JULIE KENNYA DE LIMA FERREIRA

PREVALÊNCIA DO OPERON *sfr*AABCD (SURFACTINA) EM *Bacillus subtilis*: UMA ANÁLISE PANGENÔMICA

LAVRAS-MG
2021

JULIE KENNYA DE LIMA FERREIRA

PREVALÊNCIA DO OPERON *sfrA*ABCD (SURFACTINA) EM *Bacillus subtilis*: UMA ANÁLISE PANGENÔMICA

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Microbiologia Agrícola, na área de concentração em Microbiologia Agrícola, para obtenção do título de Mestre.

Prof. Dr. Victor Satler Pylro
Orientador

Prof. Dra. Cristina Ferreira Silva e Batista
Coorientadora

LAVRAS-MG
2021

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pela própria autora.

Ferreira, Julie Kennya de Lima.
Prevalência do operon *sfr*AABCD (surfactina) em *Bacillus subtilis*: uma análise pangenômica / Julie Kennya de Lima Ferreira. - 2021.
65 p. : il.

Orientador: Victor Satler Pylro.
Coorientadora: Cristina Ferreira Silva e Batista.
Dissertação (mestrado acadêmico) - Universidade Federal de Lavras, 2021.
Bibliografia.

1. Genômica comparativa. 2. Montagem híbrida de genoma. 3. NGS. I. Pylro, Victor Satler. II. Batista, Cristina Ferreira Silva e. III. Título.

JULIE KENNYA DE LIMA FERREIRA

PREVALÊNCIA DO OPERON *sfrA*ABCD (Surfactina) EM *Bacillus subtilis*:
UMA ANÁLISE PANGENÔMICA

PREVALENCE OF THE *sfrA*ABCD OPERON (Surfactin) IN *Bacillus subtilis*:
A PANGENOMIC ANALYSIS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Microbiologia Agrícola, na área de concentração em Microbiologia Agrícola, para obtenção do título de Mestre.

APROVADA em 30 de abril de 2021.

Dr. Daniel Kumazawa Morais, CAS
Dr. Dirceu de Sousa Melo, UFLA
Dr. Luiz Fernando Wurdig Roesch, UNIPAMPA

Prof. Dr. Victor Satler Pylro
Orientador

Prof. Dra. Cristina Ferreira Silva e Batista
Coorientadora

LAVRAS-MG
2021

Em memória de Manoel Pereira Lima†,
Adalberto Gomes Lima†,
Anderson Gomes Lima†,
Iris Carvalho† e
Bethania Lemme†.

AGRADECIMENTOS

Meus agradecimentos vão, primeiramente, àqueles que me regem e que nunca, em seu infinito amor, abandonaram-me ou deixaram de zelar por mim.

À minha família, principalmente aos meus pais, por acreditarem e confiarem em mim mais do que eu poderia.

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Agradeço também à Universidade Federal de Lavras, ao Departamento de Biologia e ao Programa de Pós-Graduação em Microbiologia Agrícola.

Aos Drs. Victor Satler Pylro e Dirceu Sousa Melo e à Ma. Kellen Kauanne Pimenta de Oliveira, por toda a paciência e confidencialidade. Nada dessa pesquisa seria possível sem os três.

Aos meus amigos, Nívia Bianca P. Lopes, Clara Resende de S. Castro, Larissa Santos Bento, Michely do Vale M. Paduan, Yuri Martins de Souza, Sarah Moreira, Italo Almeida, Raquel Leite de Oliveira Alves, Adriely Melo de Souza, Rêgila Mello, Jefferson "Panda" Leonardo, Lucas "Donnovan" Rosane e todos aqueles que, de alguma forma, estiveram comigo e acompanharam de perto o meu desenvolvimento nos últimos 2 anos.

Agradecimentos especiais a uma família que me acolheu tão bem como amiga: Cintya Barbosa Moreira Frigo, Otávio Augusto da Silva Frigo, Milena Kaori Okada e Giovanna Mayumi Okada.

RESUMO

A surfactina é um biossurfactante lipopeptídico comumente produzido por bactérias do gênero *Bacillus*. Entretanto, a sua expressão é dependente de vários fatores genéticos de adaptação e comunicação. Estudos de prevalência gênica têm se mostrado importantes para auxiliar o entendimento acerca desses fatores de interação. Nesse contexto, o trabalho teve por objetivo avaliar a prevalência do *operon* de surfactina (*sfrA*ABCD) em estirpes de *Bacillus subtilis*, bem como o repertório pangenômico da espécie através da genômica comparativa. O genoma completo do *B. subtilis* RI4914 foi sequenciado por meio de abordagem híbrida, utilizando as plataformas PGM IonTorrent e GridIONTM. Para o estudo comparativo, 153 genomas completos, classificados como *B. subtilis* foram recuperados do NCBI, dos quais foram utilizados os genes *core* para análises de pangenoma e MLST. O genoma obtido apresentou alta qualidade e boa acurácia, com completude de 98,84% e contaminação de 1,10%. A análise de genômica comparativa evidenciou o perfil pangenômico aberto em *B. subtilis* e prevalência da surfactina em todos os genomas analisados, demonstrando-se como um caráter característico da espécie.

Palavras-chave: Genômica comparativa. Montagem híbrida. GridION. PGM. MLST.

ABSTRACT

Surfactin is a lipopeptide biosurfactant commonly produced by *Bacillus* genus bacteria. However, its expression is dependent on several genetic factors of adaptation and communication. Gene prevalence studies are important to help the understanding of these interaction factors. This work aimed to evaluate the surfactin operon (*sfrA*ABCD)'s prevalence in *Bacillus subtilis* strains, as well as the pangenomic repertoire of the species through comparative genomics. The complete genome of *B. subtilis* RI4914 was sequenced using a hybrid approach, with the PGM IonTorrent and the GridIONTM platforms. A total of 153 complete genomes classified as *B. subtilis* were retrieved from the NCBI for comparative study, from which the core genes were used for pangenome and MLST analyses. The genome obtained showed high quality and good accuracy, with completeness of 98.84% and contamination of 1.10%. The comparative genomics analysis showed the open pangenomic profile in *B. subtilis* and surfactin prevalence in all analyzed genomes, classifying it as a distinctive character of this species.

Keywords: Comparative genomics. Hybrid assembly. GridION. PGM. MLST.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

PARTE 1

1 INTRODUÇÃO GERAL

Estima-se que aproximadamente 13 milhões de toneladas de surfactantes sejam utilizados todos os anos nos mais diversos setores industriais, do alimentício à indústria petroquímica. A ideia de que a produção e o uso de biossurfactantes sejam tecnologias já consolidadas tem mudado ao longo dos últimos anos, principalmente em decorrência da demanda crescente pela substituição desses compostos sintéticos por análogos de baixa toxicidade e persistência no ambiente. Apesar da redução na emissão de poluentes no ambiente (p. ex. compostos químicos lançados nas barragens de rejeitos ou em processos de recuperação de petróleo), o uso de biossurfactantes ainda não é aplicado em larga escala pela indústria (KHOPADE *et al.*, 2012).

As moléculas de biossurfactantes tendem a ser menos tóxicas que os compostos sintéticos, o que torna sua aplicação mais segura e ambientalmente mais atrativa. A exemplo disso, os compostos atualmente utilizados como coletores de sulfetos metálicos durante a etapa de flotação são sabidamente tóxicos, com rápida dissociação em água, vapores inflamáveis, provocando náuseas e vômitos se ingeridos, além de serem neurotóxicos, causando irritação da pele e das mucosas (MAKKAR; ROCKNE, 2003).

Um exemplo de biossurfactante é a surfactina, composto natural produzido por bactérias do gênero *Bacillus*, especialmente *Bacillus subtilis.* Possui uma variedade estrutural alta e está classificada no grupo de lipopeptídeos microbianos, juntamente com outros 300 compostos relatados nas últimas décadas. Entre estes, incluem-se iturinas, fengicinas e liquenisinas como os mais frequentemente explorados.

Apesar de todos os prós, o mercado de biossurfactante ainda tem como desafio o aumento do custo de produção em larga escala, quando comparado aos seus análogos sintéticos. Por conta disso, o avanço de estudos convencionais e moleculares, utilizando microrganismos produtores como alvos, têm se tornado presentes para a descoberta de novos métodos que possam diminuir o custo ou até aumentar a compreensão das vias metabólicas participantes dessa produção (ZHANG *et al.*, 2016).

Os avanços científicos e tecnológicos propiciaram uma revolução nas abordagens tradicionais de exploração de recursos biológicos. Os mecanismos de busca e descoberta biotecnológica sofreram alterações significativas em função das mudanças de modelos geradas pelos avanços em biologia molecular, genômica e bioinformática, em consequência

ao desenvolvimento de novas tecnologias de sequenciamento massivo de DNA (*Next Generation Sequencing*, NGS). Métodos moleculares com potencial aplicação na área incluem as técnicas de análises de transcriptomas, pangenomas, genômica comparativa e análises filogenéticas de microrganismos de interesse, as quais se expandiram com os métodos de NGS (SGHAIER *et al.*, 2019).

Entre as análises provenientes deste estímulo, o pangenoma revela que muitas cepas de diversos habitats apresentam variações genômicas quanto ao conteúdo genético de cada indivíduo, sendo que as principais funções de uma espécie específica são conservadas como genes centrais compartilhados entre diferentes cepas, incluindo aqueles referentes à expressão de biocompostos como a surfactina. Desde a última década, isso só se tornou possível devido ao avanço do estudo de genomas de procariotos (GUIMARÃES *et al.*, 2015).

Devido à disponibilidade de muitas cepas específicas detentoras de conjuntos gênicos de alto interesse biotecnológico, muitas passaram a ser sequenciadas e investigadas a fim de identificar tais genes e funções, sejam de interesse industrial ou não, revelando mecanismos de adaptação a diferentes ambientes necessários para a sobrevivência, como a capacidade de transferência horizontal de genes (WU *et al*., 2021).

Desde que a primeira cepa de *B. subtilis* foi sequenciada em 1997, a espécie se tornou uma das mais estudadas de forma extensiva, possuindo os maiores conjuntos de genomas sequenciados. No entanto, a análise dos genes centrais e específicos das cepas de *B. subtilis* foi realizada em um número limitado de estudos, cuja maioria, atualmente, concentra-se nas cepas patogênicas do gênero, tais como *Bacillus anthracis* e *Bacillus cereus.* Por isso, é de suma importância realizar análises mais abrangentes, caracterizando melhor os genes centrais de cepas de *B. subtilis* (BORRISS *et al.*, 2018; KUNST *et al.*, 1997).

Este trabalho possui cunho de pesquisa básica, agregando o conhecimento já desenvolvido acerca da potencialidade da surfactina, com os novos métodos de genômica microbiana. A padronização/desenvolvimento e aplicação dos métodos de genômica comparativa permite melhor entendimento de processos genéticos e/ou ecológicos que modulam a produção de biocompostos por microrganismos, como a regulação da expressão gênica, as mutações e os processos evolutivos. Esses conhecimentos são essenciais para posterior otimização dos processos por meio da engenharia genética. Sendo assim, este trabalho teve por objetivo avaliar a prevalência do *operon* da surfactina (*sfrA*ABCD) em estirpes de *B. subtilis*, bem como o repertório pangenômico da espécie, através de genômica comparativa. Além disso, a genômica comparativa permitiu o desenvolvimento de

oligonucleotídeos iniciadores adequados para estudos de expressão diferencial do gene *srfA*A, que compõem o *operon* de expressão da surfactina em *B. subtilis*.

Acredita-se que a aplicação do conhecimento gerado por esses métodos impulsionará os estudos acerca da conservação genética desse *operon* em *B. subtilis*, dando suporte para a utilização da espécie como modelo de obtenção e estudo desses biocompostos, além de desenvolver novos caminhos para a manipulação biotecnológica do microrganismo. Diante disso, este trabalho teve como objetivos avaliar a prevalência do *operon* de surfactina (*sfrA*ABCD) em estirpes de *B. subtilis*, bem como o repertório pangenômico da espécie, através de genômica comparativa.

2 REFERENCIAL TEÓRICO

2.1 Biossurfactantes: *Bacillus subtilis* e a surfactina

Bacillus subtilis é uma bactéria Gram-positiva cosmopolita que demonstra adaptabilidade genética incomum, permitindo-lhe colonizar hábitats altamente diversos. *Bacillus subtilis* é uma espécie não patogênica que apresenta alta diversidade genética, mesmo entre cepas da mesma espécie (KASPAR; NEUBAUER; GIMPEL, 2019). A espécie é comumente encontrada em solos, onde participa dos trabalhos detritívoros para a ciclagem de carbono e nitrogênio no ecossistema. Correspondem à microrganismos aeróbicos facultativos, mesofílicos (25–37 ºC) e apresentam colônias com características morfológicas variadas, diferenças essas que dependem da composição do meio de cultivo do microrganismo. Geralmente, apresentam-se em formas redondas ou irregulares, de coloração marrom ou opaca (SERRA *et al.*, 2014).

Por décadas, *B. subtilis* tem sido um modelo biológico viável para a compreensão das premissas fundamentais de bactérias Gram-positivas formadoras de esporos e, na biotecnologia, já é empregada para a síntese de diversos compostos secundários, tais como vitaminas, aminoácidos, enzimas industriais e antibióticos (HARWOOD *et al.*, 2018). Além disso, a espécie ganhou ainda mais atenção após ter potencial comprovado no controle biológico, seja a partir do efeito antagônico sobre fitopatógenos, seja com o efeito inverso a partir do crescimento favorecido (SHAFI; TIAN; JI, 2017).

Após ter sido muito subestimada perante seu potencial biossintético, estirpes selvagens de várias fontes têm demonstrado arsenais complexos para a biossíntese de, principalmente, lipopeptídeos cíclicos de uso industrial. Estima-se que 4–5% do genoma de *B. subtilis* é

dedicado somente para a biossíntese de produtos naturais a partir de inúmeras fontes de carbono e nitrogênio, tais como gordura animal, efluentes industriais, óleo queimado, soro de leite e águas residuais com alto teor de amido (MADSEN *et al.*, 2015).

Geralmente, os lipopeptídeos produzidos por *B. subtilis* são formados a partir de variações de heptapeptídeos e cadeias de ácidos graxos hidroxilados. Esses biossurfactantes são classificados em três famílias bastante conhecidas por possuírem propriedades antimicrobianas, além da atividade tensoativa: surfactinas, iturinas (as quais incluem iturinas A, B, C, D e E; bacilomicinas D, F e L; e micosubtilisina) e fengicinas (GUDIÑA *et al.*, 2013). Esses compostos têm atraído cada vez mais interesses biotecnológicos e farmacêuticos por atuarem como biossurfactantes e antibióticos (COCHRANE; VEDERAS, 2016).

Surfactina é o único membro da família com conhecida atividade biológica e também o composto produzido por *B. subtilis* mais extensivamente estudado. Foi inicialmente descoberta em caldo de cultivo para essa bactéria em específico, em 1968, época em que se descobriu seu potencial inibitório de coagulação de fibrina. Entretanto, apenas nos anos 80 o composto chamou atenção como alternativa para substituir surfactantes químicos (ARIMA *et al.*,1968; BARROS *et al.*, 2007). A surfactina biológica possui uma variedade de isoformas que, dependendo de suas propriedades físico-químicas, podem apresentar mudanças no tamanho da cadeia e na sequência de aminoácidos que compõe os bioprodutos, o que demonstra a plasticidade do complexo de peptídeo sintetase não ribossomal (SEN, 2010).

A surfactina enquanto biossurfactante participa de uma ampla gama de aplicações, desde tratamentos com antibióticos (por conta de suas propriedades de permeabilização da membrana), terapia de câncer e até em processos de recuperação de óleo. A produção em larga escala industrial desse composto pode ser viável através de cultivo em lote alimentado de alta densidade celular, uma vez que excelentes rendimentos de mais de 1 $g.L^{-1}$ foram relatados em processos biotecnológicos empregando *Bacillus* spp. (WILLENBACHER *et al.*, 2015; ZHI; WU; XU, 2017).

2.2 Métodos moleculares: genômica aplicada à ecologia microbiana

Por muito tempo, os estudos de diversidade microbiana foram dominados por técnicas de isolamento de culturas puras, o que rende até os dias de hoje a base para o avanço das pesquisas de fisiologia e genética microbiana. Porém, a atualização dessas técnicas se tornou necessária a partir do momento em que se mostrou ineficaz na amostragem de uma

comunidade microbiana no solo, recuperando não mais do que 10% de microrganismos presentes nesse ambiente (AMANN; LUDWIG; SCHLEIFER, 1995).

Com isso, muitas tecnologias dentro da biologia molecular foram estabelecidas, tais como a extração de DNA e RNA, a PCR, o sequenciamento de DNA, as análises de transcriptomas, a clonagem, entre outras. Essas técnicas permitiram o acesso a novos grupos de organismos não cultiváveis presentes numa comunidade microbiana, os quais não seriam acessados apenas por isolamento, devido às particularidades e exigências nutricionais de espécies (THIES *et al.*, 2016).

A importância de se estudar fatores evolutivos para entender a organização de organismos vivos é reconhecida, o que permitiu o seu avanço com o desenvolvimento da genômica. O estudo de genomas, principalmente completos, provou ser útil para compreender o funcionamento de microrganismos, aumentando substancialmente a compreensão da organização dos genes participantes de seus processos adaptativos (ABBY; DAUBIN, 2007). O acesso à informação genômica se dá em nível computacional e tem se tornado uma parte importante na descoberta de novos produtos naturais, expressos por centenas de genes organizados em *operons* ou não. Hoje em dia, milhões de sequências genômicas de bactérias são publicamente disponibilizadas, bem como a disponibilidade *online* de *clusters* de metabólitos secundários que aguardam conexão com seus produtos naturais codificados (ATANASOV *et al.*, 2021). Com o desenvolvimento de métodos de sequenciamento de alto rendimento e a riqueza de dados de DNA disponível, uma variedade de métodos e ferramentas de manipulação genômica foram surgindo para orientar a descoberta e caracterização desses genes e compostos (ZIEMERT; ALANJARY; WEBER, 2016).

Entre esses métodos, destacam-se as análises de comparações entre genomas, as quais são capazes de revelar aspectos sem precedentes a respeito da fisiologia, além de revelar extensa diversidade intraespecífica. A ideia de comparar múltiplos genomas bacterianos para detectar e priorizar *clusters* está fortemente conectada a ideia de utilizar métodos filogenéticos para encontrar genes que expressam compostos promissores. Isso fornece importantes hipóteses acerca dos esforços de expressão gênica e a função ecológica dessas moléculas (GRAZZIOTIN *et al.*, 2021; HUG *et al.*, 2016).

Além disso, esses e outros métodos moleculares se tornaram importantes ferramentas, não apenas para detectar características genéticas ou adaptações específicas, mas também para melhorar nosso entendimento a respeito da taxonomia, auxiliando no delineamento das espécies de procariotos. Por conta disso, avanços rápidos na ciência de genomas devem ser complementados por investimentos em sistemática, a fim de desenvolver uma taxonomia mais

fidedigna e adaptada às informações genômicas (DIENE *et al.*, 2013). Conceitos tradicionais, como espécie, gênero e família, não servem à sistemática microbiana, em que os problemas de transferência horizontal de genes e mecanismos de especiação são variados e complexos. Esses estudos precisam ser refinados para que a recuperação de dados genômicos se torne mais eficaz (STAHL; JAMES, 2002).

Estudos de genômica comparativa têm revelado que cromossomos bacterianos estão sujeitos à pressão seletiva, o que molda profundamente sua organização. Todos os processos de replicação, transcrição e regulação da expressão genética impactam a forma que os genes se organizam ao longo do genoma. A organização desses genes ocorre pela maneira assimétrica que o cromossomo bacteriano é replicado, com uma fita principal e outra posterior, uma questão correlacionada com várias características evolutivas entre as duas fitas, como mutações e localização de genes essenciais (LOBRY; SUEOKA, 2002).

Além dessas questões de organização do genoma, a genômica comparativa revelou um grau anteriormente inesperado de diversidade entre genomas procarióticos. Como por exemplo, a comparação do conteúdo do gene dentro e entre espécies (ABBY; DAUBIN, 2007). Com o avanço dos métodos genômicos, ocorreu uma transição em estudos, nos quais as análises genômicas passaram de um único ou poucos genomas para centenas a milhares, englobando o que é conhecido hoje como análises de pangenoma. Estas análises fornecem uma estrutura para estimar a diversidade genômica do conjunto de dados em mãos e prever o número de sequências de genomas que seriam necessárias para caracterizar tal diversidade (VERNIKOS *et al.*, 2015).

Outra funcionalidade significante de análises pangenômicas é que estas podem prover resoluções superiores de reconstrução da filogenia de organismos de uma forma mais confiável do que outras análises baseadas em um ou vários genes. Para fins conceituais, um pangenoma completo é definido como o número total de genes não-redundantes presentes em um determinado clado, totalizando todo o repertório genômico do mesmo. Neste tipo de análise, geralmente são encontrados genes *core*, genes acessórios e genes únicos (específicos de cada cepa), organizados em *soft core*, *shell* e genes *cloud* (STICE *et al.*, 2018).

Em *soft core* são representados aqueles genes conservados presentes em ao menos 95–99% de todos os genomas analisados, *shell* são aqueles presentes em 15–95% dos genomas e *cloud* os presentes em 0–15%. Juntos, os genes *core* e *soft-core*, representam um *pool* de genes altamente conservados, que podem fornecer informações a respeito da história evolutiva das espécies estudadas, já os genes *cloud* são aqueles específicos ou únicos de cada cepa presente na análise pangenômica, os quais são compartilhados no máximo entre dois

desses. Os restantes são genes moderadamente conservados, identificados como genes *shell* que, junto com genes *cloud*, representam o subconjunto do genoma acessório, responsáveis por refletir o estilo de vida, adaptação e também a história evolutiva das cepas em estudo (SNIPEN *et al*., 2009; VERNIKOS *et al.*, 2015).

Essa modalidade de comparação genômica demonstrou que várias espécies, incluindo patógenos humanos e bactérias ambientais exibem um pangenoma aberto. O que indica que um número grande e indeterminado de genomas adicionais seria necessário para identificar todos os genes acessíveis às espécies. Em contraste, para espécies com pangenoma fechado, genomas adicionais sequenciados não fornecem novos genes para expandir o repertório (VERNIKOS *et al.*, 2015).

Lidar com todas as possibilidades disponíveis acerca de uma comunidade microbiana, geradas por todos esses e outros métodos moleculares, ajuda a trazer à luz os segredos das "caixas pretas" da ecologia microbiana, revelando a ocorrência de genes funcionais e suas correlações com o meio ambiente (DELMONT *et al.*, 2011).

2.3 Métodos moleculares: sequenciamento massivo de DNA e vantagens de abordagens híbridas

Em resposta ao rápido avanço nos estudos genômicos, os sequenciamentos de DNA de alta performance, também conhecidos como sequenciamento NGS, entram em um cenário carregando a possibilidade de produção de uma alta quantidade de dados, de forma rápida e com custo abaixo do tradicional sequenciamento de Sanger (VASUDEVAN *et al.*, 2020).

A primeira ferramenta NGS nasceu em meados dos anos 2000, utilizando nanotecnologia e pirosequenciamento, sem a obrigatoriedade de processos de clonagem ou transformações (WILLSON, 2021). Com o seu lançamento, foi documentada uma queda de 50.000 vezes o custo de sequenciamento desde o projeto Genoma Humano e, desde então, o sequenciamento de genoma têm evoluído cada vez mais rápido com muitas tecnologias capazes de produzir leituras em diferentes quantidades, qualidades e comprimentos (NEAL-MCKINNEY *et al.*, 2021).

Apesar dessa crescente evolução, algumas propriedades foram mantidas, afetando mais especificamente a forma que os dados são organizados, bem como a integridade e precisão dos genomas resultantes. Porém, infelizmente, a cada tecnologia NGS que surge, os problemas podem ser reduzidos ou novos podem ser criados, como é o caso das médias de erros de 0,1–15% que são maiores e os comprimentos das leituras que são geralmente mais

curtos que aqueles gerados pelo tradicional sequenciamento de Sanger, o que requer maiores cuidados na análise dos resultados (GOODWIN; MCPHERSON; MCCOMBIE, 2016).

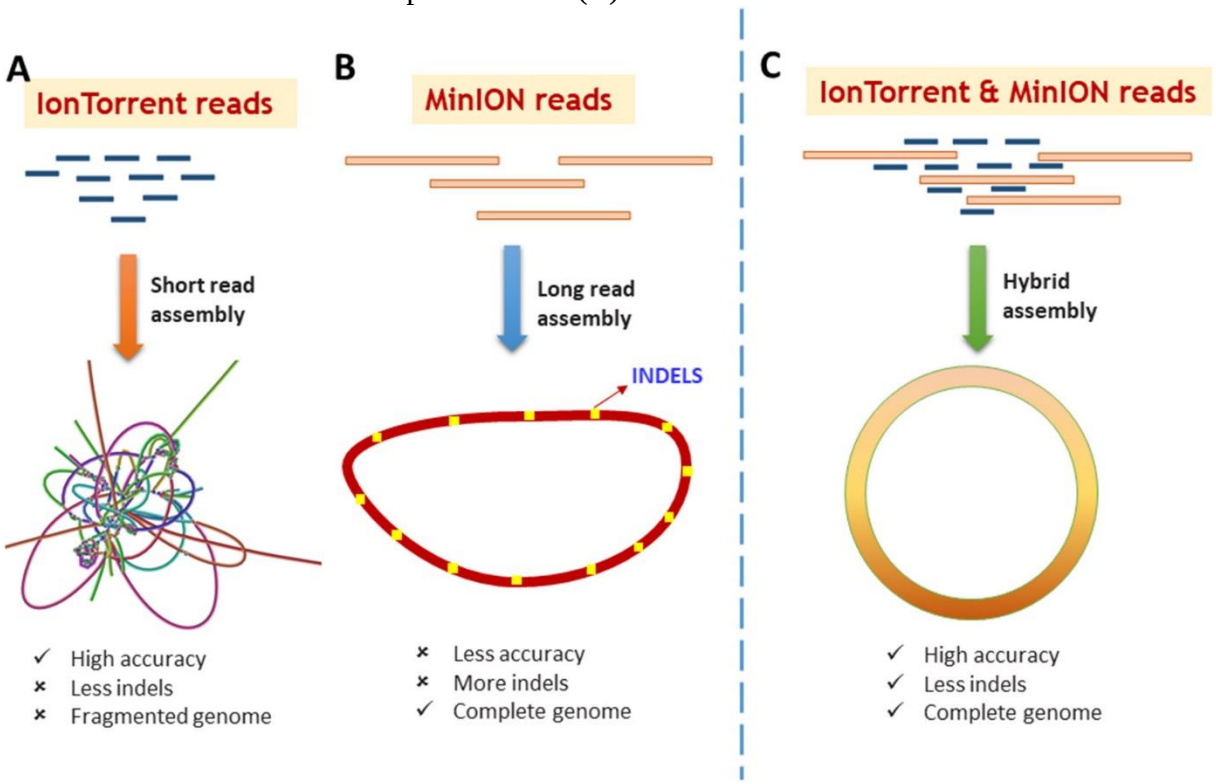
No final da última década, a tecnologia de sequenciamento de segunda geração melhorou bastante, porém sendo capaz apenas de produzir genomas fragmentados. Combinar os *contigs* desses genomas é quase impossível nas tecnologias tradicionais baseadas em leituras curtas (Figura *1.1*A), além disso, o alto conteúdo de GC e presença de regiões repetitivas podem afetar drasticamente a montagem. Foram esses obstáculos que abriram caminho para o desenvolvimento de tecnologias de sequenciamento de terceira geração (GOLDSTEIN *et al.*, 2019; DEL ANGEL *et al.*, 2018).

Embora o sequenciamento de leituras longas supere a limitação de comprimento de outras abordagens NGS, a mesma ainda possui uma média de erros considerável (GOODWIN; MCPHERSON; MCCOMBIE, 2016) (Figura *1.1*B). Para contrapor esse problema, uma das alternativas é unir plataformas que geram leituras curtas e plataformas que geram leituras longas, diminuindo os erros de cada um isoladamente e aumentando a qualidade do genoma final (Figura *1.1*C) (GOODWIN *et al.*, 2015).

O PGM IonTorrent e plataformas da Oxford Nanopore Technology são exemplos de equipamentos que podem realizar este tipo de trabalho. O PGM foi a primeira plataforma NGS sem sensibilidade óptica. Em vez de gerar uma cascata enzimática para o equipamento registrar um sinal, o PGM detecta íons de H^+ que são liberados a cada vez que um dNTP é incorporado na sequência. Essa atividade altera o pH do meio, sendo esse o sinal gerado pelo equipamento. Entretanto, a mudança de pH muitas vezes não é proporcional ao número de nucleotídeos detectado, o que diminui a acurácia da leitura (GOODWIN; MCPHERSON; MCCOMBIE, 2016; ROTHBERG *et al.*, 2011).

Já a Oxford Nanopore surge como algumas alternativas aos sequenciamentos de segunda geração, capazes de produzir sequências longas em um curto espaço de tempo. Possuem tecnologia nanopore integrada em chip eletrônico que ajuda a detectar os dNTPs de uma única fita de DNA, oferecendo um fluxo de trabalho rápido e simples com o mínimo de tempo de preparação da amostra (TYSON *et al.*, 2018). Outra vantagem, é que além da geração de mais de 15 GB de dados em tempo real, o seu tamanho e baixo custo fazem o equipamento se destacar e ganhar favoritismo entre as outras plataformas de sequenciamento (GOLDSTEIN *et al.*, 2019; VASUDEVAN *et al.*, 2020).

Figura 1.1 – Conjunto de leituras curtas resultam em *contigs* sem informação estrutural do genoma (A). Conjuntos de leituras longas resultam em um genoma completo com muitos *indels* (B). Hibridização das técnicas rende um genoma completo, de alta acurácia e poucos erros (C).



Fonte: Vasudevan *et al.* (2020)

REFERÊNCIAS

ABBY, Sophie; DAUBIN, Vincent. Comparative genomics and the evolution of prokaryotes. **Trends in Microbiology**, v. 15, n. 3, p. 135-141, 2007. DOI: https://doi.org/10.1016/j.tim.2007.01.007.

AMANN, Rudolf I.; LUDWIG, Wolfgang; SCHLEIFER, Karl-Heinz. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. **Microbiological Reviews**, v. 59, n. 1, p. 143-169, 1995. DOI: https://doi.org/10.1128/mr.59.1.143-169.1995.

ARIMA, Kei; KAKINUMA, Atsushi; TAMURA, Gakuzo. Surfactin, a crystalline peptidelipid surfactant produced by *Bacillus subtilis*: isolation, characterization and its inhibition of fibrin clot formation. **Biochemical and Biophysical Research Communications**, v. 31, n. 3, p. 488-494, 1968. DOI: https://doi.org/10.1016/0006-291X(68)90503-2.

ATANASOV, Atanas G. *et al*. Natural products in drug discovery: advances and opportunities. **Nature Reviews Drug Discovery**, v. 20, n. 3, p. 200-216, 2021. DOI: https://doi.org/10.1038/s41573-020-00114-z.

BARROS, Francisco Fábio Cavalcante *et al*. Surfactina: propriedades químicas, tecnológicas e funcionais para aplicações em alimentos. **Química Nova**, v. 30, n. 2, p. 409-414, 2007. DOI: https://doi.org/10.1590/S0100-40422007000200031.

BORRISS, Rainer *et al*. *Bacillus subtilis*, the model Gram-positive bacterium: 20 years of annotation refinement. **Microbial Biotechnology**, v. 11, n. 1, p. 3-17, 2018. DOI: https://doi.org/10.1111/1751-7915.13043.

COCHRANE, Stephen A.; VEDERAS, John C. Lipopeptides from *Bacillus* and *Paenibacillus* spp.: a gold mine of antibiotic candidates. **Medicinal Research Reviews**, v. 36, n. 1, p. 4–31, 2016. DOI: https://doi.org/10.1002/med.21321.

DEL ANGEL, Victoria Dominguez *et al*. Ten steps to get started in genome assembly and annotation. F1000Research, v. 7, ELIXIR-148, 2018. DOI: https://dx.doi.org/10.12688%2Ff1000research.13598.1.

DELMONT, Tom O. *et al*. Metagenomic mining for microbiologists. **The ISME Journal**, v. 5, n. 12, p. 1837-1843, 2011. DOI: https://doi.org/10.1038/ismej.2011.61.

DIENE, Seydina M. *et al*. The rhizome of the multidrug-resistant Enterobacter aerogenes genome reveals how new "killer bugs" are created because of a sympatric lifestyle. **Molecular Biology and Evolution**, v. 30, n. 2, p. 369-383, 2013. DOI: https://doi.org/10.1093/molbev/mss236.

GOLDSTEIN, Sarah *et al*. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. **BMC genomics**, v. 20, n. 1, p. 1-17, 2019. DOI: https://doi.org/10.1186/s12864-018-5381-7.

GOODWIN, Sara *et al*. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. **Genome Research**, v. 25, n. 11, p. 1750-1756, 2015. DOI: http://www.genome.org/cgi/doi/10.1101/gr.191395.115.

GOODWIN, Sara; MCPHERSON, John D.; MCCOMBIE, W. Richard. Coming of age: ten years of next-generation sequencing technologies. **Nature Reviews Genetics**, v. 17, n. 6, p. 333–351, 2016. DOI: https://doi.org/10.1038/nrg.2016.49 .

GRAZZIOTIN, Ana Laura et al. Comparative genomics of a novel clade shed light on the evolution of the genus *Erysipelothrix* and characterise an emerging species. **Scientific Reports**, v. 11, n. 1, p. 1-12, 2021. DOI: https://doi.org/10.1038/s41598-021-82959-x.

GUDIÑA, Eduardo J. *et al*. Potential therapeutic applications of biosurfactants. **Trends in Pharmacological Sciences**, v. 34, n. 12, p. 667-675, 2013. DOI: https://doi.org/10.1016/j.tips.2013.10.002.

GUIMARÃES, Luis Carlos *et al*. Inside the pan-genome - methods and software overview. **Current Genomics**, v. 16, n. 4, p. 245-252, 2015. DOI: https://doi.org/10.2174/1389202916666150423002311.

HARWOOD, Colin R. *et al*. Secondary metabolite production and the safety of industrially important members of the *Bacillus subtilis* group. **FEMS Microbiology Reviews**, v. 42, n. 6, p. 721-738, 2018. DOI: https://doi.org/10.1093/femsre/fuy028.

HUG, Laura A. *et al*. A new view of the tree of life. **Nature Microbiology**, v. 1, n. 5, p. 1-6, 2016. DOI: https://doi.org/10.1038/nmicrobiol.2016.48.

KASPAR, Felix; NEUBAUER, Peter; GIMPEL, Matthias. Bioactive secondary metabolites from *Bacillus subtilis*: a comprehensive review. **Journal of Natural Products**, v. 82, n. 7, p. 2038-2053, 2019. DOI: https://doi.org/10.1021/acs.jnatprod.9b00110.

KHOPADE, Abhijit *et al*. Production and stability studies of the biosurfactant isolated from marine *Nocardiopsis* sp. B4. **Desalination**, v. 285, p. 198-204, 2012. DOI: https://doi.org/10.1016/j.desal.2011.10.002.

KUNST, F. *et al*. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. **Nature**, v. 390, n. 6657, p. 249-256, 1997. DOI: https://doi.org/10.1038/36786.

LOBRY, Jean R.; SUEOKA, Noboru. Asymmetric directional mutation pressures in bacteria. **Genome Biology,** v. 3, n. 10, p. 1-14, 2002. DOI: https://doi.org/10.1186/gb-2002-3-10-research0058.

MADSEN, Jens K. *et al*. The anionic biosurfactant rhamnolipid does not denature industrial enzymes. **Frontiers in Microbiology**, v. 6, p. 292, 2015. DOI: https://doi.org/10.3389/fmicb.2015.00292.

MAKKAR, Randhir S.; ROCKNE, Karl J. Comparison of synthetic surfactants and biosurfactants in enhancing biodegradation of polycyclic aromatic hydrocarbons. **Environmental Toxicology and Chemistry: An International Journal,** v. 22, n. 10, p. 2280-2292, 2003. DOI: https://doi.org/10.1897/02-472.

NEAL-MCKINNEY, Jason M. *et al*. Comparison of MiSeq, MinION, and hybrid genome sequencing for analysis of *Campylobacter jejuni*. **Scientific reports**, v. 11, n. 1, p. 1-10, 2021. DOI: https://doi.org/10.1038/s41598-021-84956-6.

ROTHBERG, Jonathan M. *et al*. An integrated semiconductor device enabling non-optical genome sequencing. **Nature**, v. 475, n. 7356, p. 348-352, 2011. DOI: https://doi.org/10.1038/nature10242.

SEN, Ramkrishna. Surfactin: biosynthesis, genetics and potential applications. In: SEN, Ramkrishna. (ed.) **Biosurfactants**. Advances in Experimental Medicine and Biology. New York: Springer, 2010. v. 672, p. 316-323. DOI: https://doi.org/10.1007/978-1-4419-5979-9_24.

SERRA, Cláudia R. *et al*. Sporulation during growth in a gut isolate of *Bacillus subtilis*. **Journal of Bacteriology**, v. 196, n. 23, p. 4184-4196, 2014. DOI: https://doi.org/10.1128/JB.01993-14.

SGHAIER, Haitham *et al*. Genomic insights of halophilic *Planococcus maritimus* SAMP MCC 3013 and detail investigation of its biosurfactant production. **Frontiers in Microbiology**, v. 10, p. 235, 2019. DOI: https://doi.org/10.3389/fmicb.2019.00235.

SHAFI, Jamil; TIAN, Hui; JI, Mingshan. *Bacillus* species as versatile weapons for plant pathogens: a review. **Biotechnology & Biotechnological Equipment**, v. 31, n. 3, p. 446-459, 2017. DOI: https://doi.org/10.1080/13102818.2017.1286950.

SNIPEN, Lars; ALMØY, Trygve; USSERY, David W. Microbial comparative pan-genomics using binomial mixture models. **BMC Genomics**, v. 10, n. 1, p. 1-8, 2009. DOI: https://doi.org/10.1186/1471-2164-10-385.

STAHL, David A.; TIEDJE, James M. **Microbial ecology and genomics: a crossroads of opportunity**. Washington (DC): American Academy of Microbiology, 2002, 28 p. DOI: https://doi.org/10.1128/aamcol.23feb.2001.

STICE, Shaun P. *et al*. *Pantoea ananatis* genetic diversity analysis reveals limited genomic diversity as well as accessory genes correlated with onion pathogenicity. **Frontiers in Microbiology**, v. 9, p. 184, 2018. DOI: https://doi.org/10.3389/fmicb.2018.00184.

THIES, Stephan *et al*. Metagenomic discovery of novel enzymes and biosurfactants in a slaughterhouse biofilm microbial community. **Scientific Reports**, v. 6, n. 1, p. 1-12, 2016. DOI: https://doi.org/10.1038/srep27035.

TYSON, John R. *et al*. MinION-based long-read sequencing and assembly extends the *Caenorhabditis elegans* reference genome. **Genome Research**, v. 28, n. 2, p. 266-274, 2018. DOI: http://www.genome.org/cgi/doi/10.1101/gr.221184.117.

VASUDEVAN, Karthick *et al*. Highly accurate-single chromosomal complete genomes using IonTorrent and MinION sequencing of clinical pathogens. **BMC Genomics**, v. 112, n. 1, p. 545-551, 2020. DOI: https://doi.org/10.1016/j.ygeno.2019.04.006.

VERNIKOS, George *et al*. Ten years of pan-genome analyses. **Current Opinion in Microbiology**, v. 23, p. 148-154, 2015. DOI: https://doi.org/10.1016/j.mib.2014.11.016.

WATTAM, Alice R. *et al*. PATRIC, the bacterial bioinformatics database and analysis resource. **Nucleic Acids Research**, v. 42, n. D1, p. D581-D591, 2014. DOI: https://doi.org/10.1093/nar/gkt1099.

WILLENBACHER, Judit *et al*. Enhancement of surfactin yield by improving the medium composition and fermentation process. **AMB Express**, v. 5, n. 1, p. 1-9, 2015. DOI: https://doi.org/10.1186/s13568-015-0145-0.

WILLSON, Joseph. **Sequencing — the next generation**. Nature, Milestones in Genomic Sequencing, Milestone 3, 2005, 10 fev. 2021. Disponível em: https://www.nature.com/articles/d42859-020-00103-7. Acesso em: 6 abr. 2021.

WU, Hao; WANG, Dan; GAO, Feng. Toward a high-quality pan-genome landscape of *Bacillus subtilis* by removal of confounding strains. **Briefings in Bioinformatics**, v. 22, n. 2, p. 1951-1971, 2021. DOI: https://doi.org/10.1093/bib/bbaa013.

ZHANG, Junhui *et al*. Production of lipopeptide biosurfactants by *Bacillus atrophaeus* 5-2a and their potential use in microbial enhanced oil recovery. **Microbial Cell Factories**, v. 15, n. 1, p. 1-11, 2016. DOI: https://doi.org/10.1186/s12934-016-0574-8.

ZHANG, Yunbin *et al*. Antibacterial activity and mechanism of cinnamon essential oil against *Escherichia coli* and *Staphylococcus aureus*. **Food Control**, v. 59, p. 282-289, 2016. DOI: https://doi.org/10.1016/j.foodcont.2015.05.032.

ZHI, Yan; WU, Qun; XU, Yan. Genome and transcriptome analysis of surfactin biosynthesis in *Bacillus amyloliquefaciens* MT45. **Scientific Reports**, v. 7, n. 1, p. 1-13, 2017. DOI: https://doi.org/10.1038/srep40976.

ZIEMERT, Nadine; ALANJARY, Mohammad; WEBER, Tilmann. The evolution of genome mining in microbes - a review. **Natural Product Reports**, v. 33, n. 8, p. 988-1005, 2016. DOI: https://doi.org/10.1039/C6NP00025H.

PARTE 2

ARTIGO ÚNICO – GENOMA COMPLETO ALTAMENTE PRECISO DA CEPA RI4914 DE *Bacillus subtilis* E SUA GENÔMICA COMPARATIVA

Julie Kennya de Lima Ferreira[1], Cristina Ferreira Silva e Batista[2], Victor Satler Pylro[3]

Versão preliminar – Norma NBR 6022 (ABNT 2018)

Resumo

As bactérias do gênero *Bacillus* são amplamente exploradas por três motivos: fácil manuseio em laboratório, organismos modelos para o estudo de bactérias Gram-positivas e alto potencial biotecnológico quanto a prospecção de metabólitos secundários. Neste último, a surfactina se destaca por ser um biossurfactante lipopeptídico comum a este gênero, produzida principalmente por *Bacillus subtilis* e *Bacillus anthracis*. Para que a surfactina seja de fato expressa pela bactéria, a ativação de uma cascata reguladora da expressão gênica precisa ocorrer, porém o *operon* responsável por esta atividade, esteja ativo ou inativo, pode ser encontrado conservado no material genético dos indivíduos, independente de diferenças de cultivo ou habitat. Nesse contexto, o trabalho teve por objetivo avaliar a prevalência do *operon* de surfactina (*sfrA*ABCD) em estirpes de *Bacillus subtilis*, bem como o repertório pangenômico da espécie por meio de genômica comparativa, além de sequenciar o genoma da cepa RI4914 de *Bacillus subtilis* através de abordagem híbrida de sequenciamento. Para isso, as plataformas PGM IonTorrent e GridION™ foram utilizadas para o sequenciamento e 153 genomas foram recuperados do NCBI, dos quais foram realizadas as análises de pangenoma e MLST, além da análise de prevalência do *operon* de surfactina (*sfrA*ABCD). O genoma obtido apresentou alta qualidade e boa acurácia, com completude de 98,84% e contaminação de 1,10%. A análise de genômica comparativa evidenciou o perfil pangenômico aberto em *B. subtilis* e prevalência da surfactina em todos os genomas analisados, demonstrando-se como um caráter distintivo dessa espécie.

Palavras-chave: Surfactina (*sfrA*ABCD). Análise multilocus MLST. Abordagem híbrida de sequenciamento.

HIGHLY ACCURATE COMPLETE GENOME OF THE *Bacillus subtilis* STRAIN RI4914 AND ITS COMPARATIVE GENOMICS

Abstract

Bacillus genus bacteria are widely exploited for three reasons: easy laboratory handling, model organisms for the study of Gram-positive bacteria, and high biotechnological potential for prospecting secondary metabolites. In the latter, surfactin can be highlighted,

[1] Bióloga, Mestranda do Programa de Pós-Graduação em Microbiologia Agrícola, Universidade Federal de Lavras. E-mail: juliekennya1@gmail.com
[2] Doutora em Ciências dos Alimentos, Professora Associada do Departamento de Biologia, Universidade Federal de Lavras. E-mail: cristinafsb@dbi.ufla.br
[3] Doutor em Microbiologia Agrícola, Professor Adjunto do Departamento de Biologia, Universidade Federal de Lavras. E-mail: victor.pylro@gmail

characterized as a common lipopeptide biosurfactant of the genus, produced mainly by *Bacillus subtilis* and *Bacillus anthracis*. For the surfactin to be in fact expressed by the bacteria, a cascade activation regulates gene expression needs to occur, but the operon responsible for this activity, whether active or not, can be found conserved in the genetic material of the belonging, regardless of differences in cultivation or habitat. In this context, the study aimed to evaluate the prevalence of the surfactin operon (*sfrA*ABCD) in *Bacillus subtilis* strains, as well as the pangenomic repertoire of the species, through comparative genomics. In addition, to sequencing the genome of the RI4914 strain of *Bacillus subtilis* through of hybrid sequencing approach. Thereunto, PGM IonTorrent and GridIONTM platforms were used for sequencing and 153 genomes were retrieved from the NCBI. Of these, pan-genome and MLST analyzes were performed, in addition to the prevalence analysis of the surfactin operon (*sfrA*ABCD). The genome obtained showed high quality and good accuracy, with completeness of 98.84% and contamination of 1.10%. The comparative genomics analysis showed the open pangenomic profile in *B. subtilis* and surfactin prevalence in all analyzed genomes, demonstrating it as a distinctive character of this species.

Keywords: Surfactin (*sfrA*ABCD). Multilocus MLST analysis. Hybrid sequencing approach.

1 INTRODUÇÃO

As abordagens NGS (*Next-generation sequencing*) têm alcançado amplas possibilidades nos estudos de sequenciamentos de genomas, exomas, RNA, micro-RNA, sequenciamento de metilação e metagenômica. Isso colaborou para o alto crescimento, proporcional a diminuição do custo de utilização de tecnologias NGS. Com esse avanço, tornou-se possível classificar e organizar as plataformas NGS em segunda e terceira geração, sendo as plataformas de segunda geração todas aquelas que trabalhavam o sequenciamento de material genético seguindo a utilização de leituras curtas, e as plataformas de terceira geração aquelas que trabalhavam a mesma função, porém utilizando-se de leituras longas (SINGH; BHATIA, 2020).

Como exemplos amplamente utilizados desta divisão, tem-se a plataforma ION Torrent, como sequenciador de leituras curtas, e a plataforma MinION, da terceira geração de NGS. Ambas são utilizadas em diagnósticos de amostras clínicas, porém, as plataformas de terceira geração oferecem maiores vantagens significativas, como a possibilidade de sequenciamento sem a necessidade de PCR ou etapas de clonagem e também, o rendimento de um genoma cuja qualidade é superior aos obtidos por plataformas de segunda, devido a utilização de leituras longas, pois as mesmas diminuem consideravelmente a taxa de erro durante a montagem dos genomas (GULILAT *et al.*, 2019).

Atualmente, plataformas de terceira geração como o MinION, têm ganhado popularidade em países mais desenvolvidos devido ao custo, flexibilidade e praticidade que o

equipamento possui. Entretanto, estudos recentes têm apresentado resultados superiores para obtenção de genomas completos quando leituras longas e leituras curtas são utilizadas em conjunto. Essa técnica de abordagem híbrida de sequenciamento une as duas gerações de plataformas NGS a fim de gerar montagens de genomas completos fechados, de alta qualidade e sem *gaps*.

Essa abordagem surgiu como forma de contrapor as limitações presentes nas plataformas de segunda e terceira geração, como as altas taxas de erro presentes em ambas as gerações, a complexidade da montagem de genomas provenientes da alta fragmentação do material genético das plataformas de leituras curtas e o alto número de *gaps* presentes em plataformas de leituras longas. Portanto, este trabalho objetivou-se realizar abordagem de sequenciamento híbrido para montagem do genoma da cepa de *Bacillus subtilis* RI4914, incorporando leituras curtas e longas para montar um genoma completo com alta precisão e *INDELS* significativamente reduzidos. Além disso, também objetivou-se analisar o repertório pangenômico da espécie quanto a prevalência do *operon* de surfactina (*sfrA*ABCD), comum a espécie. Após a montagem, a mesma foi avaliada quanto a qualidade e integridade do genoma e o repertório pangenômico da espécie também foi avaliado.

2 MATERIAL E MÉTODOS

2.1 Obtenção e reativação do isolado de *Bacillus subtilis* RI4914

O isolado *Bacillus subtilis* RI4914, caracterizado como produtor do biossurfactante surfactina está depositado na Coleção de Culturas do Laboratório de Biotecnologia e Biodiversidade para o Meio Ambiente, do Departamento de Microbiologia de Universidade Federal de Viçosa (UFV). Esse isolado foi obtido a partir de amostras de água de produção do campo de exploração de petróleo da formação Rio Itaúna, Conceição da Barra, Espírito Santo e, previamente, caracterizado por Fernandes *et al*. (2016).

Para a reativação, o isolado foi cultivado em Ágar Triptona de Soja (TSA), a 30 ºC, por 24 h e posteriormente, uma colônia isolada foi transferida para um tubo contendo 5 mL de Caldo Triptona de Soja (TSB), em triplicata. Os tubos foram incubados sob agitação de 200 rpm, a 30 ºC, por 18 h.

2.2 Condições de cultivo do *Bacillus subtilis* RI4914 para avaliação da produção de biossurfactante

Após a reativação, uma alíquota do pré-inóculo foi transferida para Erlenmeyers de 125 mL contendo 30 mL de meio mineral, contendo glicose (tratamento 1) ou glicerol sintético (tratamento 2) como fonte de carbono, de modo a se obter uma densidade óptica a 600 nm equivalente a 0,05. A composição do primeiro tratamento foi de 13,9 $g.L^{-1}$ de KHP_4, 2,7 $g.L^{-1}$ de KH_2PO_4, 0,05 $g.L^{-1}$ de extrato de levedura, 4,24 $g.L^{-1}$ de $NaNO_3$, 40 $g.L^{-1}$ de glicose e 50 mL de solução de micronutrientes. A composição do segundo tratamento foi de 13,9 $g.L^{-1}$ de K_2HP_4, 2,7 $g.L^{-1}$ de KH_2PO_4, 0,05 $g.L^{-1}$ de extrato de levedura, 4,24 $g.L^{-1}$ de $NaNO_3$, 40 $g.L^{-1}$ de glicerol e 50 mL de solução de micronutrientes. A composição da solução de micronutrientes foi de 0,5 $g.L^{-1}$ de EDTA, 3,0 $g.L^{-1}$ de $MgSO_4.7H_2O$, 0,5 $g.L^{-1}$ de $MnSO_4.4H_2O$, 1,0 $g.L^{-1}$ de NaCl, 0,1 $g.L^{-1}$ de $CaCl.2H_2O$, 0,1 $g.L^{-1}$ $CaCl_2.2H_2O$, 0,1 $g.L^{-1}$ de $CaCl_2.6H_2O$, 0,1 $g.L^{-1}$ de $ZnSO_4.7H_2O$, 0,01 $g.L^{-1}$ de $CuSO_4.5H_2O$, 0,01 $g.L^{-1}$ de $Na_2MoO_4.2H_20$, 0,01 $g.L^{-1}$ de NaO_4Se, 0,01 $g.L^{-1}$ de $Na_2WO_4.2H_2O$ e 0,02 $g.L^{-1}$ de $NiCl_2.6H_2O$. Essa solução foi preparada, autoclavada e, posteriormente, adicionada aos meios de cultura esterilizados. Ambos os tratamentos foram incubados por 48 h, a 30 ºC e sob agitação constante de 200 rpm e avaliados em triplicata.

2.3 Teste de espalhamento de gota

Para execução do teste de espalhamento de gota, utilizou-se 500 μL do sobrenadante das amostras centrifugadas (5 min. a 12.000 rpm, em tubos de Eppendorfs), 70 mL de água destilada, uma placa de Petri (150 x 20 mm) e 20 μL de petróleo. O processo seguiu o método proposto por Morikawa, Hirata e Imanaka (2000) com modificações, em que a placa de Petri foi preenchida pela água destilada juntamente com o petróleo adicionado a toda sua superfície. A partir disso, o sobrenadante das culturas foi adicionado ao centro da placa de Petri, formando um halo de coloração mais clara no filme de óleo que, após 30 min. foram medidos com auxílio de uma régua.

2.4 Extração de ácidos nucleicos de *Bacillus subtilis* RI4914

O DNA e RNA das culturas, para ambos os tratamentos, foram coextraídos no Laboratório de Ecologia Microbiana – UFLA, utilizando RNA PowerSoil® Total RNA

Isolation Kit (MoBio Laboratories, Carlsbad, CA) e RNA PowerSoil® DNA Elution Accessory Kit, seguindo o protocolado pelo fabricante. A qualidade dos ácidos nucleicos extraídos foi avaliada por espectrofotometria com o auxílio do Nanodrop Lite (ThermoFisher Scientific, Waltham, MA, USA) e a quantificação foi realizada por fluorometria em equipamento Qubit 4.0 (ThermoFisher Scientific, Waltham, MA, USA), utilizando os kits dsDNA BR Assay Kit e Qubit RNA BR Assay Kit, para DNA e RNA, respectivamente.

2.5 Sequenciamento do genoma de *Bacillus subtilis* RI4914, montagem e anotação

Uma subamostra do DNA genômico, de aproximadamente 4 µg, foi tratado com *Rapid Sequencing Kit* (SQK-RAD004; Oxford Nanopore Technologies, UK). A biblioteca resultante foi sequenciada utilizando a plataforma GridION™, utilizando uma *flowcell Spot-ON Mk1* (FLO-MIN 106, versão R9; Oxford Nanopore Technologies, UK) e *Library Loading Bead Kit* versão R9 (EXP-LLB001; Oxford Nanopore Technologies, UK). As sequências longas (*long reads)* brutas foram obtidas com o auxílio do programa MinKNOW v3.5.6, em uma corrida de 72 h e o *basecalling* foi realizado simultaneamente, utilizando o programa Albacore v2.0.2.

Outra subamostra do mesmo DNA foi sequenciado utilizando a plataforma PGM IonTorrent (ThermoFisher Scientific, Waltham, MA, USA). A biblioteca foi preparada utilizando *Ion Plus Fragment Library Kit* e clonalmente amplificada no equipamento One Touch 2 System (ThermoFisher Scientific, Waltham, MA USA), utilizando o Ion PGM^{TM} Template Hi-Q OT2 400 Kit. A biblioteca amplificada foi carregada em um chip 316^{TM} v.2 e posteriormente sequenciada em plataforma PGM IonTorrent, utilizando o PGM Hi-Q Sequencing 400 Kit.

As sequências longas, obtidas em GridION, foram montadas por abordagem *de novo,* utilizando o programa Canu v1.5 (KOREN *et al.*, 2017), seguindo os parâmetros padrão para dados de Oxford Nanopore. A montagem de baixa qualidade obtida foi então corrigida (*polished*) com o programa Racon (VASER *et al.*, 2017), utilizando as sequências curtas provenientes do sequenciamento no PGM IonTorrent. Para isso, as sequências curtas foram mapeadas contra a montagem principal de sequências longas, utilizando a ferramenta de alinhamento Burrows-Wheeler (LI; DURBIN, 2009). O procedimento com o Racon foi realizado três vezes, para a obtenção de um genoma completo, circular e de alta qualidade.

A completude do genoma e o nível de contaminação foram estimadas com programa CheckM (PARKS *et al.*, 2015), no modo "lineage_wf". A qualidade do genoma foi estimada com Quast (GUREVICH *et al.*, 2013), utilizando *Bacillus subtilis* subsp. *subtilis* str. *168* -

NC_000964.3 como o genoma de referência. A identidade média nucleotídica com base no BLAST (ANIb) entre o genoma obtido e a sua referência foi determinada pelo programa JSpecies (RICHTER; ROSSELLÓ-MÓRA, 2009); onde ANI > 95% indica que ambos os genomas pertencem a mesma espécie (GORIS *et al.*, 2007; RICHTER *et al.*, 2016). Por fim, a anotação do genoma foi feita com o programa PATRIC versão 3.5.23 (WATTAM *et al.*, 2014) e com o *NCBI Prokaryotic Genome Annotation Pipeline* – PGAP (TATUSOVA *et al.*, 2016).

2.6 Genômica comparativa

Todos os genomas completos de *B. subtilis*, disponíveis no banco de dados GenBank (https://www.ncbi.nlm.nih.gov) em 14 de maio de 2020, foram recuperados, resultando em um total de 153 genomas (Apêndice A). As comparações das estimativas de hibridização digital DNA-DNA (dDDH) e a identidade média nucleotídica (ANI) entre os genomas de *B. subtilis* RI4914 e sua referência *B. subtilis* subsp. *subtilis* str. *168* - NC_000964.3 foram calculadas usando os *webservers* JSpeciesWS (RICHTER *et al.*, 2016) e KostasLab (RODRIGUEZ-R; KONSTANTINIDIS, 2014), respectivamente.

2.7 Análises de reconstrução filogenética

Sequências completas do gene *rpoB*, que codifica para a subunidade β da RNA polimerase, foram recuperadas dos genomas em estudo e utilizadas para as análises filogenéticas. O genoma de *Bacillus amyloliquefaciens* DSM7 (NC_014551.1) foi utilizado como grupo externo para a análise. O alinhamento das sequências foi executado com o programa Muscle (EDGAR, 2004). A análise de Inferência Bayesiana (BA) foi executada utilizando o programa MrBayes (HUELSENBECK; RONQUIST, 2001) para fins de reconstrução filogenética. Para isso, foram rodadas duas análises independentes, com quatro cadeias cada, sendo uma fria e três quentes, iniciadas com quatro diferentes árvores aleatórias, em 10.000.000 de gerações de *Markov chain Monte Carlo* - MCMC. A verossimilhança das topologias resultantes foi checada e 25% das árvores geradas foram eliminadas para garantir a permanência apenas daquelas cujas áreas apresentavam os melhores resultados de verossimilhança, finalizando assim a geração da árvore consenso. A robustez de cada nó da árvore foi obtida pela “probabilidade posterior”, que foi calculada pela frequência de cada nó na árvore consenso.

2.8 Análise de Pangenoma

A análise de pangenoma dos 153 genomas foi executada com o auxílio do Roary Package v3.11.2, utilizando os parâmetros padrão (PAGE *et al.*, 2015). Para isso, o programa Prokka v1.14.6 foi usado para anotação das montagens que foram então usadas como dados de entrada para o Roary, a fim de determinar os conjuntos de genes *cores* e de genes acessórios de todos os genomas avaliados.

2.9 Análise de Multilocus Sequence Typing (MLST)

A análise de sequência *multilocus* (MLST) foi realizada utilizando o conjunto de genes *core* gerados pela ferramenta Roary. Um total de 266 genes *cores* foram concatenados e alinhados com o programa MAFTT (KATOH; STANDLEY, 2013) e manualmente checados. As análises de MLST foram executadas por meio do método de Máxima Verossimilhança, utilizando o programa RAxML v.8.2.10, com os parâmetros padrão. A robustez de cada nó foi estimada pelo teste de Bootstrap com 1.000 repetições.

Similarmente, o *operon* responsável pela biossíntese da surfactina em *B. subtilis*, *srfA*ABCD de ~27 kb, foi recuperado dos genomas anotados e posteriormente utilizados para análise de MLST pelo método de Inferência Bayesiana.

2.10 Avaliação da expressão do gene srfAA, relacionado a produção de surfactina em *Bacilus subtilis*, nos tratamentos avaliados

A expressão do gene *srfA*A, que codifica para a biossíntese da proteína surfactina sintetase subunidade 1, foi avaliada nos tratamentos avaliados. Oligonucleotídeos iniciadores específicos para o gene *srfA*A foram desenhados, utilizando como referência a lista dos genes recuperados dos genomas de *B. subtilis*. Brevemente, as sequências anotadas pelo Prokka, como srfAA foram recuperadas dos genomas obtidos no GenBank. Posteriormente, as sequências em formato FASTA foram alinhadas com o Clustal X (LARKIN *et al.*, 2007), utilizando o programa MEGA X v 10.1.6. A ferramenta *NCBI Conserved Domain Search* (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) foi utilizada para verificação dos domínios ativos e, em seguida, a ferramenta BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) foi utilizada para verificar a identidade dos domínios com o genoma de referência – *Bacillus subtilis* subsp. *subtilis* str. 168. As áreas de domínio conservado foram respeitadas, de forma

que o produto da PCR fosse gerado sem esta parte. As sequências selecionadas foram inseridas na ferramenta Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) para obtenção dos oligonucleotídeos iniciadores. Os parâmetros utilizados foram: *PCR Product Size 80 (min)* e *150 (max), Primer melting temperatures: 57 (min), 60 (opt), 63 (max)* e *3 (Max $T_m$diff)*. Cada *primer* gerado foi testado na ferramenta OligoAnalyzer™ (https://www.idtdna.com/calc/analyzer), para avaliar as formações de *Hairpin*, *self-dimer* e *hetero-dimer*. Os parâmetros utilizados foram: Hairpin: DeltaG, quanto mais positivo, melhor; temperatura: quanto menor, melhor; Self-dimer: DeltaG, menor do que 10% do DeltaG total. Dois pares de oligonucleotídeos iniciadores foram obtidos e sintetizados, mas um deles (*srf*AA-F: AGGCGGGGATCTTTGACA e *srf*AA-R: TGAAGCGGAATCTCAATGC) apresentou melhor eficiência e foi selecionado para a análise de RT-qPCR. O gene codificador do rRNA 16S foi utilizado como controle endógeno (16S - F: CCTACGGGAGGCAGCAG e 16S - R: ATTACCGCGGCTGCTGG).

As amostras de RNA de ambos os tratamentos foram tratadas com o *TURBO DNA-free*™ *Kit* (ThermoFisher Scientific, Waltham, MA, USA) para eliminação de DNA residual, de acordo com as recomendações do fabricante para 5,0 μg de RNA. Para avaliar a integridade do RNA tratado, as amostras foram submetidas a eletroforese em gel de agarose (0,7%), corado com *GelRed Nucleic Acid Gel Stain* (Sigma Aldrich) e visualizado em fotodocumentador UV-transiluminador (UVITEC FireReader XS D-77Ls-20. M). As amostras foram novamente quantificadas utilizando-se o espectrofotômetro *NanoVue*® (GE Healthcare).

A síntese do cDNA foi realizada com o *High Capacity cDNA Reverse Transcription Kit* (ThermoFisher Scientific, Waltham, MA, USA), seguindo as recomendações do fabricante para reações sem inibidor de RNAse (Tabela 1.1), e a ciclagem de cDNA seguiu a incubação em termociclador por 10 min. a 25 ºC, 120 min. a 37 ºC e 5 min. a 85 ºC.

Tabela 1.1 – Reação de síntese de cDNA.

Componente	Volume
10X RT Buffer	2,0 μL
25X dNTP Mix (100 mM)	0,8 μL
10X RT Random Primers	2,0 μL
MultiScribe Reverse Transcriptase	1,0 μL
Nuclease-free H2O	4,2 μL
Total per reaction	10,0 μL

Fonte: Da autora

As reações de amplificação foram realizadas em termociclador Rotor-Gene Q Real-Time PCR (Qiagen), com sistema de detecção SYBR® Green. O volume final da reação para cada amostra foi de 15 µL, sendo 7,5 µL de Master Mix SYBR Green (QuantiNova SYBR Green PCR Kit - Qiagen), 1,5 µL de cDNA na diluição 1/5 e 1,5 µL de cada oligonucleotídeo iniciador e 3,0 µL água livre de RNase. No controle sem DNA, água livre de RNase foi adicionada para um volume final de 15 µL.

Três repetições biológicas de cada tratamento foram utilizadas, com triplicatas técnicas para cada uma das repetições. As condições de ciclagem utilizadas foram: 5 min. a 95 °C para ativação da enzima, seguidos de 40 ciclos de 5 s a 95 °C e 10 s a 60 °C, com uma rampa de *melting* para se avaliar a especificidade da reação de 55 °C a 95 °C, com o aumento de 1 °C a cada 5 s. A eficiência de amplificação dos oligonucleotídeos iniciadores do gene alvo e de referência foi determinada por meio de curva de diluição (Tabela 2.2). A análise da expressão relativa foi calculada utilizando o método descrito por Pfaffl (2001), utilizando o gene codificador do rRNA 16 como referência.

3 RESULTADOS

3.1 Sequenciamento, montagem e anotação do genoma do *Bacillus subtilis* RI4914

Um total de 3,48 milhões de sequências longas (20.69 GB) foram obtidas a partir de sequenciamento na plataforma GridION^{TM}, com o tamanho das sequências variando de 70 a 79.827 pb. As sequências maiores que 1.000 pb foram utilizadas para montagem *de novo*. Similarmente, um total de 4,14 milhões de sequências curtas, de alta qualidade (1.02 Gb, > Q20), foram obtidas a partir do sequenciamento em plataforma IonTorrent, com tamanhos variando de 25 a 381 pb. A montagem inicial com o Canu gerou um genoma completo, mas de baixa qualidade [alta quantidade de CDSs (regiões de codificação)]. Após a correção utilizando as sequências curtas, o genoma final consistiu em um único cromossomo contínuo e circular, sem plasmídeos detectados e de alta qualidade. O genoma apresentou tamanho de 4.100.952 pb e conteúdo G+C de 43.48% (Figura 2.1A).

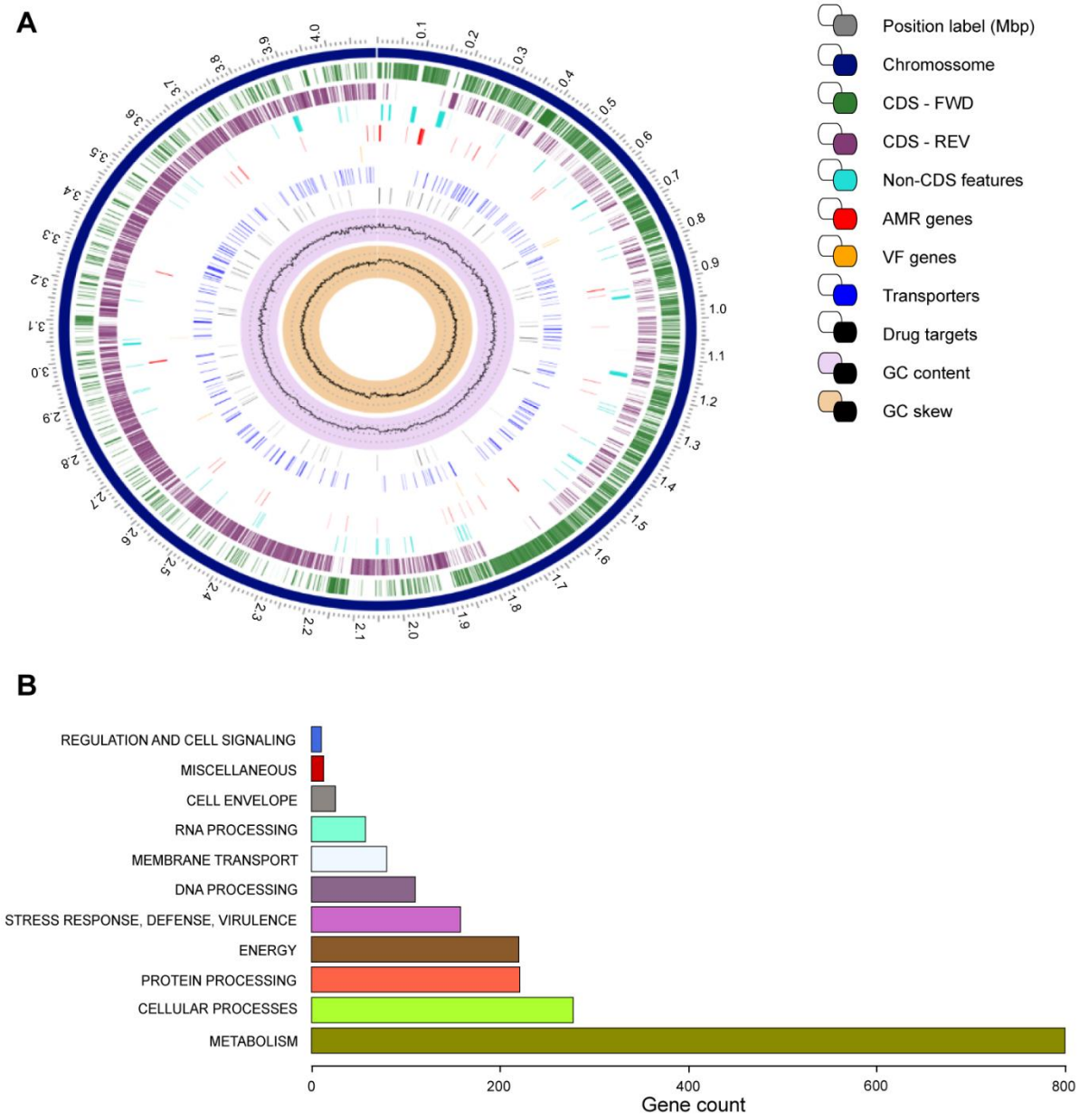
A completude do genoma e contaminação estimada pelo CheckM (PARKS *et al.*, 2015), foi de 98,84% e 1,10%, respectivamente, sendo classificado como um genoma completo e de baixa contaminação. A avaliação da montagem estimada pelo Quast (GUREVICH *et al.*, 2013) é apresentada no Apêndice B. A média da identidade nucleotídica

baseada no BLAST (ANIb) entre o genoma obtido e sua referência foi de 98,26%, indicando que ambos pertencem a mesma espécie (GORIS *et al.*, 2007; RICHTER *et al.*, 2016).

O dDDH calculado a partir da distância Genoma/Genoma foi de 0,0157. A distância foi transformada em valores análogos a DDH, usando um modelo linear generalizado (GLM), inferido a partir de banco de dados de referências empíricas, composta por valores reais de DDH e sequências genômicas. O valor obtido foi de 86,70% [84,1–88,9%], baseado na fórmula recomendada pelo programa utilizado (identidades/comprimento de HSP). A probabilidade de DDH > 70% (*i.e.*, mesma espécie) e de DDH > 79% (*i.e.*, mesma subespécie) foram de 94,53% e 59,84% (a partir de regressão logística), respectivamente. Além disso, a diferença entre a porcentagem de conteúdo G+C foi de 0,03, reforçando que ambos os genomas, o de referência e o em estudo, pertencem a mesma espécie.

Um total de 4,693 sequências codificadoras (CDS) e 116 RNAs não-codificadores, incluindo 86 tRNA e 30 rRNA (10 cópias completas do *operon* ribossomal) foram identificados com o auxílio do *webserver* PATRIC. As funções foram classificadas em classes de subsistemas pelo PATRIC (Figura 2.1B). A maioria das CDSs anotadas foram classificadas como relacionadas ao “metabolismo”, seguido de “processos celulares” e processamento de proteínas. Além disso, o PATRIC anotou 48 genes relacionados à resistência a antibióticos, incluindo uma β-lactamase de classe D (Apêndice C).

Figura 2.1 – (A) Visão gráfica circular do cromossomo de *Bacillus subtilis* RI4914. (B) Classe de subsistemas classificados pelo PATRIC. Da parte mais externa ao centro do círculo: CDS (*foward*), CDS (*reverse*), caracteres de *non-CDS*, mutações pontuais associadas à resistência, transportadores, *drug targets*, conteúdo GC e viés GC.

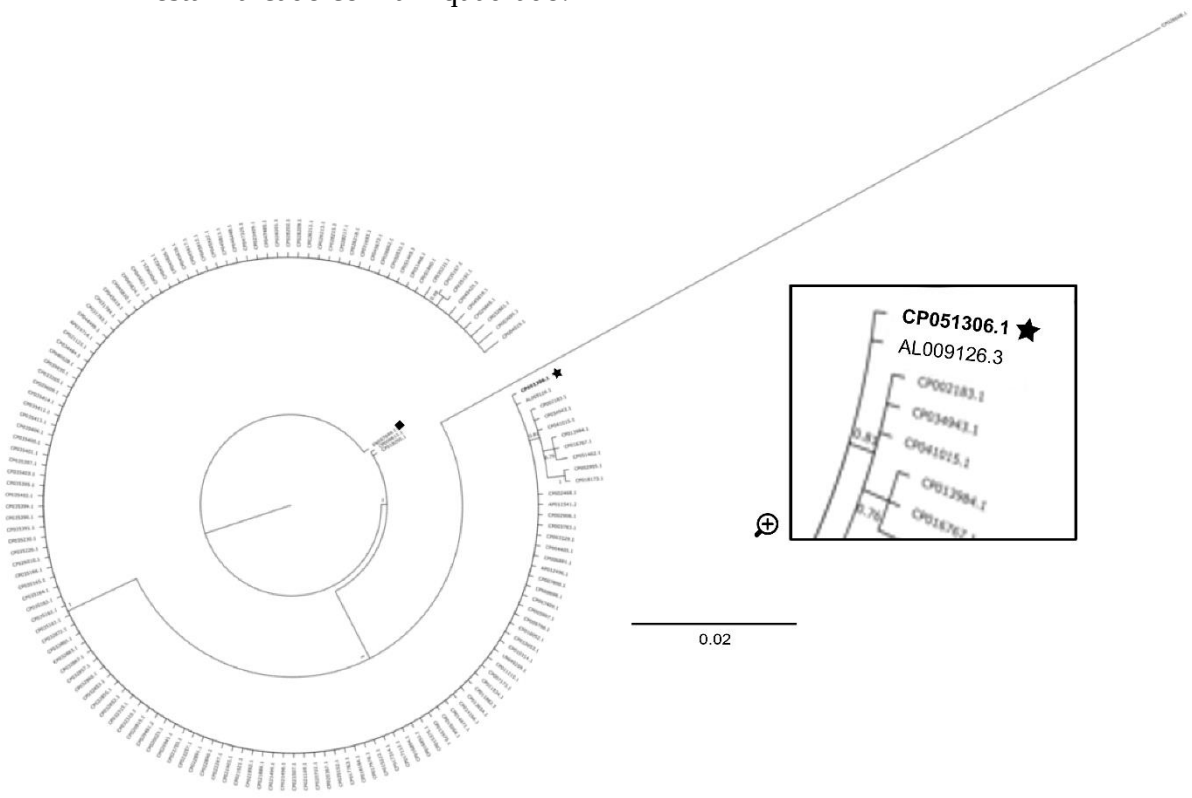


Fonte: Da autora

3.2 Reconstrução filogenética de *Bacillus subtilis*

A confirmação da classificação dos 153 genomas obtidos foi realizada por meio de análises de reconstrução filogenética. A hipótese de Inferência Bayesiana para a reconstrução filogenética com base na sequência do gene *rpoB* é ilustrada na Figura 2.2. O gene *rpoB* é universal, altamente conservado entre bactérias e se apresenta em única cópia, sendo reconhecido como um bom marcador genético para comparações filogenéticas (ADÉKAMBI; DRANCOURT; RAOULT, 2009).

Figura 2.2 – Árvore obtida análise de Inferência Bayesiana, a partir de sequências do gene *rpoB*. Os valores nos ramos representam probabilidade posterior. A sequência recuperada *Bacillus subtilis* RI4914 está marcada com uma estrela (em destaque) e a de *Bacillus amyloliquefaciens* DSM7, grupo externo da análise, está marcado com um quadrado.



Fonte: Da autora

O gene codificador de rRNA 16S apresenta diversas cópias na maioria dos genomas bacterianos e algumas dessas podem ser de pseudogenes, o que pode aumentar a frequência de variações nas diferentes cópias de um mesmo organismo. Como esperado, o gene *rpoB* apresentou alta conservação nos representantes de *B. subtilis* avaliados e, provavelmente, as poucas variações observadas são relativas aos erros durante os processos

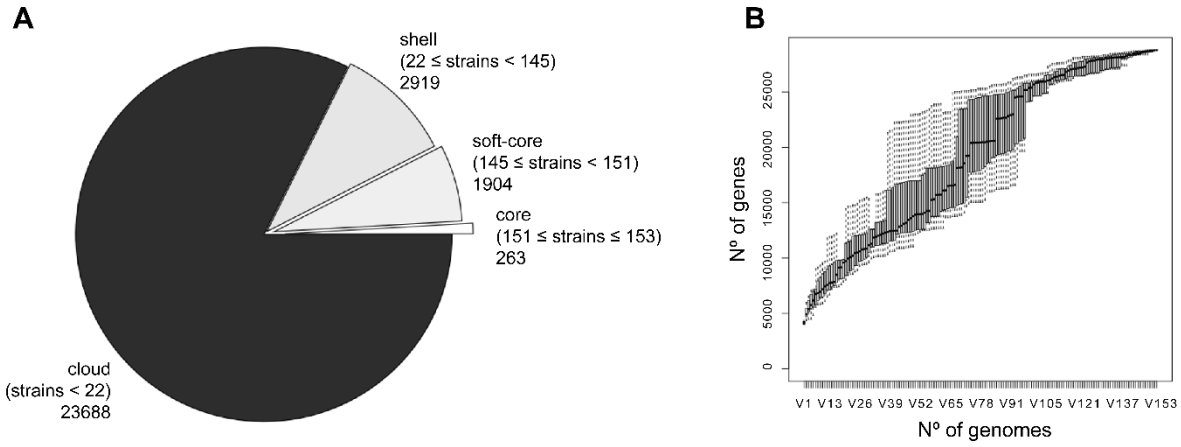
de montagem dos genomas. As sequências de genes *rpoB* não foram recuperadas de 2 dois dos 153 genomas analisados. O genoma de *Bacillus subtilis* subsp. *subtilis* strain D12-5 (CP014858) apresentou problemas de anotação do gene, enquanto em *Bacillus subtilis* subsp. *subtilis* strain BS155 (CP029052) se apresentou fora do quadro de leitura (erro de *frameshift*). Essas falhas podem ocorrer por problemas na anotação e/ou em decorrência de baixa qualidade da montagem do genoma.

Em resumo, a partir da análise de Inferência Bayesiana foi possível confirmar que a estirpe RI4914 de *B. subtilis* está bastante próxima da referência *Bacillus subtilis* subsp. *subtilis* str. 168 (NCBI: AL009126.3), isolada por Rasmussen *et al.* (2009).

3.3 Análise pangenômica

A análise pangenômica englobou todos os genomas completos, sem *gaps* (1 *contig*), classificados como *B. subtilis*, disponíveis no NCBI em 12 de maio de 2020. Nos 153 genomas avaliados foi recuperado um total de 28.774 genes, incluindo 263 genes *core* (genes compartilhados por pelo menos 151 genomas). O genoma acessório ficou composto por 28.508 genes, dos quais 1.904 foram classificados como *soft-core*, 2.919 como *shell* e 23.688 genes *cloud* (Figura 2.3A). O pangenoma de *B. subtilis* foi classificado como aberto, já que o número de genes únicos contidos no pangenoma continuou a crescer, à medida que um novo genoma era adicionado a análise (Figura 2.3B).

Figura 2.3 – (A) Pangenoma de *Bacillus subtilis* exibindo o número de *core*, *soft-core*, *shell* e *cloud* genes. (B) Pangenoma aberto de *Bacillus subtilis*. O número de genes no pangenoma aumenta com o número de novos genomas sequenciados.



Fonte: Da autora

3.4 Análise de Multilocus Sequence Typing (MLST)

Os 263 genes *core* concatenados e alinhados com MAFTT foram utilizados para análise *multilocus* e ilustrados conforme a Figura 2.4. A cepa de *B. subtilis* RI4914 teve melhor agrupamento com as cepas *B. subtilis* JAAA, isolada de água estuarina (SHENG-JIE *et al.*, 2020), *B. subtilis* SRCM103581, isolado de alimento fermentado de soja (KIM, 2017) e *B. subtilis* PJ-7, isolado de Cheonggukjang, um alimento fermentado coreano (HEO *et al.*, 2019). Para análise de agrupamento, a utilização dos genes *core* se mostrou mais eficiente do que as análises utilizando *housekeepings*, como o gene *rpoB*, quanto à capacidade discriminatória.

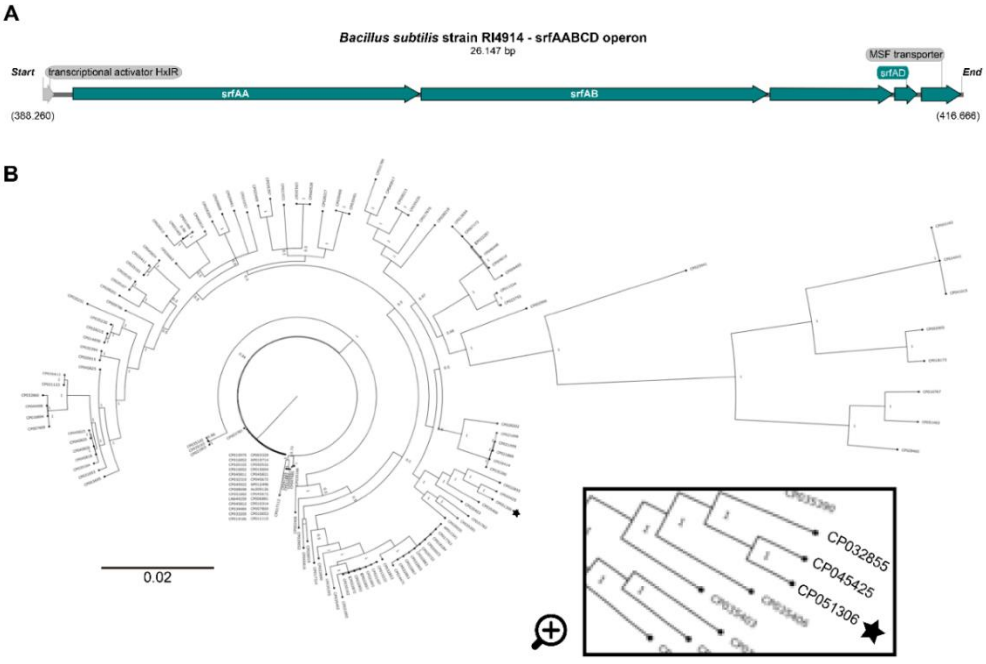
Figura 2.4 – Árvore da análise de tipagem molecular MLST para os genes *core* obtidos. A cepa do estudo, *Bacillus subtilis* RI4914, está em destaque e marcada com uma estrela.

Fonte: Da autora

3.5 Avaliação genômica da prevalência do *operon* codificador da produção de surfactina em *Bacillus subtilis*

O *operon srf*A completo – *srf*AA *(Surfactina sintetase subunidade* 1), *srf*AB (*Surfactina sintetase subunidade* 2), *srf*AC (*Surfactina sintetase subunidade* 3), and *srf*AD (*Surfactina sintetase subunidade* thioesterase) (Figura 2.5A), que codifica para a produção de surfactina, um lipopeptídeo sintetizado por sintetases de peptídeos não ribossomais (NRPS), foi recuperado de 146 dos 153 genomas avaliados. Isso demonstra que essa função é comum à essa espécie de *Bacillus.* A hipótese de Inferência Bayesiana para esse *operon* é fornecida na Figura 2.5B. O *operon srfA* da estirpe de *B. subtilis* RI4914 foi mais proximamente relacionado ao de *B. subtilis* JAAA (CPO45425.1), isolado de água estuarina (SHENG-JIE *et al.*, 2020) e ao de *Bacillus subtilis* subsp. *subtilis* strain PJ-7 (CP032855.1), isolado de Cheonggukjang, um alimento fermentado coreano (HEO *et al.*, 2019). Observa-se que as sequências do *operon srfA* são conservadas entre os genomas analisados, porém mais estudos são necessários para verificar se as variações observadas são reais ou se são artefatos de montagem.

Figura 2.5 – (A) Representação do *operon* da Surfactina na cepa RI4924 de *Bacillus subtilis*. (B) Árvore obtida por análise Bayesiana a partir dos genes *core* concatenados. Os valores em cada ramo representam a probabilidade obtida por análise BA. O *B. subtilis* RI4914 está marcado com uma estrela.

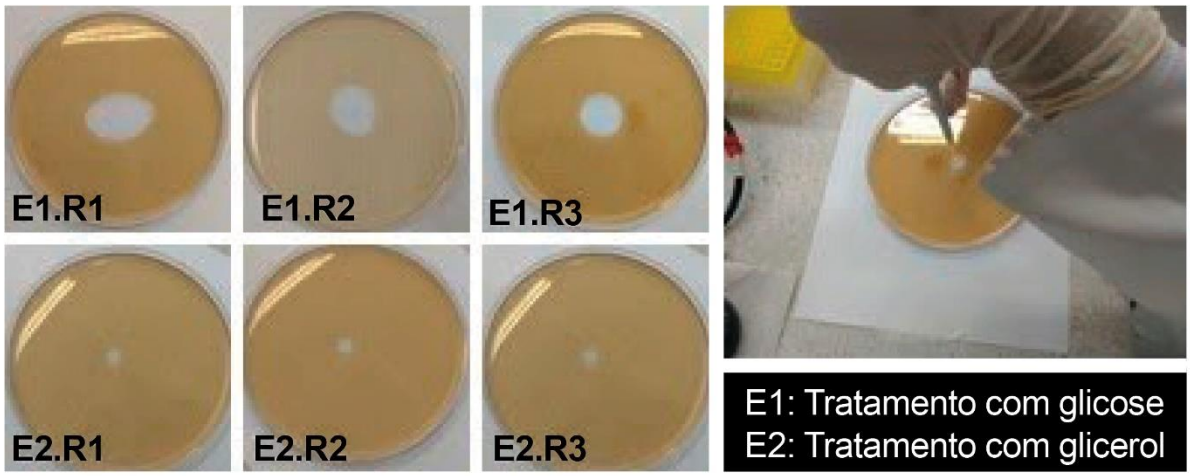


Fonte: Da autora

3.6 Produção de surfactina por *Bacillus subtilis* RI4914 e análise da expressão diferencial do gene srfAA, primeiro gene do *operon* de síntese da surfactina em *Bacillus subtilis*

No ensaio de espalhamento de gota, o tratamento contendo glicose como fonte de carbono apresentou os maiores halos (Figura 2.6). As repetições apresentaram resultado homogêneo, sugerindo que nesse tratamento houve maior produção de biossurfactante, quando comparado ao tratamento contendo glicerol. É importante ressaltar que esses testes foram realizados com os sobrenadantes brutos, isto é, sem passar por processo de purificação e, possivelmente, o uso de sobrenadantes purificados devam gerar halos maiores (Tabela 2.2).

Figura 2.6 – Teste de espalhamento de gota para cada tratamento (E1: Glicose/ E2: Glicerol) e suas respectivas repetições (R1, R2 e R3).



Fonte: Da autora

Tabela 2.2 – Medidas dos halos formados pelo espalhamento de gota em cada repetição de ambos os tratamentos.

	Glicose (cm)	Glicerol (cm)
Repetição 1	3,9 x 3,9	0,9 x 1,0
Repetição 2	3,9 x 4	1,1 x 0,9
Repetição 3	3,3 x 3,8	0,9 x 0,9

Fonte: Da autora.

Para avaliar se o aumento da produção de biossurfactantes no tratamento contendo glicose se deu em decorrência do aumento de produção de surfactina, avaliou-se a expressão diferencial do primeiro gene do *operon srf*AABCD (*srf*AA), por meio de RT-

qPCR. Dois pares de oligonucleotídeos iniciadores com algo no gene *srf*AA foram desenhados e sintetizados para este estudo, além de um par para o gene rDNA 16S, que foi utilizado como controle endógeno (Tabela 1.). Porém, nos testes *in vitro*, apenas um dos pares apresentou melhor qualidade para a quantificação. A porcentagem de eficiência foi similar e satisfatória para ambos os oligonucleotídeos iniciadores (iguais ou superiores a 100%), entretanto, um deles apresentou valores de R^2 muito baixos, sem separação dos pontos de curva e as amplificações foram tardias, no mesmo Ct que os NTCs (controle negativo). Isso indica que os oligonucleotídeos não estavam anelando no gene alvo, mas sim com ele mesmo, gerando dímeros (Tabela 1.).

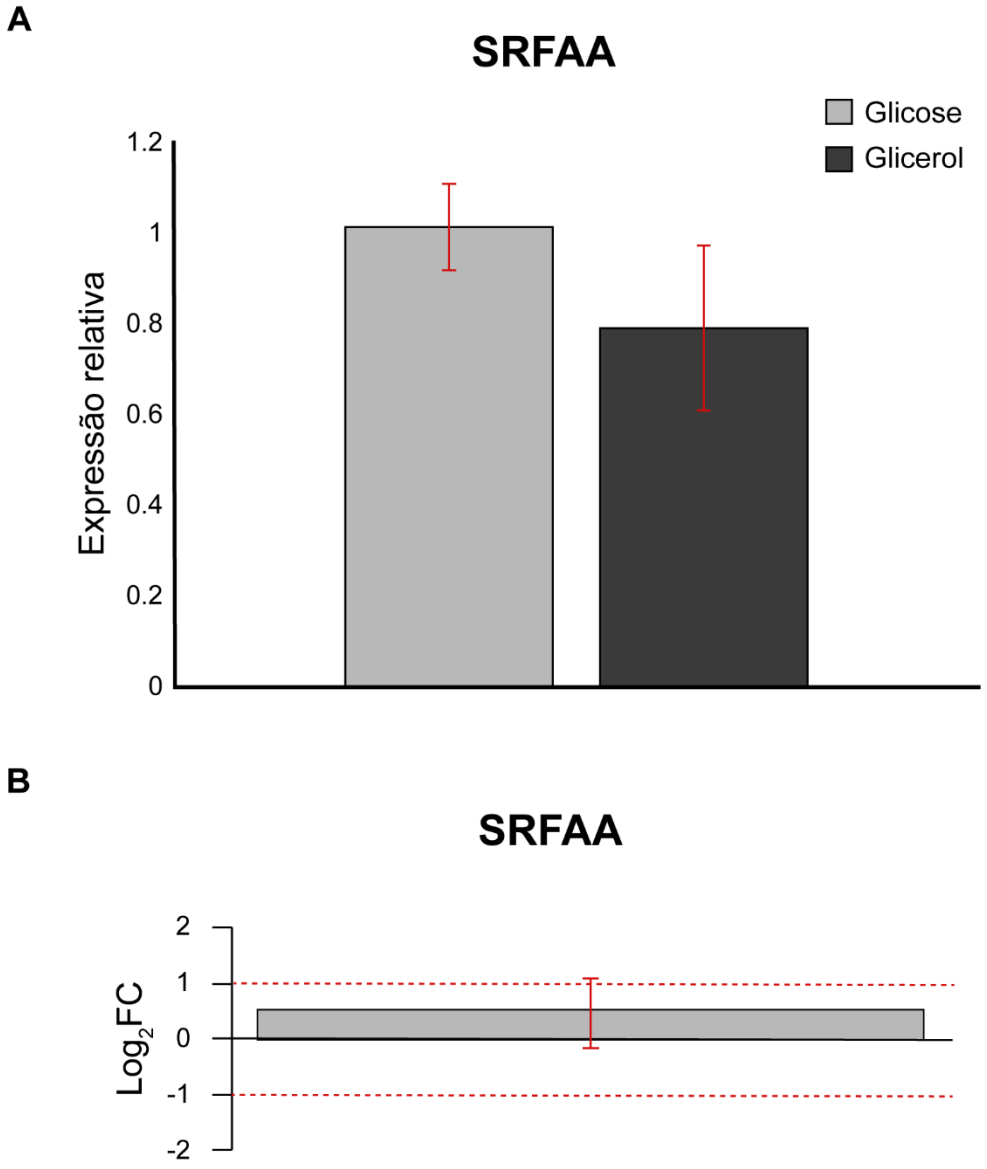
Tabela 1.2 – Resultados do teste de eficiência de amplificação dos primer de referência (16S) e os primer de srfAA testados. O utilizado para execução da qPCR foi o 2srfAA.

Gene	Concentração	R^2	Eficiência
rRNA 16S **F: CCTACGGGAGGCAGCAG** **R: ATTACCGCGGCTGCTGG**	**1 nmol**	**0,99408**	**86%**
1*srf*AA F: AGGCGGGGATCTTTGACA R: TGAAGCGGAATCTCAATGC	2 nmol	0,02015	100%
2*srfA*A **F: ATGGCTTCATTCGTTCGGA** **R: GACGGTTCTCTTCAAGCC**	**2 nmol**	**0,98867**	**112%**

Nota: os pares de oligonucleotideos destacados em **negrito** foram utilizados nas análises.
Fonte: Da autora.

A partir do resultado obtido pela RT-qPCR, observou-se aumento absoluto na expressão do gene srfAA, no tratamento contendo glicose, quando comparado ao tratamento contendo glicerol (Figura 2.7A). Porém, apesar da preferência por glicose para a produção deste biossurfactante ser maior, tornando a glicose uma boa candidata para ensaios de otimização da produção, a diferença de expressão relativa não foi significante a 95% de confiança, para um *p*-valor de 0,0854, de acordo com o modelo linear estatístico misto proposto por Steibel *et al.* (2009) (Figura 2.7B).

Figura 2.7 – Expressão relativa do gene srfAA da produção de surfactina por *Bacillus subtilis* RI4914, quando cultivado em glicose e em glicerol (A). Log_2 de *Fold Change* entre os tratamentos glicose e glicerol (B).



Fonte: Da autora

4 DISCUSSÃO

4.1 Montagem híbrida do genoma do *Bacillus subtilis* RI4914

O desenvolvimento de plataformas de sequenciamento massivo de DNA, conhecidas como *Next Generation Sequencing* (NGS), permitiu o avanço científico e tecnológico (GOODWIN *et al.*, 2016), com grande implicação na área de sequenciamento de genomas completos (*Whole Genome Sequencing* – WGS). Atualmente, duas abordagens

de NGS têm sido utilizadas de forma complementar, permitindo a obtenção de genomas completos e de alta qualidade (baixa taxa de erros) (VASUDEVAN *et al.*, 2020). Neste trabalho, utilizou-se as plataformas GridION^{TM} (Oxford Nanopore) para obtenção de sequências longas e o PGM Ion Torrent (ThermoFisher) para a obtenção de sequências curtas. Após a montagem, com as sequências do GridION^{TM}, foi possível obter um único cromossomo contíguo, porém o mesmo estava sujeito a uma alta taxa de erros, o que foi confirmado pelo inesperado alto número de CDSs em decorrência de erros do tipo *frameshift*. Isso dificulta a acurácia da análise, bem como a identificação de genes marcadores ou SNPs (*single-nucleotide polymorphism*), podendo gerar resultados equivocados em análises de genômica comparativa (VASUDEVAN *et al.*, 2020). Sendo assim, a aplicação da abordagem híbrida de montagem permitiu a correção desses erros, por meio do mapeamento de sequências curtas de alta qualidade ao genoma completo.

Em alguns casos, as leituras longas podem apresentar *gaps* pela falta de complementaridade das sequências, em decorrência de erros que a plataforma gera. A hibridização com sequências curtas de alta qualidade torna a montagem mais satisfatória, uma vez que as leituras curtas participam em sobreposição às leituras longas, reduzindo *indels* e *gaps* (WICK *et al.*, 2017). Vasudevan *et al.* (2020) também observaram melhora na qualidade de genomas bacterianos de interesse clínico (maior completude e baixa contaminação), após mapeamento de sequências curtas obtidas de PGM IonTorrent em montagens realizadas com sequências longas obtidas pela plataforma MinION (Oxford Nanopore). Essa estratégia também se mostrou melhor do que a aplicação de diversas rodadas de polimento, apenas usando os dados brutos de sequências longas (arquivos no formato FAST5).

Neste estudo, foram necessárias três rodadas de polimento para obtenção de um genoma circular completo, o que forneceu uma completude de 98,84% e taxa contaminação de 1,10%, gerando um genoma final completo de alta qualidade.

Goldstein *et al.* (2019) também utilizaram sequenciamento em diferentes plataformas: Illumina para sequências curtas e MinION para sequências longas, individualmente e de forma híbrida, para determinação da melhor eficiência de montagem. Esses autores também obtiveram melhores resultados quando as sequências de Illumina corrigiram as montagens do MinION e pontuaram a capacidade de sequências longas em desambiguar regiões genômicas repetitivas, porém biologicamente importantes, como *clusters* de genes biossintéticos de metabólitos secundários. De forma similar, Risse *et al.*

(2015) obtiveram resultados similares quando utilizaram abordagem híbrida de montagem, obtendo genomas completos sem *gaps* e com baixa taxa de erros.

Entretanto, Deschamps *et al.* (2016) obtiveram desempenho igualmente eficiente quando as sequências longas de MinION foram corrigidas com leituras curtas ou sem correção (apenas sequências longas). Por conta disso, expressaram que a escolha de montagem híbrida ou não híbrida deve ser ditada baseada no tamanho e na complexidade do modelo a ser sequenciado. Esses autores pontuam, por exemplo, que em genomas grandes e complexos, sequências curtas podem alinhar de forma incorreta na montagem, abrindo margem para que o próprio processo de correção possa criar leituras e mapeamentos em múltiplas regiões do genoma.

4.2 Reconstrução filogenética de *Bacillus subtilis*

Neste trabalho, avaliou-se a filogenia de *B. subtilis*, por meio da comparação das sequências do gene *rpoB* recuperadas de todas os representantes dessa espécie com genoma completo depositado no NCBI. A partir dos resultados, pode-se observar o alto grau de conservação do gene nessa espécie (Figura 2.2). O *B. subtilis* RI4914, foco deste estudo, apresentou 100% de similaridade do gene *rpoB* com a espécie-tipo, *Bacillus subtilis* subsp. *subtilis* str. 168 (NCBI: AL009126.3). Estes resultados corroboram a correta classificação desses genomas como pertencentes a espécie *B. subtilis*. Isso é de extrema importância, pois a inclusão de bactérias, erroneamente classificadas, interfere negativamente em estudos posteriores, como por exemplo, análises de pangenoma (WU; WANG; GAO, 2021).

Análises filogenéticas e estudos taxonômicos bacterianos podem ser realizados utilizando como alvos o gene *rDNA 16S*, que codifica para o rRNA 16S e o gene *rpoB*, que codifica para a subunidade β da RNA polimerase. O gene *rpoB* apresenta alta correlação com dados de hibridização DNA-DNA (DDH) e ANI. Similaridades de sequências desse gene maior ou igual a 97,7%, correlacionam-se significativamente com valores de DDH maiores que 70% e valores de ANI maiores que 94,3% – que indicam mesma espécie (ADÉKAMBI; DRANCOURT; RAOULT, 2009). O uso do gene *rpoB* apresenta algumas vantagens sobre o *rDNA 16S*, a saber: (1) o *rDNA 16S* pode apresentar múltiplas cópias no genoma, sendo que esse número de cópias pode ainda variar de forma intraespecífica, prejudicando a análise de frequências e anotação gênica. Isso não acontece com o gene *rpoB*, pois está presente no genoma em cópia única; (2) divergências genéticas do *rpoB* se

correlacionam melhor com a divergência genômica geral, provendo melhor suporte estatístico para reconstrução de filogenias; (3) o alto nível de conservação do gene *rDNA 16S* pode ocultar variações intra e interespecíficas e; (4) o gene *rpoB* é um gene codificador de proteína e, por conta disso, os dados gerados por esse marcador são mais legíveis e melhor interpretados quando inseridos em uma estrutura de análise filogenética (LOUCA; DOEBELI; PARFREY, 2018; VOS *et al.*, 2012; ZAW; EMRAN; LIN, 2018).

Uma desvantagem do uso do gene *rpoB* para reconstruções filogenéticas se dá pela limitação na disponibilidade de sequências em bancos de dados. O gene rDNA 16S é mais difundido, principalmente em decorrência de alta facilidade de sequenciamento do gene completo, que possui ~ 1.500 bp (3 vezes menor que o rpoB), o que o tornou popular para esse tipo de estudo. Atualmente, em decorrência do avanço nas tecnologias de sequenciamento de DNA, o número de genomas completos depositados em bancos de dados, *e.g.*, NCBI cresceu bastante (https://www.ncbi.nlm.nih.gov/genbank/statistics/), permitindo a recuperação de genes de interesse, a exemplo do *rpoB*.

Ao estudar a diversidade de *Bacillus* em ambiente marinho, KI, ZHANG e QIAN. (2009) identificaram 9 espécies por meio de análises de sequências do gene rDNA 16S. Entretanto, quando as sequências do gene *rpoB* foram sequenciadas, a resolução filogenética foi aproximadamente 4,5 vezes melhor que as previamente obtidas com o uso do *rDNA 16S*, demonstrando que o polimorfismo do gene *rpoB* em *Bacillus* pode ser utilizado para elucidar espécies desse gênero, provendo um esquema aprimorado de identificação de *Bacillus* spp.

Ao se tratar de análises para relações filogenéticas, o uso do gene *rpoB* como marcador é recomendável, em decorrência da sua maior correlação com a identidade média de aminoácidos, o que reflete uma relação em todo o nível do genoma (KONSTANTINIDIS; TIEDJE, 2005). Além disso, Adékambi *et al.* (2008) citam o gene *rpoB* como útil no delineamento de espécies e, partindo desse critério, os mesmos conseguiram confirmar a proximidade filogenética entre *Borrelia recurrentis*, bactéria transmitida por piolho e que é causadora de febre recorrente em todo o mundo, e *Borrelia duttonii*, também causadora de febre, porém transmitida por carrapatos na África Oriental. Recentemente, Grazziotin *et al.* (2021) utilizaram o gene *rpoB* como marcador para estudar o gênero *Erysipelothrix* e confirmaram que esse gene foi o mais indicado para delineamento filogenético dentro do gênero e elucidaram um possível novo táxon ao observarem que suas cepas alvo formavam um grupo monofilético altamente suportado.

4.3 Análise pangenômica de *Bacillus subtilis*

O avanço teórico e tecnológico no sequenciamento de genomas tem expandido as informações disponíveis acerca da extensa variação do conteúdo gênico de determinadas espécies, principalmente em procariotos. As análises pangenômicas englobam ferramentas para o estudo da diversidade gênica, utilizando conjuntos completos de genes centrais (chamados de *core*), acessórios e únicos (INGLIN; MEILE; STEVENS, 2018). Os genes *core* são aqueles presentes em todos os representantes de uma determinada espécie, os acessórios são aqueles que variam dentro da própria espécie, e os únicos são aqueles específicos de cada representante (DOMINGO-SANANES; MCINERNEY, 2021).

Desses conjuntos completos de genes, usa-se para fins didáticos a divisão em subgrupos: *softcore, shell* e *cloud*, como forma de organizar a frequência dos genes *core* e acessórios. O *soft-core* diz respeito àqueles genes que estão presentes em 95–99% dos genomas isolados para estudo, *shell* são aqueles que estão presentes em 15–95% dos genomas e os genes *cloud* são os que fazem parte de 0–15% dos genomas isolados. Dentre estes, o *soft-core* mais os genes *core* (presentes em pelo menos 99% de todos os genomas estudados) formam o que são conhecidos como o genoma *core* e aqueles organizados em *shell* e *cloud* formam juntos o genoma acessório (BEZUIDT *et al.*, 2016). Esta organização pode ser verificada na Figura 2.3A.

Bacillus subtilis é uma espécie de bactéria Gram-positiva, de ampla distribuição e comumente descrita no solo e na água (HARWOOD *et al.*, 2018). O pangenoma de *B. subtilis* foi classificado como "aberto", visto que o número de genes totais continua a aumentar à medida que novos genomas são incluídos na análise (Figura 2.3B). Esses resultados caracterizam uma alta plasticidade gênica e boa capacidade para incorporação de novas informações no genoma por representantes de *B. subtilis*, que podem ser provenientes de transferência horizontal intra e interespecífica fornecendo à espécie maior variabilidade genética, mesmo sujeita a processos de deriva e seleção natural (ROULI *et al.*, 2015).

Essa capacidade é uma característica marcante em espécies de procariotos com potencial de colonização em diferentes espaços, como no caso de *B. subtilis*, aumentando a adaptação a esses ambientes (WU; WANG; GAO, 2021). Pode-se considerar que tais condições estão relacionadas com a funcionalidade, principalmente dos genes acessórios que são adicionados, os quais podem ser neutros ou benéficos, mas que por processos de recombinação e mutações no material genético, na maioria das vezes, conseguem fornecer

características novas, como resistência a antibióticos, mecanismos de sobrevivência em situações de estresse ou até novos aspectos de virulência (COLEMAN *et al.*, 2009).

Choi *et al.* (2020) encontraram perfil pangenômico similar em seu estudo com estirpes de *B. subtilis* isoladas de alimentos fermentados, comercializados em países do leste asiático. Esses autores usaram 61 genomas provenientes de cepas da mesma espécie, mas diferentes países (29 cepas isoladas da China, 3 do Japão e 24 da Coréia do Sul) e mais 5 isoladas especificamente para o estudo. Eles obtiveram como resultado 2.098 genes centrais entre as 61 estirpes e 3.275 entre as cinco geradas no estudo, além de 6.061 genes únicos em um pangenoma aberto. Desta forma, o resultado de seu estudo comprovou e demonstrou a diferença genética presente dentro da espécie e, consequentemente, a facilidade de permanência da mesma, por mais que fossem provenientes de diferentes regiões.

Alguns fatores podem interferir nas análises pangenômicas (WU; WANG; GAO, 2021), como a presença de múltiplos clones de uma estirpe na análise, pois esta estirpe contará com o seu arcabouço de genes específicos que provavelmente será compartilhado entre duas ou mais cepas, o que ocasionará na identificação equivocada de "genes acessórios", diminuindo a acurácia da análise. Outro fator é a presença de estirpes com alta proporção de pseudogenes (muitas vezes como um artefato de erros da montagem do genoma), estes pseudogenes podem passar despercebidos e serem considerados genes únicos. Porém, os programas utilizados para análise pangenômica podem verificar facilmente a presença desses erros e eliminá-los da análise, bem como podem identificar e excluir os representantes clonais (WU; WANG; GAO, 2021).

4.4 Avaliação de prevalência do *operon* codificador da surfactina em *Bacillus subtilis*

O *operon srfA* possui ~27 kb e codifica os quatro genes da via de produção da surfactina em bactérias – *srfA*ABCD (Figura 2.5A). Além disso, esse *operon* é requerido para outras funções celulares em *B. subtilis*, como no desenvolvimento de competência e no processo de esporulação e, sendo assim, nem sempre a sua ativação será voltada para a síntese do biossurfactante (NAKANO *et al.*, 1991). A biossíntese de surfactina é dependente de alguns fatores, como a dependência de densidade celular (*quorum sensing*), que garante a ativação prévia do sistema de transdução de sinal ComP-ComA (NAKANO *et al.*, 1991). São esses genes que precisarão ser expressos para que o *operon srf*A seja

ativado e realmente produza o metabólito, garantindo que a biossíntese não ocorra de forma constante (COUTTE *et al.*, 2010; SUN *et al.*, 2009).

A análise de genômica comparativa permitiu avaliar a prevalência desse *operon* em representantes de *B. subtilis*. Em sete dos 153 genomas analisados o *operon srfA* não foi automaticamente anotado pelo Prokka. Porém, por meio de busca manual baseada em BLAST, foi possível encontrar o *operon* nesses genomas (dados não mostrados), indicando possíveis erros de montagem ou falhas na anotação automática. É importante ressaltar que erros de montagens podem ocorrer por inúmeras razões, como por exemplo, partes que são descartadas incorretamente como erros ou repetições, outras que são inseridas em locais ou orientações erradas e, além disso, sequências repetitivas, polimorfismos, dados ausentes e erros acabam por limitar o comprimento dos *contigs* que podem ser construídos, comprometendo a análise (BAKER, 2012).

Por fim, esses resultados sugerem que a produção de surfactina é um caráter comum a *B. subtilis* e, provavelmente, estava presente em seu ancestral.

A análise de reconstrução filogenética do *operon* indica a alta conservação dos genes, o que é esperado para genes funcionais (Figura 2.5B). O *operon* de *B. subtilis* RI4914 apresenta maior proximidade aos *operons* de *B. subtilis* JAAA (CPO45425.1), seguido daquele obtido a partir do genoma do *Bacillus subtilis* subsp. *subtilis* strain PJ-7 (CP032855.1). Assim como o RI4914, o JAAA foi isolado de sistemas aquáticos, sendo isolado de água estuarina (SHENG-JIE *et al.*, 2020) e o segundo, PJ-7 foi isolado de um alimento fermentado coreano, o Cheonggukjang.

Evidências sobre prevalência gênica em bactérias pode fornecer novas perspectivas acerca da regulação metabólica na espécie, como observado por Esmaeili *et al.* (2019), que reportaram alta prevalência de genes associados a Metallo-β-lactamase em *Pseudomonas aeruginosa*, fator que colabora para a ampliação da resistência a antibióticos. De formar similar, Hsu *et al.* (2019), ao avaliarem a prevalência gênica em *Campylobacter showae,* notaram que os genes de virulência não são conservados e que sua prevalência é baseada na região de infecção, sendo que aqueles presentes em adenomas do cólon não são os mesmos presentes em inflamações do tecido intestinal humano.

4.5 Produção da surfactina por e análise da expressão diferencial do gene srfAA em *B. subtilis* RI4914

A capacidade de produção de surfactina em culturas de *B. subtilis* RI4914 suplementadas com glicose e com glicerol foi avaliada por meio do método de espalhamento de gota em óleo (YOUSSEF *et al.*, 2004) (Figura 2.6). Como esperado, os sobrenadantes das culturas em meio contendo glicose apresentaram os maiores halos de espalhamento, sugerindo que essa fonte de carbono foi capaz de proporcionar aumento na produção de surfactina. Veshareh *et al.* (2019) obtiveram resultados similares, observando que o maior diâmetro de halo formado foi proveniente do sobrenadante de *B. subtilis* MJ01, também cultivado em glicose como fonte de carbono. Da forma similar, Sohail e Jamil (2020) registraram comportamento similar de *B. tequilensis* MH142145 em 24 e 48 h de cultivo em glicose e glicerol.

Além disso, este trabalho avaliou a expressão diferencial do gene *srfA*A, responsável pela síntese da subunidade 1 da surfactina, em ambos os tratamentos de cultivo. Nenhuma diferença significativa foi observada na expressão desse gene (p = 0,0854), mas houve uma forte tendência de aumento de expressão no tratamento com a glicose (Figura 2.7), corroborando os resultados obtidos no teste do espalhamento de gota. Possivelmente, o tempo de cultivo foi insuficiente para alterar a expressão de forma significativa, mas estudos posteriores são necessários para confirmar.

Diferentes pares de oligonucleotídeos iniciadores para avaliação da expressão de genes da surfactina foram testados no decorrer deste trabalho, porém, nenhum deles apresentou capacidade de amplificação e/ou eficiência satisfatórios (dados não mostrados). Os avanços nas tecnologias de sequenciamento, associados aos métodos aplicados ao estudo de genômica comparativa, permitiu o desenho de oligonucleotídeos iniciadores adequados (condições ideais pra RT-qPCR) para estudo da expressão do gene *srfA*A em *B. subtilis*. O uso de diferentes condições de cultivos, com alta e baixa taxa de produção de surfactina foi importante para validar o método.

Diversos trabalhos têm objetivado o aumento de produção de surfactina por *B. subtilis*. O estudo realizado por Abdel-Mawgoud, Aboulwafa e Hassouna (2008), identificou o melaço como ótima fonte de carbono alternativa e $NaNO_3$ como fonte de nitrogênio para otimizar a produção de surfactina por *B. subtilis*. Já Cheng *et al.* (2018) conseguiram otimizar a síntese de surfactina enriquecendo o meio de cultura com uma mistura de pó de glicose, açúcar mascavo, melaço, farinha de soja, fermento e farinha de peixe em processos fermentativos. Por fim, Ghribi e Chaabouni (2011) obtiveram maior concentração de biossurfactantes quando o meio foi suplementado com 5 $g.L^{-1}$ de ureia como fonte de nitrogênio e quando houve suplementação de 2% de querosene no meio.

Esses estudos são importantes, pois podem demonstrar a base para otimizações na produção do biossurfactante a partir de métodos convencionais de cultivo, porém, as bases moleculares modulando essas alterações são ainda pouco conhecidas.

5 CONCLUSÃO

O estudo de genômica comparativa de *B. subtilis* indicou a alta prevalência do *operon* da surfactina (*srfA*ABCD) nas estirpes da espécie. Isso fornece suporte para a utilização de *B. subtilis* como microrganismo modelo em estudos genéticos sobre a surfactina.

A utilização de abordagem híbrida de sequenciamento e montagem foi essencial para a obtenção do genoma completo e acurado de *B. subtilis* RI4914, sendo um método que deve ser explorado com mais frequência, a fim de garantir a qualidade de estudos de genômica comparativa.

A análise de pangenoma das cepas de *B. subtilis* demonstrou um perfil pangenômico aberto, indicando alta plasticidade genética da espécie, bem como caráter não exigente para assimilação de novos genes em seu material genético.

A reconstrução filogenética do grupo, com base no gene *rpoB* recuperado dos genomas completos assinados como *B. subtilis*, demonstrou a confiabilidade do banco de dados utilizado para essa espécie.

REFERÊNCIAS

ABDEL-MAWGOUD, A. Mohammad; ABOULWAFA, M. Mabrouk; HASSOUNA, Nadia Abdel Haleem. Optimization of surfactin production by *Bacillus subtilis* isolate BS5. **Applied Biochemistry and Biotechnology**, v. 150, n. 3, p. 305-325, 2008. DOI: https://doi.org/10.1007/s12010-008-8155-x.

ADÉKAMBI, Toïdi *et al*. Complete rpoB gene sequencing as a suitable supplement to DNA-DNA hybridization for bacterial species and genus delineation. **International Journal of Systematic and Evolutionary Microbiology**, v. 58, n. 8, p. 1807-1814, 2008. DOI: https://doi.org/10.1099/ijs.0.65440-0.

ADÉKAMBI, Toïdi; DRANCOURT, Michel; RAOULT, Didier. The rpoB gene as a tool for clinical microbiologists. **Trends in Microbiology**, v. 17, n. 1, p. 37-45, 2009. DOI: https://doi.org/10.1016/j.tim.2008.09.008.

BAKER, Monya. De novo genome assembly: what every biologist should know. **Nature Methods**, v. 9, n. 4, p. 333-337, 2012. DOI: https://doi.org/10.1038/nmeth.1935.

BEZUIDT, Oliver K. *et al.* The Geobacillus pan-genome: implications for the evolution of the genus. **Frontiers in Microbiology**, v. 7, p. 723, 2016. DOI: https://doi.org/10.3389/fmicb.2016.00723.

CHENG, Yeong Hsiang *et al.* Optimization of surfactin production from *Bacillus subtilis* in fermentation and its effects on *Clostridium perfringens*-induced necrotic enteritis and growth performance in broilers. **Journal of Animal Physiology and Animal Nutrition**, v. 102, n. 5, p. 1232-1244, 2018. DOI: https://doi.org/10.1111/jpn.12937.

CHOI, Hye Jin *et al.* Comparative genomic and functional evaluations of *Bacillus subtilis* newly isolated from Korean traditional fermented foods. **Foods**, v. 9, n. 12, p. 1805, 2020. DOI: https://doi.org/10.3390/foods9121805.

COLEMAN, Jeffrey J. *et al.* The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. **PLoS Genetics**, v. 5, n. 8, p. e1000618, 2009. DOI: https://doi.org/10.1371/journal.pgen.1000618.

COUTTE, François *et al.* Effect of pps disruption and constitutive expression of srfA on surfactin productivity, spreading and antagonistic properties of *Bacillus subtilis* 168 derivatives. **Journal of Applied Microbiology**, v. 109, n. 2, p. 480-491, 2010. DOI: https://doi.org/10.1111/j.1365-2672.2010.04683.x.

DESCHAMPS, Stéphane *et al.* Characterization, correction and de novo assembly of an Oxford Nanopore genomic dataset from *Agrobacterium tumefaciens*. **Scientific Reports**, v. 6, n. 1, p. 1-11, 2016. DOI: https://doi.org/10.1038/srep28625.

DOMINGO-SANANES, Maria Rosa; MCINERNEY, James O. Mechanisms that shape microbial pangenomes. **Trends in Microbiology**, v. 29, n. 6, p. 493-503, 2021. DOI: https://doi.org/10.1016/j.tim.2020.12.004.

EDGAR, Robert C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Research**, v. 32, n. 5, p. 1792-1797, 2004. DOI: https://doi.org/10.1093/nar/gkh340.

ESMAEILI, Davoud *et al.* Alerting prevalence of MBLs producing *Pseudomonas aeruginosa* isolates. **Gene Reports**, v. 16, p. 100460, 2019. DOI: https://doi.org/10.1016/j.genrep.2019.100460.

FERNANDES, P. L. *et al*. Biosurfactant, solvents and polymer production by Bacillus subtilis RI4914 and their application for enhanced oil recovery. **Fuel**, v. 180, p. 551-557, 2016. DOI: https://doi.org/10.1016/j.fuel.2016.04.080.

GHRIBI, Dhouha; ELLOUZE-CHAABOUNI, Semia. Enhancement of *Bacillus subtilis* lipopeptide biosurfactants production through optimization of medium composition and adequate control of aeration. **Biotechnology Research International**, v. 2011, p. 1-6, 2011. DOI: https://doi.org/10.4061/2011/653654.

GOLDSTEIN, Sarah *et al*. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. **BMC genomics**, v. 20, n. 1, p. 1-17, 2019. DOI: https://doi.org/10.1186/s12864-018-5381-7.

GOODWIN, Sara; MCPHERSON, John D.; MCCOMBIE, W. Richard. Coming of age: ten years of next-generation sequencing technologies. **Nature Reviews Genetics**, v. 17, n. 6, p. 333–351, 2016. DOI: https://doi.org/10.1038/nrg.2016.49 .

GORIS, Johan *et al.* DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. **International Journal of Systematic and Evolutionary Microbiology**, v. 57, n. 1, p. 81-91, 2007. DOI: https://doi.org/10.1099/ijs.0.64483-0.

GRAZZIOTIN, Ana Laura et al. Comparative genomics of a novel clade shed light on the evolution of the genus *Erysipelothrix* and characterise an emerging species. **Scientific Reports**, v. 11, n. 1, p. 1-12, 2021. DOI: https://doi.org/10.1038/s41598-021-82959-x.

GULILAT, Markus *et al*. Targeted next generation sequencing as a tool for precision medicine. **BMC Medical Genomics**, v. 12, n. 1, p. 1-17, 2019. DOI: https://doi.org/10.1186/s12920-019-0527-2.

GUREVICH, Alexey *et al.* QUAST: Quality assessment tool for genome assemblies. **Bioinformatics**, v. 29, n. 8, p. 1072-1075, 2013. DOI: https://doi.org/10.1093/bioinformatics/btt086.

HARWOOD, Colin R. *et al*. Secondary metabolite production and the safety of industrially important members of the *Bacillus subtilis* group. **FEMS Microbiology Reviews**, v. 42, n. 6, p. 721-738, 2018. DOI: https://doi.org/10.1093/femsre/fuy028.

HEO, Jun *et al.* Genetic marker gene, recQ, differentiating *Bacillus subtilis* and the closely related *Bacillus* species. **FEMS Microbiology Letters**, v. 366, n. 16, p. fnz172, 2019. DOI: https://doi.org/10.1093/femsle/fnz172.

HSU, Tiffany *et al*. Comparative genomics and genome biology of *Campylobacter showae*. **Emerging Microbes & Infections**, v. 8, n. 1, p. 827-840, 2019. DOI: https://doi.org/10.1080/22221751.2019.1622455.

HUELSENBECK, John P.; RONQUIST, Fredrik. MRBAYES: Bayesian inference of phylogenetic trees. **Bioinformatics**, v. 17, n. 8, p. 754-755, 2001. DOI: https://doi.org/10.1093/bioinformatics/17.8.754.

INGLIN, Raffael C.; MEILE, Leo; STEVENS, Marc J.A. Clustering of pan- and core-genome of *Lactobacillus* provides novel evolutionary insights for differentiation. **BMC Genomics**, v. 19, n. 1, p. 1-15, 2018. DOI: https://doi.org/10.1186/s12864-018-4601-5.

KATOH, Kazutaka; STANDLEY, Daron M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. **Molecular Biology and Evolution**, v. 30, n. 4, p. 772-780, 2013. DOI: https://doi.org/10.1093/molbev/mst010.

KI, Jang-Seu; ZHANG, Wen; QIAN, Pei-Yuan. Discovery of marine *Bacillus* species by 16S rRNA and rpoB comparisons and their usefulness for species identification. **Journal of Microbiological Methods**, v. 77, n. 1, p. 48-57, 2009. DOI: https://doi.org/10.1016/j.mimet.2009.01.003.

KIM, Hyerim. **Analyses of microbial communities and metabolites in Korean fermented soybean foods, meju and doenjang, and *Bacillus subtilis* pan-genome**. 2017. Tese (Doutorado) – Department of Agricultural Biotechnology, College of Agriculture and Life Science, Seoul National University, Seoul, 2017. Disponível em: https://hdl.handle.net/10371/119550. Acesso em: 8 jun. 2021.

KONSTANTINIDIS, Konstantinos T; TIEDJE, James M. Towards a genome-based taxonomy for prokaryotes. **Journal of Bacteriology**, v. 187, n. 18, p. 6258-6264, 2005. DOI: https://doi.org/10.1128/JB.187.18.6258-6264.2005.

KOREN, Sergey *et al.* Canu: scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. **Genome Research,** v. 27, n. 5, p. 722-736, 2017. DOI: http://www.genome.org/cgi/doi/10.1101/gr.215087.116.

LARKIN, M.A. *et al.* Clustal W and Clustal X version 2.0. **Bioinformatics**, v. 23, n. 21, p. 2947-2948, 2007. DOI: https://doi.org/10.1093/bioinformatics/btm404.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics**, v. 25, n. 14, p. 1754-1760, 2009. DOI: https://doi.org/10.1093/bioinformatics/btp324.

LOUCA, Stilianos; DOEBELI, Michael; PARFREY, Laura Wegener. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. **Microbiome**, v. 6, n. 1, p. 1-12, 2018. DOI: https://doi.org/10.1186/s40168-018-0420-9.

MORIKAWA, Masaaki; HIRATA, Yoshihiko; IMANAKA, Tadayuki. A study on the structure–function relationship of lipopeptide biosurfactants. **Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids**, v. 1488, n. 3, p. 211-218, 2000. DOI: https://doi.org/10.1016/S1388-1981(00)00124-4.

NAKANO, M. M. *et al.* srfA is an operon required for surfactin production, competence development, and efficient sporulation in *Bacillus subtilis*. **Journal of Bacteriology**, v. 173, n. 5, p. 1770-1778, 1991. DOI: https://doi.org/10.1128/jb.173.5.1770-1778.1991.

PAGE, Andrew J. *et al.* Roary: Rapid large-scale prokaryote pan genome analysis. **Bioinformatics**, v. 31, n. 22, p. 3691-3693, 2015. DOI: https://doi.org/10.1093/bioinformatics/btv421.

PARKS, Donovan H. *et al.* CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. **Genome Research**, v. 25, n. 7, p. 1043-1055, 2015. DOI: http://www.genome.org/cgi/doi/10.1101/gr.186072.114.

PFAFFL, Michael W. A new mathematical model for relative quantification in real-time RT–PCR. **Nucleic Acids Research**, v. 29, n. 9, p. e45-e45, 2001. DOI: https://doi.org/10.1093/nar/29.9.e45.

RASMUSSEN, Simon; NIELSEN, Henrik Bjørn; JARMER, Hanne. The transcriptionally active regions in the genome of *Bacillus subtilis*. **Molecular Microbiology**, v. 73, n. 6, p. 1043-1057, 2009. DOI: https://doi.org/10.1111/j.1365-2958.2009.06830.x.

RICHTER, Michael *et al.* JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. **Bioinformatics**, v. 32, n. 6, p. 929-931, 2016. DOI: https://doi.org/10.1093/bioinformatics/btv681.

RICHTER, Michael; ROSSELLÓ-MÓRA, Ramon. Shifting the genomic gold standard for the prokaryotic species definition. **Proceedings of the National Academy of Sciences of the United States of America**, v. 106, n. 45, p. 19126-19131, 2009. DOI: https://doi.org/10.1073/pnas.0906412106.

RISSE, Judith *et al.* A single chromosome assembly of *Bacteroides fragilis* strain BE1 from Illumina and MinION nanopore sequencing data. **GigaScience,** v. 4, n. 1, p. s13742-015-0101-6, 2015. DOI: https://doi.org/10.1186/s13742-015-0101-6.

RODRIGUEZ-R, Luis M.; KONSTANTINIDIS, Konstantinos T. Nonpareil: A redundancy-based approach to assess the level of coverage in metagenomic datasets. **Bioinformatics,** v. 30, n. 5, p. 629-635, 2014. DOI: https://doi.org/10.1093/bioinformatics/btt584.

ROULI, L. *et al.* The bacterial pangenome as a new tool for analysing pathogenic bacteria. **New Microbes and New Infections**, v. 7, p. 72–85, 2015. DOI: https://doi.org/10.1016/j.nmni.2015.06.005.

SHENG-JIE, Zhang *et al*. The complete genome sequence of the algicidal bacterium *Bacillus subtilis* strain JA and the use of quorum sensing to evaluate its antialgal ability. **Biotechnology Reports**, v. 25, p. e00421, 2020. DOI: https://doi.org/10.1016/j.btre.2020.e00421.

SINGH, Aditya; BHATIA, Prateek. Comparative sequencing data analysis of Ion Torrent and MinION sequencing platforms using a clinical diagnostic haematology panel. **International Journal of Laboratory Hematology**, v. 42, n. 6, p. 833-841, 2020. DOI: https://doi.org/10.1111/ijlh.13286.

SOHAIL, Rafeya; JAMIL, Nazia. Isolation of biosurfactant producing bacteria from Potwar oil fields: Effect of non-fossil fuel based carbon sources. **Green Processing and Synthesis**, v. 9, n. 1, p. 77-86, 2020. DOI: https://doi.org/10.1515/gps-2020-0009.

STEIBEL, Juan Pedro *et al.* A powerful and flexible linear mixed model framework for the analysis of relative quantification RT-PCR data. **Genomics**, v. 94, n. 2, p. 146-152, 2009. DOI: https://doi.org/10.1016/j.ygeno.2009.04.008.

SUN, Huigang *et al.* Enhancement of surfactin production of *Bacillus subtilis* fmbR by replacement of the native promoter with the Pspac promoter. **Canadian Journal of Microbiology**, v. 55, n. 8, p. 1003-1006, 2009. DOI: https://doi.org/10.1139/W09-044.

TATUSOVA, Tatiana *et al.* NCBI prokaryotic genome annotation pipeline. **Nucleic Acids Research**, v. 44, n. 14, p. 6614-6624, 2016. DOI: https://doi.org/10.1093/nar/gkw569.

VASER, Robert *et al.* Fast and accurate de novo genome assembly from long uncorrected reads. **Genome Research**, v. 27, n. 5, p. 737-746, 2017. DOI: http://www.genome.org/cgi/doi/10.1101/gr.214270.116.

VASUDEVAN, Karthick *et al*. Highly accurate-single chromosomal complete genomes using IonTorrent and MinION sequencing of clinical pathogens. **BMC Genomics**, v. 112, n. 1, p. 545-551, 2020. DOI: https://doi.org/10.1016/j.ygeno.2019.04.006.

VESHAREH, Moein Jahanbani *et al.* Isolation and screening of *Bacillus subtilis* MJ01 for MEOR application: biosurfactant characterization, production optimization and wetting effect on carbonate surfaces. **Journal of Petroleum Exploration and Production Technology**, v. 9, n. 1, p. 233-245, 2019. DOI: https://doi.org/10.1007/s13202-018-0457-0.

VOS, Michiel *et al.* A comparison of rpoB and 16S rRNA as markers in pyrosequencing studies of bacterial diversity. **PLoS ONE**, v. 7, n. 2, p. e30600, 2012. DOI: https://doi.org/10.1371/journal.pone.0030600.

WATTAM, Alice R. *et al*. PATRIC, the bacterial bioinformatics database and analysis resource. **Nucleic Acids Research**, v. 42, n. D1, p. D581-D591, 2014. DOI: https://doi.org/10.1093/nar/gkt1099.

WICK, Ryan R. *et al.* Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. **PLoS Computational Biology**, v. 13, n. 6, p. e1005595, 2017. DOI: https://doi.org/10.1371/journal.pcbi.1005595.

WU, Hao; WANG, Dan; GAO, Feng. Toward a high-quality pan-genome landscape of *Bacillus subtilis* by removal of confounding strains. **Briefings in Bioinformatics**, v. 22, n. 2, p. 1951-1971, 2021. DOI: https://doi.org/10.1093/bib/bbaa013.

YOUSSEF, Noha H. *et al.* Comparison of methods to detect biosurfactant production by diverse microorganisms. **Journal of Microbiological Methods**, v. 56, n. 3, p. 339-347, 2004. DOI: https://doi.org/10.1016/j.mimet.2003.11.001.

ZAW, Myo T.; EMRAN, Nor A.; LIN, Zaw. Mutations inside rifampicin-resistance determining region of rpoB gene associated with rifampicin-resistance in *Mycobacterium tuberculosis*. **Journal of Infection and Public Health**, v. 11, n. 5, p. 605-610, 2018. DOI: https://doi.org/10.1016/j.jiph.2018.04.005.

APÊNDICES

Apêndice A – Lista de genomas utilizados neste trabalho, juntamente com seus respectivos códigos de acesso.

Código de acesso NCBI	Identificação
AL009126.3	*Bacillus subtilis* subsp. *subtilis* str. 168
CP002183.1	*Bacillus subtilis* subsp. *spizizenii* str. W23
CP002468.1	*Bacillus subtilis* BSn5
AP011541.2	*Bacillus subtilis* subsp. *natto* BEST195 DNA
CP002905.1	*Bacillus subtilis* subsp. *spizizenii* TU-B-10
CP002906.1	*Bacillus subtilis* subsp. *subtilis* RO-NN-1
CP003783.1	*Bacillus subtilis* QB928
CP003695.1	*Bacillus subtilis* subsp. *subtilis* str. BSP1
CP004019.1	*Bacillus subtilis* XF-1
CP003329.1	*Bacillus subtilis* subsp. *subtilis* 6051-HGW
CP004405.1	*Bacillus subtilis* subsp. *subtilis* str. BAB-1
CP006881.1	*Bacillus subtilis* PY79
AP012496.1	*Bacillus subtilis* BEST7003 DNA
CP007800.1	*Bacillus subtilis* subsp. *subtilis* str. JH642 substr. AG174
CP008698.1	*Bacillus subtilis* subsp. *subtilis* str. AG1839
CP007409.1	*Bacillus subtilis* subsp. *subtilis* str. OH 131.1
CP005997.1	*Bacillus subtilis* TOA
CP009611.1	*Bacillus subtilis* strain Bs-916
CP009796.1	*Bacillus subtilis* strain SG6
CP010052.1	*Bacillus subtilis* subsp. *subtilis* str. 168
CP010053.1	*Bacillus subtilis* strain PS832
CP010314.1	*Bacillus subtilis* subsp. *subtilis* strain 3NA
LN649259.1	*Bacillus subtilis* genome assembly BS49Ch
CP011115.1	*Bacillus subtilis* KCTC 1028
CP007173.1	*Bacillus subtilis* HJ5
CP011534.1	*Bacillus subtilis* strain UD1022
CP011882.1	*Bacillus subtilis* strain TO-A JPC
CP013654.1	*Bacillus subtilis* subsp. *subtilis* strain BSD-2
CP013984.1	*Bacillus subtilis* subsp. *inaquosorum* strain DE111
CP014166.1	*Bacillus subtilis* subsp. *subtilis* strain CU1050
CP014471.1	*Bacillus subtilis* subsp. *natto* strain CGMCC 2108
CP014858.1	*Bacillus subtilis* subsp. *subtilis* strain D12-5
CP015004.1	*Bacillus subtilis* strain SZMC 6179J
CP015975.1	*Bacillus subtilis* subsp. *subtilis* strain delta6
CP015375.1	*Bacillus subtilis* subsp. *subtilis* strain KCTC 3135
CP016852.1	*Bacillus subtilis* subsp. *subtilis* strain 168G
CP016894.1	*Bacillus subtilis* strain HJ0-6
CP017112.1	*Bacillus subtilis* strain BS16045
CP017314.1	*Bacillus subtilis* strain BS38
CP015222.1	*Bacillus subtilis* strain HRBS-10TDI13
CP017676.1	*Bacillus subtilis* strain VV2
CP018295.1	*Bacillus subtilis* strain J-5
CP018173.1	*Bacillus subtilis* strain MJ01
CP018184.1	*Bacillus subtilis* strain KH2

CP017763.1	*Bacillus subtilis* strain 29R7-12
CP020102.1	*Bacillus subtilis* strain NCIB 3610
CP020367.1	*Bacillus subtilis* strain GQJK2
CP020722.1	*Bacillus subtilis* strain Bs-115
CP021169.1	*Bacillus subtilis* strain TLO3
CP016767.1	*Bacillus subtilis* strain CW14
CP021507.1	*Bacillus subtilis* subsp. *subtilis* strain SRCM101441
CP021498.1	*Bacillus subtilis* subsp. *subtilis* strain SRCM101444
CP021499.1	*Bacillus subtilis* subsp. *subtilis* strain SRCM100757
CP021889.1	*Bacillus subtilis* subsp. *subtilis* strain SRCM100761
CP021892.1	*Bacillus subtilis* subsp. *subtilis* strain SRCM100333
CP021921.1	*Bacillus subtilis* subsp. *subtilis* strain SRCM101392
CP021903.1	*Bacillus subtilis* strain ge28
CP022287.1	*Bacillus subtilis* strain SX01705
CP022890.1	*Bacillus subtilis* strain DKU_NT_02
CP022891.1	*Bacillus subtilis* strain DKU_NT_03
CP023257.1	*Bacillus subtilis* strain TLO3
CP023755.1	*Bacillus subtilis* subsp. *subtilis* strain NCD-2
CP025941.1	*Bacillus subtilis* strain BJ3-2
CP020023.1	*Bacillus subtilis* strain ATCC 21228
CP029461.2	*Bacillus subtilis* strain QB61
CP029465.1	*Bacillus subtilis* subsp. *inaquosorum* strain KCTC 13429
CP020915.1	*Bacillus subtilis* strain 50-1
CP029052.1	*Bacillus subtilis* subsp. *subtilis* strain BS155
CP032310.1	*Bacillus subtilis* strain WB800N
CP032315.1	*Bacillus subtilis* strain MZK05
CP032852.1	*Bacillus subtilis* subsp. *subtilis* strain GFR-12
CP032855.1	*Bacillus subtilis* subsp. *subtilis* strain PJ-7
CP032853.1	*Bacillus subtilis* subsp. *subtilis* strain MH-1
CP032860.1	*Bacillus subtilis* subsp. *subtilis* strain SSJ-1
CP032857.1	*Bacillus subtilis* subsp. *subtilis* strain 2RL2-3
CP032867.1	*Bacillus subtilis* subsp. *subtilis* strain N4-2
CP032863.1	*Bacillus subtilis* subsp. *subtilis* strain N2-2
CP032861.1	*Bacillus subtilis* subsp. *subtilis* strain N1-1
CP032865.1	*Bacillus subtilis* subsp. *subtilis* strain N3-1
CP032872.1	*Bacillus subtilis* subsp. *subtilis* strain 2KL1
CP035161.1	*Bacillus subtilis* strain SRCM103862
CP035162.1	*Bacillus subtilis* strain SRCM103886
CP035163.1	*Bacillus subtilis* strain SRCM103923
CP035164.1	*Bacillus subtilis* strain SRCM104005
CP035165.1	*Bacillus subtilis* strain SRCM103881
CP035166.1	*Bacillus subtilis* strain SRCM103971
CP035167.1	*Bacillus subtilis* strain SRCM104008
CP035191.1	*Bacillus subtilis* strain SRCM104011
CP026010.1	*Bacillus subtilis* strain ATCC 11774
CP035226.1	*Bacillus subtilis* strain SRCM103517
CP035230.1	*Bacillus subtilis* strain SRCM103551
CP035231.1	*Bacillus subtilis* strain SRCM103571
CP035391.1	*Bacillus subtilis* strain SRCM103689
CP035390.1	*Bacillus subtilis* strain SRCM103641

CP035394.1	*Bacillus subtilis* strain SRCM103696
CP035402.1	*Bacillus subtilis* strain SRCM103576
CP035395.1	*Bacillus subtilis* strain SRCM103697
CP035403.1	*Bacillus subtilis* strain SRCM103581
CP035397.1	*Bacillus subtilis* strain SRCM103773
CP035401.1	*Bacillus subtilis* strain SRCM103837
CP035400.1	*Bacillus subtilis* strain SRCM103835
CP035406.1	*Bacillus subtilis* strain SRCM103612
CP035413.1	*Bacillus subtilis* strain SRCM103629
CP035411.1	*Bacillus subtilis* strain SRCM103622
CP035414.1	*Bacillus subtilis* strain SRCM103637
CP026608.1	*Bacillus subtilis* strain HDZK-BYSB7
CP029609.1	*Bacillus subtilis* subsp. *subtilis* strain G7
CP033205.1	*Bacillus subtilis* strain MBI 600
CP039935.1	*Bacillus subtilis* strain H19
CP040528.1	*Bacillus subtilis* strain PR10
CP034484.1	*Bacillus subtilis* subsp. *subtilis* NCIB 3610 = ATCC 6051 strain NCIB 3610
CP034943.1	*Bacillus subtilis* subsp. *spizizenii* ATCC 6633
CP021123.1	*Bacillus subtilis* strain SEM-9
CP041015.1	*Bacillus subtilis* strain FDAARGOS_606
AP019714.1	*Bacillus subtilis* subsp. *subtilis* NBRC 13719 DNA
CP044498.1	*Bacillus subtilis* strain ms-2
CP045425.1	*Bacillus subtilis* strain JAAA
CP031783.1	*Bacillus subtilis* strain MENO2
CP031784.1	*Bacillus subtilis* strain HMNig-2
CP045819.1	*Bacillus subtilis* strain MB9_B4
CP045820.1	*Bacillus subtilis* strain MB9_B1
CP045824.1	*Bacillus subtilis* strain MB8_B10
CP045821.1	*Bacillus subtilis* strain MB8_B7
CP045825.1	*Bacillus subtilis* strain 75
CP045823.1	*Bacillus subtilis* strain MB8_B1
CP045826.1	*Bacillus subtilis* strain 73
CP045816.1	*Bacillus subtilis* strain P5_B2
CP045817.1	*Bacillus subtilis* strain P5_B1
CP045818.1	*Bacillus subtilis* strain MB9_B6
CP045812.1	*Bacillus subtilis* strain P8_B3
CP045922.1	*Bacillus subtilis* strain P8_B1
CP045811.1	*Bacillus subtilis* strain P9_B1
CP046448.1	*Bacillus subtilis* strain ZD01
CP047325.1	*Bacillus subtilis* strain GOT9
CP023409.1	*Bacillus subtilis* strain 7PJ-16
CP047485.1	*Bacillus subtilis* strain BJQ0005
CP028201.1	*Bacillus subtilis* strain SRCM102753
CP028202.1	*Bacillus subtilis* strain SRCM102754
CP028209.1	*Bacillus subtilis* strain SRCM102745
CP028212.1	*Bacillus subtilis* strain SRCM102748
CP028213.1	*Bacillus subtilis* strain SRCM102749
CP028215.1	*Bacillus subtilis* strain SRCM102750
CP028217.1	*Bacillus subtilis* strain SRCM102751

CP028218.1	*Bacillus subtilis* strain SRCM102756
CP031693.1	*Bacillus subtilis* strain SRCM101393
CP045672.1	*Bacillus subtilis* strain 2014-3557
CP026662.1	*Bacillus subtilis* strain H1
CP050532.1	*Bacillus subtilis* subsp. *subtilis* str. SMY
CP051306.1	*Bacillus subtilis* strain RI4914
CP051462.1	*Bacillus subtilis* strain At3
CP051465.1	*Bacillus subtilis* subsp. *subtilis* strain UCMB5121
CP051466.1	*Bacillus subtilis* subsp. *subtilis* strain UCMB5021
CP051860.1	*Bacillus subtilis* subsp. *subtilis* str. 168

Fonte: Disponibilizado em GenBank

Apêndice B – Estatística genômica estimada pelo Quast.

Estatística genômica	*B. subtilis* RI4914
Genome fraction (%)	89.851
Duplication ratio	1
Largest alignment	279465
Total aligned length	3787178
NG50	4100930
NG75	4100930
NA50	77489
NA75	42356
NGA50	77489
NGA75	40907
LG50	1
LG75	1
LA50	14
LA75	31
LGA50	14
LGA75	33
Misassemblies	
# misassemblies	96
# relocations	96
# translocations	0
# inversions	0
# misassembled *contigs*	1
Misassembled *contigs* length	4100930
# local misassemblies	63
# scaffold gap ext. mis.	0
# scaffold gap loc. mis.	0
# unaligned mis. *contigs*	0
Unaligned	
# fully unaligned *contigs*	0
Fully unaligned length	0
# partially unaligned *contigs*	1
Partially unaligned length	311358
Mismatches	
# mismatches	46635
# indels	1893
Indels length	6069
# mismatches per 100 kbp	1231.21
# indels per 100 kbp	49.98
# indels (<= 5 bp)	1685
# indels (> 5 bp)	208
# N’s	0
# N’s per 100 kbp	0
Statistics without reference	
# *contigs*	1
# *contigs* (>= 0 bp)	1
# *contigs* (>= 1000 bp)	1
# *contigs* (>= 5000 bp)	1
# *contigs* (>= 10000 bp)	1

# *contigs* (>= 25000 bp)	1
# *contigs* (>= 50000 bp)	1
Largest contig	4100930
Total length	4100930
Total length (>= 0 bp)	4100930
Total length (>= 1000 bp)	4100930
Total length (>= 5000 bp)	4100930
Total length (>= 10000 bp)	4100930
Total length (>= 25000 bp)	4100930
Total length (>= 50000 bp)	4100930
N50	4100930
N75	4100930
GC (%)	43.49

Fonte: Da autora. Disponibilizado por Quast

Apêndice C – Genes previstos por PATRIC relacionados à resistência a antibióticos no genoma da cepa de *B. subtilis* RI4914.

Gene	Produto	Função	Classificação	PubMed
MurA	UDP-N-acetilglucosamina 1-carboxiviniltransferase (EC 2.5.1.7)	UDP-N-acetilglucosamina 1-carboxiviniltransferase (EC 2.5.1.7)	alvo de antibiótico em espécies suscetíveis	8994972
rpoC	Subunidade beta 'da polimerase de RNA dirigida por DNA (EC 2.7.7.6)	Subunidade beta 'da polimerase de RNA dirigida por DNA (EC 2.7.7.6)	alvo de antibiótico em espécies suscetíveis	16723576
LiaR	Sistema de resposta ao estresse do envelope celular LiaFSR, regulador de resposta LiaR (VraR)	Sistema de resposta ao estresse do envelope celular LiaFSR, regulador de resposta LiaR (VraR)	regulador modulando a expressão de genes de resistência a antibióticos	21899450; 26020679
GdpD	Glicerofosforil diéster fosfodiesterase (EC 3.1.4.46)	Glicerofosforil diéster fosfodiesterase (EC 3.1.4.46)	proteína que altera a carga da parede celular e confere resistência a antibióticos	21899450
BceR	Regulador de resposta de dois componentes BceR	Regulador de resposta de dois componentes BceR	regulador modulando a expressão de genes de resistência a antibióticos	25118291
LiaF	Proteína de membrana LiaF (VraT), inibidor específico da via de sinalização LiaRS (VraRS)	Proteína de membrana LiaF (VraT), inibidor específico da via de sinalização LiaRS (VraRS)	regulador modulando a expressão de genes de resistência a antibióticos	21899450; 26020679
EF-G	Fator de alongamento de translação G	Fator de alongamento de translação G	alvo de antibiótico em espécies suscetíveis	17980694
ANT(6)-I	Aminoglicosídeo 6-nucleotidiltransferase (EC 2.7.7.-) => ANT (6) -I	Aminoglicosídeo 6-nucleotidiltransferase (EC 2.7.7.-) => ANT (6) -I	enzima de inativação de antibiótico	19603075; 20479200; 2168151; 8293959
BceS	Sensor de histidina quinase BceS de dois componentes	Sensor de histidina quinase BceS de dois componentes	regulador modulando a expressão de genes de	25118291

			resistência a antibióticos	
Lmr(B)	Resistência à lincomicina, bomba de efluxo MFS => Lmr (B)	Resistência à lincomicina, bomba de efluxo MFS => Lmr (B)	bomba de efluxo conferindo resistência a antibióticos	15317768
GdpD	Glicerofosforil diéster fosfodiesterase (EC 3.1.4.46)	Glicerofosforil diéster fosfodiesterase (EC 3.1.4.46)	proteína que altera a carga da parede celular e confere resistência a antibióticos	21899450
S12p	Proteína ribossômica SSU S12p (S23e)	Proteína ribossômica SSU S12p (S23e)	alvo de antibiótico em espécies suscetíveis	7934937
inhA, fabI	Enoil- [acil-proteína transportadora] redutase [NADH] (EC 1.3.1.9)	Enoil- [acil-proteína transportadora] redutase [NADH] (EC 1.3.1.9)	alvo de antibiótico em espécies suscetíveis	18193820; 10869170; 8284673
EF-Tu	Fator de alongamento de translação Tu	Fator de alongamento de translação Tu	alvo de antibiótico em espécies suscetíveis	364475; 9678602
BceA	Bacitracina exporta proteína de ligação de ATP BceA	Bacitracina exporta proteína de ligação de ATP BceA	bomba de efluxo conferindo resistência a antibióticos	25118291
LiaS	Sistema de resposta ao estresse do envelope celular LiaFSR, sensor de histidina quinase LiaS (VraS)	Sistema de resposta ao estresse do envelope celular LiaFSR, sensor de histidina quinase LiaS (VraS)	regulador modulando a expressão de genes de resistência a antibióticos	21899450; 26020679
Alr	Alanina racemase (EC 5.1.1.1)	Alanina racemase (EC 5.1.1.1)	alvo de antibiótico em espécies suscetíveis	19748470; 24303782
BceB	Bacitracina exporta proteína permease BceB	Bacitracina exporta proteína permease BceB	bomba de efluxo conferindo resistência a antibióticos	25118291
Alr	Alanina racemase (EC 5.1.1.1)	Alanina racemase (EC 5.1.1.1)	alvo de antibiótico em espécies suscetíveis	19748470; 24303782
fabL	Enoil- [acil-proteína transportadora] redutase [NADPH] (EC 1.3.1.104), FabL	Enoil- [acil-proteína transportadora] redutase [NADPH] (EC 1.3.1.104), FabL	proteína de substituição alvo de antibiótico	21185310
Tet(L)	Resistência à tetraciclina, bomba de efluxo MFS => Tet (L)	Resistência à tetraciclina, bomba de efluxo MFS => Tet (L)	bomba de efluxo conferindo resistência a antibióticos	7628724; 7877638

gyrA	Subunidade A da girase de DNA (EC 5.99.1.3)	Subunidade A da girase de DNA (EC 5.99.1.3)	alvo de antibiótico em espécies suscetíveis	9293187
Ddl	D-alanina - D-alanina ligase (EC 6.3.2.4)	D-alanina - D-alanina ligase (EC 6.3.2.4)	alvo de antibiótico em espécies suscetíveis	24303782; 24033232
EbrA	Proteína EbrA de resistência a múltiplas drogas	Proteína EbrA de resistência a múltiplas drogas	bomba de efluxo conferindo resistência a antibióticos	17417881
S10p	Proteína ribossômica SSU S10p (S20e)	Proteína ribossômica SSU S10p (S20e)	alvo de antibiótico em espécies suscetíveis	26124155
	Macrolídeo 2'-fosfotransferase, putativo	Macrolídeo 2'-fosfotransferase, putativo	enzima de inativação de antibiótico	
FosB	Proteína FosB de resistência à fosfomicina	Proteína FosB de resistência à fosfomicina	enzima de inativação de antibiótico	14677948; 16399398
gidB	16S rRNA (guanina (527) -N (7)) - metiltransferase (EC 2.1.1.170)	16S rRNA (guanina (527) -N (7)) - metiltransferase (EC 2.1.1.170)	gene que confere resistência via ausência	17238915
rpoB	Subunidade beta de RNA polimerase dirigida por DNA (EC 2.7.7.6)	Subunidade beta de RNA polimerase dirigida por DNA (EC 2.7.7.6)	alvo de antibiótico em espécies suscetíveis	3050121; 15047531; 16723576
RlmA(II)	23S rRNA (guanina (748) -N (1)) - metiltransferase (EC 2.1.1.188)	23S rRNA (guanina (748) -N (1)) - metiltransferase (EC 2.1.1.188)	enzima modificadora de alvo de antibiótico	12514124; 18406425
BceB	Bacitracina exporta proteína permease BceB	Bacitracina exportadora de proteína permease BceB	bomba de efluxo conferindo resistência a antibióticos	25118291
BcrC	Undecaprenil-difosfatase BcrC (EC 3.6.1.27), transmite resistência à bacitracina	Undecaprenil-difosfatase BcrC (EC 3.6.1.27), transmite resistência à bacitracina	proteína de proteção alvo antibiótico	12486040; 15778224; 15946938
dxr	1-desoxi-D-xilulose 5-fosfato redutoisomerase (EC 1.1.1.267)	1-desoxi-D-xilulose 5-fosfato redutoisomerase (EC 1.1.1.267)	alvo de antibiótico em espécies suscetíveis	16321944
Iso-tRNA	Isoleucil-tRNA sintetase (EC 6.1.1.5)	Isoleucil-tRNA sintetase (EC 6.1.1.5)	alvo de antibiótico em espécies suscetíveis	7929087
gyrB	Subunidade B da gyrase de DNA (EC 5.99.1.3)	Subunidade B da girase de DNA (EC 5.99.1.3)	alvo de antibiótico em espécies suscetíveis	21693461; 22279180;9293

				187
	Classe D beta-lactamase (EC 3.5.2.6)	Classe D beta-lactamase (EC 3.5.2.6)	antibiotic inactivation enzyme	
MprF	L-O-lisilfosfatidilglicerol sintase (EC 2.3.2.3)	L-O-lisilfosfatidilglicerol sintase (EC 2.3.2.3)	proteína que altera a carga da parede celular e confere resistência a antibióticos	19289517;16723576
kasA	3-oxoacil- [acil-carreadora-proteína] sintase, KASII (EC 2.3.1.179)	3-oxoacil- [acil-carreadora-proteína] sintase, KASII (EC 2.3.1.179)	alvo de antibiótico em espécies suscetíveis	10428945
YkkCD	Bomba de efluxo de múltiplas drogas de ampla especificidade YkkC	Bomba de efluxo de múltiplas drogas de ampla especificidade YkkC	bomba de efluxo conferindo resistência a antibióticos	10735877
folA, Dfr	Diidrofolato redutase (EC 1.5.1.3)	Diidrofolato redutase (EC 1.5.1.3)	alvo de antibiótico em espécies suscetíveis	20169085;25288078
GdpD	Glicerofosforil diéster fosfodiesterase (EC 3.1.4.46)	Glicerofosforil diéster fosfodiesterase (EC 3.1.4.46)	proteína que altera a carga da parede celular e confere resistência a antibióticos	21899450
EbrB	Proteína EbrB de resistência a múltiplas drogas	Proteína EbrB de resistência a múltiplas drogas	bomba de efluxo conferindo resistência a antibióticos	17417881
folA, Dfr	Diidrofolato redutase (EC 1.5.1.3)	Dihydrofolate reductase (EC 1.5.1.3)	alvo de antibiótico em espécies suscetíveis	20169085;25288078
folP	Diidropteroato sintase (EC 2.5.1.15)	Diidrofolato redutase (EC 1.5.1.3)	alvo de antibiótico em espécies suscetíveis	15673783
rho	Fator de terminação de transcrição Rho	Fator de terminação de transcrição Rho	alvo de antibiótico em espécies suscetíveis	8466900
MurA	UDP-N-acetilglucosamina 1-carboxiviniltransferase (EC 2.5.1.7)	UDP-N-acetilglucosamina 1-carboxiviniltransferase (EC 2.5.1.7)	alvo de antibiótico em espécies suscetíveis	8994972
	Macrolídeo 2'-fosfotransferase, putativo	Macrolídeo 2'-fosfotransferase, putativo	enzima de inativação de antibiótico	
PgsA	CDP-diacilglicerol - glicerol-3-fosfato 3-fosfatidiltransferase	CDP-diacilglicerol - glicerol-3-fosfato 3-fosfatidiltransferase (EC 2.7.8.5)	proteína que altera a carga da parede celular e confere	22238576

(EC 2.7.8.5)	resistência a antibióticos