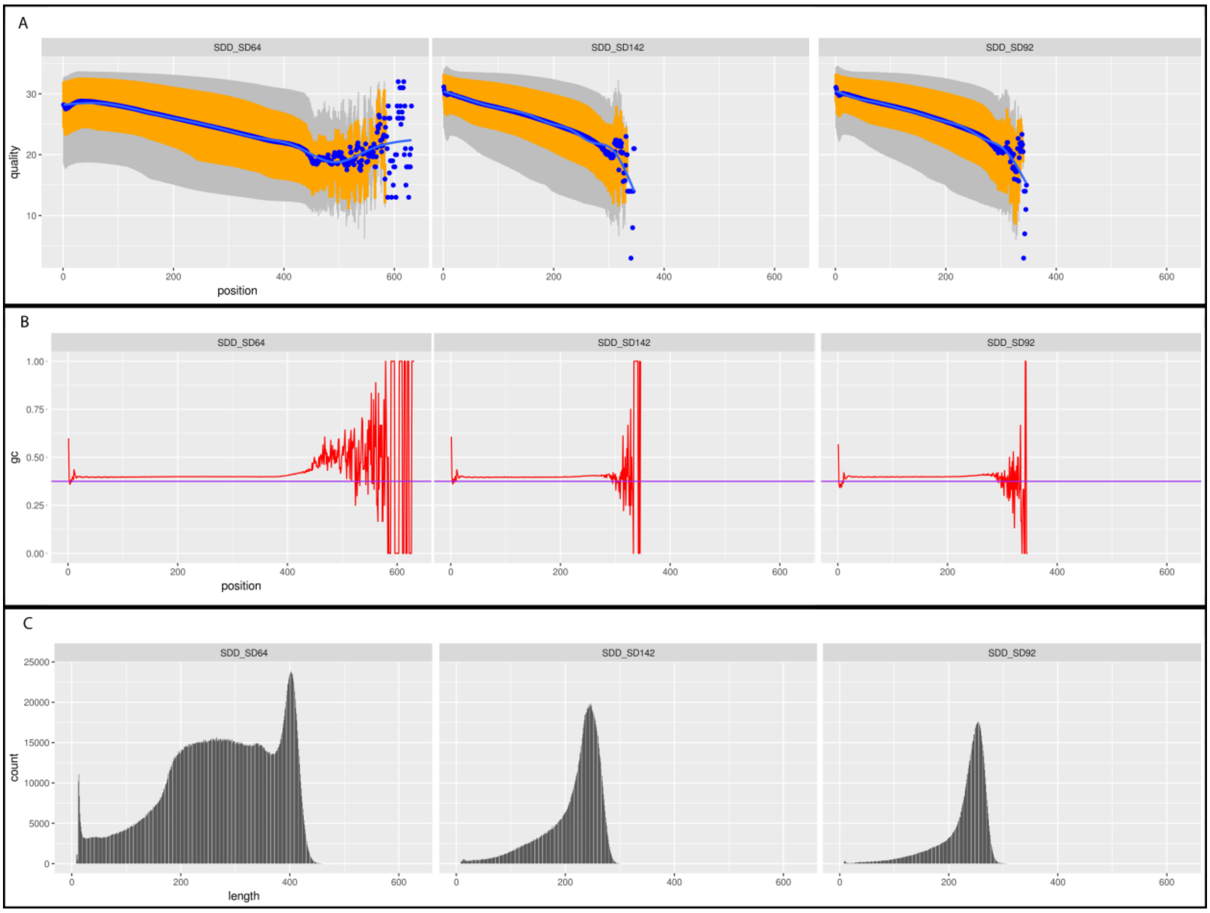


Figure 1. **QRQC analyses of quality reads obtained from Ion PGM to SD64, SD92 and SD142 strains.** A - Quality Score, boxplot analysis showing the quality of reads in the Phred scale, the blue line is the mean quality of reads and the orange is the 1st and 3rd quartiles. B – GC content percentage for each position reads. C – Sum of Read Length for each dataset.

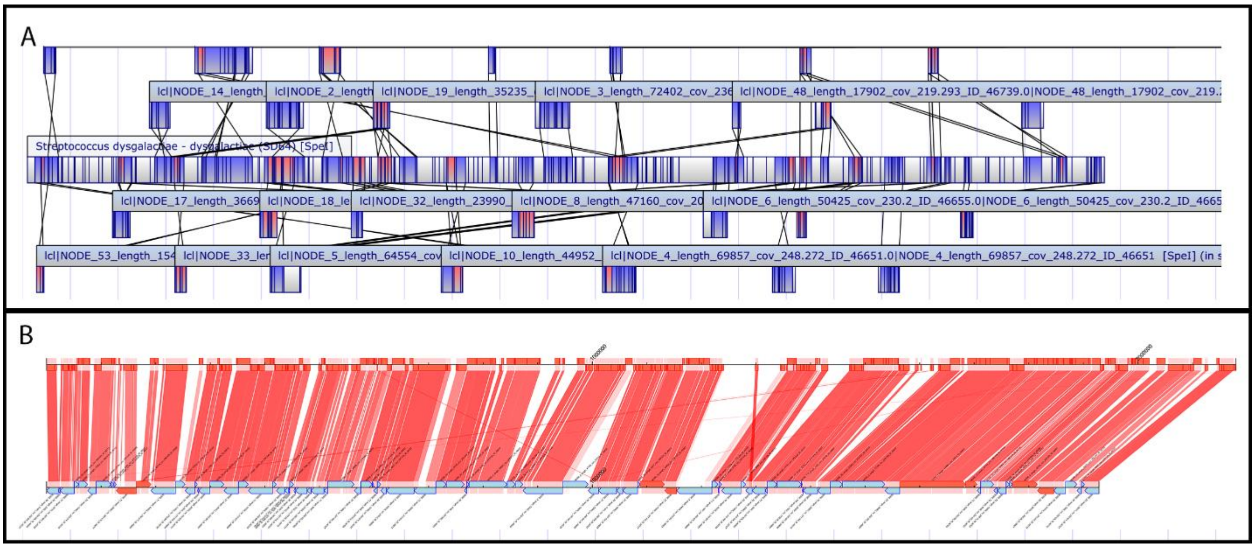

A
lcl|NODE_14_length_
lcl|NODE_2_length
lcl|NODE_19_length_35235_
lcl|NODE_3_length_72402_cov_236
lcl|NODE_48_length_17902_cov_219.293_ID_46739.0|NODE_48_length_17902_cov_219.
Streptococcus dysgalactiae - dysgalactiae (SD64) [SpeI]
lcl|NODE_17_length_3669
lcl|NODE_18_le
lcl|NODE_32_length_23990_
lcl|NODE_8_length_47160_cov_20
lcl|NODE_6_length_50425_cov_230.2_ID_46655.0|NODE_6_length_50425_cov_230.2_ID_4665
lcl|NODE_53_length_154
lcl|NODE_33_len
lcl|NODE_5_length_64554_cov
lcl|NODE_10_length_44952_
lcl|NODE_4_length_69857_cov_248.272_ID_46651.0|NODE_4_length_69857_cov_248.272_ID_46651 [SpeI] (in s
B
1000000
2000000

Figure 2. **SD64 contig alignments.** A - Alignment of the SD64 contigs with the optical map by OpGen, showing the optical map on the middle and the 36 contigs aligned along the map. B - CONTIGuator scaffold generation with alignment of SD64 strain contigs to reference complete genome sequence of *S. dysgalactiae* subsp. *equisimilis* ATCC-2713; on top the reference genome and on bottom 87 contigs aligned.

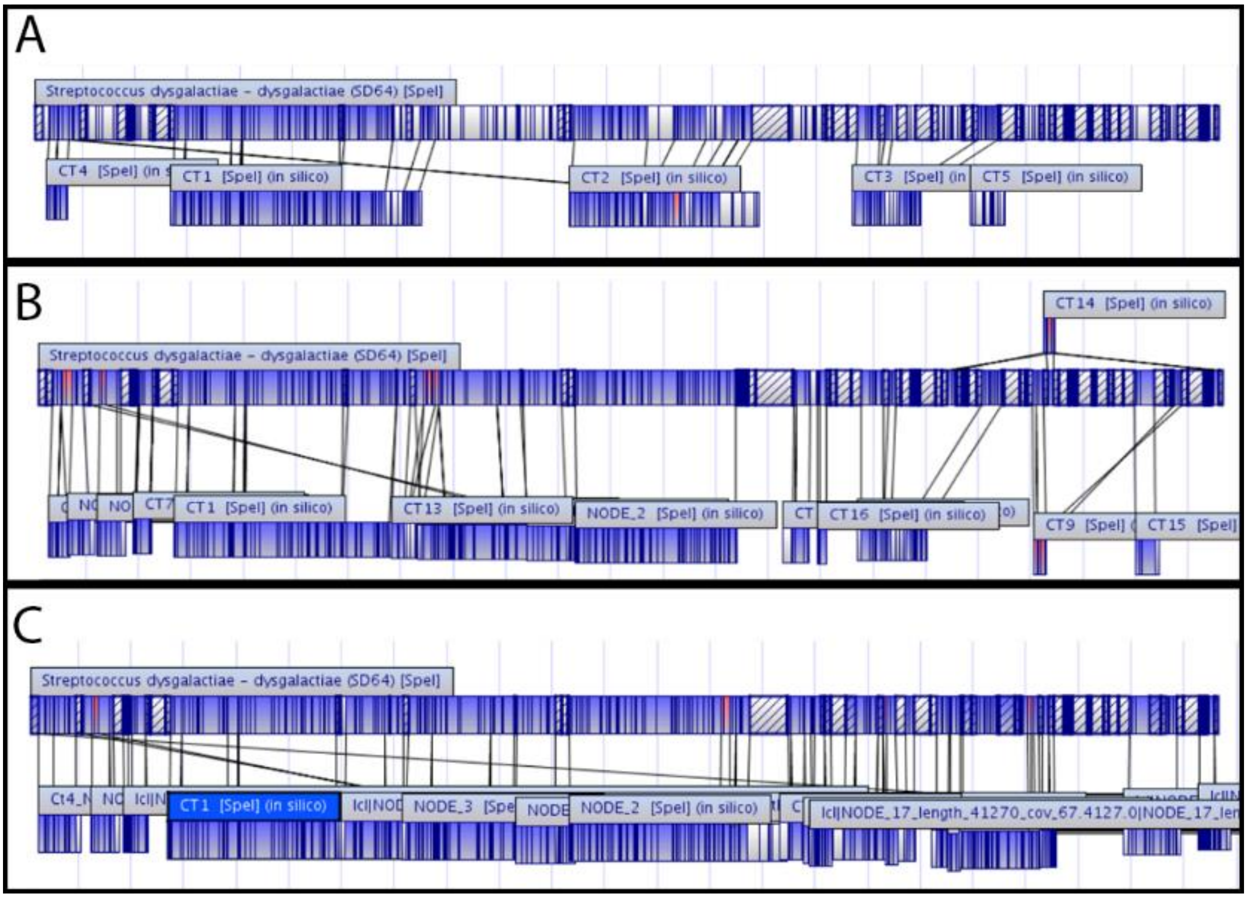


Figure 3. Whole genome coverage increasing on MapSolver while scaffolds are being constructed. A - the first gap filling run, 25 contigs were used initially to construct five super scaffold and reached the 60.89 % of WGC. B - on the second gap filling run, 9 contigs were used to construct three more scaffolds which reached the 70.51% of WGC. C – and , finally, the third gap filling run, used 3 contigs to construct one additional scaffold that lead to 83.93% of WGC.

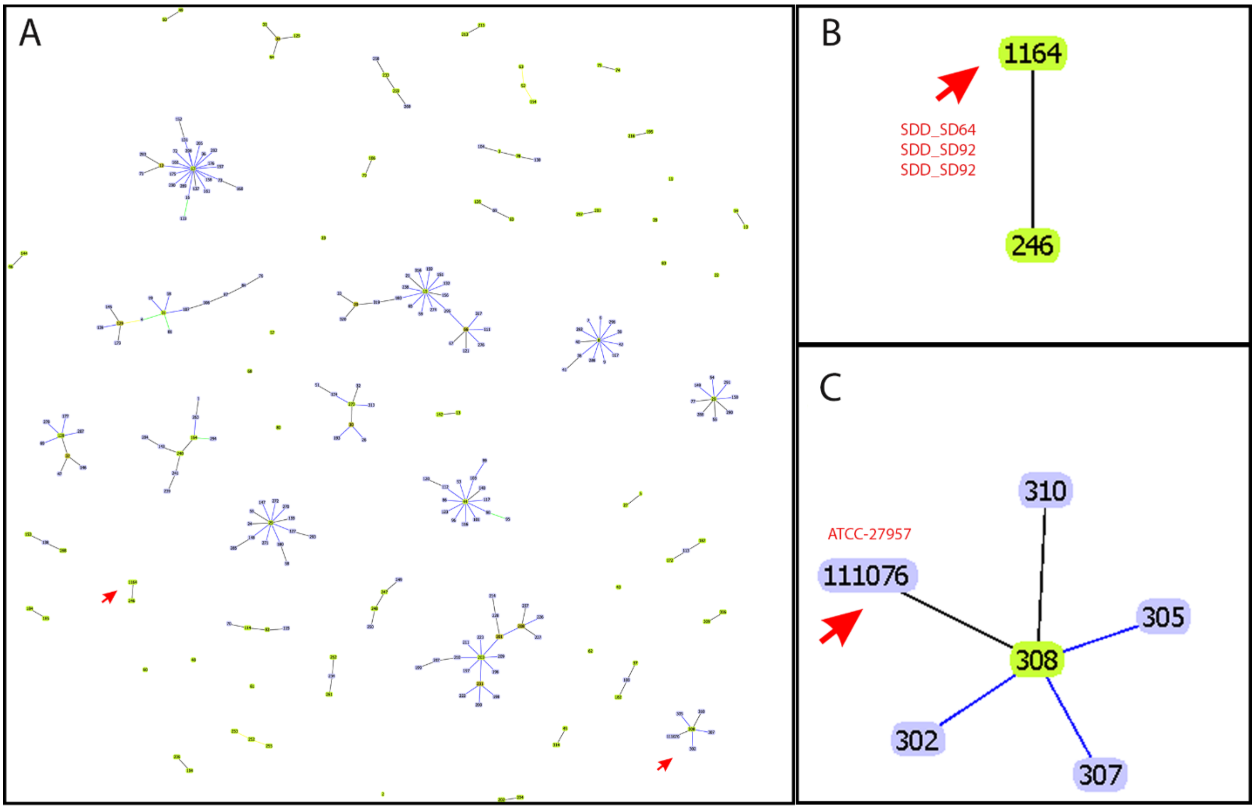


Figure 4. **GeoBURST results.** A - MLST analysis, performed in geoBURST within all the ST available online on the pubmlst database, showing the groups formed by SDD strains. B – close relationship formed with the ST of SD64, SD42, and SD192 strains and ST-246. C – Clonal Complex formed with ST of the SDD ATCC-27957 strain and ST of other bovine isolated strains.

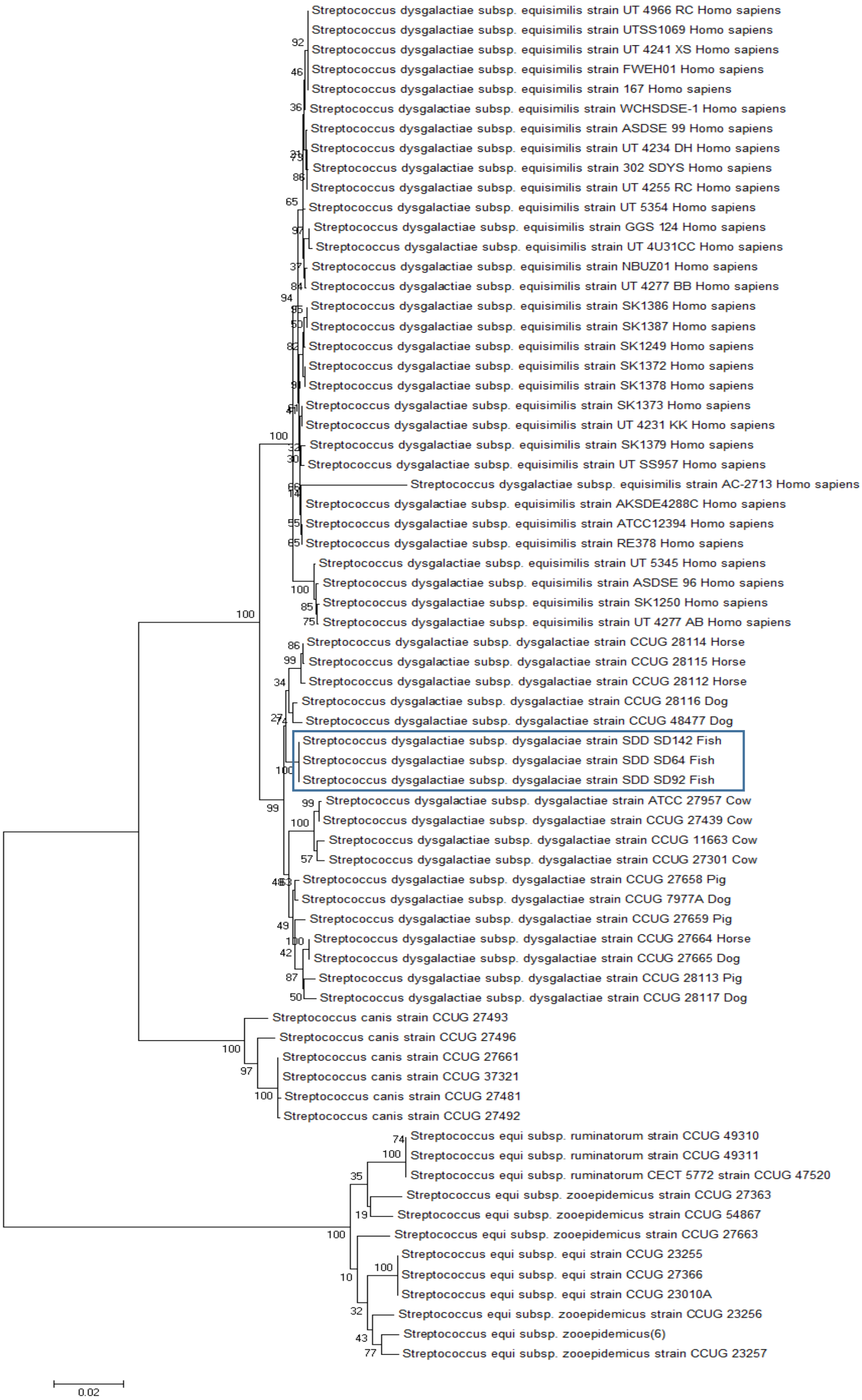

Streptococcus dysgalactiae subsp. equisimilis strain UT 4966 RC Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain UTSS1069 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain UT 4241 XS Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain FWEH01 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain 167 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain WCHSDSE-1 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain ASDSE 99 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain UT 4234 DH Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain 302 SDYS Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain UT 4255 RC Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain UT 5354 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain GGS 124 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain UT 4U31CC Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain NBUZ01 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain UT 4277 BB Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain SK1386 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain SK1387 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain SK1249 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain SK1372 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain SK1378 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain SK1373 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain UT 4231 KK Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain SK1379 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain UT SS957 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain AC-2713 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain AKSDE4288C Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain ATCC12394 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain RE378 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain UT 5345 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain ASDSE 96 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain SK1250 Homo sapiens
Streptococcus dysgalactiae subsp. equisimilis strain UT 4277 AB Homo sapiens
Streptococcus dysgalactiae subsp. dysgalactiae strain CCUG 28114 Horse
Streptococcus dysgalactiae subsp. dysgalactiae strain CCUG 28115 Horse
Streptococcus dysgalactiae subsp. dysgalactiae strain CCUG 28112 Horse
Streptococcus dysgalactiae subsp. dysgalactiae strain CCUG 28116 Dog
Streptococcus dysgalactiae subsp. dysgalactiae strain CCUG 48477 Dog
Streptococcus dysgalactiae subsp. dysgalaciae strain SDD SD142 Fish
Streptococcus dysgalactiae subsp. dysgalaciae strain SDD SD64 Fish
Streptococcus dysgalactiae subsp. dysgalaciae strain SDD SD92 Fish
Streptococcus dysgalactiae subsp. dysgalactiae strain ATCC 27957 Cow
Streptococcus dysgalactiae subsp. dysgalactiae strain CCUG 27439 Cow
Streptococcus dysgalactiae subsp. dysgalactiae strain CCUG 11663 Cow
Streptococcus dysgalactiae subsp. dysgalactiae strain CCUG 27301 Cow
Streptococcus dysgalactiae subsp. dysgalactiae strain CCUG 27658 Pig
Streptococcus dysgalactiae subsp. dysgalactiae strain CCUG 7977A Dog
Streptococcus dysgalactiae subsp. dysgalactiae strain CCUG 27659 Pig
Streptococcus dysgalactiae subsp. dysgalactiae strain CCUG 27664 Horse
Streptococcus dysgalactiae subsp. dysgalactiae strain CCUG 27665 Dog
Streptococcus dysgalactiae subsp. dysgalactiae strain CCUG 28113 Pig
Streptococcus dysgalactiae subsp. dysgalactiae strain CCUG 28117 Dog
Streptococcus canis strain CCUG 27493
Streptococcus canis strain CCUG 27496
Streptococcus canis strain CCUG 27661
Streptococcus canis strain CCUG 37321
Streptococcus canis strain CCUG 27481
Streptococcus canis strain CCUG 27492
Streptococcus equi subsp. ruminatorum strain CCUG 49310
Streptococcus equi subsp. ruminatorum strain CCUG 49311
Streptococcus equi subsp. ruminatorum CECT 5772 strain CCUG 47520
Streptococcus equi subsp. zooepidemicus strain CCUG 27363
Streptococcus equi subsp. zooepidemicus strain CCUG 54867
Streptococcus equi subsp. zooepidemicus strain CCUG 27663
Streptococcus equi subsp. equi strain CCUG 23255
Streptococcus equi subsp. equi strain CCUG 27366
Streptococcus equi subsp. equi strain CCUG 23010A
Streptococcus equi subsp. zooepidemicus strain CCUG 23256
Streptococcus equi subsp. zooepidemicus(6)
Streptococcus equi subsp. zooepidemicus strain CCUG 23257
0.02

Figure 5. Result tree of the MLSA analysis using the concatenated sequence of seven housekeeping genes (*map, pfl, ppaC, pyk, rpoB, soda* and *tuf*) subjected to phylogenetic analysis using the Minimum Evolution tree algorithm.

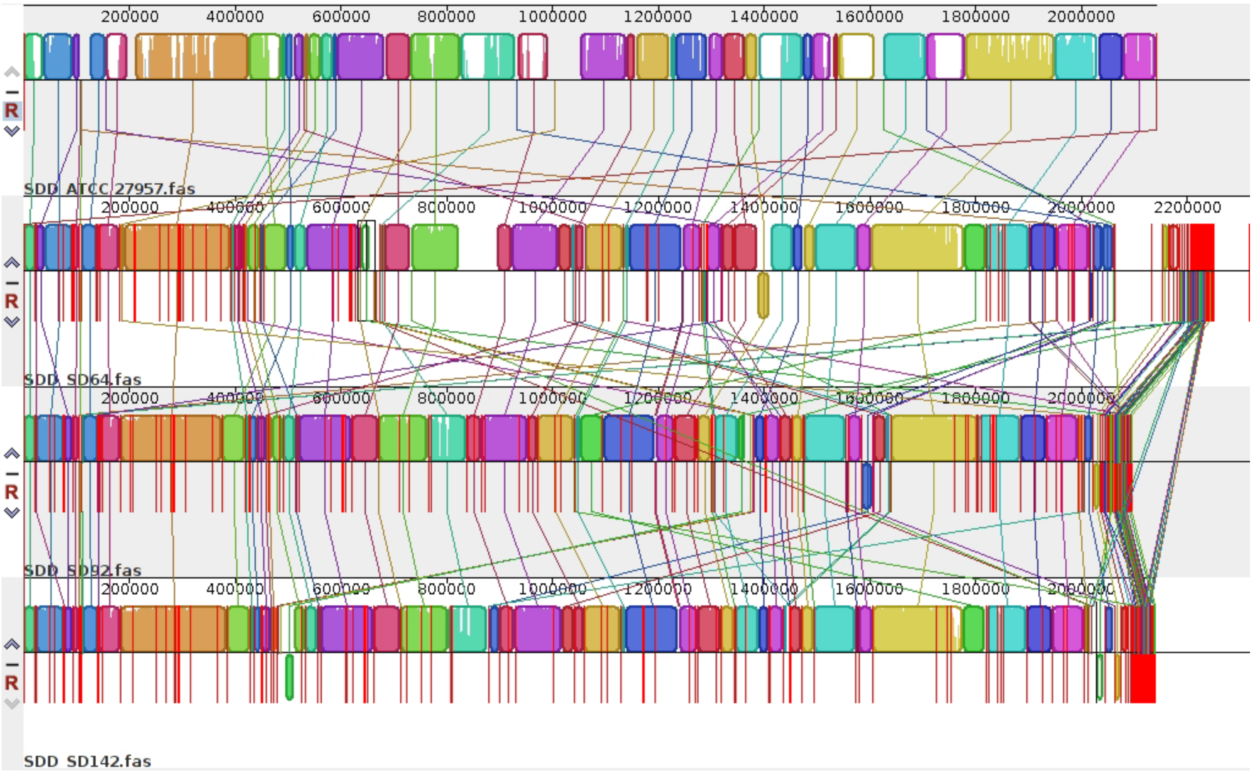
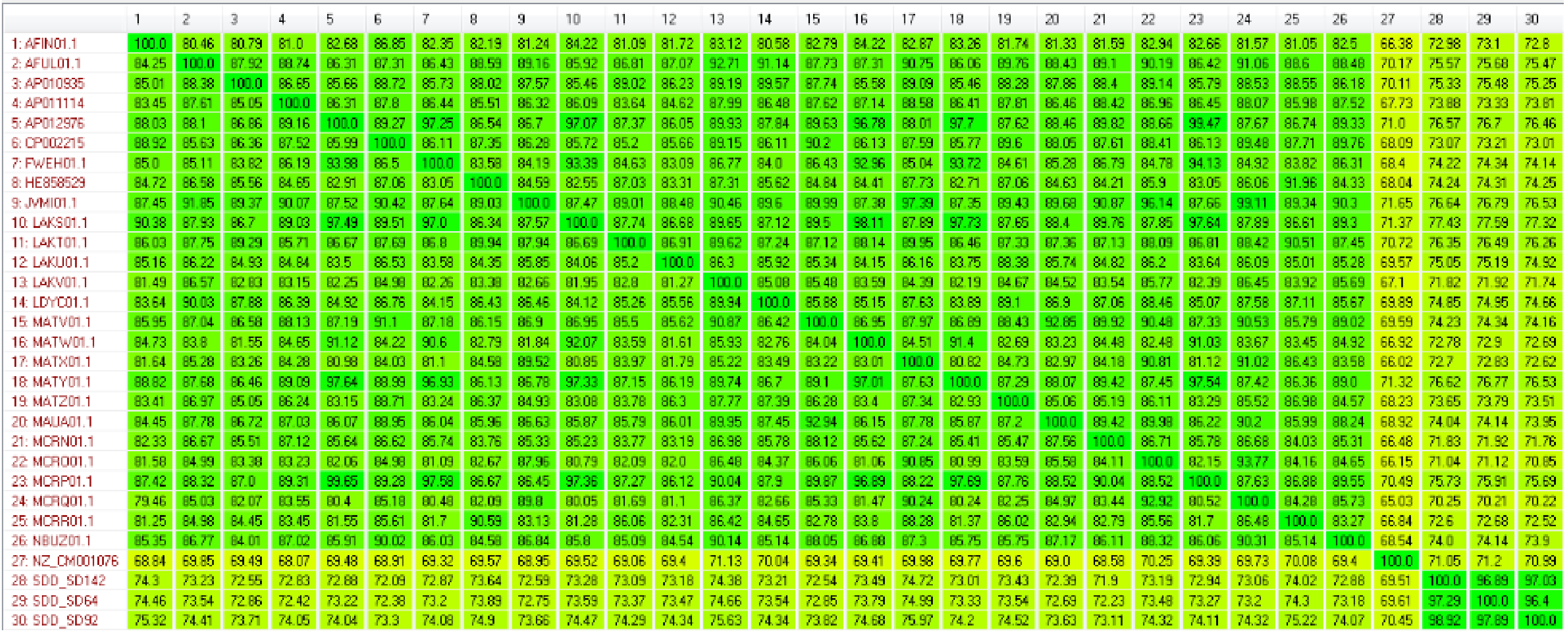


Figure 6. Locally collinear blocks (LCB) within the genomes of the tree strains of four strains of *Streptococcus dysgalactiae* subsp. *dysgalactia*e. From top to bottom ATCC-27957, SD64, SD92 and SD142. Blocks with the same color represent LCB between the genomes, where the white portions inside the blocks indicate regions of low similarity. Red vertical bars show the delimitation of the contigs and LCBs below a genome's center line are in the reverse complement orientation relative to the reference genome



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1: AFIN01.1	100.0	80.46	80.79	81.0	82.68	86.85	82.35	82.19	81.24	84.22	81.09	81.72	83.12	80.58	82.79	84.22	82.87	83.26	81.74	81.33	81.59	82.94	82.66	81.57	81.05	82.5	66.38	72.98	73.1	72.8
2: AFUL01.1	84.25	100.0	87.92	88.74	86.31	87.31	86.43	88.59	89.16	85.92	86.81	87.07	92.71	91.14	87.73	87.31	90.75	86.06	89.76	88.43	89.1	90.19	86.42	91.06	88.6	88.48	70.17	75.57	75.68	75.47
3: AP010935	85.01	88.38	100.0	86.65	85.66	88.72	85.73	88.02	87.57	85.46	89.02	86.23	89.19	89.57	87.74	85.58	89.09	85.46	88.28	87.86	88.4	89.14	85.79	88.53	88.55	86.18	70.11	75.33	75.48	75.25
4: AP011114	83.45	87.61	85.05	100.0	86.31	87.8	86.44	85.51	86.32	86.09	83.64	84.62	87.99	86.48	87.62	87.14	88.58	86.41	87.81	86.46	88.42	86.96	86.45	88.07	85.98	87.52	67.73	73.88	73.33	73.81
5: AP012976	88.03	88.1	86.86	89.16	100.0	89.27	97.25	86.54	86.7	97.07	87.37	86.05	89.93	87.84	89.63	96.78	88.01	97.7	87.62	88.46	89.82	88.66	99.47	87.67	86.74	89.33	71.0	76.57	76.7	76.46
6: CP002215	88.92	85.63	86.36	87.52	85.99	100.0	86.11	87.35	86.28	85.72	85.2	85.66	89.15	86.11	90.2	86.13	87.59	85.77	89.6	88.05	87.61	88.41	86.13	89.48	87.71	89.76	68.09	73.07	73.21	73.01
7: FWEH01.1	85.0	85.11	83.82	86.19	93.98	86.5	100.0	83.58	84.19	93.39	84.63	83.09	86.77	84.0	86.43	92.96	85.04	93.72	84.61	85.28	86.79	84.78	94.13	84.92	83.82	86.31	68.4	74.22	74.34	74.14
8: HE858529	84.72	86.58	85.56	84.65	82.91	87.06	83.05	100.0	84.59	82.55	87.03	83.31	87.31	85.62	84.84	84.41	87.73	82.71	87.06	84.63	84.21	85.9	83.05	86.06	91.96	84.33	68.04	74.24	74.31	74.25
9: JVMI01.1	87.45	91.85	89.37	90.07	87.52	90.42	87.64	89.03	100.0	87.47	89.01	88.48	90.46	89.6	89.99	87.38	97.39	87.35	89.43	89.68	90.87	96.14	87.66	99.11	89.34	90.3	71.65	76.64	76.79	76.53
10: LAKS01.1	90.38	87.93	86.7	89.03	97.49	89.51	97.0	86.34	87.57	100.0	87.74	86.68	89.65	87.12	89.5	98.11	87.89	97.73	87.65	88.4	89.76	87.85	97.64	87.89	86.61	89.3	71.37	77.43	77.59	77.32
11: LAKT01.1	86.03	87.75	89.29	85.71	86.67	87.69	86.8	89.94	87.94	86.69	100.0	86.91	89.62	87.24	87.12	88.14	89.95	86.46	87.33	87.36	87.13	88.09	86.81	88.42	90.51	87.45	70.72	76.35	76.49	76.26
12: LAKU01.1	85.16	86.22	84.93	84.84	83.5	86.53	83.58	84.35	85.85	84.06	85.2	100.0	86.3	85.92	85.34	84.15	86.16	83.75	88.38	85.74	84.82	86.2	83.64	86.09	85.01	85.28	69.57	75.05	75.19	74.92
13: LAKV01.1	81.49	86.57	82.83	83.15	82.25	84.98	82.26	83.38	82.66	81.95	82.8	81.27	100.0	85.08	85.48	83.59	84.39	82.19	84.67	84.52	83.54	85.77	82.39	86.45	83.92	85.69	67.1	71.82	71.92	71.74
14: LDYC01.1	83.64	90.03	87.88	86.39	84.92	86.76	84.15	86.43	86.46	84.12	85.26	85.56	89.94	100.0	85.88	85.15	87.63	83.89	89.1	86.9	87.06	88.46	85.07	87.58	87.11	85.67	69.89	74.85	74.95	74.66
15: MATV01.1	85.95	87.04	86.58	88.13	87.19	91.1	87.18	86.15	86.9	86.95	85.5	85.62	90.87	86.42	100.0	86.95	87.97	86.89	88.43	92.85	89.92	90.48	87.33	90.53	85.79	89.02	69.59	74.23	74.34	74.16
16: MATW01.1	84.73	83.8	81.55	84.65	91.12	84.22	90.6	82.79	81.84	92.07	83.59	81.61	85.93	82.76	84.04	100.0	84.51	91.4	82.69	83.23	84.48	82.48	91.03	83.67	83.45	84.92	66.92	72.78	72.9	72.69
17: MATX01.1	81.64	85.28	83.26	84.28	80.98	84.03	81.1	84.58	89.52	80.85	83.97	81.79	85.22	83.49	83.22	83.01	100.0	80.82	84.73	82.97	84.18	90.81	81.12	91.02	86.43	83.58	66.02	72.7	72.83	72.62
18: MATY01.1	88.82	87.68	86.46	89.09	97.64	88.99	96.93	86.13	86.78	97.33	87.15	86.19	89.74	86.7	89.1	97.01	87.63	100.0	87.29	88.07	89.42	87.45	97.54	87.42	86.36	89.0	71.32	76.62	76.77	76.53
19: MATZ01.1	83.41	86.97	85.05	86.24	83.15	88.71	83.24	86.37	84.93	83.08	83.78	86.3	87.77	87.39	86.28	83.4	87.34	82.93	100.0	85.06	85.19	86.11	83.29	85.52	86.98	84.57	68.23	73.65	73.79	73.51
20: MAUA01.1	84.45	87.78	86.72	87.03	86.07	88.95	86.04	85.96	86.63	85.87	85.79	86.01	89.95	87.45	92.94	86.15	87.78	85.87	87.2	100.0	89.42	89.98	86.22	90.2	85.99	88.24	68.92	74.04	74.14	73.95
21: MCRN01.1	82.33	86.67	85.51	87.12	85.64	86.62	85.74	83.76	85.33	85.23	83.77	83.19	86.98	85.78	88.12	85.62	87.24	85.41	85.47	87.56	100.0	86.71	85.78	86.68	84.03	85.31	66.48	71.83	71.92	71.76
22: MCRO01.1	81.58	84.99	83.38	83.23	82.06	84.98	81.09	82.67	87.96	80.79	82.09	82.0	86.48	84.37	86.06	81.06	90.85	80.99	83.59	85.58	84.11	100.0	82.15	93.77	84.16	84.65	66.15	71.04	71.12	70.85
23: MCRP01.1	87.42	88.32	87.0	89.31	99.65	89.28	97.58	86.67	86.45	97.36	87.27	86.12	90.04	87.9	89.87	96.89	88.22	97.69	87.76	88.52	90.04	88.52	100.0	87.63	86.88	89.55	70.49	75.73	75.91	75.69
24: MCRQ01.1	79.46	85.03	82.07	83.55	80.4	85.18	80.48	82.09	89.8	80.05	81.69	81.1	86.37	82.66	85.33	81.47	90.24	80.24	82.25	84.97	83.44	92.92	80.52	100.0	84.28	85.73	65.03	70.25	70.21	70.22
25: MCRR01.1	81.25	84.98	84.45	83.45	81.55	85.61	81.7	90.59	83.13	81.28	86.06	82.31	86.42	84.65	82.78	83.8	88.28	81.37	86.02	82.94	82.79	85.56	81.7	86.48	100.0	83.27	66.84	72.6	72.68	72.52
26: NBUZ01.1	85.35	86.77	84.01	87.02	85.91	90.02	86.03	84.58	86.84	85.8	85.09	84.54	90.14	85.14	88.05	86.88	87.3	85.75	85.75	87.17	86.11	88.32	86.06	90.31	85.14	100.0	68.54	74.0	74.14	73.9
27: NZ_CM001076	68.84	69.85	69.49	68.07	69.48	68.91	69.32	69.57	68.95	69.52	69.06	69.4	71.13	70.04	69.34	69.41	69.98	69.77	69.6	69.0	68.58	70.25	69.39	69.73	70.08	69.4	100.0	71.05	71.2	70.99
28: SDD_SD142	74.3	73.23	72.55	72.83	72.88	72.09	72.87	73.64	72.59	73.28	73.09	73.18	74.38	73.21	72.54	73.49	74.72	73.01	73.43	72.39	71.9	73.19	72.94	73.06	74.02	72.88	69.51	100.0	96.89	97.03
29: SDD_SD64	74.46	73.54	72.86	72.42	73.22	72.38	73.2	73.89	72.75	73.59	73.37	73.47	74.66	73.54	72.85	73.79	74.99	73.33	73.54	72.69	72.23	73.48	73.27	73.2	74.3	73.18	69.61	97.29	100.0	96.4
30: SDD_SD92	75.32	74.41	73.71	74.05	74.04	73.3	74.08	74.9	73.66	74.47	74.29	74.34	75.63	74.34	73.82	74.68	75.97	74.2	74.52	73.63	73.11	74.32	74.11	74.32	75.22	74.07	70.45	98.92	97.89	100.0

Figure 7. Heatmap Plot within the 27 genome sequences available for *Streptococcus dysgalactiae* on the NCBI databales. The strains corresponding to the SD available group are listed on the entries 1 to 26, as for the SDD group the entries on the plot correspond to 27 to 30

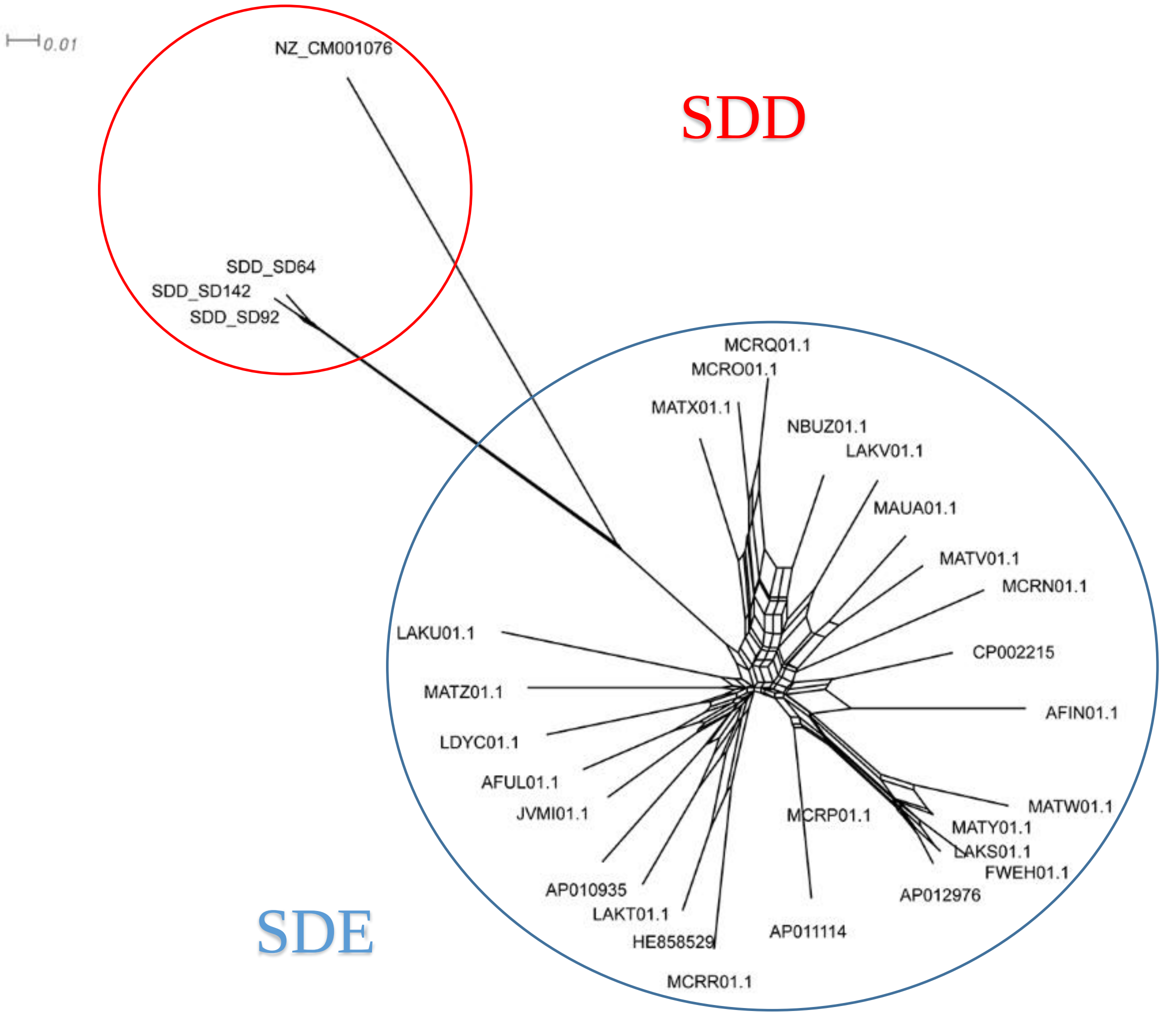


Figure 8. Phylogenomic tree showing the result of the comparison within all the genomes of both the SDD group and the SD available group strains.

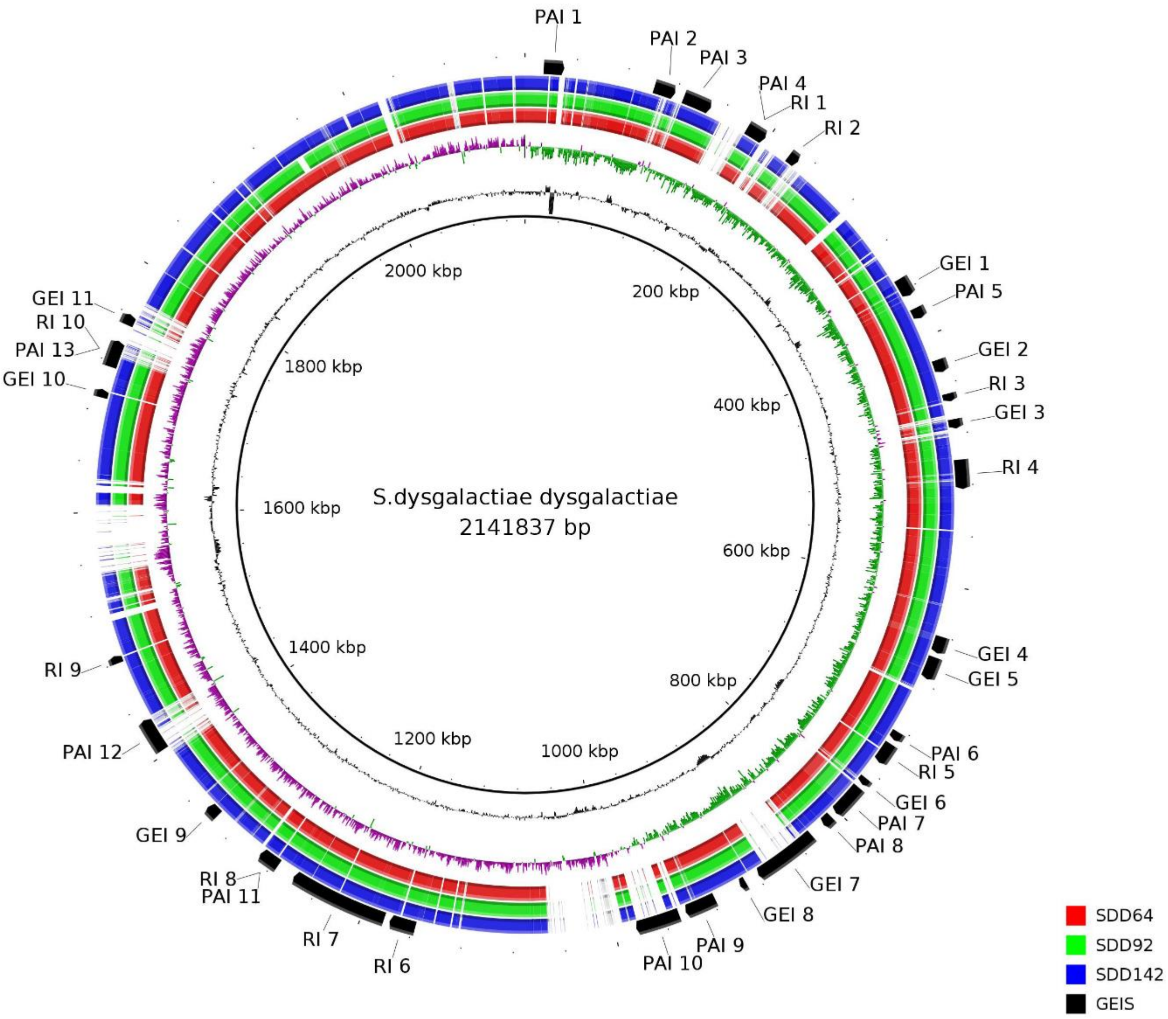


Figure 9. Putative pathogenicity and genomic islands predicted by GIPSy. Comparisons where made between *Streptococcus dysgalactiae* subsp. *dysgalactiae* ATCC- 27957 from the NCBI database and the strains SD64 (red ring), SD92 (green ring) and SD142 (blue ring).

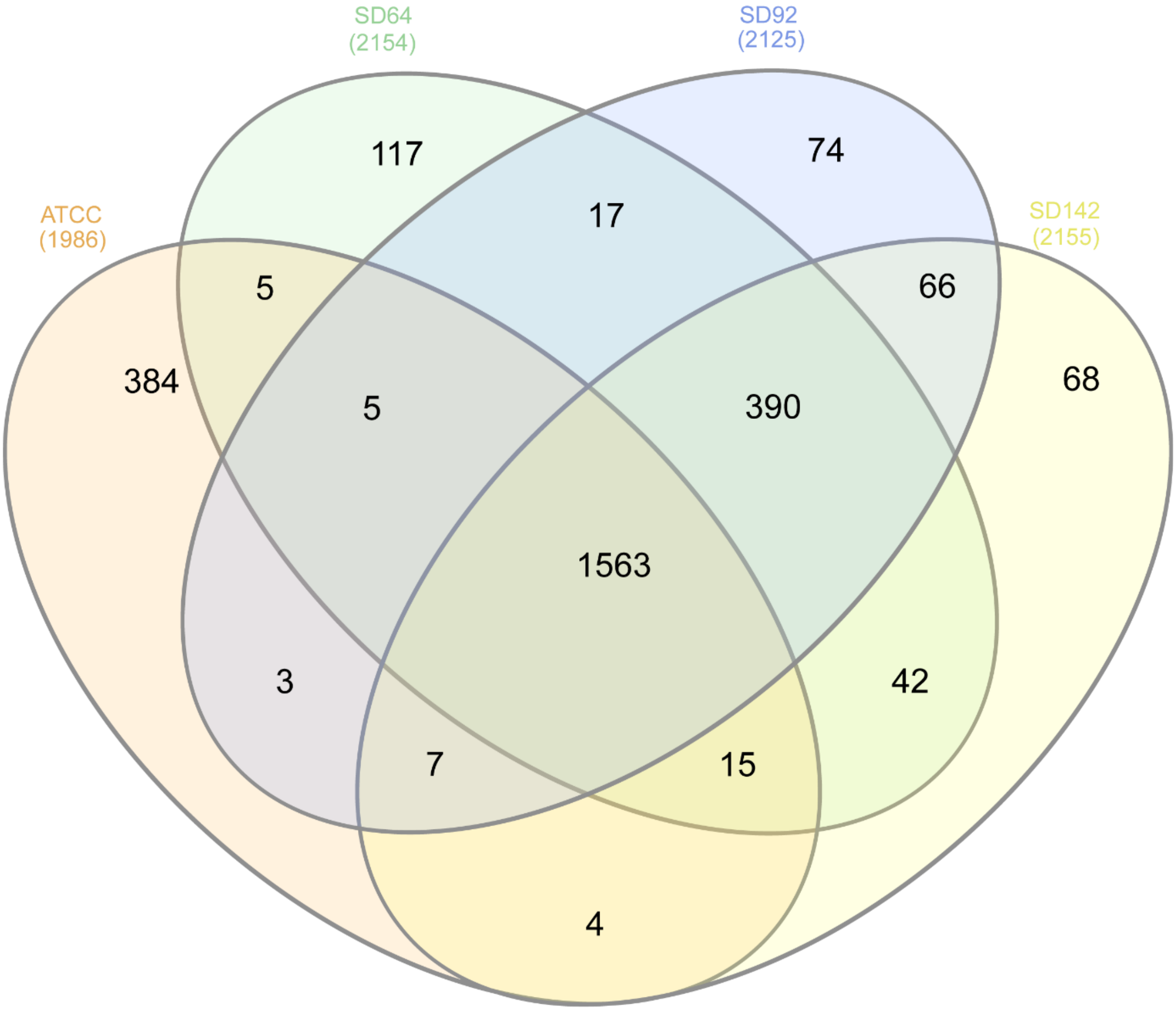


Figure 10. Venn Diagram of the orthoMCL analysis representing the core, accessory and specific genes within the all the *Streptococcus dysgalactiae* subsp. *dysgalactie* of this project and *S. dysgalactiae* subsp. *dysgalactiae* ATCC-25957.

Supplementary Table 1

Organism/Name	**Strain**	**Accession number**
Streptococcus dysgalactiae subsp. *equisimilis* AC-2713	AC-2713	HE858529.1
Streptococcus dysgalactiae subsp. *equisimilis* GGS_124	GGS_124	AP010935.1
Streptococcus dysgalactiae subsp. *equisimilis* ATCC 12394	ATCC 12394	CP002215.1
Streptococcus dysgalactiae subsp. *equisimilis* RE378	RE378	AP011114.1
Streptococcus dysgalactiae subsp. *equisimilis*	UT_4277_BB	NZ_MAUA00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*	UT_4242_AB	NZ_MATZ00000000.1
Streptococcus dysgalactiae subsp. *equisimilis* SK1249	SK1249	NZ_AFIN00000000.1
Streptococcus dysgalactiae subsp. *equisimilis* SK1250	SK1250	NZ_AFUL00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*	UT-SS1069	NZ_LAKS00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*	UT-5345	NZ_LAKV00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*	UT-5354	NZ_LAKU00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*	UT-SS957	NZ_LAKT00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*	WCHSDSE-1	NZ_LDYC00000000.1
Streptococcus dysgalactiae	302_SDYS	NZ_JVMI00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*	UT_4231_KK	NZ_MATW00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*	UT_4031CC	NZ_MATV00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*	UT_4241_XS	NZ_MATY00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*	UT_4234_DH	NZ_MATX00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*	ASDSE_96	NZ_MCRN00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*	ASDSE_99	NZ_MCRO00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*	UT_4966_RC	NZ_MCRP00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*	UT_4255RC	NZ_MCRQ00000000.1
Streptococcus dysgalactiae subsp. *equisimilis* AKSDE4288	AKSDE4288	NZ_MCRR00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*		FWEH00000000.1
Streptococcus dysgalactiae subsp. *equisimilis*		NZ_NBUZ00000000.1
Streptococcus dysgalactiae subsp. *equisimilis* 167	167	AP012976.1

Supplementary Table 3

Pathogenicity Island 1	
Gene	**Product**
tilS	tRNA(Ile)-lysidine synthase
hpt	Hypoxanthine-guanine phosphoribosyltransferase
ftsH	ATP-dependent zinc metalloprotease FtsH
-	Transposase DDE domain protein
-	hypothetical protein
yhdG	putative amino acid permease YhdG

Pathogenicity Island 2	
Gene	**Product**
hepT_1	Heptaprenyl diphosphate synthase component 2
ispE	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase
adcR	Transcriptional repressor AdcR
znuC_1	High-affinity zinc uptake system ATP-binding protein ZnuC
znuB	High-affinity zinc uptake system membrane protein ZnuB
	Putative prophage phiRv2 integrase
	hypothetical protein
	hypothetical protein
	Transposase, Mutator family
proX_1	Prolyl-tRNA editing protein ProX
	hypothetical protein
xre_1	HTH-type transcriptional regulator Xre
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein

	hypothetical protein
	hypothetical protein
	hypothetical protein
	Poxvirus D5 protein-like protein

Pathogenicity Island 3

Gene	**Product**
tyrS	Tyrosine--tRNA ligase
pbpF_1	Penicillin-binding protein 1F
rpoB	DNA-directed RNA polymerase subunit beta
rpoC	DNA-directed RNA polymerase subunit beta'
	hypothetical protein
gspE	Putative type II secretion system protein E
epsF	Type II secretion system protein F
	hypothetical protein
	hypothetical protein
	Type II secretory pathway pseudopilin
	hypothetical protein
	hypothetical protein
rsmA_1	Ribosomal RNA small subunit methyltransferase A
ackA	Acetate kinase
	hypothetical protein
	CAAX amino terminal protease self- immunity
	hypothetical protein
proC	Pyrroline-5-carboxylate reductase
pepA_1	Glutamyl aminopeptidase

Supplementary Table 3 (Continuation)

Pathogenicity Island 4/Resistance Island 1	
Gene	**Product**
tmpC_1	Membrane lipoprotein TmpC precursor
csdA	4-hydroxyphenylacetate decarboxylase activating enzyme
srlR	Glucitol operon repressor
sorC	Sorbitol operon regulator
chbA	N,N'-diacetylchitobiose-specific phosphotransferase enzyme IIA component
licB	Lichenan-specific phosphotransferase enzyme IIB component
licC	Lichenan permease IIC component
bssA	Benzylsuccinate synthase alpha subunit
fsaA	Fructose-6-phosphate aldolase 1
gldA	Glycerol dehydrogenase
	hypothetical protein
	Phosphatidylglycerophosphatase A
	hypothetical protein
mccF	Microcin C7 self-immunity protein MccF
	hypothetical protein

Resistance Island 2	
Gene	**Product**
ptsG	PTS system glucose-specific EIICBA component
mapP	Maltose 6'-phosphate phosphatase
pepA_2	Glutamyl aminopeptidase
	CAAX amino terminal protease self- immunity
	hypothetical protein
	Putative NrdI-like protein

Supplementary Table 3 (Continuation)

Genomic Island 1	
Gene	**Product**
oppF_1	Oligopeptide transport ATP-binding protein OppF
	hypothetical protein
	hypothetical protein
ppaX	Pyrophosphatase PpaX
	GTPase YlqF
yhbY	RNA-binding protein YhbY
nadD	Nicotinate-nucleotide adenylyltransferase
	putative nicotinate-nucleotide adenylyltransferase
sttH	Streptothricin hydrolase
rsfS	Ribosomal silencing factor RsfS
bsmA	Glycine/sarcosine N-methyltransferase
	hypothetical protein

Pathogenicity Island 5	
Gene	**Product**
sstT	Serine/threonine transporter SstT
ktrA	Ktr system potassium uptake protein A
ktrB	Ktr system potassium uptake protein B
rsmG	Ribosomal RNA small subunit methyltransferase G
	LemA family protein
	hypothetical protein
	hypothetical protein
arlR	Response regulator ArlR

Supplementary Table 3 (Continuation)

Genomic Island 2	
Gene	**Product**
murE	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--L- lysine ligase
ytgP_2	putative cell division protein YtgP
upp	Uracil phosphoribosyltransferase
clpP	ATP-dependent Clp protease proteolytic subunit
	hypothetical protein
	hypothetical protein
	hypothetical protein
braC	Leucine-, isoleucine-, valine-, threonine-, and alanine-binding protein precursor
livH_1	High-affinity branched-chain amino acid transport system permease protein LivH

Resistance Island 3	
Gene	**Product**
	hypothetical protein
	Transposase

Genomic Island 3	
Gene	**Product**
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	Poxvirus D5 protein-like protein
	hypothetical protein
	hypothetical protein
	hypothetical protein

	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein

Resistance Island 4

Gene	Product
msrA	Peptide methionine sulfoxide reductase MsrA
	hypothetical protein
	LysM domain protein
sph	Oleate hydratase
ybeZ	PhoH-like protein
	Uracil DNA glycosylase superfamily protein
ybeY	Endoribonuclease YbeY
dgkA	Undecaprenol kinase
era	GTPase Era
	hypothetical protein
	hypothetical protein
	Transposase
	Integrase core domain protein
	CAAX amino terminal protease self- immunity
	hypothetical protein
	hypothetical protein
	Integrase core domain protein
	hypothetical protein
	Transglutaminase-like superfamily protein
	Helix-turn-helix domain protein
mutM	Formamidopyrimidine-DNA glycosylase
coaE	Dephospho-CoA kinase
yxlF_1	putative ABC transporter ATP-binding protein YxlF
	hypothetical protein

Supplementary Table 3 (Continuation)

Genomic Island 4	
Gene	**Product**
	hypothetical protein
	hypothetical protein
amyX_1	Pullulanase
	hypothetical protein
	Phosphorylated carbohydrates phosphatase
fbp_1	Fructose-1,6-bisphosphatase class 3
fbp_2	Fructose-1,6-bisphosphatase class 3
queG	Epoxyqueuosine reductase
prfB	Peptide chain release factor 2
ftsE	Cell division ATP-binding protein FtsE
ftsX	Cell division protein FtsX
	putative metallo-hydrolase

Genomic Island 5	
Gene	**Product**
asnS	Asparagine--tRNA ligase
	hypothetical protein
	glmZ(sRNA)-inactivating NTPase
	Putative gluconeogenesis factor
whiA	Putative sporulation transcription regulator WhiA
pepD	Dipeptidase
znuA_1	High-affinity zinc uptake system binding-proteinZnuA precursor
yvoA_1	HTH-type transcriptional repressor YvoA
agaS	Putative tagatose-6-phosphate ketose/aldose isomerase
rpmE2	50S ribosomal protein L31 type B
nrnA_1	putative bifunctional oligoribonuclease and PAP phosphatase NrnA
	putative acyltransferase
yghU	Disulfide-bond oxidoreductase YghU

add2	Aminodeoxyfutalosine deaminase
	Flavodoxin
	hypothetical protein
clcA_1	H(+)/Cl(-) exchange transporter ClcA
rplS	50S ribosomal protein L19

Pathogenicity Island 6

Gene	**Product**
addA	ATP-dependent helicase/nuclease subunit A

Resistance Island 5

Gene	**Product**
sigA	RNA polymerase sigma factor SigA
	hypothetical protein
rmlD	dTDP-4-dehydrorhamnose reductase
	hypothetical protein
wfgD	UDP-Glc:alpha-D-GlcNAc-diphosphoundecaprenol beta-1,3-glucosyltransferase WfgD
tagG	Teichoic acid translocation permease protein TagG
tagH	Teichoic acids export ATP-binding protein TagH
epsE	Putative glycosyltransferase EpsE
	Rhamnan synthesis protein F
	Rhamnan synthesis protein F
	Undecaprenyl-phosphate mannosyltransferase
	hypothetical protein
	hypothetical protein
mgtA	GDP-mannose-dependent alpha-mannosyltransferase
	Sulfatase
galE	UDP-glucose 4-epimerase

Supplementary Table 3 (Continuation)

Genomic Island 6	
Gene	**Product**
iscS_1	Cysteine desulfurase
thiI	putative tRNA sulfurtransferase
	Integrase core domain protein
	Transposase
capA	Capsule biosynthesis protein CapA

Pathogenicity Island 7	
Gene	**Product**
	hypothetical protein
rplU	50S ribosomal protein L21
	hypothetical protein
rpmA	50S ribosomal protein L27
oxyR	Hydrogen peroxide-inducible genes activator
lspA	Lipoprotein signal peptidase
rluD_1	Ribosomal large subunit pseudouridine synthase D
	SNARE associated Golgi protein
pyrR	Bifunctional protein PyrR
pyrP	Uracil permease
pyrB	Aspartate carbamoyltransferase
carA	Carbamoyl-phosphate synthase small chain
carB	Carbamoyl-phosphate synthase large chain
yknX	Putative efflux system component YknX
macB	Macrolide export ATP-binding/permease protein MacB
yknZ	putative ABC transporter permease YknZ
	hypothetical protein
	hypothetical protein
	Membrane domain of glycerophosphoryl diester phosphodiesterase
	cytoplasmic glycerophosphodiester phosphodiesterase
rpsP	30S ribosomal protein S16

	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
rimM	Ribosome maturation factor RimM
trmD	tRNA (guanine-N(1)-)-methyltransferase
yumC	Ferredoxin--NADP reductase 2
	hypothetical protein
panE	2-dehydropantoate 2-reductase
lacR_1	Lactose phosphotransferase system repressor
glcR	HTH-type transcriptional repressor GlcR

Pathogenicity Island 8

Gene	**Product**
	hypothetical protein
	DegV domain-containing protein
cca	CCA-adding enzyme
yjjK_1	putative ABC transporter ATP-binding protein YjjK
	putative ABC transporter ATP-binding protein
	putative ABC transporter ATP-binding protein

Supplementary Table 3 (Continuation)

Genomic Island 7	
Gene	**Product**
mvaA	3-hydroxy-3-methylglutaryl-coenzyme A reductase
pksG	Polyketide biosynthesis 3-hydroxy-3-methylglutaryl-ACP synthase PksG
thyA	Thymidylate synthase
dhfR	Dihydrofolate reductase
	hypothetical protein
clpX	ATP-dependent Clp protease ATP-binding subunit ClpX
engB_1	putative GTP-binding protein EngB
engB_2	putative GTP-binding protein EngB
	hypothetical protein
clpC_1	putative ATP-dependent Clp protease ATP-binding subunit
	hypothetical protein
rplJ	50S ribosomal protein L10
rplL	50S ribosomal protein L7/L12
	hypothetical protein
	hypothetical protein
	hypothetical protein
hpaIIM_1	Modification methylase HpaII
	hypothetical protein
	hypothetical protein
	hypothetical protein
	CAAX amino terminal protease self- immunity
	hypothetical protein
	Type IV secretory system Conjugative DNA transfer
	hypothetical protein
	hypothetical protein
	PrgI family protein
	hypothetical protein
ltrA_2	Group II intron-encoded protein LtrA
	AAA-like domain protein
	N-acetylmuramoyl-L-alanine amidase domain-containing protein precursor

	hypothetical protein
	hypothetical protein
	Lantibiotic streptin immunity protein
srrB	Sensor protein SrrB
regX3	Sensory transduction protein regX3
scnA_1	Lantibiotic streptococcin A-FF22 precursor
scnA_2	Lantibiotic streptococcin A-FF22 precursor
mutA	Lantibiotic mutacin-2 precursor
	Lanthionine synthetase C-like protein
lagD	Lactococcin-G-processing and transport ATP-binding protein LagD
	hypothetical protein
	Fluoroquinolones export ATP-binding protein/MT2762
	ABC-2 family transporter protein
	ABC-2 family transporter protein
immR_1	HTH-type transcriptional regulator ImmR
	hypothetical protein
Int-Tn_1	Transposase from transposon Tn916
dacA_3	D-alanyl-D-alanine carboxypeptidase DacA precursor
icaB	Poly-beta-1,6-N-acetyl-D-glucosamine N-deacetylase precursor
hom	Homoserine dehydrogenase
thrB	Homoserine kinase
fgs_3	Folylpolyglutamate synthase
fgs_4	Folylpolyglutamate synthase
folE	GTP cyclohydrolase 1
folP	Dihydropteroate synthase

Supplementary Table 3 (Continuation)

Genomic Island 8

Gene	**Product**
	pheromone autoinducer 2 transporter
	hypothetical protein
	gamma-glutamyl-gamma-aminobutyrate hydrolase
	Putative glutamine amidotransferase
rex	Redox-sensing transcriptional repressor Rex
	hypothetical protein
	hypothetical protein
iscS_2	Cysteine desulfurase

Pathogenicity Island 9

Gene	**Product**
	putative ABC transporter ATP-binding protein
	putative NADH oxidase
	hypothetical protein
ldhA	L-lactate dehydrogenase 1
gyrA	DNA gyrase subunit A
	Sortase family protein
	putative lyase
znuA_2	High-affinity zinc uptake system binding-proteinZnuA precursor
	hypothetical protein
	hypothetical protein
	Periplasmic solute binding protein family protein
	Transposase
	Integrase core domain protein
	hypothetical protein
femX_2	Lipid II:glycine glycyltransferase
	Transposase
	Integrase core domain protein

	hypothetical protein
nhaS3	High-affinity Na(+)/H(+) antiporter NhaS3
guaA	GMP synthase [glutamine-hydrolyzing]
	hypothetical protein
yvoA_2	HTH-type transcriptional repressor YvoA
	putative DNA-binding protein
ffh	Signal recognition particle protein

Pathogenicity Island 10

Gene	**Product**
hssR	Heme response regulator HssR
phoR_2	Alkaline phosphatase synthesis sensor protein PhoR
	Cupin domain protein
femA	Aminoacyltransferase FemA
xerS	Tyrosine recombinase XerS
	hypothetical protein
	1,4-dihydroxy-2-naphthoate octaprenyltransferase
apbE_2	Thiamine biosynthesis lipoprotein ApbE precursor
	FMN-binding domain protein
hepT_2	Heptaprenyl diphosphate synthase component 2
	NADH dehydrogenase-like protein
	Heptaprenyl diphosphate synthase component I
graS	Sensor histidine kinase GraS
graR	Response regulator protein GraR
yxdM	ABC transporter permease protein YxdM
yxdL_2	ABC transporter ATP-binding protein YxdL
prc	Tail-specific protease precursor
	hypothetical protein
citC	[Citrate [pro-3S]-lyase] ligase
	Methylmalonyl-CoA carboxyltransferase 5S subunit
citX	Apo-citrate lyase phosphoribosyl-dephospho-CoA transferase
citF	Citrate lyase alpha chain

citE	Citrate lyase subunit beta
citD	Citrate lyase acyl carrier protein
	hypothetical protein
gcdB_1	Glutaconyl-CoA decarboxylase subunit beta
cfiA	2-oxoglutarate carboxylase large subunit
	hypothetical protein
citN	Citrate transporter
ydfH	putative HTH-type transcriptional regulator YdfH
citG	2-(5"-triphosphoribosyl)-3'-dephosphocoenzyme-Asynthase
	Putative ammonia monooxygenase
gcdB_2	Glutaconyl-CoA decarboxylase subunit beta
	Methylmalonyl-CoA carboxyltransferase 1.3S subunit
	hypothetical protein
	Methylmalonyl-CoA carboxyltransferase 5S subunit

Supplementary Table 3 (Continuation)

Resistance Island 6	
Gene	**Product**
	Transposase
	Acetyltransferase (GNAT) family protein
pyrE	Orotate phosphoribosyltransferase
pyrF	Orotidine 5'-phosphate decarboxylase
	hypothetical protein
cysB	HTH-type transcriptional regulator CysB
	CRISPR-associated protein (Cas_Csm6)
deoD	Purine nucleoside phosphorylase DeoD-type
punA	Purine nucleoside phosphorylase 1
arsC	Arsenate reductase
deoB	Phosphopentomutase
rpiA	Ribose-5-phosphate isomerase A
mnmE	tRNA modification GTPase MnmE
	CAAX amino terminal protease self- immunity
pepV	Beta-Ala-Xaa dipeptidase
	Putative NAD(P)H nitroreductase
	thiamine pyrophosphate protein
	hypothetical protein
uvrC_1	UvrABC system protein C
uvrC_2	UvrABC system protein C
ybiV	Sugar phosphatase YbiV
	hypothetical protein
ybjI_2	Flavin mononucleotide phosphatase YbjI

Supplementary Table 3 (Continuation)

Resistance Island 7	
Gene	**Product**
	putative response regulatory protein
	putative sensor-like histidine kinase
manZ_2	Mannose permease IID component
agaC_2	N-acetylgalactosamine permease IIC component 1
	putative phosphotransferase enzyme IIB component
	PTS system fructose IIA component
	hypothetical protein
	hypothetical protein
	Enterocin A Immunity
	putative hydrolase
msrB	Peptide methionine sulfoxide reductase MsrB
lepA	Elongation factor 4
ndk	Nucleoside diphosphate kinase
	hypothetical protein
yutF	putative hydrolase YutF
	Acyl-ACP thioesterase
hemN	Oxygen-independent coproporphyrinogen-III oxidase 1
	hypothetical protein
glmM	Phosphoglucosamine mutase
	YbbR-like protein
disA	DNA integrity scanning protein DisA
	UDP-N-acetylmuramate--L-alanine ligase
	cobyric acid synthase
lplJ_2	Lipoate-protein ligase LplJ
	hypothetical protein
	Transcriptional regulator PadR-like family protein
	Dihydrolipoyl dehydrogenase
	hypothetical protein
pdhC	Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex
bfmBAB	2-oxoisovalerate dehydrogenase subunit beta

acoA	Acetoin:2,6-dichlorophenolindophenol oxidoreductase subunit alpha
yjjK_2	putative ABC transporter ATP-binding protein YjjK
	hypothetical protein
axe1-6A	Carbohydrate acetyl esterase/feruloyl esterase precursor
	Ribonuclease J 2
potA_2	Spermidine/putrescine import ATP-binding proteinPotA
lsrC_2	Autoinducer 2 import system permease protein LsrC
	ABC transporter substrate binding protein
	Transposase DDE domain protein
	hypothetical protein
	tetratricopeptide repeat protein
	pheromone autoinducer 2 transporter
mutX	8-oxo-dGTP diphosphatase
	hypothetical protein
	hypothetical protein
rmlB	dTDP-glucose 4,6-dehydratase
rfbC_1	putative dTDP-4-dehydrorhamnose 3,5-epimerase
rfbC_2	putative dTDP-4-dehydrorhamnose 3,5-epimerase
rmlA	Glucose-1-phosphate thymidylyltransferase
mlr	4-methylaminobutanoate oxidase (formaldehyde-forming)
	zinc transporter ZupT
zupT	Zinc transporter ZupT
	Putative GTP cyclohydrolase 1 type 2
trmK	tRNA (adenine(22)-N(1))-methyltransferase
pdg	Ultraviolet N-glycosylase/AP lyase
dnaD_1	DNA replication protein DnaD
apt	Adenine phosphoribosyltransferase
recJ	Single-stranded-DNA-specific exonuclease RecJ
	putative oxidoreductase
rnz	Ribonuclease Z
	galactose-1-phosphate uridylyltransferase
hflX	GTPase HflX
miaA	tRNA dimethylallyltransferase
	hypothetical protein

	C4-dicarboxylate transporter/malic acid transport protein
gst	Glutathione S-transferase GST-4.5
udk_1	Uridine kinase
	putative rhodanese-related sulfurtransferase
	hypothetical protein
azr_1	NADPH azoreductase
glgP	Glycogen phosphorylase
malQ	4-alpha-glucanotransferase
malR	HTH-type transcriptional regulator MalR
malX_1	Maltose/maltodextrin-binding protein precursor
malF_1	Maltose transport system permease protein MalF
ycjP	Inner membrane ABC transporter permease protein YcjP
exuR	putative HTH-type transcriptional repressor ExuR

Supplementary Table 3 (Continuation)

Pathogenicity Island 11/Resistance Island 8

Gene	**Product**
artM	Arginine transport ATP-binding protein ArtM
	hypothetical protein
	Transposase
	Integrase core domain protein
	hypothetical protein
obg	GTPase ObgE
	hypothetical protein
pepS	Aminopeptidase PepS
corA_1	Magnesium transport protein CorA
rsuA_1	Ribosomal small subunit pseudouridine synthase A
flK	Fluoroacetyl-CoA thioesterase
naiP	Putative niacin/nicotinamide transporter NaiP
	hypothetical protein
	hypothetical protein
ybbL	putative ABC transporter ATP-binding protein YbbL
paaI	Acyl-coenzyme A thioesterase PaaI

Genomic Island 9

Gene	**Product**
nrdH	Glutaredoxin-like protein NrdH
nrdE2	Ribonucleoside-diphosphate reductase subunit alpha 2
nrdF1	Ribonucleoside-diphosphate reductase subunit beta nrdF1
clcB	Voltage-gated ClC-type chloride channel ClcB
	CAAX amino terminal protease self- immunity
	hypothetical protein
puuR	HTH-type transcriptional regulator PuuR
alaS_1	Alanine--tRNA ligase
alaS_2	Alanine--tRNA ligase

Supplementary Table 3 (Continuation)

Island Pathogenicity Island 12	
Gene	**Product**
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	Phage protein Gp19/Gp15/Gp42
	hypothetical protein
	Phage capsid family protein
	Phage capsid family protein
	hypothetical protein
	Phage Terminase
	hypothetical protein
	hypothetical protein
	hypothetical protein
	Phage portal protein, SPP1 Gp6-like
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	YopX protein
	hypothetical protein
	hypothetical protein
	hypothetical protein

	PD-(D/E)XK nuclease superfamily protein
	hypothetical protein
	recombination and repair protein RecT
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
dnaD_2	DNA replication protein DnaD
	hypothetical protein
	hypothetical protein
	Helix-turn-helix domain protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	Phage antirepressor protein KilAC domain protein
	hypothetical protein
	Helix-turn-helix domain protein
	hypothetical protein
	hypothetical protein

Supplementary Table 3 (Continuation)

Resistance Island 9	
Gene	**Product**
	VanZ like family protein
rlmN	putative dual-specificity RNA methyltransferase RlmN
	hypothetical protein
	hypothetical protein
	Transposase DDE domain protein
	hypothetical protein
qorB	Quinone oxidoreductase 2

Genomic Island 10	
Gene	**Product**
	Peptidase propeptide and YPEB domain protein
	hypothetical protein
glyS	Glycine--tRNA ligase beta subunit
	hypothetical protein
glyQ	Glycine--tRNA ligase alpha subunit
	Transposase DDE domain protein

Genomic Island 11	
Gene	**Product**
	Phage portal protein, SPP1 Gp6-like
	Phage terminase large subunit
	Terminase small subunit
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein

	hypothetical protein
	hypothetical protein
	hypothetical protein

Supplementary Table 3 (Continuation)

Pathogenicity Island 13/Resistance Island 10	
Gene	**Product**
copY	Transcriptional repressor CopY
mlhB	Monoterpene epsilon-lactone hydrolase
rbfA	Ribosome-binding factor A
infB	Translation initiation factor IF-2
rplGA	putative ribosomal protein YlxQ
	hypothetical protein
	hypothetical protein
rimP	Ribosome maturation factor RimP
	Integrase core domain protein
	Transposase
	hypothetical protein
	DNA/RNA non-specific endonuclease
	Abi-like protein
	Bacteriophage peptidoglycan hydrolase
	Phage holin protein (Holin_LLH)
	hypothetical protein
	hypothetical protein
	hypothetical protein
	gp58-like protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	Collagen triple helix repeat (20 copies)

CHAPTER III

CONCLUDING REMARKS

Next-generation sequencing (NGS) technology today represents the main source of biological data; especially as new platforms are being continuously developed for faster results by lower costs. Genome sequencing has become a powerful tool within biology, however, there are still challenges when sequencing, assembling, and closing a genome, especially if it involves a genome with high content of repetitive sequences. In the case of the strains sequenced in this work, although different NGS technologies and different bioinformatic strategies and tools were used and combined, it was not possible to reach a complete genome. In addition, although the use of an optical map helped with the construction and orientation of scaffolds, this problem could be solved by designing primers flanking the gap regions for subsequent sequencing with the Sanger technology. Therefore, a complete genome sequence of *Streptococcus dysgalactiae* subsp. *dysgalactiae* is still expected for subsequent better results.

Comparative genomic offers a great potential in order to clarify, explain and predict certain behaviors in organisms, today it is an important tool that delimits and provides the opportunity to know or predict the genetic factors among organisms even before they are experimentally proved. It is important to emphasize that our study opens the door for future analyzes on this bacterium, especially in the field of virulence, since certain factors were established on this work a posterior characterization of all the proteins involved on the virulence factors would increase the knowledge of the pathogenic potential of this bacterium. In the same way, a protein characterization and analysis of the metabolic networks could lead to a posterior construction of vaccines.

References

Abdelsalam, M., Eissa, A. E., & Chen, S. C. (2015). Genetic diversity of geographically distinct Streptococcus dysgalactiae isolates from fish. *Journal of advanced research*, 6(2), 233-238.

Becquet, C., Blachon, S., Jeudy, B., Boulicaut, J.-F., & Gandrillon, O. (2002). Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. *Genome Biology*, 3(12).

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W., Mattick, J., & Haussler, D. (2004). Ultraconserved elements in the human genome. *Science*, 304: 1321–1325.

Bentley, S. (2009). Sequencing the species pan-genome. *Nature Reviews Microbiology,*, 7(4).

Breed, R. S., Murray, E. G., & Hitchens, A. P. (1948). *Determinative Bacteriology.*

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., & Gormley, N. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME journal*, 6(8), 1621.

Costa, F. A., Leal, C. A., Leite, R. C., & Figueiredo, H. C. (2014). Genotyping of Streptococcus dysgalactiae strains isolated from Nile tilapia, Oreochromis niloticus (L.). *J. Fish Dis.*, 37, 463–469.

Davies, M. R., McMillan, D. J., Van Domselaar, G. H., Jones, M. K., & Sriprakash, K. S. (2007). Phage 3396 from a *Streptococcus dysgalactiae* subsp. *equisimilis* pathovar may have its origins in Streptococcus pyogenes. Journal of bacteriology,. *Journal of bacteriology*, 189(7), 2646-2652.

Diernhofer, K. (1932). Aesculinbouillon als Hilfsmittel für die Differenzierung von Euter-und Milchstreptokokken bei Massenuntersuchungen. Milchwirtschaftl. *Forsch*, 13, 368-378.

Edwards, D. J., & Holt, K. E. (2013). Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial informatics and experimentation*, 3(1), 2.

Farrow, J. A., & Collins, M. D. (1984). Taxonomic studies on streptococci of serological groups C, G and L and possibly related taxa. *Syst. App. Microbiol.*, 483-493.

Fleischmann, R. D., Alland, D., Eisen, J. A., Carpenter, L., White, O. P., ..., & Hickey, E. (2002). Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains. *Journal of bacteriology*, 184(19), 5479-5490.

Frost, W. (1940). *The streptococci.* Willdorf Co.

Garvie, E. I., Farrow, J. A., & Collins, M. D. (1983). Streptococcus dysgalactiae (Diernhofer) nom. rev. *Int. J. Syst. Bacteriol*, 33:404-405.

Gaviria, J. M., & Bisno, A. L. (2000). Group C and G streptococci. Streptococcal infections: clinical aspects, microbiology and molecular pathogenesis. *Oxford University Press*, 238-254.

Gresham, D., Dunham, M. J., & Botstein, D. (2008). Comparing whole genomes using DNA microarrays. *Nature reviews. Genetics,* 9(4), 291.

Hu, B., Xie, G., Lo, C. C., Starkenburg, S. R., & Chain, P. S. (2011). Pathogen comparative genomics in the next-generation sequencing era: genome alignments, pangenomics and metagenomics. *Briefings in Functional Genomics,* 10(6), 322-333.

Hughes, J. M., Wilson, M. E., Brandt, C. M., & Spellerberg, B. (2009). Human infections due to Streptococcus dysgalactiae subspecies equisimilis. *Clinical Infectious Diseases*, 49(5), 766-772.

Illumina.com. (14 de July de 2017). *Illumina.com, Paired-End vs. Single-Read Sequencing Technology*. Obtenido de https://www.illumina.com/science/technology/next-generation-sequencing/paired-end-vs-single-read-sequencing.html

Indugu, N., Bittinger, K., Kumar, S., Vecchiarelli, B., & Pitta, D. (2016). A comparison of rumen microbial profiles in dairy cows as retrieved by 454 Roche and Ion Torrent (PGM) sequencing platforms. *PeerJ*, 4, e1599.

Jensen, A. &. (2012). Delineation of Streptococcus dysgalactiae, its subspecies, and its clinical and phylogenetic relationship to Streptococcus pyogenes. *Journal of Clinical Microbiology*, 50(1), 113–126.

Lancefield, R. C. (1933). A serological differentiation of human and other groups of hemolytic streptococci. *The Journal of Experimental Medicine*, 57 (4): 571–595.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, 30:434–439.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., & Feavers, I. M. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95(6), 3140-3145.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3), 133-141.

Metzker, M. L. (2010). Sequencing technologies--the next generation. *Nature reviews. Genetics*, 11(1), 31.

Netto, L. N., Leal, C. A., & Figueiredo, H. C. (2011). Streptococcus dysgalactiae as an agent of septicaemia in Nile tilapia, Oreochromis niloticus (L.). *Journal of fish diseases*, 34(3), 251-254.

Nomoto, R., Munasinghe, L. I., Jin, D. H., Shimahara, Y., Yasuda, H., Nakamura, A., & Yoshida, T. (2004). Lancefield group C Streptococcus dysgalactiae infection responsible for fish mortalities in Japan. *Journal of fish Diseases*, 27(12), 679-686.

Nomoto, R., Unose, N., Shimahara, Y., Nakamura, A., Hirae, T., Maebuchi, K., & Yoshida, T. (2006). Characterization of Lancefield group C Streptococcus dysgalactiae isolated from farmed fish. *Journal of fish Disease*, 29(11), 673-682.

Pereira, F. L., Soares, S. C., Dorella, F. A., Leal, C. A., & Figueiredo, H. C. (2016). Evaluating the efficacy of the new Ion PGM Hi-Q Sequencing Kit applied to bacterial genomes. *Genomics*, 107(5), 189-198.

Perkel, J. (2011). Making contact with sequencing's fourth generation. *BioTechniques*, 50(2), 93-95.

Rantala, S. (2014). *Streptococcus dysgalactiae* subsp. *equisimilis* bacteremia: an emerging infection. *European journal of clinical microbiology & infectious diseases*, 33(8), 1303-1310.

Rasko, D. A., Rosovitz, M. J., Myers, G. S., Mongodin, E. F., Fricke, W. F., Gajer, P., & Henderson, I. R. (2008). The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *Journal of bacteriology*, 190(20),6881-6893.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., & Davey, M. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*.

Skerman, V., & Sneath, P. (1980). Approved list of bacterial names. *Int. J. Syst. Bacteriol.*, 30: 225–420.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., & DeBoy, R. T. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, 102(39), 13950-13955.

Touchman, J. (2010). Comparative Genomics. *Nature Education Knowledge* , 3(10):13.

Vandamme, P., Pot, B., Falsen, E., Kersters, K., & Devriese, L. A. (1996). Taxonomic Study of Lancefield Streptococcal Groups C, G, and L (Streptococcus dysgalactiae) and Proposal of S. dysgalactiae subsp. equisimilis subsp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 46(3), 774-781.

Whist, A. C., Østerås, O., & Sølverød, L. (2007). Streptococcus dysgalactiae isolates at calving and lactation performance within the same lactation. *Journal of dairy science*, 90(2), 766-778.

Willenbrock, H., Hallin, P. F., Wassenaar, T. M., & Ussery, D. W. (2007). Characterization of probiotic Escherichia coli isolates with a novel pan-genome microarray. *Genome biology*, 8(12), R267.

Woese, C. R., Winker, S., & Gutell, R. R. (1990). Architecture of ribosomal RNA: constraints on the sequence of" tetra-loops. *Proceedings of the National Academy of Sciences*, 87(21), 8467-8471.

Yang, W., & Li, A. (2009). Isolation and characterization of Streptococcus dysgalactiae from diseased Acipenser schrenckii. Aquaculture, 294(1), 14-17.

Annexes:

a. Paper entitled "Use of MALDI-TOF Mass Spectrometry for the Fast Identification of Gram-Positive Fish Pathogens " published on *Frontiers in Microbiology* journal. (https://doi.org/10.3389/fmicb.2017.01492)

b. Abstract and banner presented during the X-Meeting 2016 - 12th International Conference of the Brazilian Association of Bioinformatics and Computational Biology, entitled "*Streptococcus dysgalatiae* subsp. *dysgalactiae* an emerging fish pathogen"

frontiers
in Microbiology

ORIGINAL RESEARCH
published: 09 August 2017
doi: 10.3389/fmicb.2017.01492

Use of MALDI-TOF Mass Spectrometry for the Fast Identification of Gram-Positive Fish Pathogens

Gabriella B. N. Assis, Felipe L. Pereira, Alexandra U. Zegarra, Guilherme C. Tavares, Carlos A. Leal and Henrique C. P. Figueiredo *

AQUACEN, National Reference Laboratory for Aquatic Animal Diseases, Ministry of Agriculture, Livestock and Food Supply, Federal University of Minas Gerais, Belo Horizonte, Brazil

Gram-positive cocci, such as *Streptococcus agalactiae*, *Lactococcus garvieae*, *Streptococcus iniae*, and *Streptococcus dysgalactiae* subsp. *dysgalactiae*, are found throughout the world, particularly in outbreaks in farmed fish, and are thus associated with high economic losses, especially in the cultivation of Nile Tilapia. The aim of this study was to evaluate the efficacy of matrix-assisted laser desorption ionization (MALDI)-time of flight (TOF) mass spectrometry (MS) as an alternative for the diagnosis of these pathogens. One hundred and thirty-one isolates from Brazilian outbreaks assisted by the national authority were identified using a MALDI Biotyper from Bruker Daltonics. The results showed an agreement with respect to identification (Kappa = 1) between this technique and 16S ribosomal RNA gene sequencing for *S. agalactiae* and *L. garvieae*. However, for *S. iniae* and *S. dysgalactiae* subsp. *dysgalactiae*, perfect agreement was only achieved after the creation of a custom main spectra profile, as well as further comparisons with 16S ribosomal RNA and multilocus sequence analysis. MALDI-TOF MS was shown to be an efficient technology for the identification of these Gram-positive pathogens, yielding a quick and precise diagnosis.

Keywords: MALDI-TOF MS, *S. agalactiae*, *S. iniae*, *S. dysgalactiae* subsp. *dysgalactiae*, *Lactococcus garvieae*

OPEN ACCESS

Edited by:
Peter R. Girguis,
Harvard University, United States

Reviewed by:
Jorge Reinheimer,
National University of the Littoral, Argentina
Akira Hishinuma,
Dokkyo Medical University, Japan

***Correspondence:**
Henrique C. P. Figueiredo
figueiredoh@yahoo.com

Specialty section:
This article was submitted to Aquatic Microbiology, a section of the journal Frontiers in Microbiology

Received: *17 May 2017*
Accepted: *25 July 2017*
Published: *09 August 2017*

Citation:
Assis GBN, Pereira FL, Zegarra AU, Tavares GC, Leal CA and Figueiredo HCP (2017) Use of MALDI-TOF Mass Spectrometry for the Fast Identification of Gram-Positive Fish Pathogens. Front. Microbiol. 8:1492. doi: 10.3389/fmicb.2017.01492

INTRODUCTION

Gram-positive cocci infections pose a great threat to farmed fish worldwide (Evans et al., 2002; Agnew and Barnes, 2007; Abdelsalam et al., 2013) and especially impact warm water systems used for the cultivation of Nile tilapia, one of the major commodities of global aquaculture (FAO, 2016). Four pathogens that are highly associated with outbreaks in fish farms are *Streptococcus agalactiae*, *Lactococcus garvieae*, *Streptococcus iniae*, and *S. dysgalactiae* subsp. *dysgalactiae* (SDD) (Evans et al., 2002; Agnew and Barnes, 2007; Mian et al., 2009; Netto et al., 2011; Figueiredo et al., 2012; Abdelsalam et al., 2013). *Streptococcus agalactiae*, *S. iniae,* and *L. garvieae* cause septicemia and meningoencephalitis in several species of marine and freshwater fish (Eldar et al., 1995; Evans et al., 2002; Mian et al., 2009; Figueiredo et al., 2012; Godoy et al., 2013; Soto et al., 2015; Fukushima et al., 2017). In fish, SDD infections are characterized by a systemic multifocal inflammatory reaction and a focal necrosis of the caudal peduncle, with moderate to high mortality rates during outbreaks (Nomoto et al., 2006).

Currently, the most widely used technology for the diagnosis of these infectious diseases is the isolation of the etiological agent in blood agar medium and subsequent identification through phenotypic/biochemical tests (Vendrell et al., 2006; Figueiredo et al., 2012; Assis et al., 2016). However, the performance of these tests can lead to misidentification or a lack of species-level resolution (Brigante et al., 2006; Tavares et al., 2016). Alternative molecular methods, such as species-specific PCR (Poyart et al., 1998) and the amplification and sequencing of the 16S ribosomal RNA (rRNA) gene, are useful for diagnosis (Kolbert and Persing, 1999; Patel, 2001; Clarridge, 2004) but are expensive and time consuming, mostly in trials with large number of clinical samples.

Recently, another technology to identify microorganisms was released: matrix-assisted laser desorption ionization (MALDI)-time of flight (TOF) mass spectrometry (MS) (Clark et al., 2013; Singhal et al., 2015). In this technique, the identification of the bacterial species is done by a comparison of peptide mass fingerprints to the device database. A typical mass range of 2–20 kDa is used, which represents mainly ribosomal proteins, along with a few housekeeping proteins (Singhal et al., 2015). There are many studies demonstrating the efficiency of MALDI-TOF MS in the classification of several species in a shorter time and with a lower cost (Bilecen et al., 2015), including typing (Nagy et al., 2011; Rizzardi et al., 2013) or identification of specific markers such as methicillin resistance (Østergaard et al., 2015; Ueda et al., 2016). Furthermore, MALDI-TOF MS can be performed in a short time for a wide range pathogens in one experiment (Bizzini and Greub, 2010). Additionally, it does not need a high level of staff training, reducing the risk of laboratory-associated infections by minimizing handling of living culture materials needed for the preparation of isolates.

Thus, the aim of this study was to evaluate the efficacy of MALDI-TOF MS for the identification of four Gram-positive cocci, *S. agalactiae*, *L. garvieae*, *S. iniae*, and SDD isolated from the kidneys, brains or abscesses of diseased fish from different geographic locations between 2003 and 2016.

MATERIALS AND METHODS

Bacterial Strains

Bacterial strains were selected from the culture collection of the National Reference Laboratory for Aquatic Animal Diseases (AQUACEN) of the Brazilian Ministry of Agriculture, Livestock and Food Supply. These *S. agalactiae* ($n = 50$), *L. garvieae* ($n = 11$), *S. iniae* ($n = 47$), and SDD ($n = 23$) strains were isolated during bacteriological analyses of outbreaks in Brazilian fish farms in different years and geographical locations (Table S1). The isolation of these microorganisms was performed on chilled fish that were sent to AQUACEN for diagnosis. Swabs from brains, kidneys or abscesses were aseptically sampled and streaked onto 5% sheep blood agar (SBA) for the isolation of bacterial pathogens. These plates were incubated at 28°C for 48 h. Finally, the identification of bacterial species was carried out as previously described (Mian et al., 2009; Netto et al., 2011; Figueiredo et al., 2012; Fukushima et al., 2017).

Species Confirmation through 16S rRNA Gene Sequencing

The isolates were thawed and streaked onto 5% SBA and were incubated at 28°C for 48 h. Isolates were incubated in a lysozyme solution at 37°C overnight. Bacterial DNA was extracted with a Maxwell 16 Tissue DNA purification kit (Promega, Madison, WI, USA) according to the manufacturer's instructions. The extracted DNA was quantified using a Nanodrop spectrophotometer (Thermo Scientific, Wilmington, DE, USA). The purity of the extracted DNA was determined using the absorbance ratio at 260/280 nm. Samples with ratio of 1.8 ± 0.5 were stored at $-80°$C until use.

The 16S rRNA gene was amplified by PCR with the universal primers B37 (5′-TAC GGY TAC CTT GTT ACG A-3′) and C70 (5′-AGA GTT TGA TYM TGGC-3′) and PCR amplicons were purified according to the method described by Fox et al. (1995) for all strains used in this work. The sequencing reactions were performed using a BigDye™ Terminator Cycle Sequencing Kit (Applied Biosystems, UK) and evaluated with an ABI 3,500 Genetic Analyzer (Life Technologies, USA). Forward and reverse sequencing products were used to generate contigs with the BioEdit software (Ibis Biosciences, Carlsbad, USA) version 7.2. Their identity was evaluated using the BLAST webserver (http://www.ncbi.nlm.nih.gov/BLAST) by checking against existing sequences in the nt/nr database. A similarity of $\geq$ 97% was considered as the same species in accordance with Nguyen et al. (2016) and Větrovský and Baldrian (2013).

MALDI-TOF MS Real-Time Identification Analysis

All isolates were thawed and streaked onto 5% SBA and incubated at 28°C for 48 h. A fresh, single colony of each bacterial strain was spotted using a toothpick into a target steel plate. For each strain, 1 μl of formic acid (70%) and 1 μl of MALDI-TOF MS matrix, consisting of a saturated solution of α-cyano-4-hydroxycinnamic acid (HCCA) (Bruker Daltonics, Bremen, Germany), were applied to the spot and allowed to air-dry. Spectra were acquired using the FlexControl MicroFlex LT mass spectrometer (Bruker Daltonics) with a 60-Hz nitrogen laser, in which up to 240 laser shots are fired in spiral movements to collect 40 shot steps for each strain spot. Furthermore, parameters for mass range detection were defined to allow the identification from 1,960 to 20,137 m/z, where Ion source 1 v was 19.99 kv, Ion source 2 voltage was 18.24 kv and the lens voltage was 6.0 kv for data acquisition. Prior to measurements, calibration was preceded with a bacterial test standard (*E. coli* DH5 alpha; Bruker Daltonics). The Real Time (RT) identification score criteria used were those recommended by the manufacturer: score $\geq$ 2.000 indicates a species-level identification, score $\geq$1.700 and $<$2.000 indicates a genus-level identification, and a score $<$1.700 indicates no reliable identification. Comparisons between MALDI-TOF MS strain identifications and those of other techniques were performed with R software version 3.0.1 (R Core Team, 2013) with the agreement rates determined by the Kappa coefficient.

Creation of a Custom Main Spectra Profile

To identify possible *S. iniae* strains and to enhance the *S. dysgalactiae* discrimination at the subspecies-level in a MALDI Biotyper, Main Spectra Profiles (MSPs) were created with reference strains for each species. Fresh colonies of the *S. iniae* SI23 strain and the SDD SD64, SD92 and SD142 strains were extracted according Alatoom et al. (2011). Briefly, the strains were collected from the agar and added to 300 μl of distilled water, followed by the addition of 900 μl of ethanol. Two rounds of centrifugation for 2 min at 13,000 rpm and the complete removal of supernatant was necessary to obtain dried pellets. The pellets were suspended in 50 μl of formic acid (70%) and vortexed. Finally, 50 μl of acetonitrile was added and the mixtures were centrifuged for 2 min at 13,000 rpm. For assays, one microliter of the supernatant was spotted eight times onto a steel target. Directly after air-drying, each spot was overlaid with 1 μl of HCCA matrix. Each spot was measured three times with the same protocol/parameters described in the section above. The obtained spectra were closely analyzed in the FlexAnalysis software (Bruker Daltonics) to assess the high level of reproducibility. Finally, the spectra of each strain were uploaded to the MALDI Biotyper software version 3 (Bruker Daltonics) and assembled to generate a Main Spectra Profile (MSP) for the strains using the BioTyper MSP creation standard method. All steps were done according to the manufacturer's recommendations.

A figure illustrating the SD64 spectra was generated using R software version 3.0.1 (R Core Team, 2013), using data exported from the FlexAnalysis software (Bruker Daltonics). In addition, in order to compare the custom MSPs with the MSP preloaded on the Bruker MSP library, the BioTyper software version 3.0 (Bruker Daltonics) was used to perform a dendrogram analysis. The parameters used were distance measure = "correlation," linkage = "average," maximum number of top level nodes = "0," score oriented dendrogram "enabled," score threshold values for a single organism = "300," and score threshold values for a related organism = "0."

Streptococcus dysgalactiae Subspecies Confirmation

The SDD strains that had subspecies suggested by Costa et al. (2014) were inferred by a BLAST comparison of the 16S rRNA and *sodA* genes, and the MALDI Biotyper (Bruker Daltonics) analysis suggested a closer relationship with *S. dysgalactiae* subsp. *equisimilis* (SDE). In addition to the 16S sequencing described above, a Next-Generation Sequence (NGS) experiment was performed. Three strains (SD64, SD92, and SD142) with different pulse-field gel electrophoresis (PFGE) profiles described in previous work from our group (Costa et al., 2014) were sequenced. DNA from the SDD strain was isolated from an overnight culture using a Maxwell 16 tissue DNA purification kit using the Maxwell 16 system (both from Promega). Sequencing was conducted on the Ion Torrent Personal Genome Machine sequencing system (Life Technologies) using a 200 bp fragment library kit, according to the manufacturer's recommendations. The barcodes of the raw data were removed using an in-house script (https://github.com/aquacen/fast_sample), and assembly was performed using SPAdes v3.9.1 (Nurk et al., 2013).

SDD taxonomic classification was determined using the Jensen and Kilian (2012) method, where the analysis of the phylogenetic relationship of seven housekeeping genes (*map*, *pfl*, *ppaC*, *pyk*, *rpoB*, *sodA*, and *tuf*) through a multilocus sequence analysis (MLSA) represent an improved basis for the identification of clinically important streptococci. The concatenated sequence of these housekeeping genes is used to establish differences between species that allow a more accurate identification within the pyogenic group of streptococci. The sequence of the draft genome of SDD ATCC 27957 is available on GenBank (Accession number: CM001076) and together with the genes of 30 streptococci strains submitted with the work of Jensen and Kilian (2012) were downloaded (Accession numbers: *map*: JN632385 to JN632479; *pfl*: JN632290 to JN632384; *ppaC*: JN632195 to JN632289; *pyk*: JN632100 to JN632194; *rpoB*: JN632005 to JN632099; *sodA*: JN631910 to JN632004; *tuf*: JN631815 to JN631909).

To extract the sequences of the corresponding housekeeping genes, a homology search for each of the seven genes in the SD64, SD92, and SD142 strains was performed using the BLAST webserver (http://www.ncbi.nlm.nih.gov/BLAST), with contigs generated by assembly software. The same strategy was performed with the SDD ATCC 27957 strain. All genes for each strain were concatenated in the following order: *map-pfI-ppaC-pyk-rpoB-sodA-tuf*. Alignment and phylogeny analyses were performed using MEGA6 (Tamura et al., 2013), with the Kimura-2 model parameters, using the Minimum Evolution algorithm, and a bootstrap of 1,000 replications.

RESULTS

Species Confirmation through 16S rRNA Gene Sequencing

The sequences of the 16S rRNA PCR products, which were generated with the aforementioned forward and reverse primers, were comprised in contigs for each strain. The mean lengths of the contigs were 1,514 ± 12, 1,537 ± 14, 1,519 ± 15, and 1,515 ± 17 bp for *S. agalactiae*, *L. garvieae*, *S. iniae*, and SDD, respectively. The contigs from each strain were used as queries for the BLAST webserver, and a percentage value of the similarities for *L. garvieae* was between 98 and 100, whereas *S. agalactiae*, *S. iniae* and SDD varied between 97 and 100. For the SDD strains, it was not possible make identification at the subspecies-level. For each SDD isolate there were results referring to the SDE and SDD with the same percentage value of identity that referred to the same query coverage.

MALDI-TOF MS RT Identification of *S. agalactiae* and *L. garvieae*

For each strain-spot, 1–3 spectra were expected, according to the manufacturer's instructions for quality assurance performed by MALDI Biotyper software of acquisition. For *S. agalactiae*, 64 spectra were acquired, whereas 11 spectra were acquired for *L. garvieae*. All strains for both species were identified at the species-level (score $\geq$ 2.000). The minimal and maximal scores for *S. agalactiae* were 2.083 and 2.377 (**Table 1**), respectively, and for *L. garvieae* were 2.081 and 2.218 (**Table 2**), respectively. For

TABLE 1 | *Streptococcus agalactiae* strains identification by 16S rRNA sequencing and MALDI-TOF MS.

Strain	16S rRNA sequencing		MALDI Biotyper	
	Species	**% Identity**	**Organism best match**	**Score value**
SA001	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.330
SA005	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.318
SA007	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.371
SA009	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.302
SA016	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.357
SA020	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.296
SA030	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.289
SA033	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.211
SA053	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.206
SA073	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.259
SA075	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.189
SA079	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.251
SA081	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.327
SA085	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.207
SA095	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.275
SA097	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.227
SA102	*Streptococcus agalactiae*	99	*Streptococcus agalactiae*	2.172
SA117	*Streptococcus agalactiae*	97	*Streptococcus agalactiae*	2.162
SA132	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.322
SA136	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.220
SA159	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.364
SA172	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.339
SA184	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.306
SA191	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.207
SA201	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.221
SA209	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.309
SA212	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.377
SA218	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.331
SA220	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.351
SA245	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.192
SA256	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.083
SA289	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.167
SA330	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.317
SA333	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.294
SA341	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.296
SA343	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.276
SA346	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.254
SA374	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.363
SA375	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.360
SA623	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.248
SA627	*Streptococcus agalactiae*	100	*Streptococcus agalactiae*	2.349
SA665	*Streptococcus agalactiae*	97	*Streptococcus agalactiae*	2.281
SA719	*Streptococcus agalactiae*	98	*Streptococcus agalactiae*	2.197
SA796	*Streptococcus agalactiae*	97	*Streptococcus agalactiae*	2.359
SA808	*Streptococcus agalactiae*	97	*Streptococcus agalactiae*	2.242
SA887	*Streptococcus agalactiae*	97	*Streptococcus agalactiae*	2.185
SA929	*Streptococcus agalactiae*	99	*Streptococcus agalactiae*	2.230
SA941	*Streptococcus agalactiae*	97	*Streptococcus agalactiae*	2.257
SA959	*Streptococcus agalactiae*	97	*Streptococcus agalactiae*	2.328
SA972	*Streptococcus agalactiae*	97	*Streptococcus agalactiae*	2.183

TABLE 2 | *Lactococcus garvieae* strains identification by 16S rRNA sequencing and MALDI-TOF MS.

Strain	16S rRNA sequencing		MALDI Biotyper	
	Species	**% Identity**	**Organism best match**	**Score value**
LG002	*Lactococcus garvieae*	100	*Lactococcus garvieae*	2.166
LG005	*Lactococcus garvieae*	99	*Lactococcus garvieae*	2.195
LG009	*Lactococcus garvieae*	98	*Lactococcus garvieae*	2.084
LG010	*Lactococcus garvieae*	98	*Lactococcus garvieae*	2.218
LG011	*Lactococcus garvieae*	100	*Lactococcus garvieae*	2.213
LG015	*Lactococcus garvieae*	100	*Lactococcus garvieae*	2.142
LG018	*Lactococcus garvieae*	99	*Lactococcus garvieae*	2.110
LG019	*Lactococcus garvieae*	98	*Lactococcus garvieae*	2.114
LG020	*Lactococcus garvieae*	100	*Lactococcus garvieae*	2.184
LG021	*Lactococcus garvieae*	99	*Lactococcus garvieae*	2.165
LG022	*Lactococcus garvieae*	98	*Lactococcus garvieae*	2.081

both species a perfect agreement (Kappa = 1; CI: 1.0–1.0; and $p < 0.005$) was observed between the 16S rRNA gene sequencing and MALDI-TOF MS techniques to identify the species.

MALDI-TOF MS RT Identification of *S. iniae*

A total of 52 spectra were obtained for the 47 strains. Identification of *S. iniae* was possible in ∼53% of isolates at the genus-level (**Table 3**), and the minimal and maximal scores were 1.482 and 1.854, respectively, including 22 with no reliable identification. The genus-level was inferred by an approximation of the spectra with *S. dysgalactiae* ($n = 7$), *S. equi* ($n = 1$), and *S. pyogenes* ($n = 17$). The species identification agreement when comparing 16S rRNA gene sequencing and MALDI-TOF MS was poor (Kappa = 0.04; CI: −0.03 to 0.11; and $p = 0.063$).

To make possible the correct identification of *S. iniae* strains using the MALDI Biotyper, a custom MSP was created for this species (**Figure 1**; MSP available at http://www.renaqua.gov.br/aquacen-msp-si/). Twenty-four spectra were collected for one isolate (SI23) by the Biotyper RTC program. The spectra were analyzed in the FlexAnalysis software to identify a high level of reproducibility, and all spectra were used to create the MSP. A dendrogram generated in BioTyper software (**Figure 2**) shows the SI23 strain as a single leaf between the *S. pyogenes* and *S. dysgalactiae* clades. After the inclusion of the custom MSP of *S. iniae*, all the strains were identified at the species-level (**Table 3**), and the minimal and maximal score values were 2.013 and 2.426, respectively. A complete agreement between both tested techniques was observed (Kappa = 1; CI: 1.0–1.0; and $p < 0.005$) for species identification.

MALDI-TOF MS RT Identification of SDD

The identification of SDD isolates, using 25 spectra from 23 strains, was obtained by an approximation of *S. dysgalactiae* and SDE MSPs at the species-level. Minimal and maximal scores were 2.058 and 2.298, respectively. Of all the SDD strains, 13 were identified with proximity to the subspecies *equisimilis*, and in 10 strains, there was no discrimination of subspecies

TABLE 3 | *Streptococcus iniae* strains identification by 16S rRNA sequencing and MALDI-TOF MS (before and after custom MSP inclusion).

Strain	16S rRNA sequencing		MALDI Biotyper			
			Before custom MSP inclusion		After custom MSP inclusion	
	Species	% Identity	Organism best match	Score value	Organism best match	Score value
SI022	*Streptococcus iniae*	98	*Streptococcus pyogenes*	1.736[a]	*S. iniae* SI23	2.223
SI023	*Streptococcus iniae*	99	Not reliable identification	1.509	*S. iniae* SI23	2.089
SI024	*Streptococcus iniae*	99	*Streptococcus dysgalactiae*	1.741[a]	*S. iniae* SI23	2.165
SI025	*Streptococcus iniae*	100	Not reliable identification	1.580	*S. iniae* SI23	2.205
SI027	*Streptococcus iniae*	100	Not reliable identification	1.642	*S. iniae* SI23	2.148
SI028	*Streptococcus iniae*	100	Not reliable identification	1.683	*S. iniae* SI23	2.013
SI029	*Streptococcus iniae*	97	*Streptococcus pyogenes*	1.724[a]	*S. iniae* SI23	2.199
SI444	*Streptococcus iniae*	99	Not reliable identification	1.627	*S. iniae* SI23	2.031
SI503	*Streptococcus iniae*	97	*Streptococcus dysgalactiae*	1.737[a]	*S. iniae* SI23	2.301
SI674	*Streptococcus iniae*	98	Not reliable identification	1.664	*S. iniae* SI23	2.205
SI677	*Streptococcus iniae*	99	*Streptococcus pyogenes*	1.732[a]	*S. iniae* SI23	2.409
SI692	*Streptococcus iniae*	99	*Streptococcus equi*	1.700[a]	*S. iniae* SI23	2.332
SI696	*Streptococcus iniae*	99	Not reliable identification	1.620	*S. iniae* SI23	2.308
SI698	*Streptococcus iniae*	97	Not reliable identification	1.679	*S. iniae* SI23	2.426
SI699	*Streptococcus iniae*	98	*Streptococcus dysgalactiae*	1.821[a]	*S. iniae* SI23	2.273
SI700	*Streptococcus iniae*	98	Not reliable identification	1.605	*S. iniae* SI23	2.147
SI701	*Streptococcus iniae*	97	Not reliable identification	1.629	*S. iniae* SI23	2.272
SI702	*Streptococcus iniae*	97	Not reliable identification	1.675	*S. iniae* SI23	2.281
SI705	*Streptococcus iniae*	98	Not reliable identification	1.678	*S. iniae* SI23	2.326
SI706	*Streptococcus iniae*	99	*Streptococcus pyogenes*	1.750[a]	*S. iniae* SI23	2.122
SI711	*Streptococcus iniae*	99	*Streptococcus pyogenes*	1.733[a]	*S. iniae* SI23	2.231
SI712	*Streptococcus iniae*	97	*Streptococcus pyogenes*	1.749[a]	*S. iniae* SI23	2.124
SI713	*Streptococcus iniae*	98	*Streptococcus pyogenes*	1.713[a]	*S. iniae* SI23	2.275
SI714	*Streptococcus iniae*	99	Not reliable identification	1.641	*S. iniae* SI23	2.075
SI715	*Streptococcus iniae*	98	*Streptococcus pyogenes*	1.748[a]	*S. iniae* SI23	2.216
SI717	*Streptococcus iniae*	97	Not reliable identification	1.648	*S. iniae* SI23	2.261
SI718	*Streptococcus iniae*	98	*Streptococcus pyogenes*	1.825[a]	*S. iniae* SI23	2.249
SI720	*Streptococcus iniae*	98	*Streptococcus pyogenes*	1.829[a]	*S. iniae* SI23	2.321
SI790	*Streptococcus iniae*	97	*Streptococcus pyogenes*	1.774[a]	*S. iniae* SI23	2.204
SI791	*Streptococcus iniae*	97	*Streptococcus dysgalactiae*	1.781[a]	*S. iniae* SI23	2.255
SI792	*Streptococcus iniae*	99	Not reliable identification	1.556	*S. iniae* SI23	2.054
SI797	*Streptococcus iniae*	98	*Streptococcus pyogenes*	1.854[a]	*S. iniae* SI23	2.043
SI798	*Streptococcus iniae*	97	Not reliable identification	1.675	*S. iniae* SI23	2.293
SI819	*Streptococcus iniae*	97	*Streptococcus pyogenes*	1.787[a]	*S. iniae* SI23	2.203
SI826	*Streptococcus iniae*	98	*Streptococcus pyogenes*	1.809[a]	*S. iniae* SI23	2.244
SI831	*Streptococcus iniae*	98	Not reliable identification	1.557	*S. iniae* SI23	2.313
SI839	*Streptococcus iniae*	97	*Streptococcus dysgalactiae*	1.738[a]	*S. iniae* SI23	2.173
SI841	*Streptococcus iniae*	99	Not reliable identification	1.686	*S. iniae* SI23	2.242
SI842	*Streptococcus iniae*	97	Not reliable identification	1.614	*S. iniae* SI23	2.379
SI852	*Streptococcus iniae*	97	*Streptococcus dysgalactiae*	1.707[a]	*S. iniae* SI23	2.071
SI870	*Streptococcus iniae*	97	Not reliable identification	1.482	*S. iniae* SI23	2.228
SI875	*Streptococcus iniae*	99	Not reliable identification	1.668	*S. iniae* SI23	2.182
SI876	*Streptococcus iniae*	99	*Streptococcus dysgalactiae*	1.751[a]	*S. iniae* SI23	2.238
SI913	*Streptococcus iniae*	98	Not reliable identification	1.625	*S. iniae* SI23	2.276
SI928	*Streptococcus iniae*	99	*Streptococcus pyogenes*	1.819[a]	*S. iniae* SI23	2.031
SI954	*Streptococcus iniae*	99	*Streptococcus pyogenes*	1.802[a]	*S. iniae* SI23	2.214
SI970	*Streptococcus iniae*	99	*Streptococcus pyogenes*	1.853[a]	*S. iniae* SI23	2.241

[a] *Genus-level identification.*

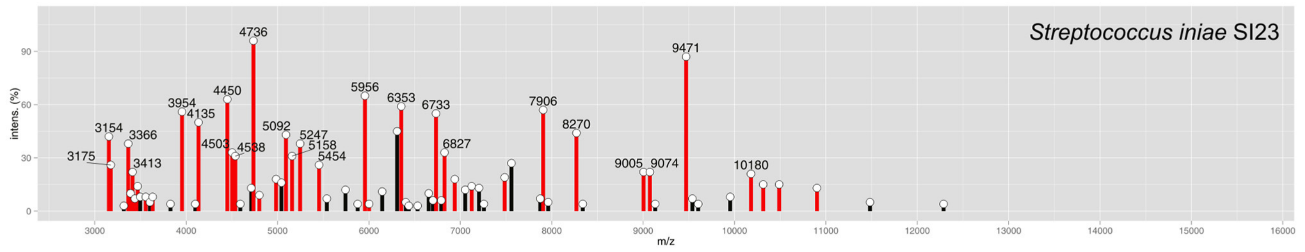


FIGURE 1 | MSP of *S. iniae* SI23 peak identification. Peaks with intensities greater than 20% are labeled. Peaks with a previously identified m/z (Kim et al., 2017) are shown in red bars.

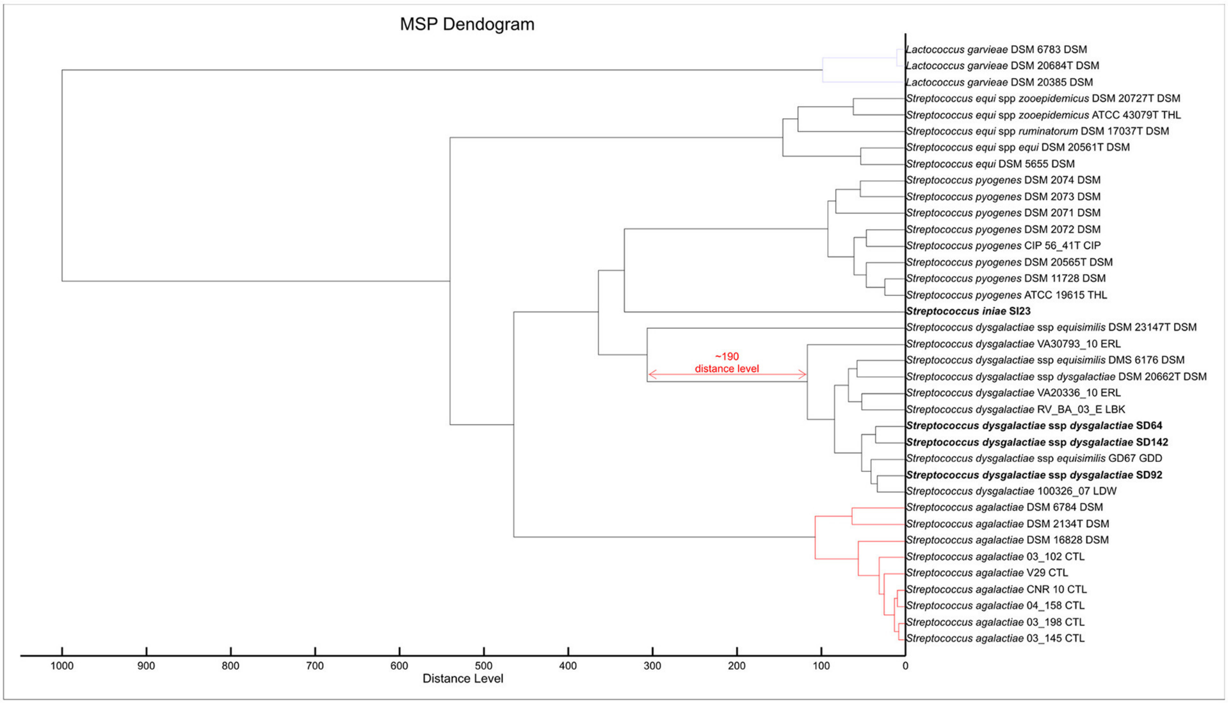

MSP Dendogram
Lactococcus garvieae DSM 6783 DSM
Lactococcus garvieae DSM 20684T DSM
Lactococcus garvieae DSM 20385 DSM
Streptococcus equi spp *zooepidemicus* DSM 20727T DSM
Streptococcus equi spp *zooepidemicus* ATCC 43079T THL
Streptococcus equi spp *ruminatorum* DSM 17037T DSM
Streptococcus equi spp *equi* DSM 20561T DSM
Streptococcus equi DSM 5655 DSM
Streptococcus pyogenes DSM 2074 DSM
Streptococcus pyogenes DSM 2073 DSM
Streptococcus pyogenes DSM 2071 DSM
Streptococcus pyogenes DSM 2072 DSM
Streptococcus pyogenes CIP 56_41T CIP
Streptococcus pyogenes DSM 20565T DSM
Streptococcus pyogenes DSM 11728 DSM
Streptococcus pyogenes ATCC 19615 THL
Streptococcus iniae SI23
Streptococcus dysgalactiae ssp *equisimilis* DSM 23147T DSM
Streptococcus dysgalactiae VA30793_10 ERL
Streptococcus dysgalactiae ssp *equisimilis* DMS 6176 DSM
Streptococcus dysgalactiae ssp *dysgalactiae* DSM 20662T DSM
Streptococcus dysgalactiae VA20336_10 ERL
Streptococcus dysgalactiae RV_BA_03_E LBK
Streptococcus dysgalactiae ssp dysgalactiae SD64
Streptococcus dysgalactiae ssp dysgalactiae SD142
Streptococcus dysgalactiae ssp *equisimilis* GD67 GDD
Streptococcus dysgalactiae ssp dysgalactiae SD92
Streptococcus dysgalactiae 100326_07 LDW
Streptococcus agalactiae DSM 6784 DSM
Streptococcus agalactiae DSM 2134T DSM
Streptococcus agalactiae DSM 16828 DSM
Streptococcus agalactiae 03_102 CTL
Streptococcus agalactiae V29 CTL
Streptococcus agalactiae CNR 10 CTL
Streptococcus agalactiae 04_158 CTL
Streptococcus agalactiae 03_198 CTL
Streptococcus agalactiae 03_145 CTL
~190
distance level
1000
900
800
700
600
500
400
300
200
100
0
Distance Level

FIGURE 2 | MSP Dendrogram analysis of the custom MSP and Bruker MSP library of *Lactococcus garvieae* and *Streptococcus* spp. Isolates from this work are in bold. *S. iniae* SI23 is alone and between the *S. pyogenes* and *S. dysgalactiae* spp. clades. SDD is in an intra-species-specific clade of *S. dysgalactiae* spp. strains. The red arrow shows a distance level of ~190 from *S. dysgalactiae* spp. *equisimilis* DSM 23147^T to other *S. dysgalactiae* strains.

(**Table 4**). The agreement between techniques was perfect when considering the species-level (Kappa = 1; CI: 1.0–1.0; and $p < 0.004$), but when considering the subspecies-level the agreement was only fair (Kappa = 0.21; CI: −0.08−0.52; $p = 0.075$). This demonstrated that both techniques were unable to identify strains at the subspecies-level.

These strains, according to previous work of our group (Costa et al., 2014), are from SDD subspecies. Therefore, an NGS experiment was done to confirm the subspecies assignments. Contigs from the assembly of the strains SD64, SD92, and SD142 (data not shown) were used for a MLSA analysis. The three strains formed a clade with SDD from work of Jensen and Kilian (2012), confirming the classification of theses strains as SDD subspecies in accordance with the methodology used (**Figure 3**).

To improve the identification by the MALDI Biotyper, custom MSPs were created for SDD (**Figure 4**; MSP available at http://www.renaqua.gov.br/aquacen-msp-sdd/). Twenty-four spectra were collected for each isolate as described above and the spectra were analyzed in the FlexAnalysis, where all spectra were used to create the MSP. A dendrogram generated in BioTyper software (**Figure 2**) shows the SD64, SD92, and SD142 strains in an intra-species-specific clade of *S. dysgalactiae* spp. **Figure 4** shows the common and exclusive peaks of custom MSPs and Bruker library MSPs, and, interestingly, the SDE DSM 23147^T shows 25 exclusive peaks.

After this inclusion (**Table 4**), all isolates matched with to the three included custom MSP for the three best matches (Table S2), with minimal and maximal scores of 2.277 and 2.579, respectively. The agreement between 16S rRNA gene sequencing and MALDI-TOF MS was poor (Kappa = 0.08, CI: −0.05−0.22; $p = 0.050$), considering that 16S rRNA gene sequencing was unable to identify subspecies, whereas with MALDI-TOF MS, they could

TABLE 4 | SDD strains identification by 16S rRNA sequencing and MALDI-TOF MS (before and after custom MSP inclusion).

Strain	16S rRNA sequencing		MALDI Biotyper			
			Before custom MSP inclusion		After custom MSP inclusion	
	Species	% Identity	Organism best match	Score value	Organism best match	Score value
SD054	*Streptococcus dysgalactiae*	97	*Streptococcus dysgalactiae* ssp. *equisimilis*	2.116	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD64	2.497
SD056	*Streptococcus dysgalactiae*	97	*Streptococcus dysgalactiae* ssp. *equisimilis*	2.298	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD64	2.480
SD061	*Streptococcus dysgalactiae*	98	*Streptococcus dysgalactiae* ssp. *equisimilis*	2.251	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD64	2.438
SD064	*Streptococcus dysgalactiae*	100	*Streptococcus dysgalactiae* ssp. *equisimilis*	2.161	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD64	2.458
SD068	*Streptococcus dysgalactiae*	98	*Streptococcus dysgalactiae* ssp. *equisimilis*	2.130	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD142	2.497
SD092	*Streptococcus dysgalactiae*	100	*Streptococcus dysgalactiae* ssp. *equisimilis*	2.058	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD64	2.320
SD120	*Streptococcus dysgalactiae*	97	*Streptococcus dysgalactiae* ssp. *equisimilis*	2.151	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD64	2.346
SD137	*Streptococcus dysgalactiae*	98	*Streptococcus dysgalactiae*	2.122	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD142	2.338
SD140	*Streptococcus dysgalactiae*	97	*Streptococcus dysgalactiae* ssp. *equisimilis*	2.130	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD142	2.531
SD142	*Streptococcus dysgalactiae*	99	*Streptococcus dysgalactiae*	2.078	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD142	2.384
SD143	*Streptococcus dysgalactiae*	98	*Streptococcus dysgalactiae*	2.164	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD64	2.479
SD145	*Streptococcus dysgalactiae*	97	*Streptococcus dysgalactiae* ssp. *equisimilis*	2.175	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD142	2.548
SD280	*Streptococcus dysgalactiae*	97	*Streptococcus dysgalactiae*	2.165	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD142	2.471
SD281	*Streptococcus dysgalactiae*	97	*Streptococcus dysgalactiae*	2.192	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD64	2.277
SD282	*Streptococcus dysgalactiae*	98	*Streptococcus dysgalactiae*	2.071	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD142	2.511
SD283	*Streptococcus dysgalactiae*	97	*Streptococcus dysgalactiae* ssp. *equisimilis*	2.201	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD142	2.513
SD284	*Streptococcus dysgalactiae*	98	*Streptococcus dysgalactiae*	2.195	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD142	2.302
SD285	*Streptococcus dysgalactiae*	97	*Streptococcus dysgalactiae*	2.073	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD142	2.523
SD286	*Streptococcus dysgalactiae*	97	*Streptococcus dysgalactiae* ssp. *equisimilis*	2.180	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD142	2.461
SD287	*Streptococcus dysgalactiae*	99	*Streptococcus dysgalactiae* ssp. *equisimilis*	2.168	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD142	2.579
SD367	*Streptococcus dysgalactiae*	97	*Streptococcus dysgalactiae*	2.177	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD142	2.432
SD370	*Streptococcus dysgalactiae*	99	*Streptococcus dysgalactiae*	2.171	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD142	2.565
SD372	*Streptococcus dysgalactiae*	98	*Streptococcus dysgalactiae* ssp. *equisimilis*	2.143	*Streptococcus dysgalactiae* subsp. *dysgalactiae* SD64	2.366

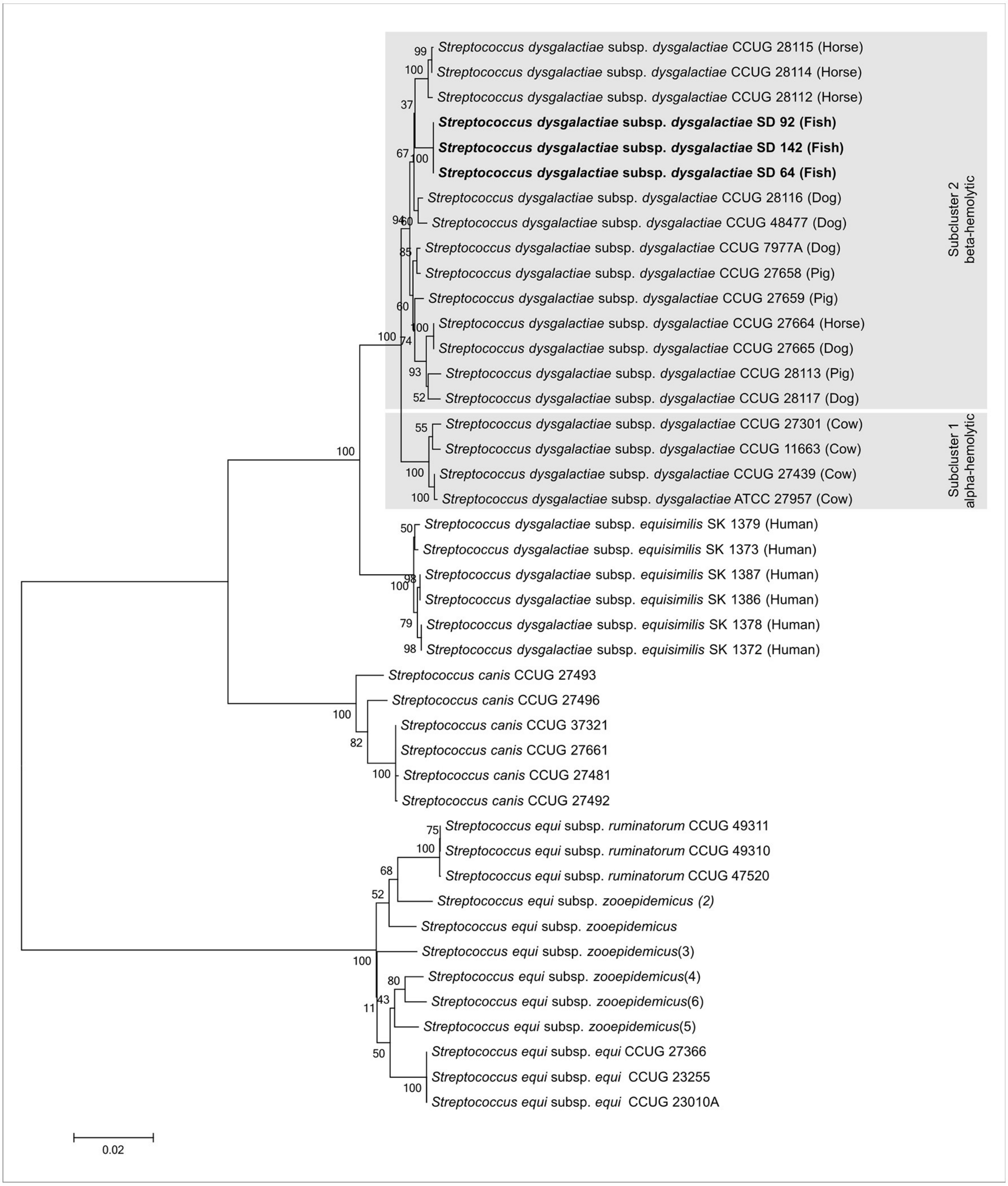

Streptococcus dysgalactiae subsp. dysgalactiae CCUG 28115 (Horse)
Streptococcus dysgalactiae subsp. dysgalactiae CCUG 28114 (Horse)
Streptococcus dysgalactiae subsp. dysgalactiae CCUG 28112 (Horse)
Streptococcus dysgalactiae subsp. dysgalactiae SD 92 (Fish)
Streptococcus dysgalactiae subsp. dysgalactiae SD 142 (Fish)
Streptococcus dysgalactiae subsp. dysgalactiae SD 64 (Fish)
Streptococcus dysgalactiae subsp. dysgalactiae CCUG 28116 (Dog)
Streptococcus dysgalactiae subsp. dysgalactiae CCUG 48477 (Dog)
Streptococcus dysgalactiae subsp. dysgalactiae CCUG 7977A (Dog)
Streptococcus dysgalactiae subsp. dysgalactiae CCUG 27658 (Pig)
Streptococcus dysgalactiae subsp. dysgalactiae CCUG 27659 (Pig)
Streptococcus dysgalactiae subsp. dysgalactiae CCUG 27664 (Horse)
Streptococcus dysgalactiae subsp. dysgalactiae CCUG 27665 (Dog)
Streptococcus dysgalactiae subsp. dysgalactiae CCUG 28113 (Pig)
Streptococcus dysgalactiae subsp. dysgalactiae CCUG 28117 (Dog)
Streptococcus dysgalactiae subsp. dysgalactiae CCUG 27301 (Cow)
Streptococcus dysgalactiae subsp. dysgalactiae CCUG 11663 (Cow)
Streptococcus dysgalactiae subsp. dysgalactiae CCUG 27439 (Cow)
Streptococcus dysgalactiae subsp. dysgalactiae ATCC 27957 (Cow)
Subcluster 2 beta-hemolytic
Subcluster 1 alpha-hemolytic
Streptococcus dysgalactiae subsp. equisimilis SK 1379 (Human)
Streptococcus dysgalactiae subsp. equisimilis SK 1373 (Human)
Streptococcus dysgalactiae subsp. equisimilis SK 1387 (Human)
Streptococcus dysgalactiae subsp. equisimilis SK 1386 (Human)
Streptococcus dysgalactiae subsp. equisimilis SK 1378 (Human)
Streptococcus dysgalactiae subsp. equisimilis SK 1372 (Human)
Streptococcus canis CCUG 27493
Streptococcus canis CCUG 27496
Streptococcus canis CCUG 37321
Streptococcus canis CCUG 27661
Streptococcus canis CCUG 27481
Streptococcus canis CCUG 27492
Streptococcus equi subsp. ruminatorum CCUG 49311
Streptococcus equi subsp. ruminatorum CCUG 49310
Streptococcus equi subsp. ruminatorum CCUG 47520
Streptococcus equi subsp. zooepidemicus (2)
Streptococcus equi subsp. zooepidemicus
Streptococcus equi subsp. zooepidemicus(3)
Streptococcus equi subsp. zooepidemicus(4)
Streptococcus equi subsp. zooepidemicus(6)
Streptococcus equi subsp. zooepidemicus(5)
Streptococcus equi subsp. equi CCUG 27366
Streptococcus equi subsp. equi CCUG 23255
Streptococcus equi subsp. equi CCUG 23010A
0.02

FIGURE 3 | Tree taxonomy analysis of SDD SD64, SD92, and SD142 strains. The strains from this work (bold) form a specific clade with other SDD strains from Jensen and Kilian (2012). Gray hatched areas are the subcluster of alpha- and beta-hemolytic strains proposed by Jensen and Kilian (2012).

be determined effectively. In contrast, considering the MLSA analysis, the agreement between this technique and MALDI-TOF MS was perfect (Kappa = 1; CI: 1.0–1.0; and $p < 0.005$).

DISCUSSION

Gram-positive cocci have been associated with acute and chronic fish diseases. They have become an increasingly important problem in the aquaculture industry in many countries (Evans et al., 2002; Vendrell et al., 2006; Agnew and Barnes, 2007; Mian et al., 2009; Netto et al., 2011; Figueiredo et al., 2012; Abdelsalam et al., 2013; Costa et al., 2014). An barrier to the better utilization of fish produced are the infectious diseases, including the control of the potential zoonotic infections caused by *S. iniae* (Keirstead et al., 2014). Thus, accelerating the diagnosis of diseases remains a big challenge. An alternative

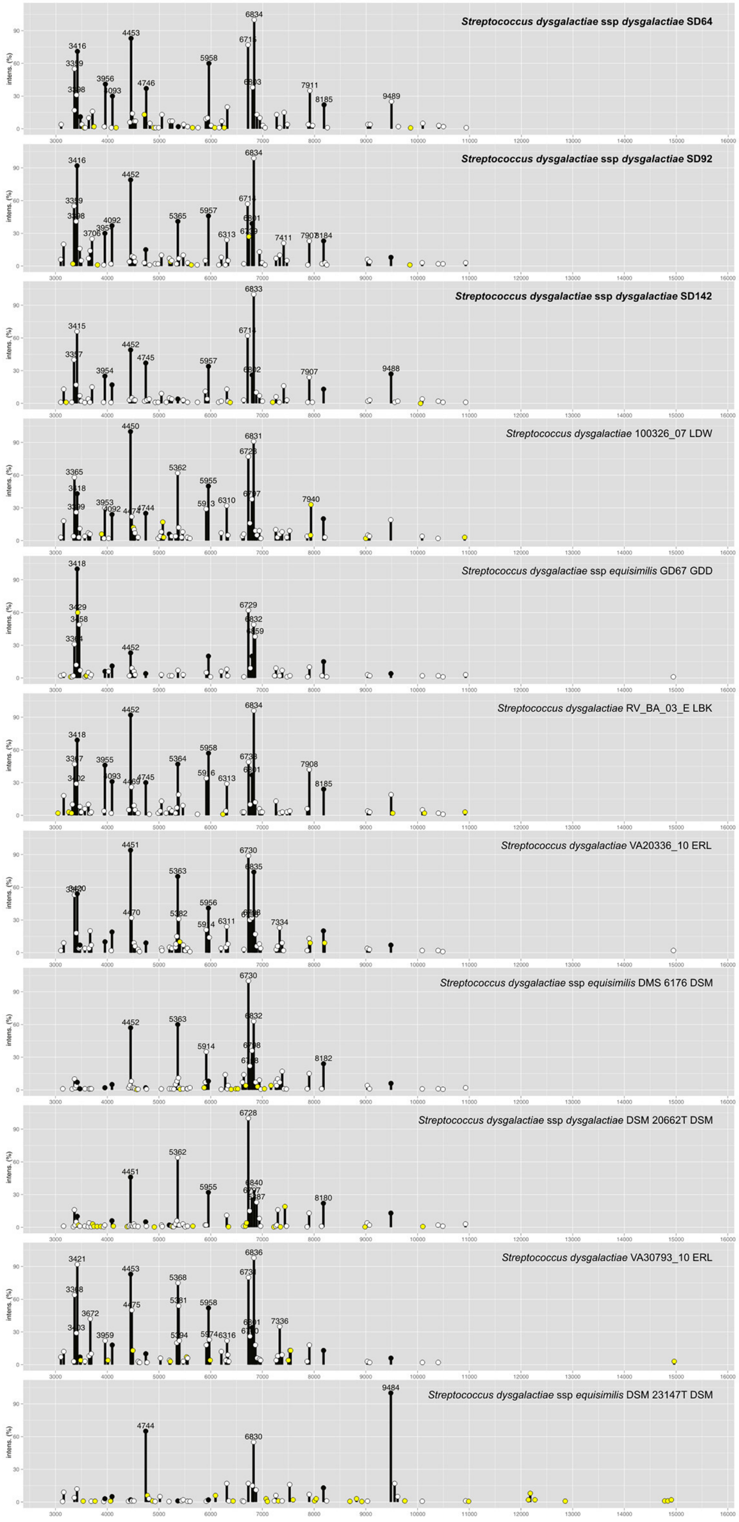


FIGURE 4 | Main Spectra Profiles of *S. dysgalactiae* subsp. *dysgalactiae* SD64, SD92, and SD142 peaks identification. The strains of this work (bold) together with the *S. dysgalactiae* group from Bruker MSP library. Peaks with intensities greater than 20% are labeled. Peaks common of all MSP are plotted in black circles. Peaks common to two or more MSPs are plotted in white circles. Peaks exclusive of each MSP are plotted in yellow circles.

for these diagnoses is species-specific PCR and 16S rRNA gene sequencing, but these techniques are expensive, time consuming and require highly technical skills. Meanwhile, the MALDI-TOF MS method can be an important technique to increase the laboratory speeds of identification of the etiological agent because it is an efficient and cost-effective method for the rapid and routine identification of bacterial isolates in the clinical microbiology laboratory (Seng et al., 2009; Seibold et al., 2010). The potential for identification at the serotype or strain level, and antibiotic resistance profiling within minutes, makes MALDI-TOF MS an on-going revolution in the clinical microbiology laboratory (Romero-Gómez et al., 2012; Østergaard et al., 2015; Sauget et al., 2016; Ueda et al., 2016).

Streptococcus agalactiae and *Lactococcus garvieae* strains were classified as the correct species in 100% of the MALDI Biotyper experiments. Both species had been cited in previous works with MALDI-TOF MS systems (Lartigue et al., 2009; Navas et al., 2013), but not with regards to strains isolated from fish. Although there are no studies about the variation of the subtype of *L. garvieae*, a large number of *S. agalactiae* subtypes are known (Jones et al., 2003). The strains obtained from fish farm outbreaks in Brazil, used in this work, are from different genomic subtypes (Godoy et al., 2013), but nevertheless they did not show divergence in RT identification using the MALDI Biotyper.

The possibility of inclusion of a custom MSP on the Bruker MALDI Biotyper makes the tool expansive and allows for its adaptation to the laboratory business independent of the equipment manufacturer. Following the example of what had previously been reported by Segawa et al. (2015), the *S. iniae* SI23 strain and SDD SD64, SD92, and SD142 strains were included as MSPs, and the results improved to 100% correct identification. Recently, Fan et al. (2017), analyzing studies performed of streptococci rapid classification, suggested an overestimated accuracy of MALDI-TOF MS systems on *Streptococcus* spp. identification, since the 16S rRNA gene sequencing analyses were only performed on discrepant results. In our analysis, all strains were identified by 16S rRNA gene sequencing or by the 16S rRNA gene in addition to housekeeping genes that were sequenced in parallel with the MALDI-TOF MS experiments, in order to achieve more confident results.

Streptococcus iniae strains, before the inclusion of a custom MSP, had matches with *S. pyogenes* and *S. dysgalactiae*, with scores lower than 2.000, suggesting a genus-level match (**Table 3**) within only ∼53% of tested isolates. The Bruker MSP library does not give MSP information about this species. The included custom MSP of SI23 showed similarities with these two species (**Figure 2**). These data corroborate with recent work from Kim et al. (2017) that shows the inclusion of *S. iniae* MSPs for the classification of *S. iniae* at the species-level, and shows the peaks list shared by *S. iniae* and *S. pyogenes*. Furthermore, 24 of 26 (∼92%) of peaks with relative intensities greater than 20 are shared between *S. iniae* ATCC 29178 (Kim et al., 2017) and *S. iniae* SI23 (**Figure 1**).

In relation to the SDD strains, during the strains' RT classification, the results were all above 2.000; however, 13 strains were classified as SDE, and the other 10 were classified as *S. dysgalactiae* species (**Table 4**). In previous work from our group (Costa et al., 2014), we suggested that the Brazilian *S. dysgalactiae* isolates were from a *dysgalactiae* subspecies, according to 16S rRNA and *sodA* genes sequencing. Because of previous work (Jensen and Kilian, 2012) based on the MLSA analysis of a combination of seven housekeeping genes and the study of their phylogenetic relationships, an identification of the tested isolates in this work as SDD was confirmed. A custom MSP was created with the chosen isolates SD64, SD92, and SD142. Each strain has a different genotype that was identified in analyses made by PFGE in a previous work from our group (Costa et al., 2014). Using the custom MSP, all the analyzed strains had a correspondence larger than 2.000 (**Table 4**), indicating a high similarity of these strains with the created MSPs. Specimens in the Bruker MSP library named SDE and *S. dysgalactiae* do not have an accessible history, and the strain identified as SDD is referenced as ATCC® 43078™, which is an isolate from a cow with mastitis (Garvie et al., 1983). Furthermore, as **Figure 2** shows, the SDE DSM 23147^T showed a distance level (i.e., similarity of selected isolates with a maximal value of divergence of 1,000) of ∼190 from another clade of *S. dysgalactiae* isolates and a different partner using MSP profiles in **Figure 4**. This characteristic suggests, taking into consideration there is no traceable information for the isolates in addition to the recent studies of *S. dysgalactiae* spp. (Jensen and Kilian, 2012; Ciszewski et al., 2016), that a reclassification, based on genomic analyses, should be done for such isolates from the Bruker MSP library.

Although the MALDI Biotyper is primarily designed for diagnoses at the species-level, in our experiments it was possible to correctly identify the subspecies of SDD, allowing for a rapid and low cost analysis when compared with other techniques to make subspecies-level identifications. MALDI-TOF MS was shown to be an efficient technology for identifying important Gram-positive cocci that cause major diseases in farmed fish.

AUTHOR CONTRIBUTIONS

GA, FP, and AZ wrote the manuscript. HF, GA, FP, GT, and CL conceived and designed the experiments. FP and AZ perform bioinformatics analyses. HF coordinated all analyses of the project. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank the Coordination for the Improvement of Higher Education Personnel (CAPES), the National Counsel of Technological and Scientific Development (CNPq) and Ministry of Agriculture, Livestock, and Food Supply (MAPA) for financing this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2017.01492/full#supplementary-material

REFERENCES

Abdelsalam, M., Asheg, A., and Eissa, A. E. (2013). *Streptococcus dysgalactiae*: an emerging pathogen of fishes and mammals. *Int. J. Veter. Sci. Med.* 1, 1–6. doi: 10.1016/j.ijvsm.2013.04.002

Agnew, W., and Barnes, A. C. (2007). *Streptococcus iniae*: an aquatic pathogen of global veterinary significance and a challenging candidate for reliable vaccination. *Vet. Microbiol.* 122, 1–15. doi: 10.1016/j.vetmic.2007.03.002

Alatoom, A. A., Cunningham, S. A., Ihde, S. M., Mandrekar, J., and Patel, R. (2011). Comparison of direct colony method versus extraction method for identification of gram-positive cocci by use of bruker biotyper matrix-assisted laser desorption ionization–time of flight mass spectrometry. *J. Clin. Microb.* 49, 2868–2873. doi: 10.1128/JCM.00506-11

Assis, G. B. N., Tavares, G. C., Pereira, F. L., Figueiredo, H. C. P., and Leal, C. A. G. (2016). Natural coinfection by *Streptococcus agalactiae* and *Francisella noatunensis* subsp. orientalis in farmed *Nile tilapia* (*Oreochromis niloticus* L.). *J. Fish Dis.* 40, 51–63. doi: 10.1111/jfd.12493

Bilecen, K., Yaman, G., Ciftci, U., and Laleli, Y. R. (2015). Performances and reliability of bruker microflex LT and VITEK MS MALDI-TOF mass spectrometry systems for the identification of clinical microorganisms. *Biomed Res. Int.* 2015:516410. doi: 10.1155/2015/516410

Bizzini, A., and Greub, G. (2010). Matrix-assisted laser desorption ionization time-of-flight mass spectrometry, a revolution in clinical microbial identification. *Clin. Microbiol. Infect.* 16, 1614–1619. doi: 10.1111/j.1469-0691.2010.03311.x

Brigante, G., Luzzaro, F., Bettaccini, A., Lombardi, G., Meacci, F., Pini, B., et al. (2006). Use of the phoenix automated system for identification of Streptococcus and Enterococcus spp. *J. Clin. Microbiol.* 44, 3263–3267. doi: 10.1128/JCM.00299-06

Ciszewski, M., Zegarski, K., and Szewczyk, E. M. (2016). *Streptococcus dysgalactiae* subsp. equisimilis isolated from infections in dogs and humans: are current subspecies identification criteria accurate? *Curr. Microbiol.* 73, 684–688. doi: 10.1007/s00284-016-1113-x

Clark, A. E., Kaleta, E. J., Arora, A., and Wolk, D. M. (2013). Matrix-assisted laser desorption ionization-time of flight massspectrometry: a fundamental shift in the routine practice of clinical microbiology. *Clin. Microbiol. Rev.* 26, 547–603. doi: 10.1128/CMR.00072-12

Clarridge, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin. Microbiol. Rev.* 17, 840–862. doi: 10.1128/CMR.17.4.840-862.2004

Costa, F. A. A., Leal, C. A. G., Leite, R. C., and Figueiredo, H. C. P. (2014). Genotyping of *Streptococcus dysgalactiae* strains isolated from Nile tilapia, *Oreochromis niloticus* (L.). *J. Fish Dis.* 37, 463–469. doi: 10.1111/jfd.12125

Eldar, A., Bejerano, Y., Livoff, A., Horovicz, A., and Bercovier, H. (1995). Experimental streptococcal meningo-encephalitis in cultured fish. *Vet. Microbiol.* 43, 33–40. doi: 10.1016/0378-1135(94)00052-X

Evans, J. J., Klesius, P. H., Gilbert, P. M., Shoemaker, C. A., Sarawi, M. A. A., Landsberg, J., et al. (2002). Characterization of β-haemolytic Group B Streptococcus agalactiae in cultured seabream, *Sparus auratus* L., and wild mullet, *Liza klunzingeri* (Day), in Kuwait. *J. Fish Dis.* 25, 505–513. doi: 10.1046/j.1365-2761.2002.00392.x

Fan, W. T., Qin, T. T., Bi, R. R., Kang, H. Q., Ma, P., and Gu, B. (2017). Performance of the matrix-assisted laser desorption ionization time-of-flight mass spectrometry system for rapid identification of streptococci: a review. *Eur. J. Clin. Microbiol. Infect. Dis.* 36, 1005–1012. doi: 10.1007/s10096-016-2879-2

FAO (2016). *The State of World Fisheries and Aquaculture (SOFIA)*. Available online at: http://www.fao.org/3/a-i5555e.pdf

Figueiredo, H. C. P., Netto, L. N., Leal, C. A. G., Pereira, U. P., and Mian, G. F. (2012). *Streptococcus iniae* outbreaks in Brazilian Nile tilapia (*Oreochromis niloticus* L.) farms. *Braz. J. Microbiol.* 43, 576–580. doi: 10.1590/S1517-83822012000200019

Fox, J. G., Yan, L. L., Dewhirst, F. E., Paster, B. J., Shames, B., Murphy, J. C., et al. (1995). *Helicobacter bilis* sp. Nov., A novel *Helicobacter* species isolated from bile, livers, and intestines of aged, inbred mice. *J. Clin. Microbiol.* 33, 445–454.

Fukushima, H. C. S., Leal, C. A. G., Cavalcante, R. B., Figueiredo, H. C. P., Arijo, S., Moriñigo, M. A., et al. (2017). *Lactococcus garvieae* outbreaks in Brazilian farms Lactococcosis in Pseudoplatystoma sp. – development of an autogenous vaccine as a control strategy. *J. Fish Dis.* 40, 263–272. doi: 10.1111/jfd.12509

Garvie, E. I., Farrow, J. A. E., and Bramley, A. J. (1983). *Streptococcus dysgalactiae* (Diernhofer) nom. rev. *Int. J. Syst. Bacteriol.* 33, 404–405. doi: 10.1099/00207713-33-2-404

Godoy, D. T., Carvalho-Castro, G. A., Leal, C. A. G., Pereira, U. P., Leite, R. C., and Figueiredo, H. C. P. (2013). Genetic diversity and new genotyping scheme for fish pathogenic *Streptococcus agalactiae*. *Lett. Appl. Microbiol.* 57, 476–483. doi: 10.1111/lam.12138

Jensen, A., and Kilian, M. (2012). Delineation of *Streptococcus dysgalactiae*, its subspecies, and its clinical and phylogenetic relationship to Streptococcus pyogenes. *J. Clin. Microbiol.* 50, 112–126. doi: 10.1128/JCM.05900-11

Jones, N., Bohnsack, J. F., Takahashi, S., Oliver, K. A., Chan, M. S., Kunst, F., et al. (2003). Multilocus sequence typing system for group B streptococcus. *J. Clin. Microbiol.* 6, 2530–2536. doi: 10.1128/JCM.41.6.2530-2536.2003

Keirstead, N. D., Brake, J. W., Griffin, M. J., Halliday-Simmonds, I., Thrall, M. A., and Soto, E. (2014). Fatal Septicemia caused by the Zoonotic Bacterium *Streptococcus iniae* during an outbreak in Caribbean reef fish. *Vet. Pathol.* 51, 1035–1041. doi: 10.1177/0300985813505876

Kim, S. W., Nho, S. W., Im, S. P., Lee, J. S., Jung, J. W., Lazarte, J. M., et al. (2017). Rapid MALDI biotyper-based identification and cluster analysis of *Streptococcus iniae*. *J. Microbiol.* 55, 260–266. doi: 10.1007/s12275-017-6472-x

Kolbert, C. P., and Persing, D. H. (1999). Ribosomal DNA sequencing as a tool for identification of bacterial pathogens. *Curr. Opin. Microbiol.* 2, 299–305. doi: 10.1016/S1369-5274(99)80052-6

Lartigue, M. F., Héry-Arnaud, G., Haguenoer, E., Domelier, A. S., Schmit, P. O., van der Mee-Marquet, N., et al. (2009). Identification of Streptococcus agalactiae isolates from various phylogenetic lineages by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *J. Clin. Microbiol.* 47, 2284–2287. doi: 10.1128/JCM.00175-09

Mian, G. F., Godoy, D. T., Leal, C. A. G., Yuhara, T. Y., Costa, G. M., and Figueiredo, H. C. P. (2009). Aspects of the natural history and virulence of *S. agalactiae* infection in *Nile tilapia*. *Vet. Microbiol.* 36, 180–183. doi: 10.1016/j.vetmic.2008.10.016

Nagy, E., Urbán, E., and Nord, C. E. (2011). ESCMID study group on antimicrobial resistance in Anaerobic Bacteria 2011 antimicrobial susceptibility of *Bacteroides fragilis* group isolates in Europe: 20 years of experience. *Clin. Microbiol. Infect.* 17, 371–379. doi: 10.1111/j.1469-0691.2010.03256.x

Navas, M. E., Hall, G., and El Bejjani, D. (2013). A case of Endocarditis caused by *Lactococcus garvieae* and suggested methods for identification. *J. Clin. Microbiol.* 51, 1990–1992. doi: 10.1128/JCM.03400-12

Netto, L. N., Leal, C. A. G., and Figueiredo, H. C. P. (2011). *Streptococcus dysgalactiae* as an agent of septicaemia in Nile tilapia, *Oreochromis niloticus* (L.). *J. Fish Dis.* 34, 251—254. doi: 10.1111/j.1365-2761.2010.01220.x

Nguyen, N. P., Warnow, T., Pop, M., and White, B. (2016). A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *Npj Biofilms Microbiomes*. 2:16004. doi: 10.1038/npjbiofilms.2016.4

Nomoto, R., Unose, N., Shimahara, Y., Nakamura, A., Hirae, T., Maebuchi, K., et al. (2006). Characterization of lancefield group C *Streptococcus dysgalactiae* isolated from farmed fish. *J. Fish Dis.* 29, 673–682. doi: 10.1111/j.1365-2761.2006.00763.x

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A., Korobeynikov, A., Lapidus, A., et al. (2013). Assembling genomes and mini-metagenomes from highly chimeric reads. *Lect. Notes Comput. Sci.* 20, 714–737. doi: 10.1007/978-3-642-37195-0_13

Østergaard, C., Hansen, S. G. K., and Møller, J. K. (2015). Rapid first-line discrimination of methicillin resistant Staphylococcus aureus strains using MALDI-TOF MS. *Int. J. Med. Microbiol.* 305, 838–847. doi: 10.1016/j.ijmm.2015.08.002

Patel, J. B. (2001). 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. *Mol. Diagn.* 6, 313–321. doi: 10.2165/00066982-200106040-00012

Poyart, C., Quesne, G., Coulon, S., Berche, P., and Trieu-Cuot, P. (1998). Identification of streptococci to species level by sequencing the gene encoding the manganese-dependent superoxide dismutase. *J. Clin. Microbiol.* 36, 41–47.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rizzardi, K., Wahab, T., and Jernberg, C. (2013). Rapid subtyping of *Yersinia enterocolitica* by matrix-assisted laserdesorption ionization-time of flight mass spectrometry (MALDI-TOF MS) for diagnostics and surveillance. *J. Clin. Microbiol.* 51, 4200–4203. doi: 10.1128/JCM.01416-13

Romero-Gómez, M. P., Gómez-Gil, R., Paño-Pardo, J. R., and Mingorance, J. (2012). Identification and susceptibility testing of microorganism by direct inoculation from positive blood culture bottles by combining MALDI-TOF and Vitek-2 Compact is rapid and effective. *J. Infect.* 65, 513–520. doi: 10.1016/j.jinf.2012.08.013

Sauget, M., van der Mee-Marquet, N., Bertrand, X., and Hocquet, D. (2016). Matrix-assisted laser desorption ionization-time of flight Mass spectrometry can detect *Staphylococcus aureus* clonal complex 398. *J. Microbiol. Methods.* 127, 20–23. doi: 10.1016/j.mimet.2016.05.010

Segawa, S., Nishimura, M., Sogawa, K., Tsuchida, S., Murata, S., Watanabe, M., et al. (2015). Identification of Nocardia species using matrix-assisted laser desorption/ionization–time-of-flight mass spectrometry. *Clin. Proteomics.* 12:6. doi: 10.1186/s12014-015-9078-5

Seibold, E., Maier, T., Kostrzewa, M., Zeman, E., and Splettstoesser, W. (2010). Identification of *Francisella tularensis* by whole-cell matrix-assisted laser desorption ionization-time of flight mass spectrometry: fast, reliable, robust, and cost-effective differentiation on species and subspecies levels. *J. Clin. Microbiol.* 48, 1061–1069. doi: 10.1128/JCM.01953-09

Seng, P., Drancourt, M., Gouriet, F., Scola, B., Fournier, P., Rolain, J. M., et al. (2009). Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Clin. Infect. Dis.* 49, 543–551. doi: 10.1086/600885

Singhal, N., Kumar, M., Kanaujia, P. K., and Virdi, J. S. (2015). MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Front. Microbiol.* 6:791. doi: 10.3389/fmicb.2015.00791

Soto, E., Wang, R., Wiles, J., Baumgartner, W., Green, C., Plumb, J., et al. (2015). Characterization of isolates of *Streptococcus agalactiae* from diseased farmed and wild marine fish from the U.S. Gulf Coast, Latin America, and Thailand. *J. Aquat. Anim. Health.* 27, 123–134. doi: 10.1080/08997659.2015.1032439

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Tavares, G. C., Costa, F. A. A., Santos, R. R., Barony, G. M., Leal, C. A. G., and Figueiredo, H. C. P. (2016). Nonlethal sampling methods for diagnosis of *Streptococcus agalactiae* infection in *Nile tilapia*, *Oreochromis niloticus* (L.). *Aquaculture* 454, 237–242. doi: 10.1016/j.aquaculture.2015.12.028

Ueda, O., Tanaka, S., Nagasawa, Z., Hanaki, H., Shobuike, T., and Miyamoto, H. (2016). Development of a novel matrix-assisted laser desorption/ionization time-of-flight mass spectrum (MALDI-TOF-MS)-based typing method to identify meticillin-resistant *Staphylococcus aureus* clones. *J. Hosp. Infect.* 90, 147–155. doi: 10.1016/j.jhin.2014.11.025

Vendrell, D., Balcázar, J. L., Ruiz-Zarzuela, I., Blas, I., Gironés, O., and Múzquiz, L. (2006). *Lactococcus garvieae* in fish: a review. *Comp. Immunol. Microbiol. Infect. Dis.* 29, 177–198. doi: 10.1016/j.cimid.2006.06.003

Větrovský, T., and Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS ONE* 8:e57923. doi: 10.1371/journal.pone.0057923

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The first complete genome sequence of *Streptococcus dysgalatiae* subsp. *dysgalactiae* an emerging fish pathogen

Alexandra A. U. Zegarra [1], Felipe L. Pereira[1], Fernanda A. Dorella[1], Alex F. Carvalho[1], Gustavo M. Barony[1], Carlos A. G. Leal[1], H. C. P. Figueiredo[1*]

1 National Reference Laboratory for Aquatic Animal Diseases of Ministry of Agriculture, Livestock and Food Supply, Belo Horizonte, MG, Brazil
* Corresponding author: figueiredoh@yahoo.com

Introduction

Streptococcus dysgalatiae subsp. *dysgalactiae* (SDD) is a Gram-positive cocci, it autoaggregates in saline, forms long chains in growth medium, it is catalase negative and α-hemolytic on blood agar. In 2002, it caused the first outbreak in southern Japanese farms. During the subsequent years fish farms in the country suffered huge losses. In Brazil, outbreaks of streptococcosis are common in the freshwater fish species Nile tilapia, *Oreochromis niloticus* (L.). In 2007, the first disease outbreak caused by SDD was spotted in Ceará state. The disease has spread worldwide and despite its increasing clinical and economic significance up until the moment, none SDD genome was fully sequenced.

Methods

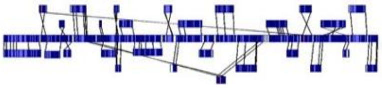
DNA isolation

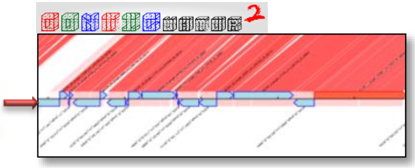
ion torrent: Single-end 200bp; Single-end 400bp; Mate pair 6kbp

illumina: Paired-end 2x150bp; Mate pair 6kbp

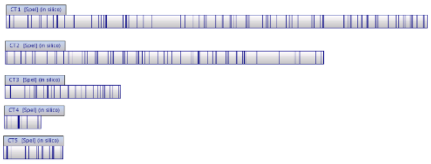
Overly fragmented assembly

Contigs of the assembly with best N50 were aligned to the map:





Construction of scaffolds:



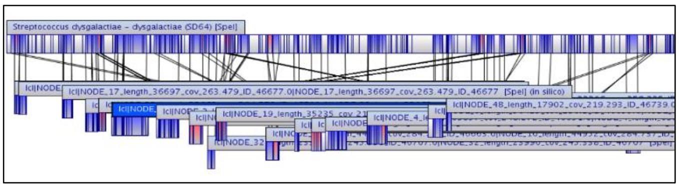
CT1: N1RC-N24-N2RC-N36RC-N12RC-N49RC-N32RC-N64RC-N19
CT2: N8RC-N3-N56RC-N54RC-N20RC-N45RC-N15-N6
CT3: N7RC-N34-N55-N13
CT4: N53RC-N31RC
CT5: N14RC-N33RC

Additional scaffolds were constructed using the output graph path and repeat resolution files of the assembly software, later, along with all these new scaffolds every contig corresponding to the rest of the assemblies previously performed, if aligned in another site of the optical map, was kept in order to execute a new assembly using the "–trustedcontigs" option of SPAdes. Furthermore, gap filling was made using CLC Genomics Workbench 7.

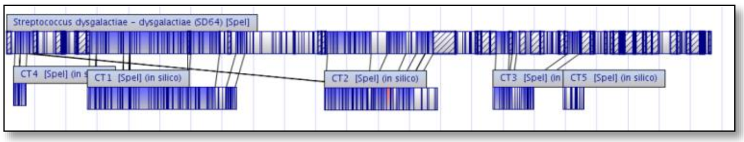
REPEAT — N — Genome completed? — Y — GenBank deposit

Results and Discussion

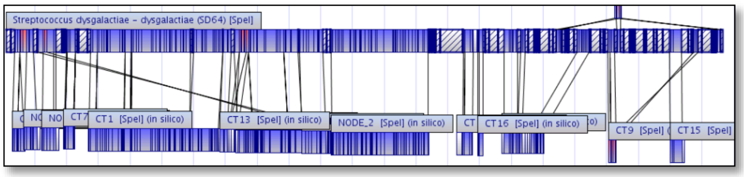
Initially, the assembly with better results gave a total of 167 contigs with an N50 value of 26,993bp and the largest contig with a 141,256bp length size and a ~44% of whole genome map (WGM) coverage:



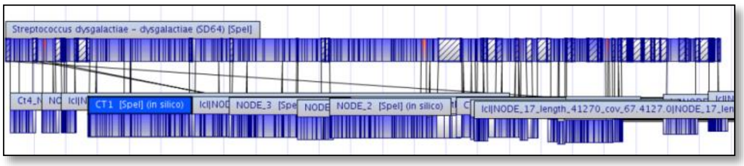
The first scaffolds constructed, along with the initial mapped contigs increased the WGM to a 60.893% coverage



Later, when the previous maped contigs were set as "-trustedcontigs" in a new assembly a 70.51% of WGM coverage was reached



The same strategy was applied to an assembly with a Mate-Pair dataset and a 83.93% of WGM coverage was reached.


Streptococcus dysgalactiae – dysgalactiae (SD64) [SpeI]
CT1 [SpeI] (in silico)
NODE_3
NODE_2 [SpeI] (in silico)
lcl|NODE_17_length_41270_cov_67.4127

Conclusion

This study empowers the use of optical mapping together with NGS strategies such as Paired-end and Mate-pair as a very effective tool in the assembly of highly repetitive genomes. As new technologies are on their way of resolving these issues, the use of optical maps propose both orientation and scaffolding construction as the main strategies in complete genome assembling.

Further results as the first SDD complete genome announcement are expected

The first complete genome sequence of *Streptococcus dysgalatiae* subsp. *dysgalactiae* an emerging fish pathogen

Alexandra A. Urrutia Zegarra, Felipe L. Pereira, Fernanda A. Dorella, Alex F. Carvalho, Gustavo Morais Barony, Carlos A. G. Leal, and Henrique C. P. Figueiredo

ABSTRACT

Streptococcus dysgalatiae subsp. *dysgalactiae* (SDD) is a Gram-positive cocci, it autoaggregates in saline, forms long chains in growth medium, it is catalase negative and α-hemolytic on blood agar. In 2002, it caused the first outbreak in southern Japanese farms. During the subsequent years fish farms in the country suffered huge losses. In Brazil, outbreaks of streptococcosis are common in the freshwater fish species Nile tilapia, *Oreochromis niloticus* (L.). In 2007, the first disease outbreak caused by SDD was spotted in Ceará state. The disease has spread worldwide and despite its increasing clinical and economic significance up until the moment, none SDD genome was fully sequenced. Therefore, considering the importance of a complete genome to characterize this fish pathogen strategy, a next-generation sequence genome initiative was managed. To obtain the SDD genome the sample was isolated from an overnight culture with the Maxwell 16 tissue DNA purification kit using the Maxwell 16 system (both from Promega, USA). A first run was conducted on the Ion Torrent PGM™ sequencing system (Life Technologies, USA) using a 200bp (~ 300- fold coverage) fragment library kit. However, as it resulted in an overly fragmented assembly, another runs were performed using a 400bp (~870-fold coverage) fragment library kit and a 400bp (~ 107 fold coverage) mate-pair kit with an insert of 6kbp. Additional runs were conducted on the Illumina® MiSEQ sequencing system using paired-end 2x150bp (~638-fold coverage) and mate-pair (~658-fold coverage), with an insert of 6kbp. Yet, as no improvements were reached in the assembly fragmentation matter an optical map was acquired. The sequences were assembled with SPAdes 3.8.0, and Newbler 2.9 software, the assembly with higher N50 was selected and aligned with the Optical Map (OpGen Inc, USA) in order to verify the orientation and start scaffolding. Additionally, CONTIGuator software and the assembly_graph text file from the assembly output were used for further scaffold construction. Initially 167 contigs were obtained with an N50 value of 26,993bp and the largest contig with a 141,256bp length size and a ~44% of whole genome map (WGM) coverage. The first scaffolds constructed were used as input in a new assembly, this strategy lead to a better N50 (28,066bp) and fewer contigs (148). The procedure was repeated and ~52% of WGM coverage was reached. Currently, 81% coverage of the WGM was reached and gap filling with CLC Genomics Workbench 7 (Qiagen, USA) still in process. The present study empowers the use of optical mapping as a tool in the assembly of highly repetitive genomes. Further results as the first SDD complete genome announcement are expected.

This study was supported by the Ministry of Agriculture, Live Stock and Food Supply; FAPEMIG; CNPq and the National Institute for Science and Technology.