ULISSES DE PÁDUA PEREIRA

SEQUENCIAMENTO GENÔMICO, ANÁLISES COMPARATIVAS E PREDIÇÃO DE ALVOS VACINAIS EM *Streptococcus agalactiae* ISOLADOS DE PEIXES, SERES HUMANOS E BOVINOS

LAVRAS - MG

2013

ULISSES DE PÁDUA PEREIRA

SEQUENCIAMENTO GENÔMICO, ANÁLISES COMPARATIVAS E PREDIÇÃO DE ALVOS VACINAIS EM *Streptococcus agalactiae* ISOLADOS DE PEIXES, SERES HUMANOS E BOVINOS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciências Veterinárias, área de concentração em Ciências Veterinárias, para a obtenção do título de Doutor.

Orientador

Dr. Henrique César Pereira Figueiredo

LAVRAS - MG

2013

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca da UFLA

Pereira, Ulisses de Pádua.
Sequenciamento genômico, análises comparativas e predição de alvos vacinais em *Streptococcus agalactiae* isolados de peixes, seres humanos e bovinos / Ulisses de Pádua Pereira. – Lavras: UFLA, 2013.
131 p. : il.

Tese (doutorado) – Universidade Federal de Lavras, 2013.
Orientador: Henrique Cear Pereira Figueiredo.
Bibliografia.

1. *Streptococcus agalactiae*. 2. Sequenciamento genômico. 3. Doenças infecciosas. 4. Predição de alvos vacinais.I. Universidade Federal de Lavras. II. Título.

CDD – 639.8

ULISSES DE PÁDUA PEREIRA

SEQUENCIAMENTO GENÔMICO, ANÁLISES COMPARATIVAS E PREDIÇÃO DE ALVOS VACINAIS EM *Streptococcus agalactiae* ISOLADOS DE PEIXES, SERES HUMANOS E BOVINOS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciências Veterinárias, área de concentração em Ciências Veterinárias, para a obtenção do título de Doutor.

APROVADA em 21 de fevereiro de 2013.

Dra. Gláucia Frasnelli Mian	UFLA
Dra. Rosane Freitas Schwan	UFLA
Dra. Patrícia Gomes Cardoso	UFLA
Dr. Vasco Ariston de C. Azevedo	UFMG

Dr. Henrique César Pereira Figueiredo
Orientador

LAVRAS – MG

2013

A Deus, pela força e obstinação;
Aos meus pais, Geraldo e Lourdes, irmãos e amigos, pelo amor, apoio,
confiança e momentos de felicidade compartilhados.

DEDICO

AGRADECIMENTOS

À Universidade Federal de Lavras e ao Departamento de Medicina Veterinária, pela oportunidade de realizar o doutorado;

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de estudos;

Aos meus pais, Geraldo e Lourdes, e irmãos, pela confiança incondicional, presença constante, ajuda financeira e pela paciência nestes longos quatro anos;

Ao professor Henrique Figueiredo, pela orientação, dedicação, oportunidade e pelos valiosos ensinamentos que contribuem para a minha formação profissional;

Ao professor Vasco Azevedo, por receber-me em sua equipe, pelos treinamentos e incontáveis ensinamentos, os quais foram essenciais para minha formação;

Aos colegas de trabalho do Laboratório LGCM, em especial Siomar, Flávia, Luís, Alfonso, Anne, Sintia, Vinícius e todos que ajudaram na execução deste projeto e compartilharam a agradável convivência nestes quase dois anos;

Ao professor Artur e toda equipe do LPDNA (Rommel, Drica, Pablo e outros), pelo importante participação na execução deste projeto;

Aos colegas de trabalho do Laboratório Aquavet;

Ao colega Lamartine (em memória) e Prof. Carlos. pela valiosa colaboração na condução do projeto;

Ao Centro de Excelência em Bioinformática - CEBio (FIOCRUz-MG);

Ao Ministerio da Pesca e Aquicultura, por financiar o projeto;

Aos amigos, José Retori, Tiago, Alessandra, Profa. Gláucia, Prof. Geraldo, Anderson e Izinara, pela amizade, conselhos e momentos vividos;

À Escola de Veterinária da UFMG, especialmente ao Departamento de

Medicina Veterinária Preventiva e o Instituto de Ciências Biológicas, por terem me recebido e fornecido as condições para a realização dos experimentos.

RESUMO GERAL

Streptococcus agalactiae (grupo B de Lancefield; GBS) é um importante patógeno para seres humanos, bovinos e peixes causando septicemia neonatal, mastite e meningo-encefalite, respectivamente. Este patógeno é responsável por significativa taxa de mortalidade em seres humanos neonatos e grandes perdas econômicas na produção de pescado e leite no Brasil e no mundo. Embora já existam genomas disponíveis de linhagens desta espécie bacteriana isoladas de seres humanos, bovinos e peixes, pouco se sabe sobre as características genômicas dos isolados de peixes. O presente trabalho foi realizado com os objetivos de sequenciar o genoma de uma linhagem de *S. agalactiae* (SA20-06) isolada de surto da doença em peixes no Brasil e fazer análises comparativas com outros 14 genomas disponíveis desta espécie de linhagens isoladas de seres humanos, bovinos e peixes. A partir dos dados destes genomas foram realizadas análises de pan-genoma, vias metabólicas, ilhas de patogenicidade e predição de potenciais alvos vacinais. O genoma completo da linhagem brasileira sequenciada apresentou menor tamanho quando comparado com tamanho do genoma de linhagens de outros hospedeiros. Grande diversidade no genoma das linhagens de *S. agalactiae* foi observada, porém o genoma das linhagens de peixe demonstrou-se menos variável. Possíveis alvos vacinais preditos por ferramentas de bioinformática são discutidos de forma global (analisando o genoma das 15 linhagens) e no subgrupo dos isolados de diferentes hospedeiros. As análises comparativas dos genomas da espécie contribuíram para o entendimento da interação patógeno hospedeiro e sugerem novos genes a serem caracterizados funcionalmente.

Palavras-chave: *Streptococcus agalactiae*. Sequenciamento genômico. Doenças infecciosas. Predição de alvos vacinais.

GENERAL ABSTRACT

Streptococcus agalactiae (Lancefield group B; GBS) is an important pathogen for humans, cattle and fish causing neonatal septicemia, mastitis and meningo-encephalitis, respectively. This pathogen is responsible for significant mortality rate in human neonates and great economic losses in fish and milkproduction in Brazil and worldwide. Although there are already available genomes of strains of this bacterial species isolated from humans, cattle and fish, fewis known about the genomic features of strains of fish. This study was conducted with the following objectives: to sequence the genome of a strain of *S. agalactiae* (SA20-06) isolated from an outbreak of disease in fish in Brazil and do comparative analyzes with other 14 available genomes of strains of this species isolated from humans, cattle and fish. From the data of these genomes were analyzed for pan-genome, metabolic pathways, pathogenicity islands and prediction of potential vaccine targets. The complete genome of Brazilian strain sequenced showed smaller size when compared to the genome of strains of other hosts. Great diversity in the genome strains of *S. agalactiae* was observed, but the fish genome strains showed to be less variable. Possible vaccine targets predicted by bioinformatics tools are discussed in the overall picture of all the strains and in the subgroup of isolates from different hosts. Comparative analyzes of the genomes of the species contributed to the understanding of host pathogen interactions and suggest new genes to be characterized functionally.

Keywords: *Streptococcus agalactiae*. Genomic sequencing. Infectious diseases. Vaccine target prediction.

SUMÁRIO

PRIMEIRA PARTE

1 INTRODUÇÃO

A aquicultura mundial é o setor da produção animal que vem apresentando maiores taxas de crescimento nas últimas décadas (FOOD AND AGRICULTURE ORGANIZATION, FAO, 2010). No Brasil, a aquicultura segue a tendência mundial de alta, com aumento de 15,3% em 2010 em relação à produção de 2009, sendo que a piscicultura continental representou 82,3% da produção total nacional. Atualmente, a tilápia é o peixe mais produzido no país, representando aproximadamente 40% da produção nacional de peixes continentais (BRASIL, , 2010).

Apesar do alto potencial para produção de organismos aquáticos, ainda há no Brasil diversos entraves que desfavorecem o desenvolvimento do setor. Dentre esses, a ocorrência de surtos de doenças infecciosas têm se mostrado um dos principais limitantes para diversos ramos da aquicultura, ocasionando sérias perdas econômicas aos produtores inviabilizando a produção em muitos casos. Protozoários, fungos e bactérias são os principais agentes etiológicos associados a casos de mortalidade em pisciculturas no país (MARTINS, et al., 2004; FIGUEIREDO et al., 2005; MIAN et al., 2009).

Nos últimos anos, bactérias do gênero *Streptococcus* têm ganhado notoriedade como principais patógenos de peixes tropicais cultivados em todo o mundo (ELDAR, 1995; EVANS et al., 2002; AGNEW; BARNES, 2007; SHEWMAKER et al., 2007). De maneira similar, a ocorrência de surtos de estreptococose em cultivos de tilápia têm se destacado no Brasil, sendo atualmente um dos maiores entraves para viabilidade das propriedades e empresas. Dados de literatura e da casuística do AQUAVET-Laboratório de Doenças de Animais Aquáticos da Universidade Federal de Minas Gerais

(UFMG) reportam a ocorrência de surtos causados pela bactéria *Streptococcus agalactiae* em tilapiculturas em 10 estados Brasileiros (MIAN et al., 2009). Casos de infecção por essa bactéria ocasionam altas taxas de mortalidade, que podem atingir até 90% do plantel (ELDAR, 1995; EVANS et al., 2002; MIAN et al., 2009). Atualmente a principal medida terapêutica empregada em caso de surtos é a antibioticoterapia. Porém, o uso desta técnica implica em potenciais riscos para o meio ambiente, resistência bacteriana, segurança alimentar da população humana, e em muitos dos casos, é empregada quando já ocorreram perdas significativas do plantel (HEUER et al., 2009).

Este patógeno pode também infectar seres humanos, sendo que *S. agalactiae* está principalmente relacionado à sepse e meningite em neonatos (BISHARAT et al., 2004; CORRÊA et al., 2010). Linhagens isoladas de diferentes hospedeiros geralmente são geneticamente distintas, e com isso acredita-se que há adaptações hospedeiro-específicas relacionadas à patogenia e virulência (DOGAN et al., 2005; SUKHNANAND et al., 2005; PEREIRA et al., 2010). Porém, há poucos estudos sobre as bases moleculares envolvidas na patogenicidade de *S. agalactiae* em peixes.

Nos últimos anos, o sequenciamento genômico tem sido utilizado de forma crescente para a caracterização de diversas espécies bacterianas. Com o advento das tecnologias de sequenciamento de próxima geração ("*Next-Generation Sequencing - NGS"*), a obtenção de dados genômicos tornou-se inquestionavelmente mais rápida e barata, com ampla aplicação principalmente para organismos procariotos. As plataformas 454 FLX (Roche, Bélgica), Illumina (Genome Analyzer), SOLiD e Ion Torrent (Applied Biosystems, EUA) possibilitam hoje o sequenciamento massivo de milhões de fragmentos de DNA simultaneamente (LOMAN et al., 2012; MARDIS, 2008; SHENDURE; JI, 2008). Essas técnicas permitiram que análises globais de genomas e de

transcriptomas se tornassem mais rápidas, de menor custo e consequentemente acessíveis (SHENDURE; JI, 2008).

A genômica comparativa pode esclarecer processos evolutivos que influenciam espécies ou populações bacterianas. Esta pode, por exemplo, identificar componentes do genoma com importantes papéis na virulência e/ou adaptação a habitats/hospedeiros e utilização de nutrientes (MONNET et al., 2010; RICHARDS et al., 2011). Além disso, estudos *in silico* em sequências genômicas nos últimos anos possibilitaram a predição de genes candidatos vacinais (vacinologia reversa) (RAPPUOLI, 2000; SANTOS, et al., 2011).

Atualmente há quinze genomas de *S. agalactiae* sequenciados, sendo dez de amostras isoladas de seres humanos (GLASER et al., 2002; TETTELIN et al., 2002; 2005), quatro de amostras isoladas de peixes e um de amostra isolada de mastite bovina. O genoma da amostra isolada de bovino revelou adaptações nicho/hospedeiro específicas, características tais compartilhadas por outros patógenos de mastite (RICHARDS et al., 2011), porém, pouco foi estudado em relação aos genomas de linhagens isoladas de peixe.

Devido ao pouco conhecimento a respeito do genoma de *S. agalactiae* isolado de peixes, seus mecanismos moleculares envolvidos na adaptação a diferentes hospedeiros (homeotérmicos e heteterotérmicos), e genes com potencial para desenvolvimento de vacinas, o sequenciamento do genoma da linhagem SA20-06 (isolada em um surto de meningoencefalite em tilápias do Nilo no estado do Paraná em 2006) deste patógeno será descrito neste trabalho. Adicionalmente, diversas análises de genômica comparativa e predição de alvos vacinais foram realizadas utilizando os genomas disponíveis de linhagens deste patógeno isoladas de peixes, seres humanos e bovinos. Estas análises utilizando linhagens isoladas destes três hospedeiros ainda não foi descrita na literatura, e no presente trabalho foram realizadas com o intuito de melhor entender as características relacionadas à adaptação ao habitat, a patogênese desta espécie

em cada um destes hospedeiros bem como predizer proteínas imunogênicas com potencial para o desenvolvimento de vacinas.

2 HIPÓTESES

a) O sequenciamento do genoma de uma linhagem de *S. agalactiae* isolada de peixe contribuirá com informações que levam à compreensão sobre a adaptação do patógeno ao hospedeiro e habitat;
b) O sequenciamento do genoma de um isolado de *S. agalactiae* de peixe irá contribuir para o pan-genoma da espécie, adicionando genes linhagem específicos relacionados à patogenicidade e virulência em peixes.

3 OBJETIVOS

3.1 Objetivos gerais

Os objetivos com o presente trabalho são os de realizar o sequenciamento, montagem e anotação do genoma de uma linhagem de *S. agalactiae* utilizando as plataformas de sequenciamento de nova geração. Posteriormente, fazer análises de genômica comparativa e procurar características nestes genomas que ajudem no entendimento de fenômenos genéticos evolutivos que justifiquem a adaptação ao hospedeiro (peixe) e *habitat*. Caracterizar, *in silico,* prováveis genes associados à virulência e patogenicidade como também predizer candidatos vacinais imunogênicos.

3.2 Objetivos específicos

a) Sequenciar o genoma de *S. agalactiae* utilizando tecnologias de sequenciamento de próxima geração: SOLiD V3 e SOLiD 5500;
b) Montar, anotar e depositar o genoma no GenBank (NCBI) utilizando os *pipelines* específicos para cada tipo de plataforma de sequenciamento;
c) Realizar análises de pan-genoma da espécie, utilizando os 15 genomas disponíveis isolados de diferentes hospedeiros;
d) Realizar a predição de ilhas de patogenicidade e análise dos genes constituintes;
e) Realizar análises *in silico* de vias metabólicas comparando os genomas de *S. agalactiae* disponíveis para averiguar possíveis adaptações ao *habitat*;

f) Predizer possíveis alvos vacinais, que estimulem tanto imunidade humoral quanto celular, utilizando pipeline de vacinologia reversa e imunoinformática.

4 REFERENCIAL TEÓRICO

4.1 Aquicultura e produção de Tilápia no Brasil

Atualmente a aquicultura é o ramo da produção de proteína de origem animal que mais cresce em todo mundo, com um crescimento na produção total de menos de um milhão de toneladas nos anos 1950 para aproximadamente 52,5 milhões de toneladas em 2008. Segundo dados da *Food and Agriculture Organization* (FAO, 2010) a aquicultura já atende metade da demanda mundial por peixes e outros animais aquáticos, sendo ainda satisfatoriamente capaz de atender a crescente procura por esse tipo de alimento. Nas últimas décadas (1970-2008) houve um incremento significativo na produção aquícola em todo o mundo, com crescimento médio de 8,3%, com destaque para a América Latina e o Caribe onde em média o crescimento foi de 21,1 %. A produção de peixes de água doce responde, atualmente, por 54,7% do volume total produzido e 41,2% da receita total obtida na aquicultura mundial (FAO, 2010).

Comparando-se a produção de 2010 (479.398t) com o montante produzido em 2008 (365.366t), fica evidente o crescimento da aquicultura no Brasil, com um incremento de 31,2% na produção durante o triênio 2008-2010. Seguindo o padrão observado nos anos anteriores, a maior parcela da produção aquícola é oriunda da aquicultura continental, na qual se destaca a piscicultura continental que representou 82,3% da produção total nacional (BRASIL, 2010). O país detém a terceira maior produção aquícola da América Latina, sendo expoente na produção de camarões e tilápias. De acordo com dados da FAO, o Brasil encontra-se na sétima posição como produtor mundial de tilápias (FAO, 2010). Esta é a espécie mais cultivada no país e sua produção apresentou aumento de aproximadamente 20% anual entre os anos de 2008 e 2010 (BRASIL, 2010).

O cultivo de tilápias em todo mundo, principalmente em regiões tropicais, vêm apresentando crescimento médio anual de 11,5%, com produção bruta inferior apenas à de carpas e salmonídeos (EL-SAYED, 1999). A indústria brasileira de tilápia cresce vigorosamente, com posição de destaque entre os países americanos. Condições climáticas favoráveis e vastos reservatórios hídricos corroboram para a expansão significativa da produção nacional. O estabelecimento da cadeia produtiva da tilápia propicia o desenvolvimento de agroindústrias de processamento, geração de milhares de empregos, maior produção de alimentos, geração de renda e intercâmbio de tecnologias, incrementando o agronegócio nacional(VERA-CALDERÓM; FERREIRA, 2004).

O consumo *per capita* de pescado *in natura* no Brasil é pequeno (9,75 kg/hab/ano) em relação a outros países do mundo como, por exemplo, nos países asiáticos BRASIL, 2010; FAO, 2010). Somente cerca de 10% da população nacional utiliza o pescado em sua alimentação sendo este hábito muito variado de acordo com a região, sendo significativamente maior nas regiões norte e nordeste quando comparado com a região sul (BRASIL, 2010).

4.2 Estreptococoses em peixes

Estreptococose é o termo genérico utilizado para designar as doenças septicêmicas de etiologia bacteriana em peixes causadas por cocos Gram positivos do gênero *Streptococcus* (MATA et al., 2004). Essa doença foi primeiramente descrita em 1956 no Japão, onde um surto de septicemia em uma fazenda comercial de truta arco-íris (*Oncorhynchus mykiss*) foi caracterizado (HOSHINA; SANO; MORIMOTO, 1958). Desde então, diversas espécies de estreptococos têm sido associadas a processos patogênicos, bem como, um número crescente de espécies de peixes e outros organismos aquáticos

cultivados têm sido descritos como susceptíveis a essa enfermidade (AGNEW; BARNES, 2007; HASSON et al., 2009; ROMALDE et al., 2008).

Pertencente à família Streptococcaceae, o gênero *Streptococcus* abrange uma ampla gama de espécies bacterianas associadas a processos etiológicos em seres humanos e animais, bem como, microrganismos saprofíticos ou não patogênicos. Atualmente, 67 espécies e 12 subespécies bacterianas são descritas como pertencentes a esse gênero (GLAZUNOVA; RAOULT; ROUX, 2009). Os *Streptococcus* são cocos Gram positivos, com diâmetro de 0,5-2,0 μm, organizados em pares ou cadeias lineares curtas, catalase negativos, não móveis e não formadores de esporos. Além das características fenotípicas, o tipo de hemólise e a sorotipagem baseada na antigenicidade de polissacarídeos capsulares (20 grupos, denominados Grupos de Lancefield) têm sido empregadas para caracterizar as diferentes espécies de *Streptococcus* (SHEWMAKER et al., 2007).

Seis espécies têm sido descritas como os principais agentes etiológicos causadores de septicemia e meningoencefalite em peixes: *Streptococcus iniae*, *Streptococcus agalactiae*, *Streptococcus parauberis*, *Streptococcus dysgalactiae*, *Streptococcus phocae* e *Streptococcus ictaluri* (HERNÁNDEZ; FIGUEROA; IREGUI 2009; NETTO; LEAL; FIGUEIREDO 2011; FIGUEIREDO et al., 2012). Dentre as principais espécies patogênicas para peixes destaca-se *S. iniae* e *S. agalactiae*, as quais causam grandes prejuízos em vários países incluindo o Brasil (EVANS et al., 2002; ZHOU et al., 2008; MIAN et al., 2009).

4.3 *Streptococcus agalactiae* - biologia

O gênero Streptococcus é formado por bactérias Gram-positivas, anaeróbias facultativas, esféricas ou ovóides com menos de 2μm de diâmetro,

que crescem em pares ou em cadeias de tamanhos variados, não formam esporos e são imóveis (HOLT, et al., 1994). Em 1933, Lancefield classificou os estreptococos, dividindo-os em grupos sorológicos distintos (A-H e K-V), com base na presença de polissacarídeos capsulares. De todos os grupos sorológicos de Lancefield, os grupos A, B, C e G são os mais comumente encontrados em humanos (BISNO; RIJN, 1995).

O *Streptococcus agalactiae*, única espécie do grupo B de Lancefield (estreptococos do grupo B - EGB), foi isolado inicialmente, em 1887, de quadros de mastite bovina (BISHARAT et al., 2004). Em geral, esta bactéria é β-hemolítica em meio ágar-sangue e apresentam-se como células esféricas ou ovóides, Gram-positivas, catalase-negativas que crescem em cadeia em meio líquido e algumas linhagens podem produzir pigmentos laranja ou amarelo. *S. agalactiae* é uma bactéria imóvel, não formadora de esporos, anaeróbia facultativa e produz ácido lático como produto final do metabolismo de carboidratos (NIZET, 2002).

S. agalactiae é um microrganismo amplamente encontrado colonizando o trato digestório e genitourinário de seres humanos e animais (YILDIRIM; LÄMMLER; WEIS, 2002; MAIONE et al., 2005). Esta bactéria também é encontrada na glândula mamaria de vários ruminantes, ocasionando mastite (BROCHET et al., 2006). Diferenças entre isolados de seres humanos e animais já foram descritas, sendo que a maioria das linhagens de seres humanos é hemolítica e não utiliza a lactose como fonte de energia. Porém, já foi descrito que linhagens isoladas de bovinos não produzem hemólise em ágar sangue e utilizam a lactose (YILDIRIM; LÄMMLER; WEIS, 2002). Adicionalmente, linhagens isoladas de peixes também já foram descritas como não hemolíticas (FIGUEIREDO et al., 2006).

São utilizados como principais carboidratos no metabolismo energético de *S. agalactiae* glicose, maltose, ribose, sacarose e trealose. Não há hodrólise

de esculina, a utilização de glicerol como fonte de energia é realizada apenas na presença de oxigênio e a via das pentoses está presente (HOLT, 1994). Embora sejam presente apenas as vias de biossíntese dos aminoácidos alanina, serina, glicina, glutamina, aspartato, asparagina e treonina, há uma grande quantidade de transportadores do tipo ABC relacionados ao transporte destes aminoácidos e também ao transporte de fontes de carbono, demonstrando uma grande diversidade metabólica da espécie o que reflete em sua capacidade de adaptação a diferentes ambientes (GLASER, et al., 2002).

4.3.1 *Streptococcus agalactiae* - A doença nos principais hospedeiros

Desde a década de 1930, a bactéria *Streptococcus agalactiae,* é uma das principais causadoras de doenças em seres humanos e animais. Adicionalmente, este microrganismo também é frequentemente encontrado como membro da microbiota comensal dos sistemas digestório e genito-urinário em seres humanos (MAIONE et al., 2005).

Este patógeno acomete principalmente seres humanos, bovinos e peixes, entretanto está associado com casos de doenças em vários outros hospedeiros como aves, camelos, caninos, equinos, felinos, rãs, hamsters, camundongos, primatas, dentre outros (ELLIOT et al., 1990; YILDIRIM et al., 2002; YILDIRIM; LÄMMLER; WEIS, 2002 ; HETZEL et al., 2003; JOHRI et al.2006; GARCIA, et al. 2008). Essa bactéria é responsável por casos de pneumonia, septicemia e meningite em seres humanos neonatos, e possui alta taxa de morbidade em gestantes e mortalidade em adultos imunocomprometidos (MAIONE, et al., 2005; JOHRI et al. 2006). Na medicina veterinária esse microrganismo tem se destacado como causador de mastite clínica e subclínica em bovinos (KEEFE, 1997; WAAGE et al., 1999). Patógeno emergente na aquicultura, essa bactéria é responsável por altas taxas de mortalidade em casos

de septicemia e meningoencefalite em peixes de água doce, marinhos e estuarinos (EVANS et al., 2002). Casos de infecção por *S. agalactiae* foram descritas em mais de 20 espécies de peixes (OLIVARES-FUSTER et al., 2008).

No Brasil, o primeiro relato da ocorrência da doença em peixes foi no ano de 2003, onde foram identificados surtos de estreptococoses em tilapiculturas no Norte do Paraná (SALVADOR et al., 2003; 2005). Posteriormente, infecções causadas por *S. agalactiae* foram relatadas em tilapiculturas nos estados do Paraná, São Paulo, Espírito Santo, Minas Gerais, Bahia, Ceará (FIGUEIREDO et al., 2006; MIAN et al., 2009). Segundo dados da casuística do AQUAVET, surtos da doença em peixes também já ocorreram nos estados de Santa Catarina, Mato Grosso, Pernambuco e Alagoas. Atualmente,esse patógeno é considerado o principal risco sanitário para criações comerciais de tilápia do Nilo (*Oreochromis niloticus*) no país. Taxas de mortalidades elevadas têm sido verificadas em surtos causados por *S. agalactiae* em tilápias do Nilo e os principais fatores de risco para ocorrência de surtos são aumento da temperatura da água (acima de 27 °C) e manejo intensivo como, por exemplo, altas densidades de estocagem (MIAN et al., 2009). Medidas de manejo tais como processos de seleção e classificação também têm sido caracterizados como desencadeadores de surtos para tilápias do Nilo cultivadas em tanques-rede.

Os casos de infecção natural por *S. agalactiae* são observados principalmente em peixes adultos, entretanto, em condições experimentais a doença pode acometer alevinos e juvenis (EVANS et al., 2002; MIAN et al., 2009). Os principais sinais clínicos verificados em peixes infectados são melanose, taquipneia, anorexia, excitabilidade, natação errática e em rodopios, rigidez dorsal, exoftalmia unilateral ou bilateral, ascite e morte súbita. Em tilápias, as principais alterações patológicas observadas em casos de

estreptococose por *S. agalactiae* são pericardite, epicardite, miocardite, endocardite, meningite e septicemia (CHEN; CHA; BOWSER, 2007).

Um dos principais métodos para controlar estreptococoses nas pisciculturas é o uso de antibioticos via oral. Entretanto, como a anorexia é umas das primeiras alterações fisiológicas induzidas pela infecção, essa medida terapêutica é limitada. Este procedimento evita a ocorrência da doença nos peixes não infectados, debela as infecções dos animais que se encontram no início da infecção e dos portadores assintomáticos, mas não cura os peixes que já estão apresentando sinais clínicos (HEUER et al., 2009; RATTANACHAIKUNSOPON; PHUMKHACHORN, 2009).

Como o tratamento pós-infecção não é efetivo nos casos de estreptococose em peixes, é desejável a utilização de métodos imunoprofiláticos. Assim, a vacinação contra a doença nas pisciculturas é a alternativa mais viável para prevenção e controle das infecções por *Streptococcus.* Diversas vacinas têm sido desenvolvidas contra diferentes *Streptococcus* patogênicos para peixes, sendo que, atualmente no mundo já existem vacinas de bacterians comercializadas contra *S. iniae* e *S. agalactiae* (EVANS et al., 2004; SHOEMAKER et al., 2006; 2010). Embora ainda não comercializadas, já existem vacinas contra *S. dysgalactiae* (Patente n° JP2007326794-A) e *S. phocae* (Patentes n° WO2005053716-A1; DK200600871-A; NO200602469-A).

No Brasil, foi lançada a vacina de bacterina AQUAVAC Strep Sa, a qual está sendo comercializada desde 2012. Apesar de ainda não haver estudos na literatura demonstrando a eficácia desta, resultados prévios demonstraram redução na mortalidade em fazendas produtoras de tilápia. Porém, ainda há necessidade de mais estudos sobre as relações entre variabilidade genética, perfil antigênico e capacidade de proteção contra desafio heterólogo induzida por esta vacina. Portanto, tais produtos devem ser testados frente a várias outras amostras de bactérias isoladas no país para uma melhor verificação de sua efetividade.

4.4 Patogenia

Para causar a doença, a bactéria precisa ter subsídios para aderir e invadir os tecidos do hospedeiro além de necessitar de mecanismos de evasão ou escape do sistema imune (MITCHELL et al., 2003). De modo geral, a patogenia de todas as doenças causadas por *S. agalactiae* ainda não é totalmente elucidada, uma vez que apenas em seres humanos este pode causar diversas enfermidades tais como meningite, endocardite, infecções na pele, cistite, artrites sépticas dentre outras doenças (MACHADO et al., 2012; ULETT, et al., 2012; IMAM, et al., 2012; SINGH et al., 2012; TIAN et al., 2012).

S. agalactiae é um dos principais patógenos causadores de mortalidade em seres humanos neonatos, uma vez que coloniza uma significativa parcela do sistema genital feminino de mulheres gestantes, favorecendo a contaminação do neonato no momento do parto podendo causar pneumonia, sepse e meningite (RAJAGOPAL, 2009). Após a contaminação do neonato, o patógeno chega ao pulmão onde ocorre a invasão das células epiteliaise endoteliais, posterior bacteremia e finalmente atinge o sistema nervoso central. Devido à importância da doença em neonatos, a grande maioria dos estudos é direcionada para a patogênese da doença em seres humanos e principalmente para a colonização no tecido da mucosa vaginal, pulmonar e celulas endoteliais (DORAN; NIZEL, 2004).

Diversos fatores de virulência e mecanismos envolvidos na patogênese da doença em seres humanos têm sido descritos e caracterizados. Os principais fatores de virulência já descritos para *S. agalactiae* são: as adesinas, como as proteínas de ligação ao fibrinogênio FbsA e FbsB, proteína de ligação a fibronectina pavA, C5a peptidase (*scpB*), BibA, proteína de ligação a laminina, e as proteínas de pílus; as invasinas β-hemolisina/citolisina (*cylE*), fator CAMP (*cfb*), proteína C-α e proteína de superfície rib; e proteínas com função de evasão

do sistema imune tais como proteína c-β, genes da cápsula, C5a peptidase, BibA e serina protease cspA (SANTI et al., 2007; RAJAGOPAL, 2009; LIN et al., 2011).

Como observado, alguns dos fatores de virulência descritos de *S. agalactiae* possuem mais de uma função, como por exemplo, as proteínas BibA e *scpB*. Esta proteína quando deletada resulta em decréscimo na capacidade da cepa em aderir a células epiteliais de cérvix e pulmão e possui menor virulência quando comparada à linhagem parental. Outro fenômeno observado na linhagem mutante é menor capacidade de resistir à fagocitose e sobrevier em sangue humano (SANTI et al., 2007). A proteína C5a peptidase primeiramente foi descrita como uma imunoevasina, a qual cliva ocomponente c5a do sistema complemento inibindo a sua ativação (JARVA, 2003). Adicionalmente, esta proteína também teve sua função comprovada na invasão de células epiteliais e ligação a fibronectina (BECKMANN et al., 2002; CHENG, et al., 2002).

Um passo importante da patogênese de *S. agalactiae* em humanos é atravessar a barreira hematoencefálica. Contudo, poucos estudos comprovaram especificamente afunção de genes neste passo da patogênese (MAISEY;DORAN; NIZET, 2008; TAZI et al., 2010). A expressão do gene *hvgA*, um alelo especifico do gene *bibA* encontrado em linhagens hipervirulentas do ST-17, foi relatado ser necessária para a virulência diferenciada destas amostras. Neste estudo foi relatado que este gene é responsável pela adesão e invasão da barreira hematoencefálica resultando no quadro de meningite (TAZI et al., 2010). Outra proteína relacionada a esta etapa da patogênese é a proteína Srr1 (serine-rich repeat glycoprotein Srr1), a qual interage diretamente com o fibrinogênio humano e esta interação foi relatada ser importante para a invasão do sistema nervoso central e subsequente progressão da doença (SEO et al., 2012).

Embora existam fatores de virulência com função já descrita para *S. agalactiae*, existe uma porção significativa de proteínas pouco caracterizadas nos genomas de linhagens desta espécie. Com o intuito de predizer novos fatores de virulência em potencial, Lin et al. (2011), utilizando padrões nas sequencias de genes analisados por algoritmos de informatica, sugeriu mais de 10 novos genes como prováveis fatores de virulência relacionados à adesão, invasão e evasão do sistema imune.

Em peixes a patogênese dos processos infecciosos causados por essa bactéria é pouco compreendida. Apesar de ser a mesma espécie bacteriana, as amostras isoladas de peixes apresentam padrão genético distinto das isoladas de seres humanos e bovinos (PEREIRA et al., 2010). Este mesmo trabalho demonstrou que, em condições experimentais, as amostras isoladas de seres humanos são capazes de infectar alevinos de tilápia do Nilo. Porém, esses isolados apresentam virulência significativamente menor que a das amostras isoladas de peixes. Esses dados sugerem que isolados de diferentes hospedeiros podem compartilhar fatores de virulência e mecanismos de adesão e invasão, envolvidos na patogênese das infecções. Esta sugestão é ainda suportada pelo fato de que o quadro clínico de bacteremia e infecção do sistema nervoso central são semelhantes em seres humanos e peixes.

4.4.1 Ilhas de patogenicidade e fatores de virulência

Fatores de virulência de *S. agalactiae* já foram descritos estarem relacionados à transferência horizontal de genes, onde blocos de genes são inseridos no genoma bacteriano (RICHADS et al., 2011). Estas regiões genômicas com potencial de serem transferidas são denomaminadas ilhas genômicas, e mais especificamente ilhas de patogenicidade quando carream genes relacionados à virulência e patogenicidade (SOARES et al., 2012). Este é

um importante mecanismo relacionado à grande variação encontrada em populações bacterianas garantindo a capacidade destes microrganismos se adaptarem a uma grande diversidade de ambientes (DOBRINDT; HACKER, 2001).

As ilhas de patogenicidade (PAIs), uma classe de ilhas genômicas (GEIs), são responsáveis pela virulência e plasticidade genômica das bactérias patogênicas, propriedade dinâmica do genoma que pode envolver adição, perda ou rearranjo do DNA. Os genes de virulência, principal característica que distingue uma linhagem patogênica de uma avirulenta, estão inseridos de forma frequente nas PAIs (KARAOLIS et al., 1998). Este termo, ilha de patogenicidade, foi descrito pela primeira vez em 1990 (HACKER et al., 1990) e atualmente, o termo é utilizado para descrever regiões cromossômicas que foram adquiridas pelo organismo por transferência horizontal de genes. Estas regiões possuem diferente percentual de conteúdo G+C e apresentam genes que codificam fatores de virulência (GAL-MOR; FINLAY, 2006). Adicionalmente, as PAIs possuem outras características tais como ocupam regiões genômicas relativamente grandes; apresentam diferente uso de códon; eventos de deleção podem ocorrer com frequências distintas; são passíveis de transferência, possuem estruturas do tipo mosaico e estão ausentes em organismos não patogênicos do mesmo gênero ou espécie filogeneticamente próxima (SCHMIDT; HENSEL, 2004).

Em *S. agalactiae*, os genes relacionados à virulência *lmb* (proteína ligadora da laminina) e *scpB* (C5a peptidase) já foram descritos como localizados em um transposon, o qual está presente na maioria das linhagens isoladas de seres humanos e em apenas aproximadamente 20% dos isolados de bovinos (FRAKEN et al., 2001; AL SAFADI et al., 2010).

Três operons de ilhas de pilus (PI) são descritos em *S. agalactiae* sendo importantes fatores de virulência relacionados à adesão e invasão em células

endoteliais (MAISEY, et al., 2008). Estas ilhas tem sido relacionadas a transferência horizontal de genes em *S. pyogenes*, *S. pneumoniae* e inclusive em *S. agalactiae* (MANDLIK et al., 2008). Pelo menos um dos operons de ilhas de pilus esta presente no genoma de linhagens de *S. agalactiae*, sendo já relacionado à presença das ilhas PI-1 e PI-2a a colonização do sistema genital feminino e a doença invasiva em adultos. Já a combinação das ilhas PI-1 e PI-2b foi relacionada à infecção em neonatos (MARTINS et al., 2012).

Outros fatores de virulência já foram descritos em ilhas de patogenicidade, como por exemplo, os genes *cylE* e *cfb* (GLASER et al., 2002). Entretanto, tem-se demonstrado que não é importante apenas a presença do gene no genoma da linhagem, uma vez que mesmo o gene estando presente a regulação de sua expressão pode ser diferenciada de acordo com diferentes linhagens de diferentes sorotipos (JIANG, et al., 2008). Suportando isso, um estudo recente demonstrou que a expressão de vários genes de fatores de virulência e a capacidade de aderir a células epiteliais é diferenciada em linhagens de distintos sorotipos (SHARMA, et al., 2012). Diante disso, sugere-se que a regulação ("on/off") da expressão destes genes ainda precisa ser mais estudada sobre diferentes contextos, inclusive utilizando "*arrays"* de expressão gênica analisando a interação patógeno-hospedeiro.

Quando se analisa as regiões genômicas caracterizadas como ilhas presentes em *S. agalactiae*, concluiu-se que esta é uma espécie bacteriana com elevada plasticidade, com variada distribuição de elementos móveis em diferentes linhagens (BROCHET et al., 2006, TETTELIN et al., 2005) e que alguns destes elementos podem ser característicos de isolados de um hospedeiro específico (RICHARDS et al., 2011). Adicionalmente pouco se sabe sobre os fatores de virulência e ilhas de patogenicidade na população de *S. agalactiae* provenientes de peixes.

4.5 Diversidade genética e estudos genômicos em *Streptococcus agalactiae*

A variabilidade genética de *Streptococcus agalactiae* isolados de seres humanos já é bem conhecida tendo-se estabelecido clones com diferentes habilidades de virulência (BISHARAT et al., 2004; SUKHNANAND et al., 2005; OLIVEIRA et al., 2006; CORRÊA et al., 2010). Apesar de pertencerem à mesma espécie bacteriana, de forma geral, linhagens isoladas de seres humanos e bovinos são consideradas populações não relacionadas (DOGAN et al., 2005; BISHARAT et al., 2004; SORENSEN et al., 2010).

Por análise de MLST (Multilocus Sequence Typing) demonstrou-se que amostras de origem bovina apresentaram padrão genético similar a isolados de origem humana hipervirulento (ST-17), sugerindo que o clone hipervirulento de humanos tem seu ancestral em isolados de origem bovina (BISHARAT et al., 2004). Contudo esta evidencia foi contestada por Sorensen et al. (2010) suportando a hipótese que isolados de seres humanos e bovinos constituem populações distintas. Entretanto, alguns estudos sugerem limitada transmissão interespécie (SUKHNANAND et al., 2005; DOGAN et al., 2005).

Adicionalmente o mesmo padrão genético de uma amostra de origem humana e bovina foi descrito (OLIVEIRA et al., 2006) e um estudo complementar a este utilizou esta mesma amostra isolada de mastite bovina demonstrou experimentalmente em camundongos que esta amostra foi mais virulenta quando comparada a cepa isolada de ser humano (CORREA, et al., 2010). Estes dados suportam a hipótese que algumas amostras de origem bovina podem quebrar a barreira interespécie e infectar humanos e que o contrário também pode ocorrer, porém sugerindo este ser um evento raro.

Poucos trabalhos de caracterização molecular foram realizados com amostras isoladas de peixes. Um estudo comparando por Amplified Fragment Length Polymorphism (AFLP) isolados de peixes, seres humanos e bovinos

demonstrou que os isolados de peixes pertenceram a genótipos distintos dos isolados de outros hospedeiros (OLIVARES-FUSTER et al., 2008). Outro estudo comparando por MLST e sorotipagem molecular de amostras de *S. agalactiae* isoladas de peixes, golfinhos, seres humanos e bovinos demonstrou que isolados de golfinho e de peixes do Kuwait possuem o mesmo padrão molecular por estas duas técnicas que amostras isoladas de seres humanos no Japão, sugerindo transmissão cruzada entre estes hospedeiros (EVANS et al., 2008). Pereira et al. (2010) demonstraram que isolados de seres humanos, bovinos e peixes possuem padrões genéticos não relacionados por Pulsed-Field Gel Electrophoresis (PFGE) e que, apesar de possuírem padrões genéticos distintos, todas as amostras de *S. agalactiae* isoladas de seres humanos e algumas de bovinos foram capazes de infectar tilápias do Nilo, causando mortalidade.

Encontra-se disponível no NCBI o genoma completo de três amostras de *S. agalactiae*, isoladas de seres humanos e pertencentes a diferentes sorotipos. Além desses, o genoma "*draft"* de outras sete cepas adicionais dessa espécie (também isoladas de seres humanos) foram sequenciados (GLASER et al., 2002; TETTELIN et al., 2002, 2005; SINGH et al., 2012). Um estudo o qual utilizou oito genomas de *S. agalactiae* isolados de seres humanos, demonstrou que o pan-genoma da espécie está "aberto" e que a cada nova amostra sequenciada de *S. agalactiae* espera-se seradicionadoem média 33 novos genesao pan-genoma da espécie (TETTELIN et al., 2005). Demonstrando de forma mais evidente a plasticidade genômica desta espécie, o genoma de uma linhagem de *S. agalactiae* isolada de mastite bovina foi sequenciado recentemente e resultou na descrição de 183 novos genes linhagem específicos e que, cerca de 85% destes genes encontravam-se em ilhas genômicas. Dentre esses genes destaca-se o operon de nisina, que contribui para a infecçãona glândula mamária em *S. uberis*; e o operon de utilização de frutose-lactose com elevada similaridade com

S. dysgalactiae subsp. *dysgalactiae*, relacionado à utilização de frutose ou lactose como fonte de energia (RICHARDS et al., 2011). Recentemente foram publicados os genomas de três linhagens de *S. agalactiae* isoladas de peixes, sendo duas destas linhagens foram isoladas na China e são pertencentes ao sorotipo Ia (LIU; ZHANG; LU, 2012; WANG et al., 2012) e uma pertencente ao sorotipo Ib isolada em Honduras (DELANNOY et al., 2012). Ainda assim, pouco se sabe sobre os mecanismos moleculares envolvidos na patogênese e virulência presentes no genoma das linhagens isoladas de peixes.

A ampla faixa de habitat e de hospedeiros de *S. agalactiae* pode favorecer a transferência horizontal de genes, resultando em um maior pan-genoma da espécie em relação a *S. pyogenes* (LEFÉBURE; STANHOPE, 2007), o que torna essencial o conhecimento do genoma de amostras de *S. agalactiae* isoladas de outros hospedeiros.

4.6 Sequenciamento de próxima geração

Após a criação do método de sequenciamento didesoxi, por Sanger, Nicklen e Coulson (1977), o genoma completo de mais de 2100 bactérias já foram finalizados (http://www.genomesonline.org/) em 15 de fevereiro de 2013. Apesar da capacidade inicial deproduzir poucos milhares pares de bases por ano, a automatização do processo e aumento no número de capilares nas últimas décadas permitiu um incremento significativo na agilidade, possibilitando a elucidação de centenas de genomas de diversos organismos. Na última década, plataformas de sequenciamento que utilizam a tecnologia de nova geração ("*Next-Generation Sequencing*") têm sido disponibilizadas no mercado. Essas técnicas inovadoras são baseadas em nanotecnologia e na construção de bibliotecas de fragmentos ou pareadas de DNA que não dependem da clonagem em vetores, abrindo grandes horizontes para estudos genéticos (SHENDURE; JI,

2008). Os principais atributos desses sequenciadores são o custo, expressivamente inferior, e a rapidez com que os dados são gerados. Esses possuem a capacidade de processar milhões de sequências em uma única corrida, enquanto os sequenciadores convencionais de capilares processam apenas 96 ou 384 sequências simultaneamente. Além disso, uma quantidade pequena de DNA é requerida para a construção de bibliotecas (apenas alguns microgramas) (MARDIS, 2008; SHENDURE; JI, 2008).

O modelo 454 FLX da Roche foi o primeiro sequenciador de nova geração, introduzido no mercado no ano de 2004. Essa plataforma utiliza o princípio do pirosequenciamento, ou seja, a liberação de uma molécula de pirofosfato enquanto a DNA polimerase incorpora os nucleotídeos (sequenciamento por síntese de DNA). Essa reação produz a quebra da oxiluciferina pela luciferase, produzindo radiação luminosa em um comprimento de onda específico (MARGULIES et al., 2005). A imagem emitida pela luciferase é então gravada enquanto um determinado nucleotídeo é adicionado. Atualmente esta plataforma possui como principal vantagem o tamanho da leitura gerada (tamanho médio de 400 pb) o que facilita o processo de montagem genômica, porém possui como principal limitação elevada taxa de erro (0,38%) em regiões de homopolímeros (LOMAN et al., 2012). Outro modelo de sequenciador de nova geração é a plataforma Illumina. Essa plataforma foi introduzida no mercado no final do ano de 2006 e assim como o sequenciamento de Sanger e pela plataforma 454, o sequenciamento é realizado por síntese usando DNA polimerase e nucleotídeos terminadores marcados com diferentes fluoróforos. A inovação dessa plataforma consiste na clonagem *in vitro* dos fragmentos de DNA em uma plataforma sólida de vidro, processo também conhecido como PCR em ponte (bridge PCR). Esta plataforma pode produzir leituras de até 150 bp e possui como principal vantagem a elevada qualidade dos

dados gerados e como limitação o tamanho reduzido das leituras geradas (METZKER, 2010).

A plataforma SOLiD ("Sequencing by Oligo Ligation and Detection") da Applied Biosystems foi lançada no mercado em outubro de 2007. Essa plataforma utiliza a incorporação de dinucleotídeos marcados por meio da DNA ligase, seguida pela excitação do fluoróforo (o sinal emitido é captado por sensores) e a incorporação dos dinucleotídeos seguintes. Essa leitura gera um código de cores que é analisado por ferramentas de bioinformática e convertida à sequência de letras. Cada corrida no sequenciamento da plataforma SOLiD leva aproximadamente 5 dias, e produz de 3 a 4 Gb de sequências com o comprimento variando de 50 a 75 pb. O sequenciamento baseado na ligação de oligonucleotídeos ocorre por meio do anelamento de um primer universal do SOLiD, seguido pela ligação de uma sonda marcada que é detectada pela máquina em cada ciclo. O mecanismo de detecção de dinucleotídeo garante uma correção de erro, melhorando a qualidade dos dados (MARDIS, 2008; METZKER, 2010).

Outra plataforma da Applied Biosystems que foi lançada em 2011 é o Ion Torrent. Esta plataforma utiliza mecanismo semelhante ao da 454 da Roche, no entanto, ao invés de usar o pirofosfato, utiliza o íon de hidrogênio que é liberado ao ser adicionado um novo nucleotídeo. Esta plataforma atualmente gera leituras de tamanho médio de até 200 bp, porém, também possui como principal limitação a significativa taxa de erro (1,5%) em regiões de homopolímeros (LOMAN et al., 2012).

O pequeno tamanho da leitura nas plataformas SOLiD e Illumina é o principal problema para a montagem *de novo* de genomas, principalmente em regiões de repetição comumente encontradas em genomas bacterianos. A possibilidade de se trabalhar nestas plataformas com bibliotecas pareadas é o ponto chave para se resolver em grande parte este problema. Estas bibliotecas

pareadas constituem-se de dois fragmentos/leituras que são sequenciados com uma distância conhecida uma leitura da outra, obtendo-se assim milhões de leituras em pares com esta distância determinada. Porém, deve haver um critério para a escolha do tamanho do inserto entre as leituras pareadas, pois este pode influenciar de forma significativa na montagem do genoma (WETZEL; KINGSFORD; POP, 2011; CAHILL et al., 2010).

Estas plataformas de sequenciamento de próxima geração vêm sendo amplamente empregadas no sequenciamento genômico de várias espécies bacterianas, como porexemplo: quatro genomas de *Corynebcterium pseudotuberculosis* (SILVA et al., 2011, 2012; CERDEIRA et al., 2011a, 2011b) sequenciados utilizando exclusivamente a plataforma SOLiD V2 e V3; 2 genomas de *Corynebcterium pseudotuberculosis* (RUIZ, et al., 2011; TROST et al., 2010) e um de *Lactobacillus rhamnosus* (PRAJAPATI et al., 2012) que utilizaram exclusivamente a plataforma 454; um genoma de *S. agalactiae* isolado de mastite bovina (RICHARDS et al., 2011) e um de *S. parauberis* isolado de peixe (NHO et al., 2011) que utilizaram a plataforma 454 e na finalização da montagem do genoma empregaram a plataforma de Sanger; e uma cepa de *Legionella pneumophila* (AMARO et al., 2012) e uma cepa de *Brucella suis* (KE et al., 2012) em que apenas a plataforma Illumina foi utilizada.

4.7 Vacinologia reversa e imunoinformática

A vacinologia tradicional utiliza principalmente a inativação/atenuação do patógeno e a identificação de antígenos protetores que podem ser utilizados nas denominadas vacinas de subunidades. Para o desenvolvimento de vacinas de subunidades há a necessidade de se cultivar o patógeno para posterior identificação dos possíveis componentes antigênicos. Porém, somente os componentes antigênicos expressos nas condições de laboratório e em

quantidade suficiente são identificados e podem ser analisados (RAPPUOLI, 2000). A produção de vacinas pelos métodos convencionais apresenta limitações que incluem: patógenos não cultiváveis; bactérias intracelulares que requerem cultivos celulares específicos que têm um alto custo; falhas na produção de bacterinas ou inativação dos microrganismos, reversão da virulência no caso de vacinas vivas atenuadas; antígenos não expressos ou expressos em pequenas quantidades em condições laboratoriais (MOVAHEDI; HAMPSON, 2008).

Atualmente, com o grande número de genomas de patógenos disponíveis em bancos de dados, pode-se primeiramente fazer predição *in silico* de genes candidatos vacinais como forma de selecionar candidatos a serem clonados, expressados e testados como vacinas (vacinologia reversa). Para isso, esta predição baseia-se nas proteínas exportadas (proteínas ancoradas na membrana ou secretadas), por estas estarem fortemente relacionadas com a interação patógeno-hospedeiro como: aderência e invasão as células/tecidos do hospedeiro, dano aos tecidos do hospedeiro, resistência e evasão ao sistema imune do hospedeiro (SANTOS et al., 2011). Como vantagem da vacinologia reversa destaca-se a utilização de todos os genes do genoma do organismo, independente de sua expressão. E como limitações deve-se ressaltar a incapacidade de predizer antígenos não protéicos, como polissacarídeos e glicolipídeos; não predizer se os alvos selecionados são imunogênicos ou não e o grande número de alvos normalmente gerados na predição (RAPPUOLI, 2000,; 2001).

A vacinologia reversa foi primeiramente testada para identificar antígenos como potenciais alvos para vacina contra *Neisseria meningitidis* (RAPPUOLI, 2001). Foram preditos aproximadamente 600 genes candidatos, dos quais 350 foram clonados, expressados em *E. coli* e testados *in vivo* em camundongos. Cinco genes selecionados por este método foram combinados em uma vacina multicomponente, a qual induziu anticorpos bactericidas contra 90%

(dependendo do adjuvante utilizado) contra várias amostras representativas da diversidade mundial deste patógeno (GIULIANE et al., 2006).

A vacinologia reversa também obteve promissores resultados quando empregada na predição de alvos vacinais contra vários outros patógenos, como *Streptococcus agalactiae* (MAIONE et al., 2005), *Mycobacterium tuberculosis* (VIZCAÍNO et al., 2010), *Porphyromonas gingivalis* (ROSS et al., 2001), *Streptococcus pneumoniae* (WIZEMANN et al., 2001), *B. anthracis* (ARIEL et al., 2002), *Chlamydia pneumoniae* (MONTIGIANI et al., 2002), *Edwardiella tarda* (SRINIVASA RAO; LIM; LEUNG, 2003) e *Leptospira interrogans* (GAMBERINI et al., 2005).

Métodos para predizer a localização subcelular de proteínas em bactérias Gram-positivas geralmente resultam na classificação em quatro compartimentos: citoplasma, membrana, parede celular, e secretada. Os resultados das análises *in silico* utilizando estes algoritmos podem ser complementados a partir de dados da composição ou homologia com proteínas de localização conhecida, estratégia utilizada pelo software PsortB (GARDY et al., 2005), e integrados na predição final de localização das proteínas. Porém, estes algoritmos não são capazes de predizer se as proteínas classificadas como de membrana possuem partes expostas na superfície celular que podem ser reconhecidas pelo sistema imune. Levando em consideração esta limitação destes algoritmos, o software SurfG+ foi desenvolvido, criando a categoria possivelmente exposta na superfície (PSE) e demonstrando melhor acurácia (BARINOV et al., 2009).

Em geral, trabalhos que utilizam a vacinologia reversa para predição de alvos vacinais resultam em um grande número de proteínas alvo, o que torna a próxima etapa de clonagem, expressão e testes *in vivo* e *in vitro* dispendiosa e demorada (SANTOS et al., 2011). Uma solução para minimizar aquantidade de alvos vacinais da etapa experimental está na imunoinformática. A imunoinformática é uma alternativa para reduzir o número de alvos vacinais a

seres testados, uma vez que tenta identificar candidatos vacinais imunogênicos, ou seja, com maior probabilidade de estimular uma resposta imune protetora no hospedeiro (LARSEN et al., 2007; LUNDEGAARD et al., 2008). A imunoinformática tem como objetivo analisar dados genômicos *in silico* (sequencias de proteínas) utilizando abordagens computacionais resultando em interpretações com significado imunológico (KLEINSTEIN, 2008). Nesta análise busca-se de peptídeos que possuem maior probabilidade de ligação às moléculas Major Histocompatibility Complex (MHC) classes I e II, moléculas estas com função essencial na montagem da resposta imune. Os peptídeos ligantes a essas classes de moléculas se diferenciam basicamente pela quantidade de aminoácidos e por se ligarem com conformação tridimensional ou linear. Peptídeos ligantes ao MHC I variam de 8 a 11 aminoácidos e são lineares. Peptídeos ligantes ao MHC II são maiores (13 a 17 aminoácidos) e denominados não lineares, ou seja, apresentam uma conformação tridimensional para se ligarem à célula apresentadora de antígenos. Essas diferenças em relação aos peptídeos ligantes a MHC I e MHC II definem o nível de dificuldade de identificá-los, sendo mais fácil e preciso identificar peptídeos lineares (LUNDEGAARD et al., 2008). Epitopos com maior probabilidade de se ligar a molécula MHC I são relacionados a montagem de uma resposta imune celular, conferindo imunidade adaptativa ao hospedeiro por meio da ativação de linfócitos T CD8+, enquanto epitopos ligantes ao MHC II são relacionados a montagem de uma resposta imune humoral por meio da ativação de linfócitos T CD4+ ou “helper” (LARSEN et al., 2007).

O software Vaxign identifica alvos vacinais predizendo a probabilidade de adesão às moléculas de MHC I e MHC II, além de excluir proteínas com similaridade a proteínas do hospedeiro (humanos e camundongos) e excluir alvos com sequencias similares em bactérias não patogênicas (HE; XIANG; MOBLEY, 2010). Com a utilizaçãoda imunoinformática uma maior acurácia é

garantida, com menor taxa de falsos positivos e reduz significativamente o número de candidatos alvos a serem testados (RAI et al., 2012).

5 CONCLUSÕES

Como conclusões destacam-se:

a) As plataformas de sequenciamento de nova geração possibilitam o sequenciamento de genomas bactérias em menor tempo e custo, e com isso, tornando viável o sequenciamento de patógenos de interesse;
b) Com dados genômicos gerados é possível realizar análises comparativas com genomas disponíveis de outros hospedeiros e, com isso, melhor entender os mecanismos moleculares envolvidos na adaptação ao *habitat* e ao hospedeiro;
c) O desenvolvimento de novas estratégias vacinais se faz necessário, e com a disponibilidade de dados genômicos e auxílio da bioinformática, novos alvos vacinais podem ser preditos e posteriormente explorados.PARTE 1 CORRIGIDA

REFERÊNCIAS

AGNEW, W.; BARNES, A. C. Streptococcus iniae: an aquatic pathogen of global veterinary significance and a challenging candidate for reliable vaccination. **Veterinary Microbiology**, Amsterdam, v. 122, n. 1–2, p. 1-15, 2007.

AL SAFADI, R. et al. Enhanced expression of lmb gene encoding laminin-binding protein in Streptococcus agalactiae strains harboring IS1548 in scpB-lmb intergenic region. **PloS One**, v. 5, n. 5, p. e10794, Jan. 2010.

AMARO, F. et al. Whole-Genome sequence of the human pathogen Legionella pneumophila Serogroup 12 Strain 570-CO-H. **Journal of Bacteriology**, Washington, v. 194, n. 6, p. 1613-1614, 2012.

ARIEL, N. et al. Search for potential vaccine candidate open reading frames in the bacillus anthracis virulence plasmid pXO1: in silico and in vitro screening. **Infection and Immunity**, Washington, v. 70, n. 12, p. 6817-6827, Dec. 2002.

BARINOV, A. et al. Prediction of surface exposed proteins in Streptococcus pyogenes, with a potential application to other Gram-positive bacteria. **Proteomics**, Weinheim, v. 9, n. 1, p. 61-73, 2009.

BECKMANN, C. et al. Identification of novel adhesins from Group B streptococci by use of phage display reveals that C5a peptidase mediates fibronectin binding. **Infection and Immunity**, Washington, v. 70, n. 6, p. 2869-2876, June 2002.

BISHARAT, N. et al. Hyperinvasive Neonatal Group B streptococcus has arisen from a bovine ancestor. **Journal of Clinical Microbiology**, Washington, v. 42, n. 5, p. 2161-2167, 2004.

BISNO, A. L.; RIJN, I. van de. Classification of streptococci. In: MANDELL, G. L.; DOUGLAS; BENNETT'S. **Principles and practive of infectious diseases**. New York: Churchill Livingstone, 1995. p. 1784-1785.

BRASIL. Ministério da Pesca e Aquicultura. **Boletim estatístico da pesca e aquicultura.** Brasília, 2010. 128 p.

BROCHET, M. et al. Genomic diversity and evolution within the species Streptococcus agalactiae. **Microbes and Infection**, v. 8, n. 5, p. 1227-1243, Apr. 2006.

CAHILL, M. J. et al. Read length and repeat resolution: exploring prokaryote genomes using next-generation sequencing technologies. **PLoS ONE**, v. 5, n. 7, p. e11518, 2010.

CERDEIRA, L. T. et al. Whole-genome sequence of corynebacterium pseudotuberculosis PAT10 strain isolated from sheep in Patagonia, Argentina. **Journal of Bacteriology**, Washington, v. 193, n. 22, p. 6420-6421, 2011a.

CERDEIRA, L. T. et al. Complete genome sequence of corynebacterium pseudotuberculosis strain CIP 52.97, isolated from a horse in kenya. **Journal of Bacteriology**, Washington, v. 193, n. 24, p. 7025-7026, 2011b.

CHEN, C. Y.; CHA, C. B.; BOWSER, P. R. Comparative histopathology of Streptococcus iniae and Streptococcus agalactiae-infected tilapia. **Bulletin of The European Association of Fish Pathologists**, v. 27, n. 1, p. 2-9, 2007.

CHENG, Q. The Group B Streptococcal C5a Peptidase is both a specific protease and an invasin. **Infection and Immunity**, Washington, v. 70, n. 5, p. 2408-2413, maio 2002.

CORRÊA, A. B. A. et al.Virulence characteristics of genetically related isolates of group B streptococci from bovines and humans. **Veterinary Microbiology**, Amsterdam, v. 143, n. 2-4, p. 429-433, July 2010.

DELANNOY, C. M. J. et al. Draft genome sequence of a nonhemolytic fish-pathogenic Streptococcus agalactiae strain. **Journal of Bacteriology**, Washington, v. 194, n. 22, p. 6341-6312, Nov. 2012.

DOBRINDT, U.; HACKER, J. Whole genome plasticity in pathogenic bacteria. **Current Opinion in Microbiology**, Oxford, v. 4, n. 5, p. 550-557, 2001.

DOGAN, B. et al. Distribution of serotypes and antimicrobial resistance genes among streptococcus agalactiae isolates from bovine and human hosts. **Journal of Clinical Microbiology**, Washington, v. 43, n. 12, p. 5899-5906, 2005.

DORAN, K. S.; NIZET, V. Molecular pathogenesis of neonatal group B streptococcal infection: no longer in its infancy. **Molecular Microbiology**, Salem, v. 54, n. 1, p. 23-31, Oct. 2004.

EL-SAYED, A.-F. M. Alternative dietary protein sources for farmed tilapia, Oreochromis spp. **Aquaculture**, v. 179, n. 1, p. 149-168, 1999.

ELDAR, A. Experimental streptococcal meningo-encephalitis in cultured fish. **Veterinary Microbiology**, Amsterdam, v. 43, n. 1, p. 33-40, Jan. 1995.

ELLIOTT, J. A.; FACKLAM, R. R.; RICHTER, C. B. Whole-cell protein patterns of nonhemolytic group B, type Ib, streptococci isolated from humans, mice, cattle, frogs, and fish. **Journal Clinical Microbiology**, v. 28, p. 628-630, 1990.

EVANS, J. J. et al. Characterization of β-haemolytic Group B Streptococcus agalactiae in cultured seabream, Sparus auratus L., and wild mullet, Liza klunzingeri (Day), in Kuwait. **Journal of Fish Diseases**, Oxford, v. 25, n. 9, p. 505-513, 2002.

EVANS, J. J. et al. Efficacy of Streptococcus agalactiae (group B) vaccine in tilapia (Oreochromis niloticus) by intraperitoneal and bath immersion administration. **Vaccine**, v. 22, n. 27–28, p. 3769-3773, 2004.

EVANS, J. J. et al. Phylogenetic relationships among Streptococcus agalactiae isolated from piscine, dolphin, bovine and human sources: a dolphin and piscine lineage associated with a fish epidemic in Kuwait is also associated with human neonatal infections in Japan. **Journal of Medical Microbiology**, London, v. 57, n. 11, p. 1369-1376, 2008.

FIGUEIREDO, H. C. P. et al. Isolation and characterization of strains of Flavobacterium columnare from Brazil. **Journal of Fish Diseases**, Oxford, v. 28, n. 4, p. 199-204, Apr. 2005.

FIGUEIREDO, H. C. P. et al. Streptococcus agalactiae associado à meningoencefalite e e infecção sistêmica em tilápia-do-Nilo (Oreochromis niloticus) no Brasil. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte, v. 58, p. 678-680, 2006.

FIGUEIREDO, H. C. P. et al. *Streptococcus iniae* outbreak in Brazilian Nile tilapia (*Oreochromis niloticus*) farms. **Brazilian Journal of Microbiology**, São Paulo, v. 43, p. 576-580, 2012.

FOOD AND AGRICULTURE ORGANIZATION. **The state of world fisheries and aquaculture**. Rome, 2010. 197 p.

FRANKEN, C. et al. Horizontal gene transfer and host specificity of beta-haemolytic streptococci: the role of a putative composite transposon containing scpB and lmb. **Molecular microbiology**, Salem, v. 41, n. 4, p. 925-935, Aug. 2001.

GAL-MOR, O.; FINLAY, B. B. Pathogenicity islands: a molecular toolbox for bacterial virulence. **Cellular Microbiology**, Oxford, v. 8, n. 11, p. 1707-1719, 2006.

GAMBERINI, M. et al. Whole-genome analysis of Leptospira interrogans to identify potential vaccine candidates against leptospirosis. **FEMS Microbiology Letters**, Amsterdam, v. 244, n. 2, p. 305-313, 2005.

GARCIA, J. C. et al. Non-infectivity of cattle Streptococcus agalactiae in Nile tilapia, Oreochromis niloticus and channel catfish, Ictalurus punctatus. **Aquaculture**, v. 281, n. 1–4, p. 151-154, 2008.

GARDY, J. L. et al. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. **Bioinformatics,** Oxford, v. 21, n. 5, p. 617-623, Mar. 2005.

GIULIANI, M. M. et al. A universal vaccine for serogroup B meningococcus. **Proceedings of the National Academic Science.** v. 103, n. 29, p. 10834-10839 July, 2006.

GLASER, P. et al. Genome sequence of Streptococcus agalactiae, a pathogen causing invasive neonatal disease. **Molecular Microbiology**, Salem, v. 45, n. 6, p. 1499-1513, 2002.

GLAZUNOVA, O. O.; RAOULT, D.; ROUX, V. Partial sequence comparison of the rpoB, sodA, groEL and gyrB genes within the genus Streptococcus. **International Journal of Systematic and Evolutionary Microbiology**, Reading, v. 59, n. 9, p. 2317-2322, 2009.

HACKER, J. et al. Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extra intestinal Escherichia coli isolates. **Microbial Pathogenesis**, London, v. 8, n. 3, p. 213-225, 1990.

HASSON, K. W. et al. Streptococcosis in farmed Litopenaeus vannamei: a new emerging bacterial disease of penaeid shrimp. **Diseases of Aquatic Organisms**, v. 86, n. 2, p. 93-106, 2009.

HE, Y.; XIANG, Z.; MOBLEY, H. L. T. Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. **Journal of Biomedicine and Biotechnology**, v. 2010, Art. 297505, 2010

HERNÁNDEZ, E.; FIGUEROA, J.; IREGUI, C. Streptococcosis on a red tilapia, Oreochromis sp., farm: a case study. **Journal of Fish Diseases**, Oxford, v. 32, n. 3, p. 247-252, 2009.

HETZEL, U. et al. Septicaemia in emerald monitors (Varanus prasinus Schlegel 1839) caused by Streptococcus agalactiae acquired from mice. **Veterinary Microbiology**, Amsterdam, v. 95, n. 4, p. 283-293, 2003.

HEUER, O. E. et al. Human health consequences of use of antimicrobial agents in aquaculture. **Clinical Infectious Diseases**, Chicago, v. 49, n. 8, p. 1248-1253, 2009.

HOLT, J. G. et al. **Bergey's manual of determinative bacteriology**. 9th ed. Baltimore: The Williams & Wilkins, 1994. p. 527-558.

HOSHINA, T.; SANO, T.; MORIMOTO, Y. A. Streptococcus pathogenic to fish. **Journal of Tokyo University of Fisheries**, v. 44, p. 57-68, 1958.

IMAM, Y. Z. et al. Streptococcus agalactiae septic arthritis of the shoulder and the sacroiliac joints: a case report. **Case reports in Rheumatology**, v. 2012, p. 720297, Jan. 2012.

JARVA, H. Complement resistance mechanisms of streptococci. **Molecular Immunology**, Elmsford, v. 40, n. 2-4, p. 95-107, Sept. 2003.

JIANG, S.-M. et al. Variation in the group B Streptococcus CsrRS regulon and effects on pathogenicity. **Journal of Bacteriology**, Washington, v. 190, n. 6, p. 1956-1965, Mar. 2008.

JOHRI, A. K. et al. Group B Streptococcus: global incidence and vaccine development. **Nature Reviews:** microbiology, London, v. 4, n. 12, p. 932-942, Dec. 2006.

KARAOLIS, D. K. R. et al. A Vibrio cholerae pathogenicity island associated with epidemic and pandemic strains. **Proceedings of the National Academy of Sciences**, v. 95, n. 6, p. 3134-3139, 1998.

KE, Y. et al. Complete genome sequence of Brucella suis field strain BCB025 of sequence type ST22. **Journal of Bacteriology**, Washington, v. 194, n. 24, p. 6959, Dec. 2012.

KEEFE, G. P. Streptococcus agalactiae mastitis: a review. **The Canadian Veterinary Journal:** revue veterinaire canadienne, Otawa, v. 38, n. 7, p. 429-437, July 1997.

KLEINSTEIN, S. H. Getting started in computational immunology. **PLoS Computational Biology**, v. 4, n. 8, p. e1000128, 2008.

LARSEN, M. V. et al. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. **BMC bioinformatics**, v. 8, p. 424, Jan. 2007.

LEFÉBURE, T.; STANHOPE, M. J. Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. **Genome Biology**, v. 8, n. 5, p. R71, Jan. 2007.

LIN, F. P.-Y. et al. Computational bacterial genome-wide analysis of phylogenetic profiles reveals potential virulence genes of Streptococcus agalactiae. **PLoS ONE**, v. 6, n. 4, p. e17964, 2011.

LIU, G.; ZHANG, W.; LU, C. Complete genome sequence of Streptococcus agalactiae GD201008-001, isolated in China from Tilapia with Meningoencephalitis. **Journal of Bacteriology**, Washington, v. 194, n. 23, p. 6653, Dec. 2012.

LOMAN, N. J. et al. Performance comparison of benchtop high-throughput sequencing platforms. **Nature Biotechnology**, v. 30, n. 5, p. 434-439, May 2012.

LUNDEGAARD, C. et al. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. **Nucleic Acids Research**, v. 36, p. W509-W512, 2008. Supplement, 2.

MACHADO, C. et al. Streptococcus agalactiae endocarditis. **Portuguese Journal of Cardiology:** an official journal of the Portuguese Society of Cardiology, v. 31, n. 9, p. 619-621, Sept. 2012.

MAIONE, D. et al. Identification of a Universal Group B Streptococcus Vaccine by Multiple Genome Screen. **Science**, v. 309, n. 5731, p. 148-150, 2005.

MAISEY, H. C.; DORAN, K. S.; NIZET, V. Recent advances in understanding the molecular basis of group B Streptococcus virulence. **Expert Reviews in Molecular Medicine**, v. 10, p. e27, Jan. 2008.

MAISEY, H. C. et al. A group B streptococcal pilus protein promotes phagocyte resistance and systemic virulence. **FASEB Journal**: official publication of the Federation of American Societies for Experimental Biology, v. 22, n. 6, p. 1715-1724, June 2008.

MANDLIK, A. et al. Pili in Gram-positive bacteria: assembly, involvement in colonization and biofilm development. **Trends Microbiology**, v. 16, p. 33-40, 2008.

MARDIS, E. R. The impact of next-generation sequencing technology on genetics. **Trends in Genetics**, v. 24, n. 3, p. 133-141, 2008.

MARGULIES, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. **Nature**, v. 437, n. 7057, p. 376-380, Sept. 2005.

MARTINS, E. R. et al. Distribution of pilus islands in Streptococcus agalactiae causing human infections: insights into evolution and implication for vaccine development. **CVI**: clinical and vaccine immunology, Dec. 2012.

MARTINS, M. L. et al. Haematological alterations of Leporinus macrocephalus (Osteichtyes: Anostomidae) naturally infected by Goezia leporini (Nematoda: Anisakidae) in fish pond. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte, v. 56, n. 5, p. 640-646, Oct. 2004.

MATA, A. I. et al. Multiplex PCR Assay for Detection of Bacterial Pathogens Associated with Warm-Water Streptococcosis in Fish. **Applied and Environmental Microbiology**, v. 70, n. 5, p. 3183-3187, 2004.

METZKER, M. L. Sequencing technologies - the next generation. **Nature Reviews Genetics**, v. 11, n. 1, p. 31-46, Jan. 2010.

MIAN, G. F. et al. Aspects of the natural history and virulence of S. agalactiae infection in Nile tilapia. **Veterinary Microbiology**, Amsterdam, v. 136, n. 1-2, p. 180-183, 2009.

MITCHELL, T. J. The pathogenesis of streptococcal infections: from tooth decay to meningitis. **Nature Reviews Microbiology**, v. 1, n. 3, p. 219-230, Dec. 2003.

MONNET, C. et al. The Arthrobacter arilaitensis Re117 genome sequence reveals its genetic adaptation to the surface of cheese. **PLoS ONE**, v. 5, n. 11, p. e15489, 2010.

MONTIGIANI, S. et al. Genomic approach for analysis of surface proteins in chlamydia pneumoniae. **Infection and Immunity**, Washington, v. 70, n. 1, p. 368-379, 2002.

MOVAHEDI, A. R.; HAMPSON, D. J. New ways to identify novel bacterial antigens for vaccine development. **Veterinary Microbiology**, Amsterdam, v. 131, n. 1-2, p. 1-13, 2008.

N. NETTO, L.; LEAL, C. A. G.; FIGUEIREDO, H. C. P. Streptococcus dysgalactiae as an agent of septicaemia in Nile tilapia, Oreochromis niloticus (L.). **Journal of Fish Diseases**, Oxford, v. 34, n. 3, p. 251-254, 2011.

NHO, S. W. et al. Complete genome sequence and immunoproteomic analyses of the bacterial fish pathogen streptococcus parauberis. **Journal of Bacteriology**, Washington, v. 193, n. 13, p. 3356-3366, 2011.

NIZET, V. Streptococcal β-hemolysins: genetics and role in disease pathogenesis. **Trends Microbiol,** v. 10, p. 575-580, 2002.

OLIVARES-FUSTER, O. et al. Molecular typing of Streptococcus agalactiae isolates from fish. **Journal of Fish Diseases**, Oxford, v. 31, n. 4, p. 277-283, 2008.

OLIVEIRA, I. C. M. et al. Genetic relatedness between group B streptococci originating from bovine mastitis and a human group B Streptococcus type V cluster displaying an identical pulsed-field gel electrophoresis pattern. **Clinical Microbiology and Infection**, Paris, v. 12, n. 9, p. 887-893, Sept. 2006.

PEREIRA, U. P. et al. Genotyping of Streptococcus agalactiae strains isolated from fish, human and cattle and their virulence potential in Nile tilapia. **Veterinary Microbiology**, Amsterdam, v. 140, n. 1–2, p. 186-192, 2010.

PRAJAPATI, J. B. et al. Whole-genome shotgun sequencing of Lactobacillus rhamnosus MTCC 5462, a strain with probiotic potential. **Journal of Bacteriology**, Washington, v. 194, n. 5, p. 1264-1265, 2012.

RAI, J. et al. Immunoinformatic evaluation of multiple epitope ensembles as vaccine candidates: E coli 536. **Bioinformation**, v. 8, n. 6, p. 272-275, Jan. 2012.

RAJAGOPAL, L. Understanding the regulation of Group B Streptococcal virulence factors. **Future Microbiology**, London, v. 4, n. 2, p. 201-221, Mar. 2009.

RAPPUOLI, R. Reverse vaccinology. **Current Opinion in Microbiology**, Oxford, v. 3, n. 5, p. 445-450, 2000.

RAPPUOLI, R. Reverse vaccinology, a genome-based approach to vaccine development. **Vaccine**, v. 19, n. 17–19, p. 2688-2691, 2001.

RATTANACHAIKUNSOPON, P.; PHUMKHACHORN, P. Prophylactic effect of Andrographis paniculata extracts against Streptococcus agalactiae infection in Nile tilapia (Oreochromis niloticus). **Journal of Bioscience and Bioengineering**, Osaka, v. 107, n. 5, p. 579-582, 2009.

RICHARDS, V. P. et al. Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted Streptococcus agalactiae. **Infection, Genetics and Evolution**, Amsterdam, v. 11, n. 6, p. 1263-1275, 2011.

ROMALDE, J. L. et al. Streptococcus phocae, an emerging pathogen for salmonid culture. **Veterinary Microbiology**, Amsterdam, v. 130, n. 1–2, p. 198-207, 2008.

ROSS, B. C. et al. Identification of vaccine candidate antigens from a genomic analysis of Porphyromonas gingivalis. **Vaccine**, v. 19, n. 30, p. 4135-4142, 2001.

RUIZ, J. C. et al. Evidence for reductive genome evolution and lateral acquisition of virulence functions in two Corynebacterium pseudotuberculosis strains. **PLoS ONE**, v. 6, n. 4, p. e18551, 2011.

SALVADOR, R. et al. Isolation and characterization of Streptococcus spp. group B in Nile tilapias (Oreochromis niloticus) reared in hapas nets and earth nurseries in the northern region of Parana State, Brazil. **Ciência Rural**, Santa Maria, v. 35, n. 6, p. 1374-1378, dez. 2005.

SALVADOR, R. et al. Isolamento de Streptococcus spp de tilápias do nilo (Oreochromis niloticus) e qualidade da água de tanques rede na Região Norte do Estado do Paraná , Brasil Isolation of Streptococcus spp from nile tilapia (Oreochromis niloticus) and quality of water. **Semina Ciências Agrárias**, Londrina, v. 24, n. 1, p. 35-42, 2003.

SANGER F.; NICKLEN S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 74, n. 12, p. 5463–5467, 1977.

SANTI, I. et al. BibA: a novel immunogenic bacterial adhesin contributing to group B Streptococcus survival in human blood. **Molecular Microbiology**, Salem, v. 63, n. 3, p. 754-767, Feb. 2007.

SANTOS, A. et al. The reverse vaccinology- a contextual overview. **The IIOAB Journal**, v. 2, n. 4, p. 8-15, 2011.

SCHMIDT, H.; HENSEL, M. Pathogenicity islands in bacterial pathogenesis. **Clinical Microbiology Reviews**, v. 17, n. 1, p. 14-56, 2004.

SEO, H. S. et al. Binding of glycoprotein Srr1 of Streptococcus agalactiae to fibrinogen promotes attachment to brain endothelium and the development of meningitis. **PLoS pathogens**, v. 8, n. 10, p. e1002947, Oct. 2012.

SHARMA, P. et al. Role of pili proteins in adherence and invasion of Streptococcus agalactiae to the lung and cervical epithelial cells. **The Journal of Biological Chemistry**, Bethesda, v. 228, n. 6, p. 4023-4034, Dec. 2012.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. **Nature Biotechnology**, New York, v. 26, n. 10, p. 1135-1145, Oct. 2008.

SHEWMAKER, P. L. et al. Streptococcus ictaluri sp. nov., isolated from Channel Catfish Ictalurus punctatus broodstock. **International Journal of Systematic and Evolutionary Microbiology**, Reading, v. 57, n. 7, p. 1603-1606, 2007.

SHOEMAKER, C. A. et al. Efficacy of a Streptococcus iniae modified bacterin delivered using Oralject™ technology in Nile tilapia (Oreochromis niloticus). **Aquaculture**, v. 255, n. 1-4, p. 151-156, May 2006.

SHOEMAKER, C. A. et al. Protection against heterologous Streptococcus iniae isolates using a modified bacterin vaccine in Nile tilapia, Oreochromis niloticus (L.). **Journal of Fish Diseases**, Oxford, v. 33, n. 7, p. 537-544, 2010.

SILVA, A. et al. Complete Genome Sequence of Corynebacterium pseudotuberculosis I19, a Strain Isolated from a Cow in Israel with Bovine Mastitis. **Journal of Bacteriology**, Washington, v. 193, n. 1, p. 323-324, 2011.

SINGH, P. et al. Whole-genome shotgun sequencing of a colonizing multilocus sequence type 17 Streptococcus agalactiae strain. **Journal of Bacteriology**, Washington, v. 194, n. 21, p. 6005, Nov. 2012.

SOARES, S. C. et al. PIPS: Pathogenicity Island Prediction Software. **PLoS ONE**, v. 7, n. 2, p. e30848, 2012.

SØRENSEN, U. B. S. et al. Emergence and global dissemination of host-specific Streptococcus agalactiae clones. **mBio**, v. 1, n. 3, e00178, Aug. 2010.

SRINIVASA RAO, P. S.; LIM, T. M.; LEUNG, K. Y. Functional genomics approach to the identification of virulence genes involved in edwardsiella tarda pathogenesis . **Infection and Immunity**, Washington, v. 71, n. 3, p. 1343-1351, Mar. 2003.

SUKHNANAND, S. et al. Molecular subtyping and characterization of bovine and human Streptococcus agalactiae isolates. **Journal of Clinical Microbiology**, Washington, v. 43, n. 3, p. 1177-1186, 2005.

TAZI, A. et al. The surface protein HvgA mediates group B streptococcus hypervirulence and meningeal tropism in neonates. **Journal of Experimental Medicine**, New York, v. 207, n. 11, p. 2313-2322, Oct. 2010.

TETTELIN, H. et al. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V Streptococcus agalactiae. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 99, n. 19, p. 12391-12396, 2002.

TETTELIN, H. et al. 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**, Washington, v. 102, n. 39, p. 13950-13955, 2005.

TIAN M. et al. Rare case of diabetic hand ulcer caused by *Streptococcus agalactiae*. **The International Journal of Lower Extremity Wounds**. V. 11, N. 3, P.174-176, AUG./SEPT. 2012.

TROST, E. et al. The complete genome sequence of Corynebacterium pseudotuberculosis FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. **BMC Genomics**, v. 11, p. 728, Jan. 2010.

ULETT, K. B. et al. Group B streptococcus cystitis presenting in a diabetic patient with a massive abdominopelvic abscess: a case report. **Journal of Medical Case Reports**, v. 6, n. 1, p. 237, Jan. 2012.

VERA-CALDERÓN, L. E.; FERREIRA, A. C. M. Estudo da economia de escala na piscicultura em tanque-rede, no estado de São Paulo. Informações Econômicas, São Paulo, v. 34, n. 1, p. 7-17, 2004.

VIZCAÍNO, C. et al. Computational prediction and experimental assessment of secreted/surface proteins from *Mycobacterium tuberculosis* H37Rv. **PLoS Computational Biology,** v. 6, n. 6, p. e1000824, 2010.

WAAGE, S. et al. Bacteria associated with clinical mastitis in dairy heifers. **Journal of Dairy Science**, Champaign, v. 82, n. 4, p. 712-719, Apr. 1999.

WANG, B. et al. Complete genome sequence of Streptococcus agalactiae ZQ0910, a pathogen causing meningoencephalitis in the GIFT strain of Nile tilapia (Oreochromis niloticus). **Journal of Bacteriology**, Washington, v. 194, n. 18, p. 5132-5133, Sept. 2012.

WETZEL, J.; KINGSFORD, C.; POP, M. Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies. **BMC bioinformatics**, v. 12, p. 95, Jan. 2011.

WIZEMANN, T. M. et al. Use of a whole genome approach to identify vaccine molecules affording protection against Streptococcus pneumoniae infection. **Infection and Immunity**, Washington, v. 69, n. 3, p. 1593-1598, Mar. 2001.

YILDIRIM, A. Ö. et al. Pheno- and genotypic properties of streptococci of serological group B of canine and feline origin. **FEMS Microbiology Letters**, Amsterdam, v. 212, n. 2, p. 187-192, 2002.

YILDIRIM, A. Ö.; LÄMMLER, CH; WEIS, R. Identification and characterization of Streptococcus agalactiae isolated from horses. **Veterinary Microbiology**, Amsterdam, v. 85, n. 1, p. 31-35, 2002.

ZHOU, S. M. et al. Identification and genetic characterization of Streptococcus iniae strains isolated from diseased fish in China. **Journal of Fish Diseases**, Oxford, v. 31, n. 11, p. 869-875, 2008.

SEGUNDA PARTE – ARTIGOS

ARTIGO 1 Complete genome sequence of *Streptococcus agalactiae* strain sa20-06, a fish pathogen associated to meningoencephalitis outbreaks

(Artigo submetido à revista "Standards in Genomic Sciences")

Ulisses de Pádua Pereira[1,4], Anderson Rodrigues dos Santos[2], Syed Shah Hassan[2], Flávia Figueira Aburjaile[2], Siomar de Castro Soares[2], Rommel Thiago Jucá Ramos[3], Adriana Ribeiro Carneiro[3], Luís Carlos Guimarães[2], Sintia Silva de Almeida[2], Carlos Augusto Almeida Diniz[2], Maria Silvanira Barbosa[3], Pablo Gomes de Sá[3], Amjad Ali[2], Syeda Marriam Bakhtiar[2], Fernanda Alves Dorella[2], Adhemar Zerlotini[5,6], Flávio Marcos Gomes Araújo[5], Laura Rabelo Leite[5], Guilherme Oliveira[5], Anderson Miyoshi[2], Artur Silva[3], Vasco Azevedo[2], Henrique César Pereira Figueiredo[1*]

[1] AQUAVET- Laboratory of Aquatic Animal Diseases, Department of Preventive Veterinary Medicine, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil;

[2] Institute of Biologic Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil;

[3] Institute of Biologic Sciences, Federal University of Pará, Belém, PA, Brazil;

[4] Department of Veterinary Medicine, Federal University of Lavras, Lavras, MG, Brazil;

[5] Center for Excellence in Bioinformatics - FIOCRUZ-MG, Belo Horizonte, MG, Brazil;

[6] Bioinformatics Multiuser Laboratory - Embrapa, Campinas, SP, Brazil.

Corresponding Author: Henrique C. P. Figueiredo, DVM, Ph.D. (henrique@dmv.ufla.br)

Keywords: *Streptococcus agalactiae*, fish pathogen, genome sequencing.

Abstract

Streptococcus agalactiae (Lancefield group B; GBS) is the causative agent of meningoencephalitis in fish, mastitis in cows, and neonatal sepsis in humans. Meningoencephalitis is a major health problem for tilapia farming and is responsible for high economic losses worldwide. Despite its importance, the genomic characteristics and the main molecular mechanisms involved in virulence of *S. agalactiae* isolated from fish are still poorly understood. Here, we present the genomic features of the 1,820,886 bp long complete genome sequence of *S. agalactiae* SA20-06 isolated from a meningoencephalitis outbreak in Nile tilapia (*Oreochromis niloticus*) from Brazil, and its annotation, consisting of 1,710 protein-coding genes (excluding pseudogenes), 7 rRNA operons, 79 tRNA genes and 62 pseudogenes.

Resumo

Streptococcus agalactiae (grupo B de Lancefield; GBS) é o agente causador de meningoencefalite em peixes, mastite em vacas, e sepse neonatal em seres humanos. Meningoencefalite é um grande problema sanitário para a criação de tilápias e é responsável por grandes perdas econômicas em todo o mundo. Apesar de sua importância, as características genômicas e os principais mecanismos moleculares envolvidos na virulência de *S. agalactiae* isoladas de peixes são ainda pouco compreendidos. Assim, no presente trabalho, são apresentadas as características do genoma completo (com tamanho de1.820.886 bp) da linhagem SA20-06 isolada de um surto de meningoencefalite em tilápia do Nilo (*Oreochromis niloticus*), do Brasil. Foram encontrados no genoma desta linhagem 1.710 genes codificadores de proteínas (excluindo pseudogenes), 7 operons de rRNA, 79 genes tRNA e 62 pseudogenes.

Introduction

Streptococcus agalactiae, also referred as Group B *Streptococcus* (GBS), is a Gram-positive pathogen with a broad host range. GBS is the most common cause of life-threatening bacterial infections in human newborns [1] and is an important etiological agent of clinical and sub-clinical bovine mastitis [2]. In fish, *S. agalactiae* infection causes septicemia and meningoencephalitis, mainly in warm water species from freshwater, marine, or estuarine environments [3]. Currently, *S. agalactiae* is an emerging pathogen associated with severe economic losses due to high mortality rates in fish farms worldwide [4,5]. The pangenome of the species (obtained from only eight human strain genomes) is considered open and it is expected that, for every new GBS genome sequenced, approximately 33 new strain-specific genes will be identified [6]. Since, the first genome of *S. agalactiae* strain isolated from bovine mastitis was published and 183 strain-specific genes were described, and about 85% of these genes have been clustered into eight genome islands, strongly suggesting that these genes were acquired through lateral gene transfer from other bacteria of genus *Streptococcus*, which are also etiologic agents of bovine mastitis [2]. However, the molecular mechanisms of virulence and other genomic features of strains isolated from fish isolates remain unclear, and thus, the genome sequencing of different strains isolated from other hosts are still required to better understand the global complexity of this bacterial species.

Classification and Features

The genus Streptococcus comprises a heterogeneous group of bacteria that have an important role in medicine and industry. These microorganisms are Gram-positive, cocci shape of 0.6-1.2µm diameter, not motile, do not form spores, are catalase-negative and grow in pairs or chains [7]. Rebecca C. Lancefield, in her work in the early 1930´s, systematized the classification of streptococci based on the presence and type of surface antigen: cell wall

polysaccharide or lipoteichoic acid [8]. *S. agalactiae* is classified as Lancefield group B (GBS) based on the presence of a polysaccharide in the cell wall. This polysaccharide is composed of galactose, N-acetylglucosamine, rhamnose and glucitol phosphate [7]. Currently, ten serotypes are described for this species (Ia, Ib, II-IX) and occasionally some strains can be non-serotypeable [9].

Major human and animal streptococcal pathogens belong to the so designated pyogenic group of β-hemolityc streptococci [10]. In this context, the β-hemolityc bacteria *S. agalactiae*, deserves attention for causing diseases in a broad range of homeothermic and heterothermic hosts [4], although this bacteria is also a common member of the gastrointestinal tract microbiota [11].

At the end of the XIX century, GBS was initially described as an etiological agent of mastitis in cows, being reported as causing disease in humans only 50 years later [12]. In fish, *S. agalactiae* was recognized as a pathogen in 1966 [13]. Sporadically, this pathogen has also been associated with illness in many others hosts, such as chickens, camels, dogs, horses, cats, frogs, hamsters, mice, monkeys, and nutria [14].

S. agalactiae is a facultative anaerobic bacteria, that uses glucose as an energy source, and is also able to use different carbon sources such as cellobiose, beta-glucoside, trehalose, mannose, lactose, fructose, mannitol, N-acetylgalactosamine, and glucose (Table 1). This pathogen is limited in the synthesis of most amino acids precursors. Only the biosynthetic pathways for alanine, serine, glycine, glutamine, aspartate, asparagine and threonine are present [15]. The adaptation to oxygen radicals stress of this pathogen is related to superoxide dismutase (*sodA* gene) which converts superoxide anions to molecular oxygen and hydrogen peroxide, which, in turn, is metabolized by catalases and/or peroxidases [16]. Although GBS does not synthetize catalase to remove toxic H_2O_2, it is 10-fold more resistant to oxygen metabolites than the

catalase-producing *S. aureus*. This is due to the presence of several enzymes that might detoxify H_2O_2 that have been identified in the genome of *S. agalactiae* such as NADH peroxidase, NADH oxidase and thiol peroxidase [15]. This diversity of metabolic and adaptation mechanisms reflects the ability to survive in various environments and hosts.

The phylogenetic tree was constructed using 16S rRNA sequences of available *S. agalactiae* genomes and other species from the same genus (Figure1). The tree shows that all *S. agalactiae* strains are grouped together, and the SA20-06 strain is more similar to the A909 human isolate and to the GD201008-001 fish isolate from China.

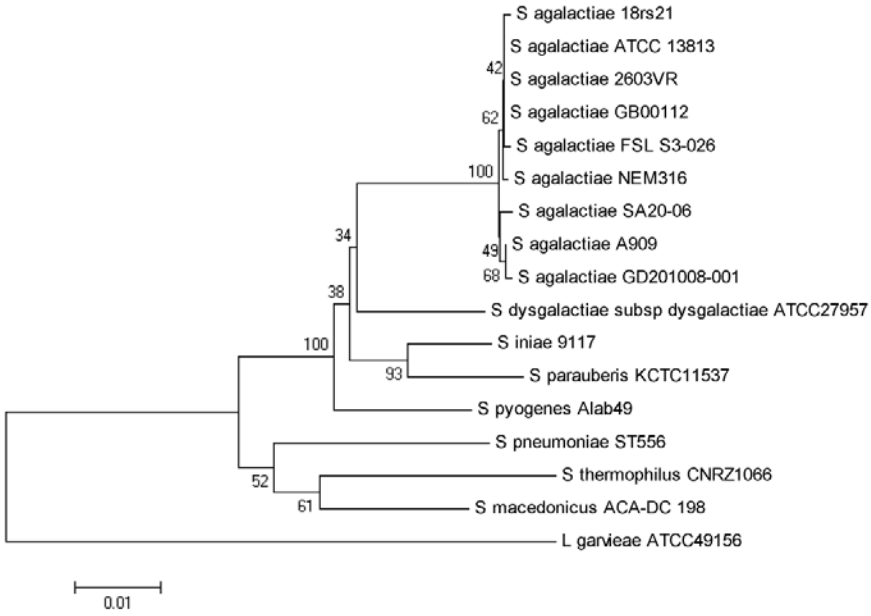

Figure 1. Phylogenetic tree highlighting the position of *S. agalactiae* strain SA20-06 in relation to other selected strains of the species and others from the

genus *Streptococcus*. The tree was based on 1,410 characters of the 16S rRNA gene sequence aligned using ClustalW2 (Larkin et al., 2007). The tree was inferred under the maximum likelihood criterion using MEGA5 software (Tamura et al., 2011) and rooted with 16S rRNA sequence of fish pathogen *Lactococcus garvieae* (a member of the *Streptococcaceae* family). The branches were mapped by the expected number of substitutions per site. The numbers above the branches are support values from 1,000 bootstrap replicates. The strains and their corresponding GenBank accession numbers (and, when applicable, draft sequence coordinates) for 16S rRNA genes are: *S. agalactiae* 18rs21, NZ_AAJO01000124; *S. agalactiae* ATCC13813, NR_040821; *S. agalactiae* 2603VR, NC_004116; *S. agalactiae* GB00112, AKXO01000029; *S. agalactiae* FSL_S3-026, AEXT01000002; *S. agalactiae* NEM316, AL766845; *S. agalactiae* SA20-06, NC_019048; *S. agalactiae* A909, NC_007432; *S. agalactiae* GD201008-001, CP003810; *S. dysgalactiae* subsp *dysgalactiae* ATCC 27957, CM001076; *S. iniae* 9117, NZ_AMOO01000003; *S. parauberis* KCT 11537, NC_015558; *S. pyogenes* alab49, NC_017596; *S. pneumonia* ST556, NC_017769; *S. thermophiles* CNRZ1066, NC_006449; *S. macedonicus* ACA-DC 198, NC_016749; *L. garvieae*, AP009332.

Table 1. Classification and general features of *S. agalactiae* SA20-06 according to the MIGS recommendations [19].

MIGS ID	Property	Term	Evidence code
	Classification	Domain *Bacteria*	TAS [20]
		Phylum *Firmicutes*	TAS [20]
		Class *Bacilli*	TAS [20]
		Order Lactobacillales	TAS [20]
		Family *Streptoccaceae*	TAS [20]
		Genus *Streptococcus*	TAS [20]
		Species *Streptococcus agalactiae*	TAS [20]
		Strain SA20-06	TAS [4]
	Gram stain	Positive	TAS [20]
	Cell shape	Spherical or ovoid	TAS [20]
	Motility	non-motile	TAS [20]
	Sporulation	non-sporulating	TAS [20]
	Temperature range	mesophile	TAS [20]
	Optimum temperature	28°C (fish isolates)	IDA
	Salinity	usually grows in 4% of NaCl, but not in 6.5%	TAS [20]
MIGS-22	Oxygen	Facultative anaerobie	TAS [20]
	Carbon source	cellobiose, beta-glucoside, trehalose, mannose, lactose, fructose, mannitol, N-acetylgalactosamine, and glucose	TAS[15]
	Energy source	Chemoorganotroph with fermentative metabolism	TAS [20]
MIGS-6	Habitat	Host	TAS [4]
MIGS-15	Biotic relationship	Symbiotic (pathogen)	TAS [4]
MIGS-14	Pathogenicity	Cows, human, fishes and other animals	TAS [12,14]
	Biosafety level	2	TAS [21]
	Isolation	Kidney of Nile tilapia	TAS [4]
MIGS-4	Geographic location	Parana state, Brazil	TAS [4]
MIGS-5	Sample collection time	2006	TAS [4]
MIGS-4.1	Latitude	not reported	
MIGS-4.2	Longitude	not reported	
MIGS-4.3	Depth	not reported	
MIGS-4.4	Altitude	not reported	

Evidence codes - IDA: Inferred from Direct Assay; TAS: Traceable Author Statement (i.e., a direct report exists in the literature); NAS: Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from the Gene Ontology project [22]. If the evidence is IDA, then the property was directly observed for a live isolate by one of the authors or an expert mentioned in the acknowledgements.

Genome sequencing and annotation

Genome project history

This strain was selected for sequencing based on the high mortality rates shown for this pathogen in fish farms worldwide and on the lack of information for the genomic characteristics of *S. agalactiae* isolated from fish and the molecular mechanisms involved in virulence in this host. The genome project is deposited in the Genomes OnLine Database [23] and the *Streptococcus agalactiae* SA20-06 complete genome sequence and annotation data were deposited in the DDBJ/EMBL/GenBank under the accession number CP003919 (RefSeq NC_019048). Sequencing, assembly steps, finishing and annotation were performed by the teams from the Laboratory of Cellular and Molecular Genetics (LGCM), Minas Gerais, Brazil; Genomics and Proteomics Network of the State of Pará (RPGP), Pará, Brazil and Center for Excellence in Bioinformatics (CEBio-FIOCRUZ-MG), Minas Gerais, Brazil. A summary of the project information is shown in Table 2.

Table 2. Genome sequencing project information.

MIGS ID	Property	Term
MIGS-31	Finishing quality	Finished
MIGS-28	Libraries used	Two mate-paired libraries (mean size 50 or 60 bp, DNA insert size of 1-2Kb)
MIGS-29	Sequencing platforms	SOLiD v3 plus and SOLiD 5500
MIGS-31.2	Sequencing coverage	5700-fold
MIGS-30	Assemblers	CLC Genome Workbench, Velvet, Edena
MIGS-32	Gene calling method	Glimmer
	Genbank ID	CP003919 (chromosome)
	Genbank Date of Release	November 02, 2012
	GOLD ID	Gc02347
	Project relevance	Animal and human pathogen

Growth conditions and DNA isolation

Streptococcus agalactiae SA20-06 was obtained from the AQUAVET (Laboratory of Aquatic Animal Diseases) bacterial collection, streaked onto 5% sheep blood agar and incubated at 28^{o}C for 48 h. After that, cells were grown in 150mL brain-heart-infusion broth (BHI-HiMedia Laboratories Pvt. Ltda, India) under agitation (150 rpm), at 28^{o}C. Genomic DNA was obtained by using phenol-chloroform-isoamylic alcohol extraction protocol using micro-wave oven [24].

Genome sequencing and assembly

The genome sequencing of *S. agalactiae* SA20-06 was performed using the SOLiD v3 Plus and SOLiD 5500 platforms (Applied Biosystems) with two mate-paired libraries (both with 1-2 kb insert size), which generated 50,223,637 and 283,953,694 reads of 50 bp and 60 bp in size, respectively. After

sequencing, the reads were subjected to quality filtering using the qualityFilter.pl script (a homemade script), in which reads with an average Phred quality of less than 20 were removed, and error sequence correction was performed with SAET software (Life Technologies).

After quality analysis, 210,004,694 reads were used in the assembly, which generated a genome coverage corresponding to ~5,700x genome coverage based on the reference genome of 2,127,839 bp size of *S. agalactiae* strain A909 (NC_007432). The genome sequence of SA20-06 was assembled based on the hybrid strategy using CLC Genome Workbench 4.9, Velvet [25] and Edena [26] software. A total of 872 contigs were generated, with N_{50} of 5,221 bp and the smallest contig having 201 bp. Due to the hybrid assembly methodology, the redundant contigs were removed using the Simplifier software [27]**.** The contigs were mapped against the reference genome (strain A909) using BLASTn, and the results were analyzed using G4ALL (**http://g4all.sourceforge.net/**) software, to extend the contigs and identify overlaps of a minimum of 30 bp between the ends of the contigs, thus yielding larger contigs.

These contigs were later subjected to a finishing process using CLC Genomics Workbench software. At this step, the contigs were ordered and oriented by mapping against the reference genome, yielding a preliminary scaffold with gaps that were removed with recursive rounds of short reads mapping against the scaffold [28].

Genome annotation

For structural annotation, the following software programs were employed: Glimmer 3, to predict genes [29]; RNAmmer, to predict rRNAs [30]; and tRNAscan-SE, to predict tRNAs [31]. Functional annotation was performed by similarity analyses using public databases of National Center for Biotechnology Information (NCBI) non-redundant database, Swiss-Prot and

InterProScan analysis [32]. Genome visualization and manual annotation were carried out using Artemis [33].

Genome properties

The complete genome of *S. agalactiae* strain SA20-06 comprises a single circular chromosome of 1,820,886 bp in length with 1,710 putative predicted genes (excluding pseudogenes), 35.56% G+C content, 7 rRNA operons, 79 tRNA genes and 62 pseudogenes (Figure 2 and Table 3). The distribution of genes into the COG functional categories is presented in Table 4.

Table 3. Genome Statistics.

Attribute	Value	% of Total[a]
Genome size (bp)	1,820,886	100.00%
DNA coding region (bp)	1,547,993	85.01%
DNA G+C content (bp)	647,477	35.56%
Number of replicons	1	
Extrachromosomal elements	0	
Total genes[b]	1,872	100.00%
RNA genes	100	5.34%
rRNA operons	7	
Protein-coding genes	1,772	94.66%
Pseudo genes	62	3.31%
Genes with function prediction	1,515	80.93%
Genes in paralog clusters	430	22.97%
Genes assigned to COGs	1,469	78.47%
Genes assigned Pfam domains	1,547	82.64%
Genes with signal peptides	302	16.13%
Genes with transmembrane helices	447	23.88%

a) The total is based on either the size of the genome in base pairs or the total number of protein coding genes in the annotated genome.

b) Also includes 62 pseudogenes.

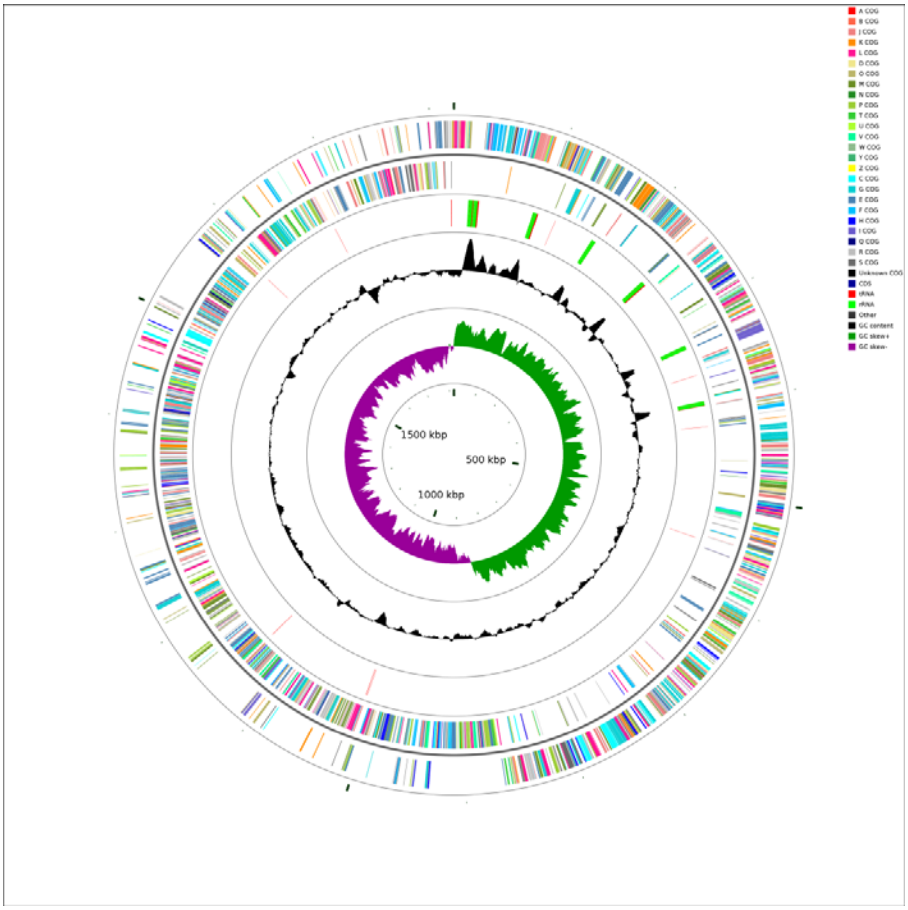

Figure 2. Graphical circular map of the genome performed with CGview comparison tool [34]. From outer to inner circle: Genes on forward strand (color by COG categories), Genes on reverse strand (color by COG categories), RNA genes (tRNAs red, rRNAs green, other RNAs black), GC content, GC skew.

Table 4. Number of genes associated with the general COG functional categories.

Code	value	%age	Description
J	146	9.2	Translation, ribosomal structure and biogenesis
A	0	0.0	RNA processing and modification
K	118	7.44	Transcription
L	86	5.42	Replication, recombination and repair
B	0	0.0	Chromatin structure and dynamics
D	17	1.07	Cell cycle control, cell division, chromosome partitioning
Y	0	0.0	Nuclear structure
V	36	2.27	Defense mechanisms
T	66	4.16	Signal transduction mechanisms
M	92	5.8	Cell wall/membrane biogenesis
N	6	0.38	Cell motility
Z	0	0.0	Cytoskeleton
W	0	0.0	Extracellular structures
U	21	1.32	Intracellular trafficking and secretion
O	53	3.34	Posttranslational modification, protein turnover, chaperones
C	46	2.9	Energy production and conversion
G	150	9.45	Carbohydrate transport and metabolism
E	134	8.44	Amino acid transport and metabolism
F	75	4.73	Nucleotide transport and metabolism
H	52	3.28	Coenzyme transport and metabolism
I	43	2.71	Lipid transport and metabolism
P	86	5.42	Inorganic ion transport and metabolism
Q	19	1.2	Secondary metabolites biosynthesis, transport and catabolism
R	192	12.10	General function prediction only
S	149	9.39	Function unknown
-	403	21.53	Not in COGs

Conclusions

Further analysis of the SA20-06 genome is now under way. These are being conducted with the objective to identify specific factors that might explain the differences in pathogenesis of disease, mainly in heterothermic hosts.

Acknowledgement

This work was supported by Ministério da Pesca e Aquicultura, Furnas Centrais Elétricas, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). We also acknowledge support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Rede Paraense de Genômica e Proteômica.

References

1. Rajagopal L. Understanding the regulation of Group B *Streptococcal* virulence factors. *Future Microbiol* 2009;**4**:201–221. PubMed http://dx.doi.org/10.2217/17460913.4.2.201

2. Richards VP, Lang P, Bitar PDP, et al. Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted *Streptococcus agalactiae*. *Infect Genet Evol* 2011;**11**:1263–1275. PubMed http://dx.doi.org/10.1016/j.meegid.2011.04.019

3. Evans JJ, Klesius PH, Gilbert PM, et al. Characterization of β-haemolytic Group B *Streptococcus agalactiae* in cultured seabream, Sparus auratus L., and wild mullet, Liza klunzingeri (Day), in Kuwait. *Journal of Fish Diseases* 2002;**25**:505–513. http://dx.doi.org/10.1046/j.1365-2761.2002.00392.x

4. Mian GF, Godoy DT, Leal CAG, Yuhara TY, Costa GM, Figueiredo HCP. Aspects of the natural history and virulence of S. agalactiae infection in Nile tilapia. *Vet Microbiol* 2009;**136**:180–183. PubMed http://dx.doi.org/10.1016/j.vetmic.2008.10.016

5. Duremdez R, Al-Marzouk A, Qasem JA, Al-Harbi A, Gharabally H. Isolation of Streptococcus agalactiae from cultured silver pomfret, Pampus argenteus (Euphrasen), in Kuwait. *J Fish Dis* 2004;**27**:307–10. PubMed http://dx.doi.org/10.1111/j.1365-2761.2004.00538.x

6. Tettelin H, Masignani V, Cieslewicz MJ, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome." *Proc Natl Acad Sci U S A* 2005;**102**:13950–13955. PubMed http://dx.doi.org/10.1073/pnas.182380799

7. Schuchat A. Epidemiology of group B streptococcal disease in the United States: shifting paradigms. *Clin Microbiol Rev* 1998;**11**:497–513. PubMed

8. Lancefield RC. A serological differentiation of specifc types of bovine hemolytic streptococci (GROUP B). *J Exp Med* 1934;**59**:441–58. PubMed

9. Slotved H-C, Kong F, Lambertsen L, Sauer S, Gilbert GL. Serotype IX, a Proposed New *Streptococcus agalactiae* Serotype. *J Clin Microbiol* 2007;**45**:2929–36. PubMed http://dx.doi.org/10.1128/JCM.00117-07

10. Carvalho M da G, Facklam R, Jackson D, Beall B, McGee L. Evaluation of three commercial broth media for pigment detection and identification of a group B *Streptococcus* (*Streptococcus agalactiae*). *J Clin Microbiol* 2009;**47**:4161–3. PubMed http://dx.doi.org/10.1128/JCM.01374-09

11. Schrag S, Gorwitz R, Fultz-Butts K, Schuchat A. Prevention of Perinatal Group B Streptococcal Disease: revised guidelines from CDC. *MMWR Recomm* 2002; 51:1–22. PubMed

12. Bisharat N, Crook DW, Leigh J, et al. Hyperinvasive Neonatal Group B Streptococcus Has Arisen from a Bovine Ancestor. *J Clin Microbiol* 2004;**42**:2161–2167. PubMed http://dx.doi.org/10.1128/JCM.42.5.2161-2167.2004

13. Robinson JA, Meyer FP. Streptococcal Fish Pathogen. *J Bacteriol* 1966;**92**:512. PubMed

14. Pereira UP, Mian GF, Oliveira ICM, Benchetrit LC, Costa GM, Figueiredo HCP. Genotyping of *Streptococcus agalactiae* strains isolated from fish, human and cattle and their virulence potential in Nile tilapia. *Vet Microbiol* 2010;**140**:186–192. PubMed http://dx.doi.org/10.1016/j.vetmic.2009.07.025

15. Glaser P, Rusniok C, Buchrieser C, et al. Genome sequence of Streptococcus agalactiae, a pathogen causing invasive neonatal disease. *Mol Microbiol* 2002;**45**:1499–1513. PubMed http://dx.doi.org/10.1046/j.1365-2958.2002.03126.x

16. Poyart C, Pellegrini E, Gaillot O, Boumaila C, Baptista M, Trieu-Cuot P. Contribution of Mn-cofactored superoxide dismutase (SodA) to the virulence of *Streptococcus agalactiae*. *Infect Immun* 2001;**69**:5098–106. PubMed http://dx.doi.org/10.1128/IAI.69.8.5098-5106.2001

17. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wal-lace IM, Wilm A, Lopez R, et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23:2947-2948. PubMed http://dx.doi.org/10.1093/bioinformatics/btm404

18. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011;**28**:2731–9. PubMed http://dx.doi.org/10.1093/molbev/msr121

19. Field D, Garrity G, Gray T, et al. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* 2008;**26**:541–7. PubMed http://dx.doi.org/10.1038/nbt1360

20. Whiley RA, Hardie JM. The Firmicutes. In: Garrity GM, Boone DR, Castenholz RW (eds), Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 3, Springer, New York, 2001, p. 655-735.

21. Moura H, Woolfitt AR, Carvalho MG, et al.MALDI-TOF mass spectrometry as a tool for differentiation of invasive and noninvasive *Streptococcus pyogenes* isolates. *FEMS Immunol Med Microbiol* 2008;**53**:333–42. PubMed http://dx.doi.org/10.1111/j.1574-695X.2008.00428.x

22. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9. PubMed http://dx.doi.org/10.1038/75556

23. Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated

metadata. *Nucleic Acids Res 2010; 38:D346-D354.*PubMed http://dx.doi.org/10.1093/nar/gkp848

24. Bollet C, Gevaudan MJ, de Lamballerie X, Zandotti C, de Micco P. A simple method for the isolation of chromosomal DNA from gram positive or acid-fast bacteria. *Nucleic Acids Res* 1991;**19**:1955. PubMed

25. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* 2008;**18**:821–9. PubMed http://dx.doi.org/10.1101/gr.074492.107

26. Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res* 2008;**18**:802–9. PubMed http://dx.doi.org/10.1101/gr.072033.107

27. Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol* 2010;**11**:R41. PubMed http://dx.doi.org/10.1186/gb-2010-11-4-r41

28. Ramos RTJ, Carneiro AR, Azevedo V, Schneider MP, Barh D, Silva A. Simplifier: a web tool to eliminate redundant NGS contigs. Bioinformation 2012;8(20): 996-999.

29. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res 1999;27(23):4636-41. PubMed http://dx.doi.org/10.1093/nar/27.23.4636

30. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res 2007;35(9):3100-8.*PubMed http://dx.doi.org/10.1093/nar/gkm160

31. Lowe TM, Eddy SR. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res* 1997;**25**:955–964. PubMed http://dx.doi.org/10.1093/nar/25.5.0955

32. Zdobnov EM, Apweiler R. InterProScan -- an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 2001;**17**:847–848. PubMed http://dx.doi.org/10.1093/bioinformatics/17.9.847

33. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. Artemis: sequence visualization and annotation. Bioinformatics 2000;16(10):944-5. PubMed http://dx.doi.org/10.1093/bioinformatics/18.suppl_1.S225

34. Grant JR, Arantes AS, Stothard P. Comparing thousands of circular genomes using the CGView Comparison Tool. *BMC genomics*2012;**13**:202. PubMed http://dx.doi.org/10.1186/1471-2164-13-202

ARTIGO 2 **Pan-genome upgrade, comparative metabolic pathways and prediction of vaccine targets in *Streptococcus agalactiae* strains isolated from human, bovine and fish**

(Artigo submetido à revista "Plos One")

Ulisses de Pádua Pereira[1,6], Siomar de Castro Soares[3], Jochen Blom[4], Carlos Augusto Gomes Leal[1,2], Rommel Thiago Jucá Ramos[5], Luís Carlos Guimarães[3], Letícia de Castro Oliveira[3], Sintia Silva de Almeida[3], Syed Shah Hassan[3], Anderson Rodrigues dos Santos[3], Anderson Miyoshi[3], Artur Silva[5], Andreas Tauch[4], Debmalya Barh[7], Vasco Azevedo[3], Henrique César Pereira Figueiredo[1,2*]

[1] AQUAVET - Laboratory of Aquatic Animal Diseases, Department of Preventive Veterinary Medicine, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil;

[2] AQUACEN – National Reference Laboratoty for Aquatic Animal Diseases – Ministry of Fisheries and Aquaculture, Belo Horizonte, MG. Brazil;

[3] Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil;

[4]Center for Biotechnology, Bielefeld University, Bielefeld, Nordrhein-Westfalen, Germany;

[5] Institute of Biological Sciences, Federal University of Pará, Belém, PA, Brazil;

[6] Department of Veterinary Medicine, Federal University of Lavras, Lavras, MG, Brazil;

[7] Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal, India.

Corresponding Author: Henrique C. P. Figueiredo, DVM, Ph.D. (figueiredoh@yahoo.com).

E-mail addresses:

Ulisses de Pádua Pereira – upaduapereira@gmail.com

Siomar de Castro Soares – siomars@gmail.com

Jochen Blom - jblom@cebitec.uni-bielefeld.de

Carlos Augusto Gomes Leal – leal.cag@gmail.com

Rommel Thiago Jucá Ramos – rommelramos@ufpa.br

Luís Carlos Guimarães – luisguimaraes.bio@gmail.com

Letícia de Castro Oliveira - letcastrol@gmail.com

Síntia Almeida - sintiaalmeida@gmail.com

Syed Shah Hassan – hassan_chemist@yahoo.com

Anderson Rodrigues dos Santos – anderson2010@gmail.com

Anderson Miyoshi – miyoshi@icb.ufmg.br

Artur Silva – asilva@ufpa.br

Andreas Tauch- tauch@cebitec.uni-bielefeld.de

Debmalya Barh – dr.barh@gmail.com

Vasco Azevedo – vasco@icb.ufmg.br

Henrique César Pereira Figueiredo – figueiredoh@yahoo.com

ABSTRACT

Streptococcus agalactiae (Lancefield group B; GBS) is a major pathogen that causes meningoencephalitis in fish, mastitis in cows, and neonatal sepsis and meningitis in humans. Currently, there are available genome sequences of ten strains isolated from human, four isolated from fish and one isolated from bovine. However, genomic features correlated to niche, host adaptability and presence of cross-reactive vaccine targets ofthispathogen need to be more understood, mainly in the genome of fish strains. Therefore, the goals of this work were analyze comparatively the genome of 15 *S. agalactiae* strains by phylogenomic, pan-genomic, pathogenicity islands, and metabolic pathways analyses and also identify cross-reactive vaccine targets using reverse vaccinology and immunoinformatic strategies. The analysis revealed that the pan-genome of the species is open and the strains isolated from fish showed more clonal-like behavior when compared to the human isolates. A high diversity of pathogenicity islands and metabolic pathways was observed, and genomic features related to virulence in each host and potential cross-reactive vaccine target were discussed. Thus, this work throws new insights in respect to virulence of *S. agalactiae* in different hosts and vaccine developing against this pathogen.

Keywords:*Streptococcus agalactiae*, fish, human, bovine, pan-genomics, vaccine targets.

RESUMO

Streptococcus agalactiae (Gupo B de Lancefield; GBS) é um patógeno principal que causa meningoencefalite em peixes, mastite em vacas, e sepse neonatal e meningite em seres humanos. Atualmente, há disponível sequências genômicas de dez linhagens isoladas de seres humanos, quarto de linhagens isoladas de peixe e uma isolada de bovino. Entretanto, as características genômicas correlacionadas ao nicho, adaptabilidade ao hospedeiro e presença de alvos vacinais deste patógeno necessitam ser mais compreendidas, principalmente no genoma de linhagens isoladas de peixes. Portanto, os objetivos deste trabalho foram analizar comparativamente o genoma de 15 linhagens de *S. agalactiae*por analises de filogenômica, ilhas de patogenicidade e vias metabólicas, e também identificar alvos vacinais usando estratégias de vacinologia reversa e imunoinformática. Estas analálises revelaram que o pan-genoma da espécie é aberto e que as linhagens isoladas de peixes demonstraram ser mais conservadas quando comparadas com as linhagens isoladas de seres humanos. Uma elevada diversidade de ilhas de patogenicidade e vias metabólicas foi observada, e as caracteristicas genômicas relacionadas a virulencia em cada hospedeiro e o salvos vacinais identificados foram discutidos no trabalho. Com isso, este trabalho apresenta novas idéias em relacao a virulencia de *S. agalactiae* nos diferentes hospedeirose tambem sobre o desenvolvimento de vacinas contra este patógeno.

Palavras-chave: *Streptococcus agalactiae*, peixe, ser humano, bovino, pan-genômica, alvos vacinais.

Introduction

Streptococcus agalactiae (Lancefield group B; GBS) is a major bacterial pathogen that causes diseases affecting human, bovine and fish [1-3]. In humans, it is frequently associated with neonatal sepsis and meningitis, but can also affect immunocompromised adults besides also being a common colonizer of the gastrointestinal and genitourinary tracts [4,5]. In dairy cattle, GBS is an important pathogen of clinical and subclinical mastitis, affecting the milk quality and production [6-9]. In fish, *S. agalactiae* is an emergent pathogen that causes meningoencephalitis with considerable mortality in wild and cultured species worldwide [2,10-12]. Sporadically, this bacterium has been associated with illness in many others hosts, such as chickens, camels, dogs, horses, cats, frogs, hamsters, mice, monkeys, and crocodile [13-16].

There are currently ten serotypes (Ia, Ib, II-IX) described for this species and occasionally some strains can be classified as non-serotypeable [17]. For human and bovine many serotypes have been described causing disease [18-21], however, only three serotypes (Ia, Ib and III) has been associated to GBS disease in fish farms [12,22]. In addition, fish strains of *S. agalactiae* have been shown to be less diverse than human and bovine strains by molecular techniques such as pulsed-field gel electrophoresis (PFGE) and multi locus sequence typing (MLST) [12,23,24].

The genetic diversity of *S. agalactiae* populations isolated from different hosts has been studied and most of the works indicate that human, bovine and fish strains are frequently not related to each other [9,12,23,25]. However, some studies showed that human and bovine or human and piscine strains sporadically have the same or related genetic profile, suggesting that, cross-species transmission can occur [9,12,26,27]. To support this cross transmission, an infection study showed that one bovine strain was more virulent in newborn mice experimentally challenged than the human strain [18]. Other works showed

that human [23,28] and bovine [23] strains can infect and kill Nile tilapia (*Oreochromis niloticus*) in experimental challenge.

Several virulence factors have already been reported to be important in the pathogenesis of *S. agalactiae*, such as adhesins *fbsA* and *fbsB* (fibrinogen-binding protein A and B), *bibA* (immunogenic bacterial adhesin), *lmb* (laminin-binding protein), *scpB* (C5a peptidase), *pavA* (fibronectin-binding protein); pili proteins and *srr1* (serine-rich repeat glycoprotein Srr1); invasins *cfb* (CAMP factor), *cylE* (β-hemolysin/cytolysin) and *iagA* (invasion-associated gene); and immune system evasins such as capsule genes, *scpB* (C5a peptidase) and *cspA* (serine protease CspA) [29-31]. But the presence of theses virulence genes may vary according to the strain and origin host [9,20,32,33], for example, the genes *scpB* and *lmb* are present in almost all human isolates and they are found in less than 50% of bovine isolates [34]. However, little is known about the presence and importance of these virulence factors in strains isolated from fish [35].

Antibiotic therapy is largely used in human health and animal production as preventive and curative measure in GBS infections, however, it has many limitations [36-40]. Thus, immunoprophylaxis strategy is the most viable option to reduce the damage caused by this pathogen in human health and animal production [41,42]. Several types of vaccine formulations have been tested to use in human such as inactive bacteria, capsular carbohydrates and recombinant proteins exposed on bacterial surface. However, there were relatively few clinical trials with GBS vaccines in the last years [36]. Capsular polysaccharide (CPS) serotypes II and III based vaccines have elicited effective immune response, however there is little or no cross protection against different serotypes [43]. Therefore further studies are needed to investigate immune interference when more than two GBS CPS types are simultaneously administered [44]. Another factor that limits the CPS vaccination is the increasing number of non-

serotypeable strains [45] making the use of antigenic conserved protein in vaccine development highly desirable [46].

Currently, three distinct pili island (PI) are characterized in *S. agalactiae* and studies have shown that these genes have key role in bacterial adhesion and invasion in host during pathogenesis [47,48]. Additionally, promising results have been described in this context using a combination of these proteins located on PIs [41,46,49]. Additional studies have been performed using pilus proteins and it was suggested to utilize these target to design a universal vaccine against GBS infection in humans [42,46,49].

There are few studies of vaccines against mastitis by *S. agalactiae*, because this is usually easy to control through measures of hygiene and treatment [20]. However, a work that used the conserved Sip protein along with one *S. aureus* protein, showed higher efficacy of this chimeric vaccine than the inactivate vaccine of each one of these pathogens in experimental challenge in lactating mice [50]. In fish, bacterin vaccines against GBS have already been tested, and the vaccine protection was demonstrated to be variable between strains of the same serotype and different genotypes, suggesting that capsular serotype are not the only ones responsible for immunogenicity [22].

The first genomes of *S. agalactiae* sequenced was the genome of strains serotype III *S. agalactiae* NEM316 [51] and serotype V *S. agalactiae* 2603 V/R [52]. Later, Tettelin et al. [53] sequenced the genome of additional six strains (representative of five major disease-causing serotypes) and performed the pan-genome analysis of the *S. agalactiae* species using the genome sequences of eight human strains and showed that the pan-genome of this species is open, and it is expected that, for every new GBS genome sequenced, approximately 33 new strain-specific genes will be identified. After few years, the first genome of *S. agalactiae* strain isolated from bovine mastitis was published where 183 strain-specific genes were described, and about 85% of those genes were shown

to be clustered into eight genomic islands acquired through lateral gene transfer (LGT), suggesting environmental (metabolic pathways to fructose, lactose and nitrogen metabolism) and host (acquisition of potential virulence factor) adaptations of the pathogen [3]. Recently, the genome sequence of four additional *S. agalactiae* strains isolated from fish in China, Honduras and Brazil were reported [35,54,55] or deposited in GenBank. However, the genomic features of the fish strains remain poorly studied and complete comparative genome analysis of the species are necessary in order to better understand about molecular mechanisms involved in niche and host adaptation of these strains. Additionally, pan-genome analysis approaches can help in vaccine development [42].

Therefore, aiming to better understand the environment and host adaptations of *S. agalactiae*, the pan-genome analysis upgrade of this species was performed in this work using fifteen genome sequences of strains isolated from different hosts with further analyses of: comparative metabolic pathways; prediction of potential gene candidates to vaccine development and/or diagnostic; and, phylogenomic correlations with others bacteria of the *Streptococcus* genus. This work gives new insights about the genome structure of *S. agalactiae* in a comparative view at strain and host levels.

Materials and Methods

Genome sequences

The complete and draft genome sequences of 15 *S. agalactiae* strains were retrieved from the NCBI database (http://www.ncbi.nlm.nih.gov/genbank/): 10 strains isolated from humans (three complete genomes), 1 strain isolated from cow (draft genome) and 4 strains isolated from fishes (2 complete genomes) (Table 1). The human strains were isolated in USA, Italy, Canada and United Kingdom or unknown country. The clinical descriptions of these strains were

septicemia and meningitis or they were isolated from body sites not causing disease (colonizing the oral cavity or vagina). The bovine strain was isolated from clinical mastitis in the USA, and the fish strains were isolated from meningoencephalitis outbreak in China, Brazil and Honduras (Table 1).

Streptococcus genus phylogenomic analyses

The Gegenees (version 1.1.4) software was used to retrieve the GenBank sequences of all the genome sequences from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/) and, subsequently, to perform the phylogenomic analyses. Briefly, Gegenees fractionated the genome into small sequence fragments and performed an all-versus-all similarity search to set the minimum content shared by all the genomes. Next, to generate the percentages of similarity, the minimum shared content was subtracted from all the genomes, and the variable content of all strains was compared. These percentages were plotted in a heat map chart with a spectrum ranging from red (low similarity) to green (high similarity) [56]. After that, the distance matrix exported from Gegenees was used as an input file for the SplitsTree (version 4.12.6) software to generate a phylogenomic tree using the UPGMA grouping method [57,58].

Pan-genome, core genome and singleton analyses

The pan-genome analysis were performed for all of the following three datasets: A) all strains, using *S. agalactiae* strain A909 as a reference; B) human strains, using *S. agalactiae* strain A909 as a reference; and C) fish strains, using *S. agalactiae* strain SA20-06 as a reference; similarly as performed by Soares et al. [59]. The EDGAR (version 1.2) software [60] was used to calculate the pan-genome, core genome and singletons of the *S. agalactiae* species. EDGAR is a framework for the comparative analysis of prokaryotic genomes that performs

homology analyses based on a specific cutoff that is automatically adjusted to the query dataset. Briefly, the genome sequences of *S. agalactiae* were retrieved from GenBank, the homogenization of the genome annotations was conducted by GenDB (version 2.4) [61], an EDGAR project was created and homology calculations based on BLAST Score Ratio Values (SRVs) at protein level were performed. This results in a value ranging from 0 to 1 [62], which is multiplied by 100 and rounded in a percentage value of homology. Finally, a sliding window on the SRV distribution pattern is used to automatically calculate the SRV cutoff with EDGAR [60] . For this study, a SRV cutoff of 30 was estimated, and consequently genes that present a Bidirectional Best Hit higher than this cutoff were considered to be orthologous genes.

The subset of genes presenting orthologs in all the selected strains was considered the core genome. All the genes of subject strain X was compared with all genes of query strain Y, and only genes with orthologs in both strains were members of core XY. The resulting core XY was then compared with all genes of query strain Z to generate the core XYZ, and the comparisons continued in a reductive manner. The genes that represent the pan-genome was calculated in a similar way, but in an additive manner: the initial pan-genome was composed of all genes of strain X, and the non-orthologous genes of strain Y were added to create the pan-genome XY, the non-orthologous genes of strain Z were added to pan-genome XY to generate the pan-genome XYZ, and the subsequent strains in the same way. Finally, the singleton genes were calculated as genes that were present in only one strain and thus did not present orthologs in any other *S. agalactiae* genome of the analysis.

The developments of the core genome and singletons were calculated as described previously [53,63]. The α value of the pan-genome analysis estimate whether the pan-genome is open ($\alpha < 1$; meaning that for each newly sequenced genome, there will be new genes and the pan-genome will increase) or closed (α

> 1; where addition of new genomes will not significantly affect the pan-genome).

The core genome of all the 15 strains, and the differential core genes subset of human and fish strains were classified by Cluster of Orthologous Genes (COG) functional categories as follows: 1. Information storage and processing; 2. Cellular processes and signaling; 3. Metabolism; and 4. Poorly characterized. To perform this analysis, the query sets of core genes were submitted to BLAST protein (BLASTp) similarity searches against the COG database and the proteins with *E*-values higher than 10^{-6} were discarded. The best BLAST results for each protein were considered for the COG functional category information retrieval.

Pathogenicity island prediction

The pathogenicity islands (PAIs) of the genomes of *S. agalactiae* were assessed using PIPS: Pathogenicity Island Prediction Software (version 1.1.2). Summarizing, PIPS takes into consideration common features of PAIs, such as G+C content, codon usage deviation, high concentrations of virulence factors and hypothetical proteins, the presence of transposases and tRNA flanking sequences, and the absence of the query region in non-pathogenic organisms of the same genus or related species [64]. *Streptococcus thermophilus* strain CNRZ1066 was selected as the non-pathogenic organism of the same genus [65] , and separate predictions were performed for each genome. The sizes of the islands were compared with those of all the other strains via ACT: Artemis Comparison Tool (version 10.2.0) [66] and BRIG (BLAST Ring Image Generator) [67]. To perform this analysis, only complete genomes of *S. agalactiae* were used. Accurate prediction of pathogenicity islands in draft genomes is limited, as it was suggested in PIPS software work [64].

After this curation of PAIs, the genes of all the islands in each strain were assessed for their presence or absence in all the other strains using the pan-

genome data generated by EDGAR. The overall number of genes in the PAIs of the subject strain that were shared by the query strains was expressed as a percentage and plotted in a heat map. The number used in each PAI was standardized with the same number used by Glaser et al. [51] (PiSa with roman numbers of I to XIV). Newly predicted PAIs were named in arabic numbers (PiSa with arabic numbers of 1 to 11). The PAI´s comparison maps were visualized using the software BLAST Ring Image Generator (BRIG) [67].

Metabolic pathways prediction

The two main files used for reconstructing the *S. agalactiae* metabolic pathways were the artificial genome sequence file in FASTA format and the artificial genome annotation file in GBK format. These artificial files were generated from pan-genome or core genes multi-FASTA files of EDGAR by perl script, where unique artificial files in FASTA and GBK formats were generated. Posteriorly, Pathway/Genome Database (PGDB) for pan-genome of all 15 strains and core genes datasets of *S. agalactiae*(core genome of 15 strains and core genes subset of human or fish strains) were computationally predicted using Pathway Tools software version 16.0 [68], developed by SRI International. The MetaCyc, a highly curated, nonredundant reference database of small-molecule metabolism, is used as a reference database for the PathoLogic component of the Pathway Tools software [69]. The global pan-genome (15 genomes) metabolic pathways were used with reference genome to comparative analysis with the core genes datasets (core genome of 15 strains and core genes subsets of human or fish strains).

Antigenic candidate genes prediction

The prediction of cross-reactive vaccine candidates was performed according to the criteria described by Soares et al. [70]. The potential targets have to present

the following attributes: (rule-I) to be exposed to the host immune system, like secreted, surface-exposed and membrane proteins [71]; (rule-II) to show MHC I and II binding properties with adhesion probability greater than 0.51 and no similarity to human proteins when analyzed by Vaxign software [72]; and (rule-III) protein conservation among different genomes of *S. agalactiae* [70,72] into all 15 strains, only in human strains or only in fish strains.

The subcellular localization of proteins (rule-I) was predicted by an *in silico* analysis using the SurfG+ 1.0 tool [73], where a multi-FASTA file of pan-genome of 15 genomes of *S. agalactiae* was analyzed. SurfG+ pipeline searches protein motifs, including SignalP, LipoP and TMHMM, which are related to subcellular localization. It also creates novel HMMSEARCH profiles to predict cell wall retention signals. After that, it looks for retention signals, lipoproteins, SEC pathway export motifs, and transmembrane motifs. The proteins were characterized as cytoplasmic (CYT) if none of these motifs were found in their sequence. Moreover, SurfG+ has the ability to better distinguish between MEM (membrane proteins) and PSE (potentially surface exposed proteins), which may assist in the choice of possible vaccine targets. Thereby, proteins could be classified into four different subcellular locations: CYT, MEM, PSE, or SEC (secreted proteins).

In order to apply the rule II, the proteins predicted by SurfG+ as SEC, PSE and MEM were analyzed by the Vaxign software. The software evaluates the adhesion probability, protein conservation between different genomes in OrthoMCL database, and excludes sequences of nonpathogenic strains or host-similar proteins. *S. agalactiae* proteins with high similarity to human proteins were excluded from the analysis [72].

As one of the aims of this work was to search conserved vaccine candidates, the proteins predicted by EDGAR software into core genome of all strains, core genes subset of human strains and core genes subset of fish strains were

considered to suite the rule-III. However, the conservation of proteins (rule-III) does not exclude the antigenic target, once that a universal vaccine described by Maione et al. [42] had in its constitution proteins of accessory genome. Additionally, some antigenic proteins can be good vaccine targets against more prevalent serotypes in specific geographic region or host.

Results

Phylogenomics of the *Streptococcus* genus and *S. agalactiae* strains isolated from different hosts

According to the generated phylogenomic tree (Figure 1), all *S. agalactiae* strains are found to be grouped together in a distinct cluster and close to other pathogenic species such as *S. pyogenes*, *S. dysgalactiae*, *S. uberis*, *S. parauberis*. However, other pathogenic species such as *S. pneumoniae*, *S. pseudopneumoniae*, *S. suis* and *S. mutans* were allocated in a different cluster. *S. infantarius*, *S. macedonicus*, *S. pasterianus* and *S. gallolyticus* were grouped together forming a significantly close cluster. The non-pathogenic *S. thermophillus* [65] and non-casual pathogen *S. salivarius* [74] rooted the phylogenetic tree.

In the *S. agalactiae* cluster, the similarity ranged from 66 to 100% according to the heat map, the *S. agalactiae* ATCC13813 rooted this species cluster and the *S. agalactiae* FSLS3-026 bovine strain formed a separate branch. The four *S. agalactiae* fish strains grouped together in two different groups, one with the two serotype Ib strains SA20-06 and STIR-CD-17, and other one with the two serotype Ia strains (GD201008-001 and ZQ0910), and this latter group was rooted by strain *S. agalactiae* A909 serotype Ia. In human strains, the distribution is not related with the serotypes, once the serotype Ia strains A909 and 515 are grouped in different branches in the heat map, and the same occurred in the two serotype V strains 2603V/R and CJB111. Although COH1

and GB00112 serotype III strains grouped together, the other serotype III strain NEM316 grouped in the different branch.

The pan-genome of the species *S. agalactiae*

The pan-genome analysis of the *S. agalactiae* species (total number of non-redundant genes) was performed by EDGAR software (Figure 2). The resulting pan-genome of *S. agalactiae* contains a total of 5,143 genes, which is 2.35-fold larger than the average of genes in the 15 strains (2,182). A similar result was observed when the pan-genome was calculated using only human strains (4,730 genes), 2.09-fold the average total number of genes in human strains. Using fish strains, a significantly different scenario emerged, and the pan-genome was demonstrated to be more limited (2,176 genes), only 1.13-fold the average number of genes in fish strains. The pan-genome calculated for all the 15 strains of *S. agalactiae* was inferred as open, with $\alpha = 0.62$ ($\alpha = 1-\gamma$).

Core genome of the species *S. agalactiae*

The core genome of the species is the group of genes from the pan-genome that are shared by all strains utilized in pan-genome analysis. The core genome of *S. agalactiae*, using all the 15 strains, contains 1,111 genes, that represent less than a quarter of the pan-genome of the species (5,143 genes). The extrapolation of the curve can be calculated using the formula $n = \kappa * \exp[-x/\tau] + tg(\theta)$, where "n" is the expected group of genes for a given number of genomes, "x" is the number of genomes and the other terms are constants defined to fit the specific curve. According to this formula, the core genome subset of genes probably will decay with the addition of new genomes and tend to converge to ~993 genes (19.03% of the pan-genome of the species) (Figure 3).

The core genome analysis was also performed on human and fish strains separately (Figure 3). The core genes of *S. agalactiae* human strains contain

1,297 genes, and tend to stabilize in ~1,150 genes. The core genes of *S. agalactiae* fish strains consist of 1,523 genes, and tend to stabilize in ~1,507 genes. Then, considering that the core genome of species consists in 1,111 genes (calculated using the 15 genomes), and the pan-genome of human strains have 1,297 genes, one can calculate that the core genome of human strains have 186 orthologous genes that are shared by all strains of this host and are absent in one or more strains isolated from fish. Additionally, using the same concept, the core genome of fish strains (1,523 genes) contain 412 genes that are shared by all strains of this host and not present in one or more strains isolated from humans. Although this number is high, none of the fish strains differential core genes were absent in all human strain genomes. On the other hand, 4 genes of core genome of human strains (SAK_1318, SAK_1319, SAK_1991, SAK_2052) were absent in all strains isolated from fish. Additionally, there were 108 and 56 genes on core genes subset of fish and human strains, respectively, that are absent from the genome of bovine strain.

The classification by similarity analyses using the Cluster of Orthologous Groups (COG) of core genome of all 15 strains and the differential core genes subset of human and fish strains showed that a large proportion of genes of the core genome of all strains are classified under the categories metabolism, poorly characterized and information storage and processing (Figure 4). When analyzing the differential core genes subset of human strains, most of the genes were assigned in the categories metabolism, poorly characterized and cellular processes and signaling. When analyzing the differential core genes subset fish strains, a higher proportion of genes was classified as poorly characterized (Figure 4). Finally, the core genes subset of human strains had a larger proportion of genes classified in the category metabolism than core genes subset of the fish strains.

Singletons: strains-specific genes predicted in *S. agalactiae*

The singleton genes are those which are present in only one of the strains included in the pan-genome analysis. To calculate the expected number of new genes that each sequenced genome will add, it was used the formula $n = \kappa * \exp[-x/\tau] + tg(\theta)$. This analysis was performed for the three datasets: A) all genomes, B) only genomes of strains isolated from humans and C) only genomes of strains isolated from fish. Using this formula for all genomes, it is expected that each newly sequenced genome, will add approximately 80 genes to the pan-genome of the species (Figure 3). In a similar way, the analysis of singleton genes subset of human strains shows that ~103 genes are expected to be added. In contrast, this analysis when performed for fish strains, revealed a discrepant scenario where non-significant number of new genes will be added (Figure 3).

PAIs prediction in the *S. agalactiae* genomes

The 14 pathogenicity islands previously described in NEM316 strain genome [51] were correctly predicted here, the same number of this PAIs was conserved (I to XIV) and few small changes in the size of some PAIs were observed (Figure 5). When we analyzed all the 5 complete genomes we found 11 new PAIs (PiSa 1 to 11), high diversity on the pathogenicity islands content in different genomes and deletions of islands segments of distinct sizes between strains, showing the high genomic plasticity of this pathogen (Figure 5 and Figure 6). Three of 11 additional islands have been shown to harbor genes with importance in GBS pathogenicity (PiSa 1, PiSa 6 and PiSa 7) and, in general, the gene contents of the other 8 PAIs are poorly characterized.

Some virulence factor genes are found in PAIs of all five genomes analyzed, such as the *srr-1* gene that is found in PAI PiSa 7, and the genes *iagA* and *cylE*

that are harbored in PAI PiSa VI. These genes have already been described to be important in adhesion and invasion of host tissues [29,30].

The PAIs PiSa III, VII and VIII are present only in NEM316 strain, and represent the same island in three distinct positions inside the genome. In these island there are encoded some LPXTG proteins with unknown function, ABC transporter proteins and the protein Clp (Figure 5A). Clp proteins have been associated with modulation of virulence in *S. pneumoniae* mainly in colonization of host cells [75]. The PAI PiSa X of NEM316 strain has a region deleted in other strains. In this region, proteins with LPXTG domain proteins of unknown function were found. The virulence factor genes *scpB* (C5a peptidase) and *lmb* (laminin-binding surface protein) are present in PAI PiSa XII of human strains (NEM316, A909 and 2603 V/R) and absent in fish strains. These genes have already been described to be related to mobile genetic elements [76,77] and this can explain the exchange of this sequence segment of fish strains genome, suggesting the non-essential importance of these genes in pathogenesis of GBS disease in fish.

The *cfb* gene (CAMP factor protein) is present in the island PiSa XIII, and is conserved in all the 15 strains tested. Also in this island, the gene *gbs2018* was found. However, different alleles of this gene have been found in each strain. The PAI PiSa 11 is present only in A909 strain. The gene content of this island is mainly composed of phage related proteins which have unknown function.

The tail of the PiSa XI was larger in the NEM316 genome when compared to the other genomes. In this additional fragment, there are encoded eight poorly characterized proteins (hypothetical proteins). The PAI PiSa 6 of the strain 2603 V/R has an exclusive region that is absent in other genomes. The capsular genes are present in this island, and this region of deletion corresponds to the genes *cpsM* and *cpsO,* which are absent in serotypes Ia, Ib and III. The acquisition of capsular genes of other bacteria has already been suggested [78], and recently it

was reported the possibility of capsular switch between serotype III and IV among strains belonging to the hypervirulent CC17 [79].

Generally, a higher degree of conservation was observed in pathogenicity islands of the strains isolated from fish (Figure 5B). However, significant deletions of segments of pathogenicity islands were found when we compare the fish strains serotype Ia GD201008-001 and serotype Ib SA20-06. In the PAI PiSa 1, the important virulence gene *bac* (C protein beta-antigen) is present in serotype Ia fish strains and absent in fish strains serotype Ib. In human strains A909 and H36B this gene is also present, and is related with immune system evasion [80]. Moreover, there are two transposase genes flanking the *bac* gene in GD201008-001 which could explain deletion of this DNA fragment in SA20-06 strain or insertion in GD201008-001 strain.

The PAI PiSa IV was almost totally conserved in serotype Ia and Ib of fish and human strains. Unexpectedly, inside this island, the gene *bca* (C protein alpha-antigen) was absent in serotype Ib SA20-06 fish strain. This gene has been reported to help in the adherence and invasion of pathogen in host cells [81] and can be important in virulence of serotype Ia fish strains. Interestingly, while we tested the presence of *bac* and *bca* genes in *S. agalactiae* isolated from fish (serotypes Ia and Ib) by PCR, we could not detect both of these genes (data not shown).

Regarding the PAI PiSa VI of SA20-06, it was found to harbor bacteriocin genes which were absent in human strains and GD201008-001 fish strain. The island PiSa 3 start with CRISPRII operon type IC only in the SA20-06 strain, and these genes have important function in protection against mobile genetic elements in prokaryotes [82]. A significant deletion was observed in PiSa VI in serotype Ib fish strain SA20-06 (and human strains NEM316 and 2603 V/R), and this island was only conserved in serotype Ia human strain A909 and fish strain GD201008-001. The collagen-like protein gene is one of the genes lost by this deletion in

SA20-06 strain, and this protein has been related to promote adhesion in respiratory epithelial cells in group A *Streptococcus* [83]. The PAI PiSa IX was totally conserved only in serotype Ia strains isolated from fish (GD201008-001) and human (A909). The segment deleted in SA20-06 genome consists mainly in poorly characterized proteins (hypothetical proteins) and phage proteins.

The PAI PiSa XIV has been partially deleted in SA20-06 strain when compared with GD201008-001 strain. In this segment a LPXTG domain protein gene (A964_1958 - LPXTG-motif cell wall anchor domain protein) is located, which is flanked by phage and recombinase genes. This cell wall anchored protein is also present only in ZQ0910 strain, suggesting that it may be conserved in the serotype Ia fish strains.

Unexpectedly, the pilus island was not predicted as a pathogenicity island in the *S. agalactiae* genomes. When we analyzed these segments of genomes, we could not observe codon usage or G+C content deviation, and, moreover phage or transposases proteins were absent. The only PAI features of these segments were the absence of this region in non-pathogenic *S. thermophilus* and similarity of some proteins with virulence factors. However, these pili islands may have been acquired by LGT and lost these characteristics due to codon adaptation, which makes their prediction as PAIs limited [64]. Clearly, the pili islands will be discussed in this work due to their relevance in virulence and vaccine studies and, also, to their previous prediction as pathogenicity islands [84]. The comparative viewer tool of EDGAR software was used in order to visualize the pili islands in all the 15 genomes, and additional BLASTp analysis were performed to confirm the presence of each pili protein. However when the pilus island is absent in draft genomes we can not exactly know whether this island is really absent or was not represented in the genome assembly. Pilus island 1 is only present in NEM316, A909 and 2603 V/R human strains. The pilus island 2a was only found in NEM316 and 2603 V/R human strains and pilus island 2b was

present in SA20-06 and GD201008-001 fish strains and in human strain A909 (Figure 6).

Metabolic pathways prediction

In order to identify metabolic features that are conserved in all strains or only in human, bovine or fish strains separately, the metabolic pathways analysis was performed comparing the core genome of all strains; or core genes subset of human or fish strains; or of bovine genome against the metabolic pathways of pan-genome of all the 15 strains. In the pan-genome file, we have identified 192 pathways and 1075 enzymatic reactions. In core genome of all strains and core genes subset of human strains and fish strains 115, 130 and 132 pathways and 686, 726 and 747 enzymatic reactions were found, respectively. The bovine strains showed 170 pathways and 953 enzymatic reactions. In general, the following pathways were conserved in all datasets: degradation of proteins, carbohydrates (ribose, D-mannose, fructose, glucose and xylose), aspartate, arginine, alanine and ethanol; conversion of acetyl-CoA in ethanol; the biosynthesis of purines, pyrimidines and secondary metabolites; metabolism of inorganic nutrients and tRNA´s; glycolysis and pentose phosphate pathway.

Some differences in pathways were observed, being some pathways conserved in all fish strains and bovine strain and not in all human strains such as pyruvate conversion to (S)-lactate, sucrose degradation; some steps of fatty acids biosynthesis and elongation and steps of some amino acids biosynthesis. In contrast, some pathways were conserved between all human and bovine strains, but not in all fish strains, such as acetaldehyde biosynthesis II, glutamate biosynthesis II and III, and a different enzyme used by all fish strains in glycerol degradation.

The enzyme galactokinase (SAI_0497) was only found in serotype Ib fish strains SA20-06 and STIR-CD-17 and in human strains H36B (also Ib serotype). This

gene has a premature stop codon in A909 strain, probably making the metabolic pathway incomplete or non-functional. This shows that the galactose degradation I pathway is only completely present in these strains. However, the other galactose degradation pathway (lactose and galactose degradation I) is only completely present in bovine and human ATCC13813 strains (Figure 7).

The melibiose degradation pathway was present in all fish strains and only in two human strains (H36B and A909), resulting in alpha-D-galactose and beta-D-glucose. The starch degradation was conserved in all human strains, in bovine strain and two serotypes Ia fish strains (ZQ0910 and GD201008-001). The conversion of raffinose to alpha-D-galactose and sucrose and the conversion of an alpha-D-galactoside to alpha-D-galactose and an organic molecule were present in A909 and H36B human strains (serotypes Ia and Ib) and in all fish strains.

Prediction of cross-reactive vaccine candidates for *S. agalactiae*

The SurfG+ analyses of 5,143 predicted proteins of pan-genome of all the 15 strains of *S. agalactiae* identified 156 SEC proteins, 398 PSE proteins, 773 MEM proteins and 3,816 CYT proteins (rule I). The 1,327 proteins classified as SEC, PSE or MEM were evaluated by Vaxign, and 150 presented adhesion probability greater than 0.51 (Table S1). Eight proteins (SAK_1493, SAI_0146, SAK_0722, SAK_0477, SAK_0477, SAK_2135, SAK_0084, and SAK_0084) were shown to be similar to mammalian proteins and were excluded from the analyses. Thirty six proteins were present in the core genome of the 15 *S. agalactiae* strains. A total of 41 candidates were found to be conserved in genomes of human isolates (Table 2). From those, five proteins (SAK_0050, SAK_1293, SAK_1319, SAK_066, and SAK_1074) were exclusively verified in the strains of this host. No exclusive vaccine targets were found in fish isolates.

Seven proteins (SAK_2073, SAK_0337, SAK_1994, SAK_1009, SAK_0556, SAK_2007, and SAK_0854) selected by the procedure did not present any similarity with proteins of known functions. Seventeen proteins (SAK_1426, SAK_1656, SAK_0604, SAK_0457, SAK_0166, SAK_1870, SAK_1625, SAL_1416, SAK_1503, SAK_0321, SAK_0301, SAK_0442, SAK_1580, SAK_1235, SAK_0553, SAK_1293, and SAK_1074) were found to be similar to known proteins, but with uncharacterized functions in *S. agalactiae*. Eleven proteins (SAK_0050, SAK_1271, SAK_1497, SAK_0932, SAK_1394, SAK_1158, SAK_1087, SAK_1784, SAK_1109, SAK_1927, and SAK_0685) were found to be similar to virulence or metabolic proteins of other pathogenic *Streptococcus* species, but, with uncertain function in *S. agalactiae*. In addition, some virulence factors or immunogenic proteins fully or partially characterized in *S. agalactiae* were predicted as candidates, such as Sip (SAK_0065), laminin binding protein (Lmb; SAK_1319), penicillin- binding protein (PBP; SAK_0370 and SAK_0222), immunodominant A antigen (SAK_2105), and zoocin A (SAK_0064).

Discussion

Phylogenomic analysis of the *Streptococcus* genus

In bacteria, the classical phylogenetic analysis performed by the use of rRNA sequences is largely utilized; however, it is often impossible to define a precise evolutionary relationship, mainly in closely related species [85]. Currently, this type of analysis is more reliable when performed using a large number of genes or whole genomes, once that it makes possible to have less interference of variable mutation rates, horizontal gene transfer (HGT) or misalignments [86] and not to use only conserved regions that could generate tendentious results [56]. Also, phylogenetic methods that use a large set of sequences have become standard for phylogeny studies [87].

Nevertheless, according to the 16S rRNA gene tree of the 12 *Streptococcus* species, the pathogenic species *S. agalactiae*, *S. equi*, *S. pyogenes*, *S. dysgalactiae* grouped in close branches [88], as noted in the phylogenomic analysis in this work. There was observed a conservation of close relationship in the group mitis species *S. mitis* and *S. pneumoniae* in all analysis methods (16S rRNA, and phylogenomic analysis in this work). Moreover, some conservation degree was observed between *S. infantarius*, *S. macedonicus*, *S. pasteurianos* and *S gallolyticus* (equinus group); *S. dysgalactiae* subsp *equisimilis* and *S. pyogenes* (pyogenic group); and *S. mitis*, *S. oralis*, *S. pneumoniae* and *S. pseudopneumoniae* (mitis group). These results suggest that there is some conservation degree in the variable genome regions in some species of these groups and that this conservation was extensive to more than two species in equinus and mitis groups and limited only between two species of pyogenic group (*S. dysgalactiae* subsp *equisimilis* and *S. pyogenes*).

Genomic analysis in *S. agalactiae*species – host adaptability

At the *S. agalactiae* species level, many studies using different molecular methods have shown high genetic diversity in this species population isolated from human, bovine an fish [4,9,10,33], and this fact has also been suggested here once that, into *S. agalactiae* cluster, the similarity ranged from 66 up to 100 % between 15 strains. In spite of different molecular methods, MLST is the most used technique to determine the evolutionary relationship in this bacterial species and has been correlated with host adaptability and virulence [26,89]. Although most studies has been concluded that *S. agalactiae* isolated from different hosts belong to distinct populations, some works have demonstrated that human and bovine strains can infect fish [23,28] and that bovine strain was virulent in experimental challenge in mice [18]. Evans et al. [28] showed that serotype Ia ST-7 strain isolated from human caused mortality in experimental

challenge in fish. In the work of Pereira et al. [23], it was demonstrated that human strain serotype V and ST-26 and bovine isolate non-serotypeable and ST-256 infected Nile tilapia resulting in mortality. Additionally in this work, all human strains experimentally inoculated in fish can be recovered 48h post infection and this observation was limited to only few bovine strains [23]. In the heat map (Figure 1), fish strains are grouped close to human strains, and bovine strain grouped separately of the human and fish strains. Strains of serotype Ia and ST-7 have been described causing diseases in humans [1,90] and fish [12,28], and in the phylogenomic analysis shown here, a close relationship can be observed between the A909 human strain (serotype Ia and ST-7) and the two fish strains (GD201008-001 and ZQ0910) (Figure 1). Although the bovine strain has been grouped separately from human and fish strains, this bovine strain belongs to the ST-67 which is considered host adapted [26] . A high proportion of genes acquired through lateral gene transfer were observed in the genome of this strain, showing the importance of performing more studies using bovine strains of other ST´s and clonal complexes.

Particularly in fish strains, serotype Ia strains grouped together forming a sub-group and the same occurred to serotypes Ib strains. Taking the number of genes into account, along with the genome size and STs of strains of each serotype we can suggest that strains of serotype Ib and ST-260 or ST-253 (or other similar STs) are host adapted to fish and that serotype strains Ia and ST-7 (and other serotypes and ST combinations) can infect both human and fish hosts. Additionally, some STs has been described to adaptability to host, such as CC-67 which is considered adapted to bovine host [26].

Although the *Streptococcus* genus is a heterogenic and complex group and controversial results have been found when limited methods are used, the phylogenomic analysis presented here is a better alternative to understand the phylogenetic relationships intra and interspecies in the genus.

The pan-genome of 15 *S. agalactiae* strains shown here, isolated from three different hosts, is considered open as described by Tettelin et al. [53] who used eight strains isolated from human to accomplish his analysis, and suggested that new genes will still be found after sequencing more strains. In this previous work, it was proposed that each newly sequenced genome, will possibly add ~33 new genes to the pan-genome of the species. However, here we demonstrated that each newly sequenced genome will increase the pan-genome size by ~80 new genes. When we performed the pan-genome analysis in EDGAR software using the same eight strains used by Tettelin et al. [53] (data not shown), we found that ~114 new genes will be added as a result of newly sequenced genome and not 33 as previously suggested. In general, the conclusion that the pan-genome is open is in agreement with the previous work, however, these differences in the prediction of the number of new genes that will be added to the pan-genome is mainly due methodological difference. To identify the orthologous genes, Tettelin et al. [53] used the intuitive cutoff based on the result of DNA and protein sequence alignments, which possibly contributed to the occurrence of a different result. On the other hand, using EDGAR, the orthology thresholds were generated automatically based on BLAST Score Ratio Values (SRVs), estimating the appropriate cutoff for every genome datasets independent of the degree of conservation in the species or genus genome groups [60].

The pan-genome of all strains proved to be more than two times larger than the mean number of genes of the 15 genomes, and a similar result was observed when only human strains were analyzed. Using only the fish strains, the pan-genome was only 1.13-fold larger than the mean amount of genes of these isolates, suggesting more limited diversity in this pathogen population. Supporting this fact, strains from fish were found to have one of the smallest

numbers of singleton genes (Table 1) and have also presented a higher degree of conservation in core genes subset (Figure 3).

With respect to the core genome of all the 15 strains and subtractive core genes subset of human strains, the large proportion of genes related to the COG category "Metabolism" suggesting a high degree of conservation in the basal metabolic pathways in the species. The "Metabolism" category consists in genes involved in the production and conversion of energy, as well as the transport and metabolism of carbohydrates, amino acids, nucleotides, coenzymes, lipids, inorganic ions and secondary metabolites. In the subtractive core genes subset of fish strains, the larger proportion of genes were classified as "Poorly characterized", indicating that even though this core genes subset is bigger, the function of the most of these genes is not known (Figure 4).

Some features of the core genes subset of fish strains were observed as related to energetic metabolism such as sucrose and melibiose degradation. In core genes subset of human strains, serotype Ia fish strains and bovine strain, the starch degradation pathway was conserved. In fish strains, the pathways for utilization of sucrose (conserved in all fish strains), raffinose (Ib fish strains) and starch (Ia fish strains) suggest an adaptive advantage, once that these sugars are present in organic matter derived from plants [91]. In *S. mutans*, the presence of these carbohydrates (sucrose and starch) resulted in dynamic remodeling of the transcriptional profile, increasing the expression of genes related to virulence [92]. In a recent work in *S. suis*, the gene *ccpA* has already been described as modulating carbohydrate metabolism (mainly glucose and sucrose) and was also strongly related to the regulation of virulence genes [93]. This gene is conserved in other *Streptococcus* species including *S. agalactiae*, and is present in all fish strains, bovine strain and five human strains (2603V/R, NEM316, ATCC13813, GB00112 and A909). Thereby, the utilization of these sugars in *S. agalactiae*

may be related not only to energy obtainment but also the regulation of virulence.

The pilus island 2b was found in all fish strains, in bovine strain and four human strains (A909, COH1, ATCC13813 and GB00112). Pilus structure in *S. agalactiae* was described in the last decade, and it has been shown to have a pivotal role in adhesion and invasion in host cells [94,95]. The serotype Ib fish strains show a the different size in the backbone protein (also observed in human strain ATCC13813 and bovine strain) and the sortase B is absent in STIR-CD-17 strain, when compared with A909 human strain (Figure 6). Even so ancillary pilus proteins can be secreted by sortase A or even anchored in the cell wall, from where it may interact with the host [96]. Although in these genomes the backbone protein is smaller than in A909 strain genome, the functionality of these small proteins remains unclear. Interesting, fish strains only harbor this pilus island; it has been described as rare event in human strains and has been related with hypervirulent lineages (particularly the clonal complex 17- CC17); and, it is also related to tropism [97]. We have found only the pilus island 2b in the majority of fish strains isolated from Brazilian fish farms (data not shown) and this indicates the importance of this pilus type in the pathogenicity and virulence of GBS in fish. Taken together, the absence of pilus island 1 was observed as an unusual event, more related with serotype Ia [97], although here we showed that it can also occur in serotype Ib fish strains. In bovine strains related with ST-67 this event was also observed, suggesting independent evolution between strains of different hosts [8].

Additionally, the pilus island 2a was completely found in six human strains (2603 V/R, NEM316, 18RS21, H36B, 515 and CJB111), and pilus island 1 was observed as conserved in seven human strains (A909, 2603 V/R, 18RS21, NEM316, CJB111, COH1 and GB00112). The combination of these two pilus islands (PI-1 and PI-2a) was observed in four human strains (NEM316, 2603

V/R, 18RS21 and CJB111) and it was associated with maternal colonization and invasive disease in adults and the combination of pilus island 1 and 2b (observed in human strains A909 and COH1) was related with neonatal infection [97].

The CRISPR loci type II-A system (*cas9*, *cas1*, *cas2* and *csn2* genes) was conserved in all the 15 strains of *S. agalactiae*. Interestingly, the CRISPR loci type I-C system was only found in fish strains serotype I-b, and in SA20-06 genome, the complete operon of these genes is located in the PAI PiSa 3. CRISPR system has been associated with immunity against phages, plasmids and more generally against mobile genetic elements (MGEs) [82]. CRISPR type II-A was showed to be highly dynamic and active in *S. agalactiae* species and contributes to the MGEs diversity in the population [98,99]. However, the type I-C system is present in few strains, frequently fragmented and has only been described as a complete operon in GBS strains isolated from a frog and belonging to ST-260, the same ST of fish strain STIR-CD-17 and other fish strains [98]. The *cas* genes of CRISPR type I-C system have high similarity with *cas* genes of *S. pyogenes* and *S. dysgalactiae* subsp. *equisimillis* and therefore have probably been laterally transferred between these species. Additionally, this type of CRISPR system is more conserved among unrelated strains suggesting that it can be adaptively advantageous in specific environmental conditions [98], such as aquatic environment.

Bacteriocin genes (SaSA20_0544 and SaSA20_0545, in SA20-06 genome strain) were only found in serotype I-b fish strains located inside the PAI PiSa VI, which have approximately 65% of similarity with genes *blpM* and *blpN* of *S. pneumoniae*. These genes have been reported as related to intraspecies [100] and interspecies competition [101]. This feature can be very useful to survival of GBS fish strains in aquatic environment where there is a vast diversity of microorganisms. Upstream to these bacteriocin genes, also in the PAI PiSa VI, two abortive infection related proteins (SaSA20_0542 and SaSA20_0543) were

found to be encoded in SA20-06 fish strain, which have approximately 63% of similarity with the abortive infection AbiGI and AbiGII proteins of *S. macedonicus* ACA-DC 198 strain, respectively. The AbiGI protein was found in the other serotype Ib fish strain STIR-CD-17 (a draft genome) with high similarity with the protein of SA20-06 (SaSA20_0542). Although AbiGI and AbiGII proteins were present in the genome of 2603 V/R strain in the PAI XII and are also found in genomes of other human strains and bovine strain, these AbiG proteins have high similarity with proteins of *S. dysgalactiae* subsp. *equisimilis* and S. *suis* and low similarity with AbiG proteins of SA20-06, suggesting a different source of acquisition in fish and human or bovine strains. In *Lactococcus lactis* the abortive infection mechanism is considered the most efficient group of antiphage system that inhibit the multiplication of various lactococcal phage groups [102-104] and has been suggested to originate from horizontal gene transfer from a species of low GC content [104]. However, the gene under the locus tag SaSA20_0543 (AbiGII) has a premature stop codon in SA20-06 genome suggesting that this antiphage mechanism may be non-functional in this strain, but this adaptive advantage can exist (in a functional way) in other fish strains. In a similar way, in the bovine strain FSLS3-026, the nisin operon was identified, but there was no nisin production due to truncation of the gene *nsaB* by an insertion element [3].

Another gene shared only by the two serotype Ib fish strains is the allele 6 of the gene *gsb2018*. This gene is flanked by *secE* and *nusG* genes and the majority of its alleles has been described as adhesins and named *bibA* gene [31] . One allele of this gene is named *hvgA* and has been associated with the high virulence of CC17 strains in humans [105]. The allele 6 of this gene has only been described in a trout isolate [1] suggesting a need for a functional characterization to assay the putative importance of this allele in fish disease caused by serotype Ib GBS

strains. Supporting this, we screened serotype Ia and Ib *S. agalactiae* fish strains and this allele was present in all 26 strains analyzed (data not shown).

In serotype Ia fish strains, many of the proteins only found in these two strains are poorly characterized and annotated as hypothetical proteins. However, the protein annotated as LPXTG cell wall surface protein (A964_1958, in GD201008-001 genome, present in the PAI PiSa XIV) can interact with the host once that it is anchored in the bacterial wall. Although not yet characterized, this protein may have a function in virulence and/or be a good vaccine target against outbreak diseases caused by this serotype in fish farms.

In general, the majority of the singleton genes from all strains are hypothetical proteins (mainly in complete genomes) or proteins with partial sequences (draft genomes). This fact suggests a limitation on pan-genome analysis when draft genomes are used. However, in SA20-06 fish strain, a gene (SaSA20_0706) showed low similarity with fibrinogen binding protein (*fbsB*) of other *S. agalactiae* strains, although it has the same flanking genes in the genomes. This gene was reported as important in adhesion and invasion in epithelial cells [106] and it suggest that this different can probably be related with virulence in fish.

Additionally, the presence of virulence factor genes *srr-1*, *cylE*, *cfb* and *iagA* (harbored into PAIs) and *sodA*, *hylB*, pavA and *fbsA* (not related to PAIs) in the genome of all the 15 strains analyzed in this work suggest that these virulence genes have significant importance in pathogenesis of GBS independently of host. Even though these genes are present and show in general high similarity, some them have broad variation of gene size (*srr-1*, *cylE* and *fbsA*) and certainly this fact causes interference in functionality of this genes. The genes srr-1 and iagA have already proven function in cross of blood-brain barrier in human strains, suggesting an importance of functional characterization of these genes in fish pathogenesis.

Prediction of cross-reactive vaccine candidates

The pathogenesis of GBS disease in human newborns and fish shows certain degree of similarity with acute infections characterized by septicaemia and meningitis [2,43]. The main treatment applied in both hosts is the administration of antibiotics to mother and to diseased fish, respectively. However, this approach has many complications, limitations, and failure problems [22,107]. The use of vaccines comprises the most attractive alternative to treat the disease. Clinical trials of conjugate vaccines with purified capsular polysaccharides have demonstrated to be safe and immunogenic in humans, but, they just elicit protection against the same serotype [43]. A similar issue was verified in fish vaccine trials, where in general, low protection rates were verified in heterologous challenge with GBS isolates from distinct PFGE types [22]. Therefore, the development of broad protective vaccines is highly needed to control GBS infections.

Little success has been achieved in developing universal GBS vaccines by conventional approaches based on cultivation of bacteria and identification of immunogenic antigens [43]. Some exceptions are Sip and GBS pilus proteins, which have presented satisfactory results of cross-protection. However, low exposure levels and absence in some GBS strains, respectively, may compromise the applicability of those antigens to universal vaccines development [31,46]. The reverse vaccinology, based on the *in silico* identification of novel immunogenic candidates, opened a new way to the vaccine development. It improved the efficiency and time needed to antigen discovery [108] . A previous study evaluated the pan genome of eight GBS strains and performed the prediction of vaccine candidates [42] . They analyzed an initial number of 589 proteins. From those, 312 were successfully expressed; however, just four of them elicited protection. The authors used just as *in silico* parameter for the screening that was the genes encoding putative surface-

associated and secreted protein; it could explain poor results obtained. Herein, the prediction of vaccine candidates in the pan genome of 15 *S. agalactiae* strains from different hosts was performed. Our approach was able to identify 36 potential targets for all strains, and 41 for human isolates from an initial broad of 1327 proteins. In addition to the evaluation of genes of membrane associated and secreted proteins, the adhesion probability of proteins to MHC I and II receptors was estimated. These analyses may provide higher accuracy to potential immune response induced by the candidates.

The *in silico* screening identified some well characterized virulence factors or immunogenic proteins such as Sip (group B streptococcal surface immunogenic protein) and LmB proteins in the broad of candidates; they were previously proved to be good vaccine targets [42,43]. Therefore, the method used to predict the vaccine candidates in the present work seems to be feasible. Future studies have to be carried out to evaluate the protection elicited by those candidates in mice and fish models. Finally, the prediction selected 35 proteins without known function in *S. agalactiae* (Table 2). Those could be associated with the pathogenesis, and their characterization could provide new insights about GBS infection.

Conclusions

With the data presented here we can conclude that the strains isolated from fish and humans are closely related phylogenetically than compared between the strain of cattle and human strains. This may be related to the fact that this pathogen causes a disease limited to mammary gland in cattle, while in fish and human, the pathogen is associated with a systemic disease. Comparative genomic studies conducted here showed important characteristics of the population of *S. agalactiae* in general and in each host separately, opening doors to new studies of functional characterization of proteins and vaccine

development. However, these differences certainly do not answer all questions in relation to the real importance of each one of these features at the time of infection in host. Studies of expression arrays can assist in the understanding of host-pathogen interaction.

Acknowledgement

This work was supported by Ministério da Pesca e Aquicultura, Furnas Centrais Elétricas, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). We also acknowledge support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Rede Paraense de Genômica e Proteômica.

References

1. Brochet M, Couvé E, Zouine M, Vallaeys T, Rusniok C et al. (2006) Genomic diversity and evolution within the species *Streptococcus agalactiae*. Microbes Infect 8: 1227-1243.

2. Mian GF, Godoy DT, Leal CAG, Yuhara TY, Costa GM et al. (2009) Aspects of the natural history and virulence of *S. agalactiae* infection in Nile tilapia. Vet Microbiol 136: 180-183.

3. Richards VP, Lang P, Bitar PDP, Lefébure T, Schukken YH et al. (2011) Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted *Streptococcus agalactiae*. Infect Genet Evol 11: 1263-1275.

4. Gherardi G, Imperi M, Baldassarri L, Pataracchia M, Alfarone G et al. (2007) Molecular epidemiology and distribution of serotypes, surface proteins, and

antibiotic resistance among group B streptococci in Italy. J Clin Microbiol 45: 2909-2916.

5. Park SE, Jiang S, Wessels MR (2012) CsrRS and environmental pH regulate group B *streptococcus* adherence to human epithelial cells and extracellular matrix. Infect Immun 80: 3975-3984.

6. Karlsmose S, Kunstmann L, Rundsten CF, Krogh K, Larsen HKD et al. (2012) External quality assurance system (EQAS) for identification of mastitis pathogens in Denmark from 2006 to 2011. Prev Vet Med in press.

7. Merl K, Abdulmawjood A, Lämmler C, Zschöck M (2003) Determination of epidemiological relationships of *Streptococcus agalactiae* isolated from bovine mastitis. FEMS Microbiol Lett 226: 87-92.

8. Sørensen UBS, Poulsen K, Ghezzo C, Margarit I, Kilian M (2010) Emergence and global dissemination of host-specific *Streptococcus agalactiae* clones. MBio 1(3):e00178-10.

9. Dogan B, Schukken YH, Santisteban C, Boor KJ (2005) Distribution of serotypes and antimicrobial resistance genes among *Streptococcus agalactiae* isolates from bovine and human hosts. J Clin Microbiol 43: 5899-5906.

10. Chen M, Li L, Wang R, Liang W, Huang Y et al. (2012) PCR detection and PFGE genotype analyses of streptococcal clinical isolates from tilapia in China. Vet Microbiol 159: 526-530.

11. Evans JJ, Klesius PH, Shoemaker CA (2004) Efficacy of *Streptococcus agalactiae* (group B) vaccine in tilapia (*Oreochromis niloticus*) by intraperitoneal and bath immersion administration. Vaccine 22: 3769-3773.

12. Evans JJ, Bohnsack JF, Klesius PH, Whiting AA, Garcia JC et al. (2008) Phylogenetic relationships among *Streptococcus agalactiae* isolated from piscine, dolphin, bovine and human sources: a dolphin and piscine lineage

associated with a fish epidemic in Kuwait is also associated with human neonatal infections in Japan. J Med Microbiol 57: 1369-1376.

13. Yildirim AO, Lämmler C, Weiss R (2002) Identification and characterization of *Streptococcus agalactiae* isolated from horses. Vet Microbiol 85: 31-35.

14. Yildirim AO, Lämmler C, Weiss R, Kopp P (2002) Pheno- and genotypic properties of streptococci of serological group B of canine and feline origin. FEMS Microbiol Lett 212: 187-192.

15. Elliott JA, Facklam RR, Richter CB (1990) Whole-cell protein patterns of nonhemolytic group B, type Ib, streptococci isolated from humans, mice, cattle, frogs, and fish. J Clin Microbiol 28: 628-630.

16. Bishop EJ, Shilton C, Benedict S, Kong F, Gilbert GL et al. (2007) Necrotizing fasciitis in captive juvenile *Crocodylus porosus* caused by *Streptococcus agalactiae*: an outbreak and review of the animal and human literature. Epidemiol Infect 135: 1248-1255.

17. Slotved H, Kong F, Lambertsen L, Sauer S, Gilbert GL (2007) Serotype IX, a Proposed New *Streptococcus agalactiae* Serotype. J Clin Microbiol 45: 2929-2936.

18. Corrêa ABA, Américo MA, Oliveira ICM, Silva LG, de Mattos MC et al. (2010) Virulence characteristics of genetically related isolates of group B streptococci from bovines and humans. Vet Microbiol 143: 429-433.

19. Duremdez R, Al-Marzouk A, Qasem JA, Al-Harbi A, Gharabally H (2004) Isolation of *Streptococcus agalactiae* from cultured silver pomfret, *Pampus argenteus* (Euphrasen), in Kuwait. J Fish Dis 27: 307-310.

20. Duarte RS, Miranda OP, Bellei BC, Brito MAVP, Teixeira LM (2004) Phenotypic and molecular characteristics of *Streptococcus agalactiae* isolates recovered from milk of dairy cows in Brazil. J Clin Microbiol 42: 4214-4222.

21. Palmeiro JK, Dalla-Costa LM, Fracalanzza SEL, Botelho ACN, da Silva Nogueira K et al. (2010) Phenotypic and genotypic characterization of group B streptococcal isolates in southern Brazil. J Clin Microbiol 48: 4397-4403.

22. Chen M, Wang R, Li L, Liang W, Li J et al. (2012) Screening vaccine candidate strains against *Streptococcus agalactiae* of tilapia based on PFGE genotype. Vaccine 30: 6088-6092.

23. Pereira UP, Mian GF, Oliveira ICM, Benchetrit LC, Costa GM et al. (2010) Genotyping of *Streptococcus agalactiae* strains isolated from fish, human and cattle and their virulence potential in Nile tilapia. Vet Microbiol 140: 186-192.

24. Oliveira ICM, De Mattos MC, Areal MFT, Ferreira-Carvalho BT, Figuiredo AMS et al. (2005) Pulsed-field gel electrophoresis of human group B streptococci isolated in Brazil. J Chemother 17: 258-263.

25. Sukhnanand S, Dogan B, Ayodele MO, Zadoks RN, Craver MPJ et al. (2005) Molecular subtyping and characterization of bovine and human *Streptococcus agalactiae* isolates. J Clin Microbiol 43: 1177-1186.

26. Bisharat N, Crook DW, Leigh J, Harding RM, Ward PN et al. (2004) Hyperinvasive neonatal group B streptococcus has arisen from a bovine ancestor. J Clin Microbiol 42: 2161-2167.

27. Oliveira ICM, de Mattos MC, Pinto TA, Ferreira-Carvalho BT, Benchetrit LC et al. (2006) Genetic relatedness between group B streptococci originating from bovine mastitis and a human group B *Streptococcus* type V cluster displaying an identical pulsed-field gel electrophoresis pattern. Clin Microbiol Infect 12: 887-893.

28. Evans JJ, Klesius PH, Pasnik DJ, Bohnsack JF (2009) Human *Streptococcus agalactiae* isolate in Nile tilapia (*Oreochromis niloticus*). Emerg Infect Dis 15: 774-776.

29. Lin FP, Lan R, Sintchenko V, Gilbert GL, Kong F et al. (2011) Computational bacterial genome-wide analysis of phylogenetic profiles reveals potential virulence genes of *Streptococcus agalactiae*. PLoS One 6: e17964.

30. Seo HS, Mu R, Kim BJ, Doran KS, Sullam PM (2012) Binding of glycoprotein Srr1 of *Streptococcus agalactiae* to fibrinogen promotes attachment to brain endothelium and the development of meningitis. PLoS Pathog 8: e1002947.

31. Santi I, Maione D, Galeotti CL, Grandi G, Telford JL et al. (2009) BibA induces opsonizing antibodies conferring in vivo protection against group B *Streptococcus*. J Infect Dis 200: 564-570.

32. Dore N, Bennett D, Kaliszer M, Cafferkey M, Smyth CJ (2003) Molecular epidemiology of group B streptococci in Ireland: associations between serotype, invasive status and presence of genes encoding putative virulence factors. Epidemiol Infect 131: 823-833.

33. Corrêa ABDA, Oliveira ICMD, Pinto TDCA, Mattos MCD, Benchetrit LC (2009) Pulsed-field gel electrophoresis, virulence determinants and antimicrobial susceptibility profiles of type Ia group B streptococci isolated from humans in Brazil. Mem Inst Oswaldo Cruz 104: 599-603.

34. Zadoks RN, Middleton JR, McDougall S, Katholm J, Schukken YH (2011) Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans. J Mammary Gland Biol Neoplasia 16: 357-372.

35. Liu G, Zhang W, Lu C (2012) Complete genome sequence of *Streptococcus agalactiae* GD201008-001, isolated in China from tilapia with meningoencephalitis. J Bacteriol 194: 6653.

36. Heath PT (2011) An update on vaccination against group B *streptococcus*. Expert Rev Vaccines 10: 685-694.

37. Edmond KM, Kortsalioudaki C, Scott S, Schrag SJ, Zaidi AKM et al. (2012) Group B streptococcal disease in infants aged younger than 3 months: systematic review and meta-analysis. Lancet 379: 547-556.

38. Cañada-Cañada F, Muñoz de la Peña A, Espinosa-Mansilla A (2009) Analysis of antibiotics in fish samples. Anal Bioanal Chem 395: 987-1008.

39. Mitchell JM, Griffiths MW, McEwen SA, McNab WB, Yee AJ (1998) Antimicrobial drug residues in milk and meat: causes, concerns, prevalence, regulations, tests, and test performance. J Food Prot 61: 742-756.

40. Millanao B A, Barrientos H M, Gómez C C, Tomova A, Buschmann A et al. (2011) [Injudicious and excessive use of antibiotics: public health and salmon aquaculture in Chile]. Rev Med Chil 139: 107-118.

41. Nuccitelli A, Cozzi R, Gourlay LJ, Donnarumma D, Necchi F et al. (2011) Structure-based approach to rationally design a chimeric protein for an effective vaccine against Group B *Streptococcus* infections. Proc Natl Acad Sci U S A 108: 10278-10283.

42. Maione D, Margarit I, Rinaudo CD, Masignani V, Mora M et al. (2005) Identification of a universal Group B *streptococcus* vaccine by multiple genome screen. Science 309: 148-150.

43. Johri AK, Paoletti LC, Glaser P, Dua M, Sharma PK et al. (2006) Group B *Streptococcus*: global incidence and vaccine development. Nat Rev Microbiol 4: 932-942.

44. Paoletti LC, Madoff LC (2002) Vaccines to prevent neonatal GBS infection. Semin Neonatol 7: 315-323.

45. Baker CJ, Edwards MS (2003) Group B streptococcal conjugate vaccines. Arch Dis Child 88: 375-378.

46. Margarit I, Rinaudo CD, Galeotti CL, Maione D, Ghezzo C et al. (2009) Preventing bacterial infections with pilus-based vaccines: the group B *streptococcus* paradigm. J Infect Dis 199: 108-115.

47. Maisey HC, Hensler M, Nizet V, Doran KS (2007) Group B streptococcal pilus proteins contribute to adherence to and invasion of brain microvascular endothelial cells. J Bacteriol 189: 1464-1467.

48. Konto-Ghiorghi Y, Mairey E, Mallet A, Duménil G, Caliot E et al. (2009) Dual role for pilus in adherence to epithelial cells and biofilm formation in *Streptococcus agalactiae*. PLoS Pathog 5: e1000422.

49. Sharma P, Lata H, Arya DK, Kashyap AK, Kumar H et al. (2013) Role of Pilus Proteins in Adherence and Invasion of *Streptococcus agalactiae* to the Lung and Cervical Epithelial Cells. J Biol Chem 288: 4023-4034.

50. Xu H, Hu C, Gong R, Chen Y, Ren N et al. (2011) Evaluation of a novel chimeric B cell epitope-based vaccine against mastitis induced by either *Streptococcus agalactiae* or *Staphylococcus aureus* in mice. Clin Vaccine Immunol 18: 893-900.

51. Glaser P, Rusniok C, Buchrieser C, Chevalier F, Frangeul L et al. (2002) Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. Mol Microbiol 45: 1499-1513.

52. Tettelin H, Masignani V, Cieslewicz MJ, Eisen JA, Peterson S et al. (2002) Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc Natl Acad Sci U S A 99: 12391-12396.

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

54. Delannoy CMJ, Zadoks RN, Lainson FA, Ferguson HW, Crumlish M et al. (2012) Draft genome sequence of a nonhemolytic fish-pathogenic *Streptococcus agalactiae* strain. J Bacteriol 194: 6341-6342.

55. Wang B, Jian J, Lu Y, Cai S, Huang Y et al. (2012) Complete genome sequence of *Streptococcus agalactiae* ZQ0910, a pathogen causing meningoencephalitis in the GIFT strain of Nile tilapia (*Oreochromis niloticus*). J Bacteriol 194: 5132-5133.

56. 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: e39107.

57. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23: 254-267.

58. Kloepper TH, Huson DH (2008) Drawing explicit phylogenetic networks and their integration into SplitsTree. BMC Evol Biol 8: 22.

59. Soares SC, Silva A, Trost E, Blom J, Ramos R et al. (2013) The Pan-Genome of the Animal Pathogen *Corynebacterium pseudotuberculosis*Reveals Differences in Genome Plasticity between the Biovar *ovis* and *equi* Strains. PLoS One 8: e53818.

60. Blom J, Albaum SP, Doppmeier D, Pühler A, Vorhölter F et al. (2009) EDGAR: a software framework for the comparative analysis of prokaryotic genomes. BMC Bioinformatics 10: 154.

61. Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T et al. (2003) GenDB--an open source genome annotation system for prokaryote genomes. Nucleic Acids Res 31: 2187-2195.

62. Lerat E, Daubin V, Moran NA (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. PLoS Biol 1: E19.

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

64. Soares SC, Abreu VAC, Ramos RTJ, Cerdeira L, Silva A et al. (2012) PIPS: pathogenicity island prediction software. PLoS One 7: e30848.

65. Bolotin A, Quinquis B, Renault P, Sorokin A, Ehrlich SD et al. (2004) Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*. Nat Biotechnol 22: 1554-1558.

66. Carver TJ, Rutherford KM, Berriman M, Rajandream M, Barrell BG et al. (2005) ACT: the Artemis Comparison Tool. Bioinformatics 21: 3422-3423.

67. Alikhan N, Petty NK, Ben Zakour NL, Beatson SA (2011) BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12: 402.

68. Karp PD, Paley S, Romero P (2002) The Pathway Tools software. Bioinformatics 18 Suppl 1: S225-32.

69. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA et al. (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 38: D473-9.

70. Soares SC, Trost E, Ramos RTJ, Carneiro AR, Santos AR et al. (2012) Genome sequence of *Corynebacterium pseudotuberculosis* biovar *equi* strain 258 and prediction of antigenic targets to improve biotechnological vaccine production. J Biotechnol in press.

71. Rappuoli R (2001) Reverse vaccinology, a genome-based approach to vaccine development. Vaccine 19: 2688-2691.

72. He Y, Xiang Z, Mobley HLT (2010) Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. J Biomed Biotechnol 2010: 297505.

73. Barinov A, Loux V, Hammani A, Nicolas P, Langella P et al. (2009) Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other Gram-positive bacteria. Proteomics 9: 61-73.

74. Shewmaker PL, Gertz REJ, Kim CY, de Fijter S, DiOrio M et al. (2010) *Streptococcus salivarius* meningitis case strain traced to oral flora of anesthesiologist. J Clin Microbiol 48: 2589-2591.

75. Kwon H, Ogunniyi AD, Choi M, Pyo S, Rhee D et al. (2004) The ClpP protease of *Streptococcus pneumoniae* modulates virulence gene expression and protects against fatal pneumococcal challenge. Infect Immun 72: 5646-5653.

76. Al Safadi R, Amor S, Hery-Arnaud G, Spellerberg B, Lanotte P et al. (2010) Enhanced expression of lmb gene encoding laminin-binding protein in *Streptococcus agalactiae* strains harboring IS1548 in *scpB-lmb* intergenic region. PLoS One 5: e10794.

77. Franken C, Haase G, Brandt C, Weber-Heynemann J, Martin S et al. (2001) Horizontal gene transfer and host specificity of beta-haemolytic streptococci: the role of a putative composite transposon containing *scpB* and *lmb*. Mol Microbiol 41: 925-935.

78. Cieslewicz MJ, Chaffin D, Glusman G, Kasper D, Madan A et al. (2005) Structural and genetic diversity of group B *streptococcus* capsular polysaccharides. Infect Immun 73: 3096-3103.

79. Bellais S, Six A, Fouet A, Longo M, Dmytruk N et al. (2012) Capsular switching in group B *Streptococcus* CC17 hypervirulent clone: a future challenge for polysaccharide vaccine development. J Infect Dis 206: 1745-1752.

80. Nagano N, Nagano Y, Nakano R, Okamoto R, Inoue M (2006) Genetic diversity of the C protein beta-antigen gene and its upstream regions within clonally related groups of type Ia and Ib group B streptococci. Microbiology 152: 771-778.

81. Bröker G, Spellerberg B (2004) Surface proteins of *Streptococcus agalactiae* and horizontal gene transfer. Int J Med Microbiol 294: 169-175.

82. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E et al. (2011) Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol 9: 467-477.

83. Chen S, Tsai Y, Wu C, Liao S, Wu L et al. (2010) Streptococcal collagen-like surface protein 1 promotes adhesion to the respiratory epithelial cell. BMC Microbiol 10: 320.

84. Danne C, Dramsi S (2012) Pili of gram-positive bacteria: roles in host colonization. Res Microbiol 163: 645-658.

85. Merkl R, Wiezer A (2009) GO4genome: a prokaryotic phylogeny based on genome organization. J Mol Evol 68: 550-562.

86. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311: 1283-1287.

87. Oshima K, Nishida H (2007) Phylogenetic relationships among mycoplasmas based on the whole genomic information. J Mol Evol 65: 249-258.

88. Lal D, Verma M, Lal R (2011) Exploring internal features of 16S rRNA gene for identification of clinically relevant species of the genus *Streptococcus*. Ann Clin Microbiol Antimicrob 10: 28.

89. Lin FC, Whiting A, Adderson E, Takahashi S, Dunn DM et al. (2006) Phylogenetic lineages of invasive and colonizing strains of serotype III group B Streptococci from neonates: a multicenter prospective study. J Clin Microbiol 44: 1257-1261.

90. Jones N, Bohnsack JF, Takahashi S, Oliver KA, Chan M et al. (2003) Multilocus sequence typing system for group B *streptococcus*. J Clin Microbiol 41: 2530-2536.

91. Hugouvieux-Cotte-Pattat N, Charaoui-Boukerzaza S (2009) Catabolism of raffinose, sucrose, and melibiose in *Erwinia chrysanthemi* 3937. J Bacteriol 191: 6960-6967.

92. Klein MI, DeBaz L, Agidi S, Lee H, Xie G et al. (2010) Dynamics of *Streptococcus mutans* transcriptome in response to starch and sucrose during biofilm development. PLoS One 5: e13478.

93. Tang Y, Wu W, Zhang X, Lu Z, Chen J et al. (2012) Catabolite control protein A of *Streptococcussuis* type 2 contributes to sugar metabolism and virulence. J Microbiol 50: 994-1002.

94. Dramsi S, Caliot E, Bonne I, Guadagnini S, Prévost M et al. (2006) Assembly and role of pili in group B streptococci. Mol Microbiol 60: 1401-1413.

95. Maisey HC, Quach D, Hensler ME, Liu GY, Gallo RL et al. (2008) A group B streptococcal pilus protein promotes phagocyte resistance and systemic virulence. FASEB J 22: 1715-1724.

96. Mandlik A, Swierczynski A, Das A, Ton-That H (2008) Pili in Gram-positive bacteria: assembly, involvement in colonization and biofilm development. Trends Microbiol 16: 33-40.

97. Martins ER, Andreu A, Melo-Cristino J, Ramirez M (2013) Distribution of Pilus Islands in *Streptococcus agalactiae* That Cause Human Infections: Insights into Evolution and Implication for Vaccine Development. Clin Vaccine Immunol 20: 313-316.

98. Lopez-Sanchez M, Sauvage E, Da Cunha V, Clermont D, Ratsima Hariniaina E et al. (2012) The highly dynamic CRISPR1 system of *Streptococcus agalactiae* controls the diversity of its mobilome. Mol Microbiol 85: 1057-1071.

99. Westra ER, Brouns SJJ (2012) The rise and fall of CRISPRs--dynamics of spacer acquisition and loss. Mol Microbiol 85: 1021-1025.

100. Dawid S, Roche AM, Weiser JN (2007) The *blp* bacteriocins of *Streptococcus pneumoniae* mediate intraspecies competition both *in vitro* and *in vivo*. Infect Immun 75: 443-451.

101. Lux T, Nuhn M, Hakenbeck R, Reichmann P (2007) Diversity of bacteriocins and activity spectrum in *Streptococcus pneumoniae*. J Bacteriol 189: 7741-7751.

102. Moineau S (1999) Applications of phage resistance in lactic acid bacteria. Antonie Van Leeuwenhoek 76: 377-382.

103. Haaber J, Samson JE, Labrie SJ, Campanacci V, Cambillau C et al. (2010) Lactococcal abortive infection protein AbiV interacts directly with the phage protein SaV and prevents translation of phage proteins. Appl Environ Microbiol 76: 7085-7092.

104. O'Connor L, Coffey A, Daly C, Fitzgerald GF (1996) AbiG, a genotypically novel abortive infection mechanism encoded by plasmid pCI750 of *Lactococcus lactis* subsp. *cremoris* UC653. Appl Environ Microbiol 62: 3075-3082.

105. Tazi A, Disson O, Bellais S, Bouaboud A, Dmytruk N et al. (2010) The surface protein HvgA mediates group B *streptococcus* hypervirulence and meningeal tropism in neonates. J Exp Med 207: 2313-2322.

106. Gutekunst H, Eikmanns BJ, Reinscheid DJ (2004) The novel fibrinogen-binding protein FbsB promotes *Streptococcus agalactiae* invasion into epithelial cells. Infect Immun 72: 3495-3504.

107. Lancaster L, Saydam M, Markey K, Ho MM, Mawas F (2011) Immunogenicity and physico-chemical characterisation of a candidate conjugate vaccine against group B *streptococcus* serotypes Ia, Ib and III. Vaccine 29: 3213-3221.

108. Seib KL, Zhao X, Rappuoli R (2012) Developing vaccines in the era of genomics: a decade of reverse vaccinology. Clin Microbiol Infect 18 Suppl 5: 109-116.

109. Singh P, Springman AC, Davies HD, Manning SD (2012) Whole-genome shotgun sequencing of a colonizing multilocus sequence type 17 *Streptococcus agalactiae* strain. J Bacteriol 194: 6005.

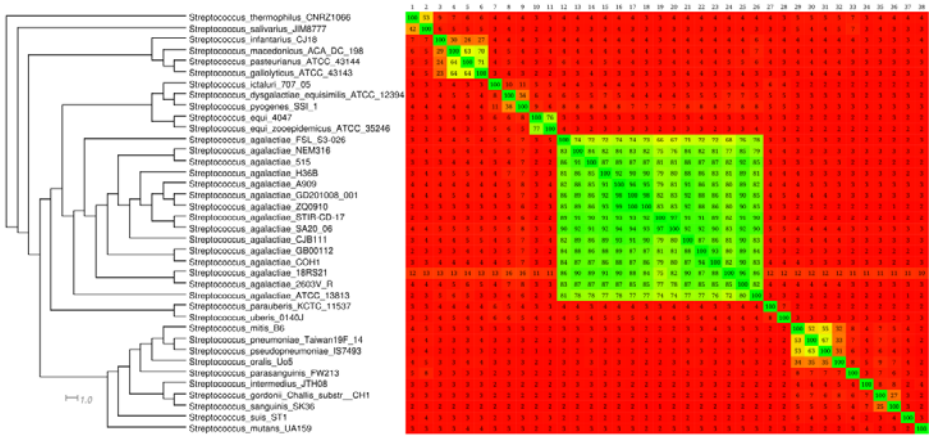

Figure 1. Phylogenomic tree and heatmap analyses of the genus *Streptococcus*.

Note: All file of the complete genomes from the genus *Streptococcus* were retrieved from the NCBI ftp site. Comparisons between the variable content of all 15 strains were plotted as percentages of similarity on the heatmap using Gegenees (version 1.1.4). The percentage of similarity was used to generate a phylogenomic tree with SplitsTree (version 4.12.6). Numbers from 1 to 38 (upper in the heatmap) represent species from *S. thermophilus* CNRZ1066 to *S. mutans* UA159. Percentages were plotted with a spectrum ranging from red (low similarity) to green (high similarity).

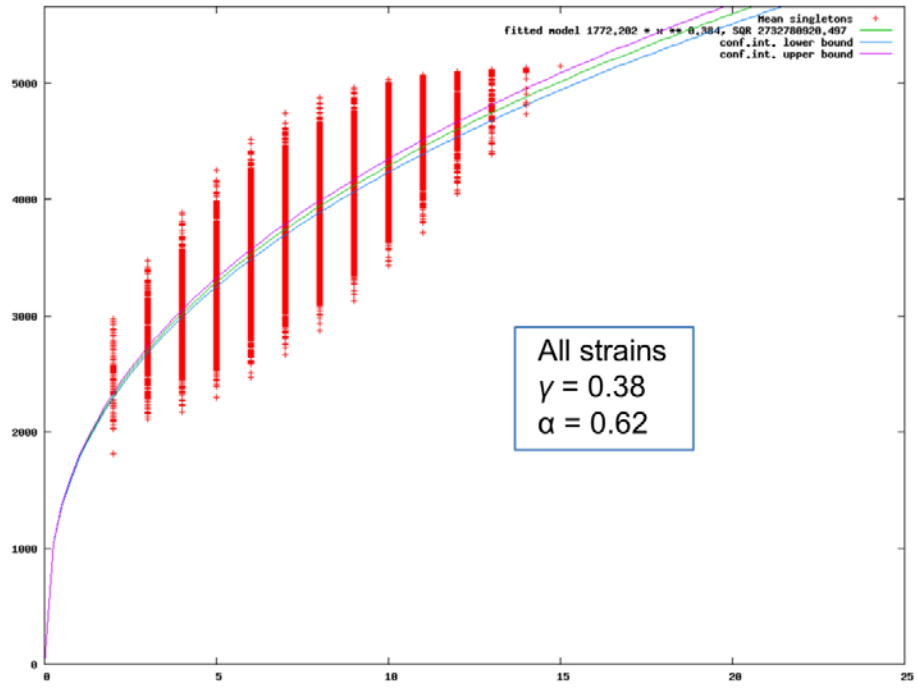

Figure 2. Pan-genome development of fifteen strains of *S. agalactiae*.

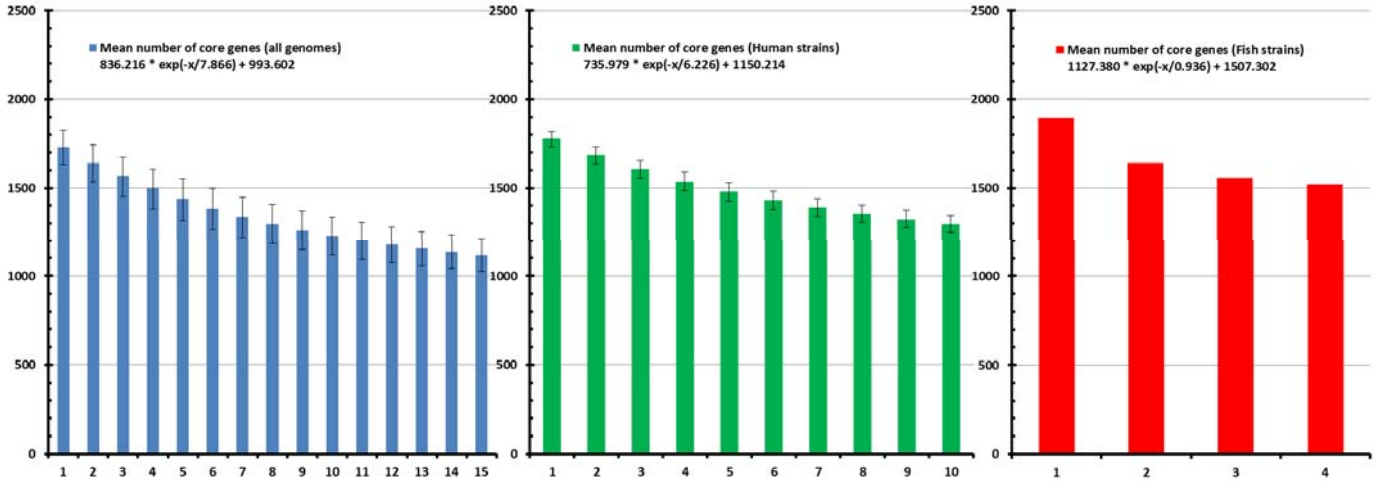

Figure 3. Core genome and singleton development of *S. agalactiae*

Note: Upper-left, the core genome development using permutations of all 15 strains of *S.agalactiae*; upper-center, the core genome development of the *S.agalactiae* human strains; upper-right, the core genome development of the *S.agalactiae* fish strains; lower-left, the singleton development using permutations of all 15 strains of *S.agalactiae*; lower-center, the singleton development of the *S.agalactiae* human strains; lower-right, the singleton development of the *S.agalactiae* fish strains.

…continues …

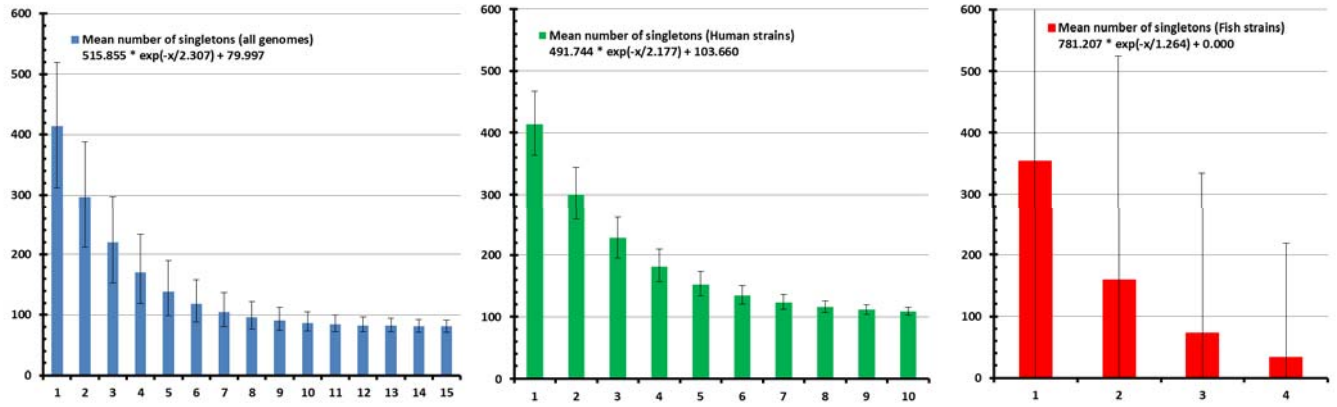

Figure 3. Core genome and singleton development of *S. agalactiae*

Note:Upper-left, the core genome development using permutations of all 15 strains of *S.agalactiae*; upper-center, the core genome development of the *S.agalactiae* human strains; upper-right, the core genome development of the *S.agalactiae* fish strains; lower-left, the singleton development using permutations of all 15 strains of *S.agalactiae*; lower-center, the singleton development of the *S.agalactiae* human strains; lower-right, the singleton development of the *S.agalactiae* fish strains.

Table 1. General information about the 15 *S. agalactiae* strains used in this work.

Strain	Serotype	ST by MLST	Host	Isolate source or Clinical description	Country of isolation	Status of Genome	Genome size (or total nucleotides in the contigs)	Number of genes	Singletons	GenBank accession N°	Reference
A909	Ia	7	Human	Septicemia	USA	Complete	2,127,839	2,136	11	CP00114	Tettelin, et al., 2005
2603V/R	V	110	Human	Septicemia	Italy	Complete	2,160,267	2,276	82	AE009948	Tettelin, et al., 2002,2005
NEM316	III	23	Human	Meningitis	Unknown	Complete	2,211,485	2,235	19	AL732656	Glaser, et al., 2002; Tettelin, et al., 2005
18RS21	II	19	Human	Septicemia	USA	Draft	2,193,092	2,421	123	AAJO01000000	Tettelin, et al., 2005
515	Ia	23	Human	Septicemia (cerebrospinal fluid)	USA	Draft	2,088,229	2,287	113	AAJP01000000	Tettelin, et al., 2005
ATCC13813	II	61	Human	Oral cavity	United Kingdom	Draft	2,243,750	2,243	151	AEQQ00000000	-
CJB111	V	1	Human	Septicemia	USA	Draft	2,105,032	2,208	95	AAJQ01000000	Tettelin, et al., 2005
COH1	III	17	Human	Septicemia	USA	Draft	2,205,442	2,412	126	AAJR01000000	Tettelin, et al., 2005
GB00112	III	17	Human	Vagina	Canada	Draft	2,033,051	1,994	19	AKXO00000000	Singh, et al., 2012
H36B	Ib	6	Human	Septicemia	USA	Draft	2,201,150	2,396	137	AAJS01000000	Tettelin, et al., 2005
FSLS3-026	III	67	Bovine	Mastitis	USA	Draft	2,455,848	2,422	105	AEXT01000000	Richards, et al., 2011
SA20-06	Ib	553	Fish	Meningoencephalitis (kidney)	Brazil	Complete	1,820,886	1,872	7	CP003919	Pereira et al., 2013
GD201008-001	Ia	7	Fish	Meningoencephalitis	China	Complete	2,063,112	2,088	8	CP003810	Liu, et al., 2012
STIR-CD-17	Ib	260	Fish	Meningoencephalitis (heart)	Honduras	Draft	1,805,303	1,737	26	ALXB00000000	Delannoy, et al., 2012
ZQ0910	Ia	7	Fish	Meningoencephalitis	China	Draft	1,814,715	2,003	20	AKAP00000000	Wang, et al., 2012

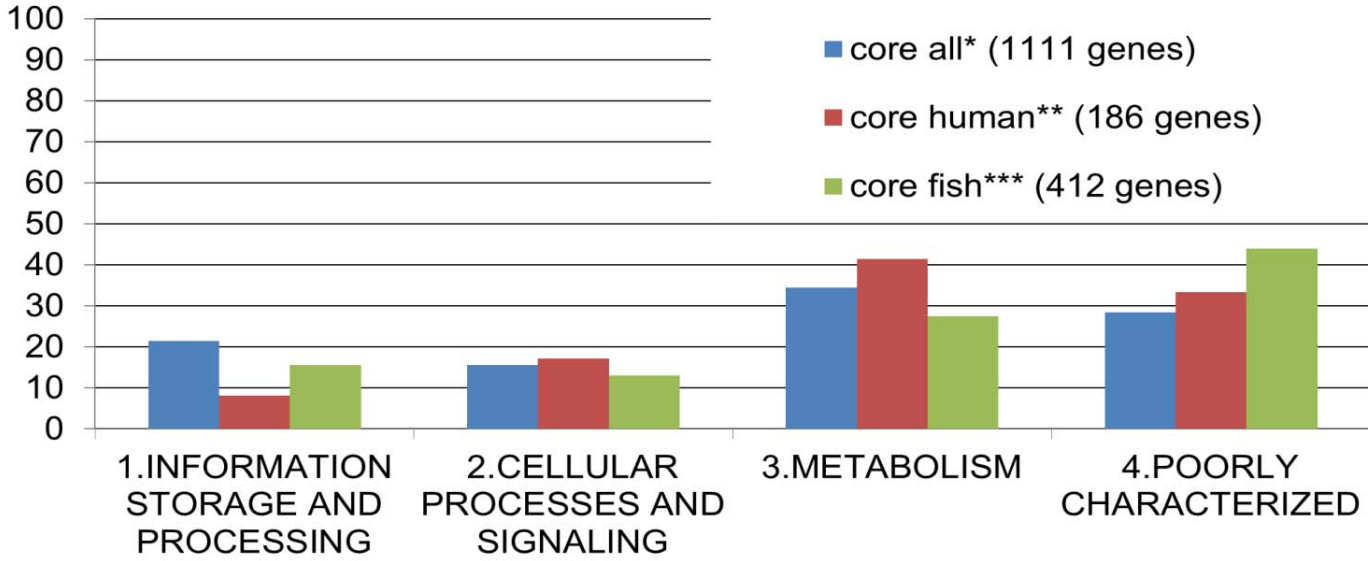

Figure 4. Core genes of the *S. agalactiae* strains classified by COG functional category.

Note: *Core all, the genes composing the core genome of all 15 strains; **core human, the genes of the subtractive core genome of the *S.agalactiae* human strains, which were absent in core genes of the all strains; ***core fish, the genes of the subtractive core genome of the *S.agalactiae* fish strains, which were absent in core genome of the all strains.

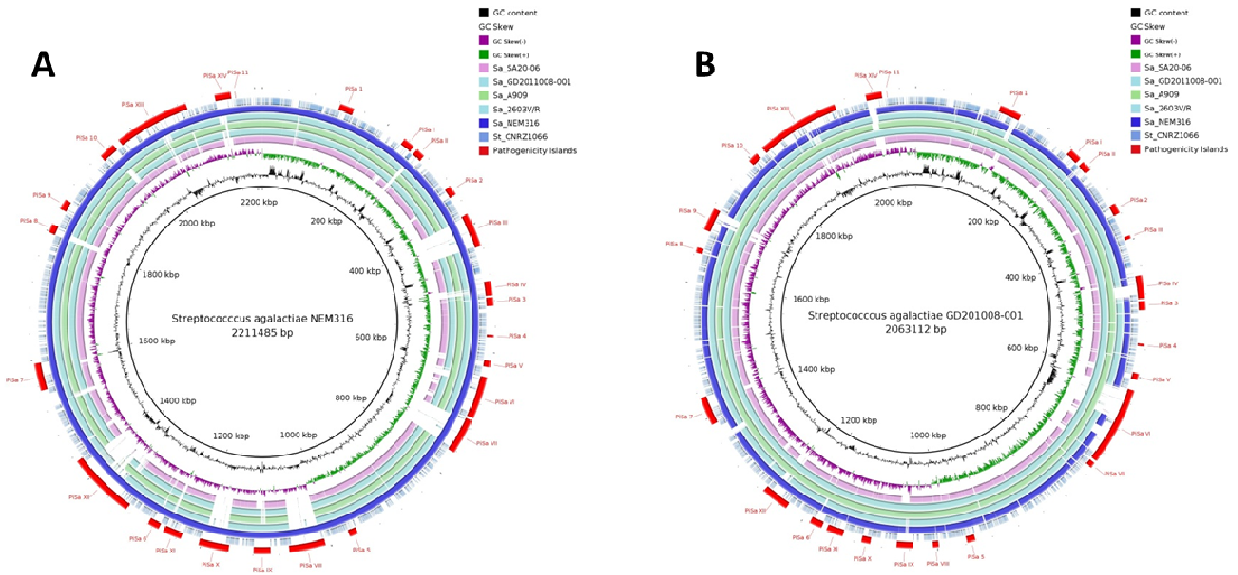

Figure 5. Pathogenicity islands distribution in *S. agalactiae* genomes

Note: A – All complete genomes of *S. agalactiae* were aligned using the genome of NEM316 strain as a reference. B - All complete genomes of S. agalactiae were aligned using the genome of GD201008-001 strain as a reference. From the inner to outer circle in A and B: the fish strains SA20-06 and GD201008-001; the human strains A909, 2603 V/R and NEM316; and, the non-pathogenic *S. thermophilus* CNRZ1066.

	2603V/R	A909	NEM316	SA20-06	GD201008-001
A909-PiSa 1	77	100	67	63	100
NEM316-PiSa I	92	96	100	50	96
A909-PiSa II	94	100	63	63	100
A909-PiSa 2	100	100	100	60	100
NEM316-PiSa III*	14	13	100	13	13
A909-PiSa IV	35	100	38	81	86
Sa20-06-PiSa 3	60	53	60	100	67
A909-PiSa 4	52	100	6	6	6
2603V/R-PiSa V	100	43	11	1	11
A909-PiSa VI	58	100	52	25	81
2603V/R-PiSa 5	100	30	30	27	30
A909-PiSa IX	61	100	61	26	100
NEM316-PiSa X	27	25	100	23	25
A909-PiSa XI	88	100	88	92	92
2603V/R-PiSa 6	100	76	81	76	76
2603V/R-PiSa XII	100	31	67	22	30
A909-PiSa 7	97	100	100	72	100
NEM316-PiSa 8	23	23	100	23	31
A909-PiSa 9	50	100	50	46	96
2603V/R-PiSa 10	100	20	32	28	18
A909-PiSa XIII	76	100	80	72	90
2603V/R-PiSa XIV	100	50	50	35	50
A909-PiSa 11	0	100	0	0	0
2603V/R-PI-1	100	86	86	0	0
NEM316-PI-2a	83	0	100	17	17
Sa20-06-PI-2b	13	88	13	100	88

Figure 6. Heatmap of the pathogenicity islands detected in the complete genomes of *S. agalactiae* and comparison of the predicted gene contents. The genomic islands were identified with the software PIPS, and the deduced similarities are shown as percentages. *

Note: Represent the pathogenicity island III, VII and VIII that are triplicated in NEM316 genome.

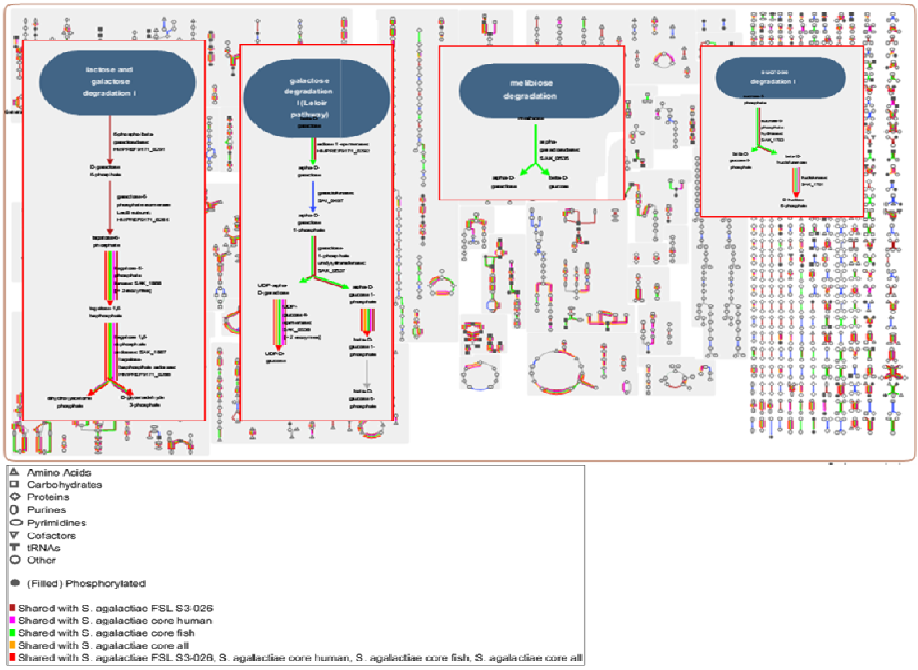

Figure 7. *Streptococcus agalactiae* metabolic pathway overview

Note: The galactose, melibiose and sucrose degradation pathways are highlighted.

Table 2. Subcellular location and adhesion probability of conserved antigenic proteins identified by reverse vaccinology strategy.

Locus _tag	Product name	Subcellular localization	Adhesion probability	Presence in core genome of all strains	Presence in core genes of human strains	Presence in core genes of fish strains	Similar to Human proteins
SAK_2105	transglycosylase-like domain-containing protein	PSE	0,751	YES	YES	YES	NO
SAK_0050	PcsB protein	SEC	0,748	NO	YES	NO	NO
SAK_0370	pencillin-binding protein	PSE	0,700	YES	YES	YES	NO
SAK_0442	hypothetical protein	SEC	0,695	YES	YES	YES	NO
SAK_0065	group B streptococcal surface immunogenic protein	PSE	0,678	YES	YES	YES	NO
SAK_2073	hypothetical protein	PSE	0,666	YES	YES	YES	NO
SAK_0337	hypothetical protein	SEC	0,662	YES	YES	YES	NO
SAK_1271	hypothetical protein	PSE	0,642	YES	YES	YES	NO
SAK_0064	zoocin A	SEC	0,642	YES	YES	YES	NO
SAK_0553	hypothetical protein	SEC	0,638	YES	YES	YES	NO
SAK_1293	hydrophobic W repeat-containing protein	SEC	0,632	NO	YES	NO	NO
SAK_1497	polar amino acid ABC transporter permease/substrate-binding protein	PSE	0,617	YES	YES	YES	NO
SAK_0932	foldase PrsA	PSE	0,613	YES	YES	YES	NO
SAK_1394	RND family efflux transporter MFP subunit	SEC	0,607	YES	YES	YES	NO

“Table 2, continues”

Locus _tag	Product name	Subcellular localization	Adhesion probability	Presence in core genome of all strains	Presence in core genes of human strains	Presence in core genes of fish strains	Similar to Human proteins
SAK_1319	laminin-binding surface protein	PSE	0,589	NO	YES	NO	NO
SAK_1994	hypothetical protein	SEC	0,588	YES	YES	YES	NO
SAK_1009	hypothetical protein	SEC	0,584	YES	YES	YES	NO
SAK_1656	amino acid ABC transporter amino acid-binding protein	PSE	0,575	YES	YES	YES	NO
SAK_0604	GDSL family lipase/acylhydrolase	SEC	0,574	YES	YES	YES	NO
SAK_1158	hypothetical protein	PSE	0,568	YES	YES	YES	NO
SAK_1087	phosphate ABC transporter substrate-binding protein	PSE	0,567	YES	YES	YES	NO
SAK_1784	CHAP domain-containing protein	SEC	0,565	YES	YES	YES	NO
SAK_0556	hypothetical protein	SEC	0,565	YES	YES	YES	NO
SAK_0457	GDXG lipolytic enzyme family protein	PSE	0,561	YES	YES	YES	NO
SAK_0166	ribose ABC transporter ribose-binding protein	PSE	0,561	NO	YES	NO	NO
SAK_1870	mannosyl-glycoproteinendo-beta-N-acetylglucosamidase	SEC	0,556	YES	YES	YES	NO
SAK_1625	polar amino acid ABC transporter polar amino acid-binding protein	PSE	0,556	YES	YES	YES	NO
SAK_2007	hypothetical protein	SEC	0,554	YES	YES	YES	NO
SAK_0854	hypothetical protein	PSE	0,548	YES	YES	YES	NO

“Table 2, conclusion”

Locus _tag	Product name	Subcellular localization	Adhesion probability	Presence in core genome of all strains	Presence in core genes of human strains	Presence in core genes of fish strains	Similar to Human proteins
SAK_ 1235	hypothetical protein	PSE	0,545	YES	YES	YES	NO
SAL_ 1416	PTS system, fructose specific IIABC components	PSE	0,541	YES	YES	YES	NO
SAK_ 0222	penicillin-binding protein 1B	PSE	0,539	YES	YES	YES	NO
SAK_ 1109	hypothetical protein	SEC	0,539	YES	YES	YES	NO
SAK_ 1503	cell wall surface anchor family protein	PSE	0,536	YES	YES	YES	NO
SAK_ 1580	hypothetical protein	SEC	0,534	YES	YES	YES	NO
SAK_ 1927	phosphate ABC transporter substrate-binding protein	PSE	0,532	YES	YES	YES	NO
SAK_ 0321	lipoprotein	PSE	0,532	YES	YES	YES	NO
SAK_ 0301	quaternary amine ABC transporter amino acid-binding protein	PSE	0,531	YES	YES	YES	NO
SAK_ 0685	zinc ABC transporter substrate-binding protein	SEC	0,517	YES	YES	YES	NO
SAK_ 1074	ABC transporter substrate-binding protein	SEC	0,514	NO	YES	NO	NO
SAK_ 1426	ferrichrome ABC transporter substrate-binding protein	PSE	0,512	YES	YES	YES	NO