UNIVERSIDADE FEDERAL DA BAHIA
RENORBIO
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

BRENA MOTA MOITINHO SANT'ANNA

Análises genômicas de bactérias endofíticas B*acillus velezensis* 629 e S*erratia marcescens* 1274

SALVADOR - BA

2018

BRENA MOTA MOITINHO SANT'ANNA

Análises genômicas de bactérias endofíticas B*acillus velezensis* 629 e S*erratia marcescens* 1274

Tese apresentada ao Programa de Pós-graduação em Biotecnologia da Rede Nordeste de Biotecnologia (RENORBIO) do Ponto Focal Bahia da Universidade Federal da Bahia, como requisito para obtenção do título de Doutor em Biotecnologia.

Orientador: Prof. Dr. Milton Ricardo de Abreu Roque

SALVADOR - BA

2018

Ficha catalográfica elaborada pelo Sistema Universitário de Bibliotecas (SIBI/UFBA),
com os dados fornecidos pelo(a) autor(a).

Mota Moitinho Sant´Anna, Brena
Análises genômicas de bactérias endofíticas Bacillus velezensis 629 e Serratia marcescens 1274 / Brena Mota Moitinho Sant´Anna. -- Salvador, 2018.
128 f.

Orientador: Milton Ricardo de Abreu Roque.
Tese (Doutorado - Doutorado em Biotecnologia - RENORBIO) -- Universidade Federal da Bahia, Universidade Federal da Bahia, Programa de Pós-Graduação em Biotecnologia da Renorbio, 2018.

1. Bactérias endofíticas. 2. Bacillus velezensis. 3. Serratia marcescens. 4. Análises genômicas. I. Ricardo de Abreu Roque, Milton. II. Título.

TERMO DE APROVAÇÃO

A TESE:

"Análises genômicas de bactérias endofíticas *Bacillus velezensis* 629 e *Serratia marcescens* 1274 "

ELABORADA POR:

BRENA MOTA MOITINHO SANT'ANNA

Foi aprovada por todos os membros da banca examinadora e aceita pelo Programa de Pós-Graduação em Biotecnologia da Renorbio como requisito parcial à obtenção do título de DOUTORA EM BIOTECNOLOGIA

Salvador, Bahia, vinte e nove de maio de 2018

BANCA EXAMINADORA:

Prof. Dr. Milton Ricardo de Abreu Roque
(Orientador)
Universidade Federal da Bahia

Dr. Artur Trancoso Lopo de Queiroz
Fundação Oswaldo Cruz

Prof. Dr. Luís Gustavo Carvalho Pacheco
Universidade Federal da Bahia

Prof. Dr. Eric Roberto Guimarães Rocha Aguiar
Universidade Federal da Bahia

Luciano Kalabri Silva

Prof. Dr. Luciano Kalabric Silva
Fundação Oswaldo Cruz

Prof. Dr. Pedro Milet Meirelles
Universidade Federal da Bahia

Agradecimentos

Inicio minhas palavras agradecendo a Deus, não de forma clichê, mas da maneira mais verdadeira que posso me expressar! Pois, sinto-me abençoada por chegar até aqui e estar sempre rodeada de pessoas do bem e boas energias que me impulsionam a crescer. Meus pais são os maiores responsáveis por tudo que eu alcancei até hoje! Reforçaram tanto a importância dos estudos que eu estou nessa até hoje! Meus irmãos, que tripé é esse que formamos? Somos tão fortes juntos que toda conquista que temos é dos três! Esta também é de vocês! Continuei abençoada quando Celso Jr., meu amado companheiro, entrou em minha vida e abraçou todos os sonhos que hoje são nossos! Tudo em minha vida é melhor com ele e não há palavras que contemplem a importância dele nessa trajetória. Amo vocês infinitamente!

Eu tenho um grande amigo em meu orientador, parceria tão forte que já são muitos anos caminhando juntos e crescendo juntos! Milton é uma pessoal incrível, que me permitiu conhecer da forma mais verdadeira a ciência e a vida caminhando juntas, sem abdicar de uma delas. Obrigada por sempre me apresentar o melhor caminho e por me acalmar nos momentos mais perturbadores que a ciência trouxe. O Laboratório Bioprospector foi uma conquista! Nesses quatro anos, um mundo de histórias, pesquisas, apertos e acertos foram construídos ali. Aprendi e compartilhei tudo que chegou a mim, e agradeço a todos desse grupo, em especial: Aldi, a parceira de sempre, irmã que ganhei na ciência; Carla Sampaio, Carlitcha, pesquisadora incrível e amiga que quero sempre; Minha amiga Pati, parceira de muitos anos na ciência, que sempre me apoiou e me ensinou muito. Meus queridos amigos pesquisadores dessa equipe que estão crescendo incrivelmente, Ícaro, Felipe, Priscila, Susanna, Lui, Carla Silveira e Amanda. Aprendi muito compartilhando ideias com vocês.

Das equipes e laboratórios parceiros, um carinho especial para muitos: minha querida amiga e parceira, de vida e pesquisa, Fúlvia. Obrigada por cada conversa e apoio que sempre nos fortaleceram; meu grande amigo e irmão científico Luiz Franco, sempre estaremos na mesma equipe; minha parceira e amiga Carol, com quem dividi muitos desafios quando decidimos seguir a mesma linha de pesquisa, foi muito importante estarmos juntas; ao amigo Matheus Brito que chegou com o desafio que conseguimos transformar em trabalho, aprendi muito contigo também! Faço questão de citar também os colegas de curso e os amigos que sempre pude contar e de alguma forma fizeram parte dessa caminhada: Mille, Hendor, Dayana, Rodrigo, Diana, Mooana, Cris, Xande, Leila, Bethânia, Diego, Roberta e minha eterna amiga Sueli, que nos deixou os melhores ensinamentos de paz, perseverança e serenidade.

Não seria possível crescer em minha pesquisa com todos os ensinamentos e apoio que recebi de dois professores em especial: Prof. Artur Lopo, foi essencial nos meus primeiros passos na bioinformática e continuou me trazendo soluções e ensinamentos até a última linha. Eu tenho certeza que ganhei mais um amigo, você foi um porto seguro para mim, Muito Obrigada! Prof. Luis Pacheco, tenho certeza que cada encontro e reunião que tivemos foram fundamentais para meu crescimento, a sua inteligência e serenidade são admiráveis, obrigada por estar sempre disposto a ajudar e me permitir aprender contigo.

Estendo meus agradecimentos a professores parceiros que tiveram envolvimento em várias etapas, prof. Jorge Souza e Phellippe Marbach no projeto inicial e sequenciamento dos genomas; prof. Rommel Ramos e Artur Silva que me receberam na UFPA em Belém-PA e me apresentaram novos conhecimentos e realidades do mundo acadêmico científico durante minha temporada na missão de estudos associada à “Rede de Cooperação Acadêmica em Genômica Funcional e Biologia de Sistemas de Microrganismos”; Assim como toda sua equipe no Laboratório de Genômica e Bioinformática da UFPA que foram maravilhosos na recepção: Luis, Kenny, Marcus, Alan, Amália, Wellington e Fabrício; lembrarei sempre de vocês. Agradeço também ao prof. Vasco Azevedo que por mais de uma vez abriu as portas do seu laboratório na UFMG e me permitiu contato direto com sua equipe, em especial o amigo Marcus Canário, sempre em contato me auxiliou em vários momentos.

Quando o assunto é viagem, não tem como não agradecer a tio Arnaldinho que me hospedou em sua casa e me tratou como uma filha por todos os dias que passei em Belém! Será sempre meu tio querido do coração. E em Belo Horizonte, um agradecimento muito especial aos tios Oliveiro e Sônia e sua família que mais de uma vez, com o carinho de sempre, me fizeram sentir em casa quando estive lá.

E no meu dia a dia, quem sempre me fortaleceu são os que sempre estiveram comigo e continuam: amig@s, sogr@s, cunhad@s, ti@s, prim@s, afilhados e sobrinhas, Muito Obrigada! Família Mota, Família Moitinho, Família SantAnna, Extramigos, Superamigos, Galera Forrozin e Amigas de sempre, a minha vida faz sentido por ter vocês comigo!

Foram muitos os que participaram dessa trajetória. Sinto-me abençoada por todas as oportunidades e amigos que pude contar nesses quatro anos de doutorado. Muitíssimo Obrigada a Todos Vocês!

RESUMO

Microrganismos endofíticos são capazes de colonizar tecidos internos de plantas sem causar danos, varias espécies endofíticas estão relacionadas com controle biológico de fitopatógenos e agentes promotores de crescimento de plantas. As interações entre endofíticos e hospedeiros são complexas e pouco se conhece sobre as bases moleculares desta interação. Estudos genômicos destas bactérias endofíticas representam uma estratégia para elucidar os mecanismos envolvidos no processo endofítico. Dessa forma, dois genomas bacterianos de endofíticos, *Bacillus velezensis* 629 e *Serratia marcescens* 1274, foram sequenciados, montados e anotados para permitir a mineração genômica. Os estudos realizados envolveram a analise dos genomas e a genômica comparativa de cada isolado, com ferramentas para analise de taxonomia genômica, pangenoma, ilhas genômicas e metabólicos secundários entre os endofíticos. Uma inconstância na classificação taxonômica é percebida devido às novas ferramentas aplicadas para análises genômicas no grupo *Bacillus*. Por outro lado, para a determinação do pangenoma fechado, o número de genomas depositados ainda não foi suficiente paras as espécies *B. velezensis* e *Serratia marcescens.* O genoma dos dois isolados indicaram características benéficas na relação endofítico-hospedeiro, como potencial para biocontrole, promoção de crescimento e assimilação de nutrientes. A capacidade de colonização do tecido vegetal interno está relacionada em bactérias gram-negativas com o sistema de secreção tipo VI (T6SS). Os processos metabólicos comuns, como mediação de fitormônios, produção de antimicrobianos e assimilação de nutrientes, podem ser desencadeados por diferentes vias e as estratégias de colonização serão específicas da relação hospedeiro-endofítico. As análises comparativas neste trabalho não determinam que as interações entre as bactérias endofíticas e a planta hospedeira apresentam uma assinatura gênica comum aos endofíticos.

Palavras-chave: Bactérias endofíticas, *Bacillus velezensis, Serratia marcescens*, Análises genômicas.

ABSTRACT

Endophytic microorganisms can colonize the internal tissues of plants without causing damage, several endophytic species are related to biological control of plant pathogens and plant growth promoting. Endophytic and host interactions are complex and molecular basis of this interaction is unknown or poor described. Genomic studies of these endophytic bacteria represent a strategy to elucidate the mechanisms involved in the endophytic process. Thus, two endophytic bacterial genomes, *Bacillus velezensis* 629 and *Serratia marcescens* 1274, were sequenced, assembled and annotated to allow genomic mining. Genome analysis and comparative genomics of each isolate were carried out, with tools for taxonomy, pangenoma, genomic islands and secondary metabolic for endophytes. An inconsistency in the taxonomic classification to *Bacillus* group is perceived due to new tools applied to genomic analysis. On the other hand, for determining closed pangenoma, the number of genomes deposited has not been enough for *B. velezensis* and *Serratia marcescens* species. The genome of both isolates indicated beneficial endophytic-host relationships, as potential for biocontrol, plant growth promotion and nutrient assimilation. The ability to colonize internal plant tissue, is related in gram-negative bacteria with type VI secretion system (T6SS). Metabolic processes, such as plant hormone mediation, antimicrobial production and nutrient assimilation, can be triggered by different pathways and colonization strategies that will be specific to the host-endophytic relationship. The comparative analyzes in this work do not determine that interactions between the endophytic bacteria and host plants shows a common gene signature for all endophytes.

Keywords: Endophytic bacterium, *Bacillus velezensis*, *Serratia marcescens*, Genomic analysis.

LISTA DE FIGURAS

Revisão de literatura

Capítulos:

Manuscrito 2

Manuscrito 3

Manuscrito 4

Genomic Islands (GIs) in strain 1274 in green.

Fig.2 Comparative analysis in Venn diagram showing the shared genes in predicted Genomic Islands of six representative *S. marcescens* 1274 (endophytic), RSC-14 (endophytic), FS14 (endophytic), B3R3 (phytopathogenic), CAV1492 (clinical) and SM39 (clinical). The cluster number in each component is displayed in the bar-plots and protein-coding genes shared only three endophytic bacteria are presented in the featured frame.

Fig. 3 Genetic architecture of T6SS clusters from *S. marcescens* 1274. The conserved core gene components of the T6SS are indicated in red and the *TagB*, associated pentapeptide repeat protein, is only in the first cluster.

Fig.4 The synteny analysis across gene *FimB* in *Serratia marcescens* strain 1274 and together with other strains RSC14, FSC14, B3R3, CAV1492 and SM39 (as depicted by the program SyntTax). A, B and C represent the occurrence of that architecture in the endophytic genome 1274 and just below the other genomes. Highlighting the *FimB* gene present in *S. marcescens* plant-associated, strains RSC14, FSC14 and B3R3.

LISTA DE TABELAS

Capítulos:

Manuscrito 2

Manuscrito 4

LISTA DE ABREVIATURAS

ANI - Identidade Média de Nucleotídeos / Average Nucleotide Identity

CAT - Catalase

CBAS - Coleção de Bactérias do Ambiente e Saúde / Collection of Bacteria from Environment and Health

CDS - Coding DNA Sequence

DBG - De Bruijn Graph

DDBJ - DNA DataBank of Japan

dDDH - Hibridização DNA-DNA digital / digital DNA-DNA Hibridization

ENA - European Nucleotide Archive

GGDC - Cálculo de Distância entre Genomas / Genome to Genome Distance Calculator

GI - Ilhas Genômicas / Genomic Island

GOLD - Genomes OnLine Database

GPX - Glutathione Peroxidase

GSS - Glutathione Synthetase

GSTs - Glutathione S-Transferase

IAA - Ácido Indolacético / Indole-3-Acetic Acid

IHGSC - International Human Genome Sequencing Consortium

MAMPs - Padrões Moleculares Associados a Microrganismos / Microbe Associated Molecular Pattern

MP - Multiparanoid (Method)

NCBI - National Center for Biotechnology Information

NGS - Sequenciamento de Nova Geração / Next Generation Sequencing

NIH - National Institutes of Health

Nrps - Non-ribosomal Peptide Synthetase

OLC - Overlap Layout Consensus

PGM - Personal Genome Machine

PKS - Polyketide Synthase

RAST - Rapid Annotation using Subsystem Technology

ROS - Espécies Reativas de Oxigênio / Reactive Oxygen Species

SOD - SuperOxide Dismutase

T3SS - Type III Secretion System

T4SS - Type IV Secretion System

T6SS - Type VI Secretion System

WDCM - World Data Centre for Microorganisms

WGS - Whole Genome Shotgun

SUMÁRIO

1. Apresentação

Os microrganismos capazes de colonizar os tecidos internos de plantas, sem causar danos ou doenças, são denominados endofíticos. Diversos benefícios à planta hospedeira são atribuídos à presença do endofítico, como promoção do crescimento (por regulação de fitormônios e/ou assimilação e fixação de compostos), resistência a estresse e controle biológico de fitopatógenos. As bactérias endofíticas podem pertencer a diversos gêneros, como *Azoarcus, Azospirillum, Bacillus, Burkholderia, Enterobacter, Gluconacetobacter, Herbaspirillum, Klebsiella, Methylobacterium, Pseudomonas, Serratia, Stenotrophomonas* e *Variovorax*. O gênero *Bacillus* é amplamente conhecido na associação com plantas, e promovem diversos benefícios como o aumento à resistência a doenças. Bactérias do gênero *Serratia* estão associadas a diversos ambientes e estilos de vida, com isolados encontrados como endofíticos benéficos até patógenos de plantas e animais.

Uma das abordagens para o estudo de endofíticos pode incluir o sequenciamento e mineração genômica direcionada ao processo de interação bactéria-hospedeiro. Uma montagem de genoma minuciosa seguida de análises genômicas permite a busca de novas compreensões referentes às bases moleculares que envolvem o comportamento endofítico. As análises genômicas podem incluir diversas vertentes, incluindo análises filogenômicas; determinação de pangenoma, core-genoma, genes acessórios e únicos dentro de um grupo; predição e estudo de ilhas genômicas; e comparação genômica direcionada a um grupo ou comportamento. Este trabalho foi direcionado a partir do sequenciamento de ultima geração de duas bactérias endofíticas, *Bacillus velezensis* 629 (isolada do cacau, *Theobroma cacao*) e *Serratia marcescens* 1274 (isolada do sisal, *Agave sisalana*), seguido da montagem, anotação e estudos *in silico* dos genomas. Os dois isolados são oriundos do projeto “Genômica de bactérias endofíticas: pesquisa e formação de recursos humanos” (Linha 1 - MEC/MCTI/CAPES/CNPQ/FAPS nº 61/2011), no qual representou apoio inicial no isolamento e sequenciamento das bactérias, além do uso de servidores que permitiram a montagem e análises dos genomas.

Dessa forma o trabalho foi desenvolvido em duas etapas, que consistiu no aprofundamento de uso de ferramentas de bioinformática voltadas para montagem e análises genômicas. A primeira etapa incluiu o estudo de montadores, parâmetros e tratamento de dados dos genomas dos isolados *Bacillus velezensis* 629 e *Serratia marcescens* 1274, representando parte fundamental para dados de qualidade nas análises genômicas. O "Apêndice A" detalha os processos realizados e serviu como base para o trabalho "GATOOL: a fast and user-friendly Genome Assembly web TOOL for Ion Torrent data" (Apêndice B). A ferramenta desenvolvida representa um formato "amigável" para montagem de genomas e análises de taxonomia genômica por ANI (Average Nucleotide Identity) estão sendo incluídas na atualização.

As etapas seguintes, claramente divididas nos capítulos desta tese são referentes às análises genômicas voltadas à compreensão do processo endofítico em bactérias. Essas análises foram geradas a partir dos genomas sequenciados dos isolados bacterianos *B. velezensis* 629 e *S. marcescens* 1274, e estão apresentados como manuscritos (publicados ou submetidos para publicação) derivados desse estudo.

2. Revisão de literatura

2.1. Microrganismos endofíticos

Microrganismos endofíticos colonizam tecidos internos de plantas sem causar nenhum efeito negativo ou sintomas de doenças. Bactérias são capazes de colonizar as raízes, caules e folhas de plantas, podendo habitar dentro de células, em espaços intercelulares ou no sistema vascular vegetal (Ryan et al., 2008). Bactérias endofíticas promovem efeitos benéficos para as plantas hospedeiras, como agentes para remediação de ambientes contaminados (Newman & Reynolds, 2005), promoção de crescimento de plantas (Mitter et al., 2013), resistência ao estresse (Sziderics et al., 2007) e controle biológico de patógenos (Krishnan et al., 2015). Além disso, são fontes de moléculas naturais, como antibióticos (Doley, 2015; Mitter et al., 2013).

Associações de microrganismos endofíticos com seus hospedeiros são variadas e complexas e o entendimento dessas interações ainda não é bem elucidado (Newman & Reynolds, 2005). Entre os microrganismos endofíticos, as bactérias têm chamado atenção por sua versatilidade em variados habitats e por possuírem as mais altas densidades dentre os grupos microbianos que interagem com as plantas (Berg et al., 2005; Mendes & Azevedo, 2007). A capacidade dos endofíticos colonizarem o interior de plantas pode favorecer sua ação e representa uma vantagem em relação ao controle biológico com microrganismos, pois potencializa os seus efeitos sobre a planta hospedeira e reduz as consequências das variações populacionais, decorrentes da interação com outros microrganismos e com o meio ambiente (Guo et al., 2008).

Bactérias endofíticas, e as associadas à rizosfera, possuem efeitos benéficos similares na planta hospedeira. Contudo, as endofíticas podem interagir mais intimamente com o hospedeiro, com menor competição por fontes de carbono e maior proteção contra mudanças ambientais em relação à bactéria do solo rizosférico (Reinhold-Hurek, 1998). Também conhecidos como endofíticos competentes, estes microrganismos possuem propriedades de colonização oportunista na raiz e de adaptação ao ambiente interno da planta, mantendo o equilíbrio harmonioso com o hospedeiro (Hardoim et al., 2008). Contudo, interações entre microrganismos e o hospedeiro são variáveis e as bases moleculares da interação endofítico-planta ainda não são bem descritas, principalmente na distinção entre bactérias nocivas e benéficas (Ryan et al., 2008; Mitter et al., 2013).

A capacidade para colonizar plantas eficientemente é uma característica essencial para bactérias empregadas como agentes de controle biológico e promotores de crescimento. Chin-A-Woeng et al. (2000) demonstraram que rizobactérias quando colonizam extensivamente as raízes de plantas são bons agentes de controle de patógenos. No entanto, a falta de conhecimento sobre os mecanismos genéticos envolvidos em colonização endofítica está entre as razões para que poucos microrganismos endofíticos se transformem em produtos comerciais para serem utilizados no campo. A capacidade de bactérias utilizarem certos metabólitos da planta pode ser um pré-requisito para o estabelecimento endofítico bem sucedido, como uso de fontes de carbono disponíveis (Malfanova et al., 2013).

As plantas são naturalmente associadas a microrganismos de várias formas e as interações que ocorrem em consequência das respostas de defesa da planta na presença dos microrganismos direcionam a caracterização quanto a patogenicidade ou não. O sistema de defesa das plantas induz respostas por reconhecimento de moléculas como “padrões moleculares associados a microrganismos” (ou associados à patógenos, MAMPs/PAMPs) ou na presença de fatores de virulência dos organismos invasores (Reinhold-Hurek, 2011). Para desviar da defesa vegetal e se estabelecer como endofítico, diversas características funcionais de microrganismos foram preditas a partir de análises metagenômicas por Sessitsch (2012). Essas características incluem a presença de flagelos, enzimas degradadoras de polímeros vegetais, sistemas de secreção de proteínas tipos VI, aquisição e armazenamento de ferro, *quorum sensing*, desintoxicação de espécies reativas de oxigênio (ROS) e degradação de compostos aromáticos. Reinhold-Hurek (2011) acrescenta que características de superfície, como a composição de lipopolissacáridos ou pili de tipo IV, além das já citadas enzimas de degradação de polímeros vegetais, como celulases ou pectinases, são necessárias para uma eficiente colonização endofítica. O pilus do tipo IV de retração é utilizado não só para ligação, mas particularmente, para motilidade de contração em superfícies sólidas, o que depende da retração mediada por *PilT* do pili (Reinhold-Hurek, 2015).

Propriedades como essas podem definir a capacidade endofítica de um microrganismo e são distribuídas em diferentes grupos bacterianos como *Azoarcus, Bacillus, Burkholderia, Enterobacter, Pseudomonas, Serratia* e *Stenotrophomonas* (Ali et al., 2014; Buschart et al., 2012; Cai et al., 2017; Malfanova et al., 2013). Neste estudo, duas espécies bacterianas foram exploradas quanto a suas características endofíticas *Bacillus velezensis* e *Serratia marcescens.*

2.2. *Bacillus velezensis*

Espécie bacteriana gram-positiva, do gênero *Bacillus,* que pode ser encontrada em diversos habitats no meio ambiente, incluindo solo, água e plantas (Fan et al., 2017; Ruiz-García et al., 2005). Isolados associados a plantas podem deter características como promoção do crescimento de plantas, produção de metabólitos antimicrobianos e colonização do tecido interno de plantas (Cai et al., 2017 Liu et al. 2017, Borriss, 2011). Comumente encontrada como antagonista de fitopatógenos, é uma espécie promissora no desenvolvimento como agente de biocontrole em plantas, já que a maioria dos isolados de *B. velezensis* está associada a plantas, porém poucos descritos como endofíticos (Cai et al., 2017; Dunlap et al., 2016).

Apesar disso, uma confusão na nomenclatura e taxonomia desse grupo foi destaque nos últimos anos. Em 2016, Dunlap e colaboradores revelaram uma conclusão por análises filogenômicas que colocam as espécies *Bacillus methylotrophicus*, *Bacillus amyloliquefaciens* subsp. plantarum e *Bacillus oryzicola* como sinônimos heterotípicos de *Bacillus velezensis*. A partir daí, diversas reclassificações referentes a isolados desse grupo foram apresentadas.

2.3. *Serratia marcescens*

A *Serratia marcescens*, espécie bacteriana da família *Enterobacteriaceae*, é um bacilo gram-negativo que apresenta uma diversidade de nichos e comportamentos, incluindo isolados como patógenos humanos (Iguchi et al. 2014) e uma grande variedade de isolados ambientais, como endofíticos (Khan, 2017; Li, 2015) e fitopatogênicos (Wang, 2015). Esse grupo bacteriano caracteriza-se por seu comportamento ubíquo, podendo estar presente tanto na superfície quanto no interior de tecidos vegetais.

Duas cepas endofíticas, *S. marcescens* RSC-14 e *S. marcescens* FS14 possuem genoma completo depositado no banco de dados do NCBI. Ambas possuem propriedades benéficas a favor das suas plantas hospedeiras, tanto na promoção de crescimento e biocontrole contra fitopatógenos à tolerância a ambientes contaminados por cádmio (Cd) (Khan, 2017; Li, 2015).

2.4. Ciências genômicas

A ciência genômica se inicia com o surgimento da tecnologia de sequenciamento de DNA, que desenvolveu os primeiros métodos na década de 1970, sendo um dos pioneiros baseado no didesoxinucleotídeos (Sanger et al., 1977). Apesar de todo avanço, apenas em 1995 os primeiros genomas de organismos celulares foram completamente sequenciados, as bactérias *Haemophilus influenzae* e *Mycoplasma genitalium* (Fleischmann et al., 1995; Fraser et al., 1995). O Projeto Genoma Humano, iniciado em 1989, com cerca de 3 bilhões de bases, durou 13 anos e envolveu vários laboratórios e centros de pesquisa de diversos países (Venter et al., 2001). Esse grande projeto foi desenvolvido em torno de dois objetivos chave: construir mapas genéticos e físicos dos genomas humano e de camundongos; e sequenciar os genomas de leveduras e organismos menores como teste para sequenciar o genoma humano, maior e mais complexo (International Human Genome Sequencing Consortium-IHGSC, 2001). Desde então a ciência genômica e a bioinformática estão em desenvolvimento com avanços e novas abordagens.

Atualmente, os domínios *Bacteria* e *Archaea* apresentam maior número de projetos genoma concluídos, principalmente devido ao menor tamanho e a complexidade do genoma. Contudo, a quantidade de projetos "genoma" de *Bacteria* (101.009) e *Eukarya* (40.786) são os maiores até o início de 2018, de acordo com dados coletados em abril de 2018 no *Genomes Online Database* - GOLD (https://gold.jgi.doe.gov/statistics) (Figura 1). O GOLD é um sistema de gerenciamento de dados manualmente curados, que cataloga projetos de sequenciamento com metadados associados de todo o mundo (Mukherjee, 2016).

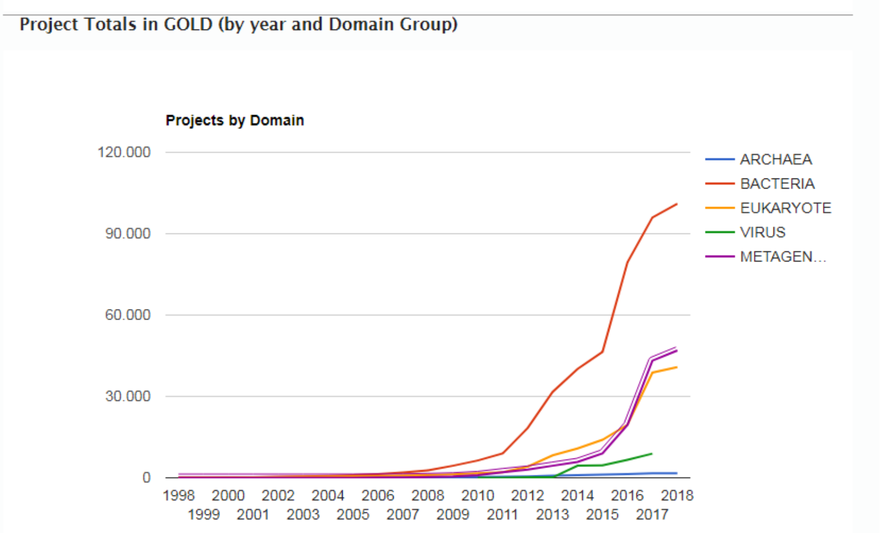


Figura 1. Distribuição dos projetos genoma de Archaea, Bacteria, Eukarya, Vírus e Metagenoma registrados no Genomes Online Database (GOLD) (https://gold.jgi-psf.org/distribution) até abril de 2018.

A partir de 2001, quando foi publicado o *draft* do genoma humano por Venter e colaboradores (2001) os esforços para métodos de sequenciamento mais baratos e mais rápidos aumentaram consideravelmente. Dessa forma o surgimento dos métodos de sequenciamento de nova geração (NGS – *Next Generation Sequencing*) foi impulsionado logo após a conclusão do projeto genoma humano (Grada & Weinbrecht, 2013) e aplicações a partir de NGS tornaram-se inesgotáveis, especialmente nas grandes áreas que envolvem a biotecnologia.

Na última década, várias plataformas NGS foram desenvolvidas e fornecem sequenciamento de baixo custo e alto rendimento. Os sequenciadores de alto rendimento surgiram em 2005 com a introdução do pirosequenciamento da Roche/454 (Margulies et al., 2005) seguido por outras plataformas, como Illumina (Solexa) e SOLiD (Applied Biosystems), que produziram maiores números de *reads*. Em 2010, a plataforma Ion Torrent PGM (e Proton mais adiante) destaca-se por menor custo, maior velocidade e equipamento mais compacto, usando tecnologia de semicondutores, que não depende da detecção óptica de nucleotídeos incorporados por fluorescência. No mesmo ano, o PacBio se destaca com fragmentos longos e ideais para montagem *de novo* de genomas (van Dijk, 2014). Mais recentemente, a Oxford Nanopore Techniques (ONT) desenvolveu o sequenciamento por nanoporos, como o MinION. Os últimos dois métodos não incluem um passo de amplificação durante a preparação da biblioteca, permitindo o sequenciamento de uma única molécula (Bleidorn, 2015). As tecnologias de sequenciamento apresentam diferenças que incluem, basicamente, o mecanismo de processamento, o tamanho, volume e qualidade dos fragmentos gerados e o custo operacional.

A plataforma Ion Torrent representa um sequenciamento de semicondutores iônicos que se baseia nas mudanças de concentração do íon de hidrogênio. Sempre que um nucleotídeo é incorporado numa cadeia de DNA, um íon de hidrogênio (ou próton) é liberado e detectado em temo real. Os poços onde ocorre o sequenciamento são preenchidos com um nucleotídeo de cada vez e a detecção de alterações no pH permite inferir se, e quantas bases foram incorporadas numa sequência de leitura (Bleidorn, 2015).

2.4.1. Bioinformática

Com os avanços na biologia molecular e os primeiros sequenciamentos de DNA houve a necessidade do tratamento dos dados gerados por meio de computadores, de forma automatizada. Em 1977, o primeiro pacote de programas específicos para uso genômico foi desenvolvido, o Staden Package (Staden, 2000). Atualmente a bioinformática é ferramenta essencial na interpretação de dados e criação de hipóteses em torno deles. A criação de bancos de dados para armazenar e gerenciar informações genômicas progrediu junto aos primeiros genomas completos sequenciados (Hutchison III, 2007).

O banco de dados GenBank foi criado em 1982 pelo National Institutes of Health (NIH), vinculado ao National Center for Biotechnology Information (NCBI), centralizando todas as sequências biológicas em um repositório acessível à comunidade (Bilofsky & Burks, 1988). Atualmente o GenBank é parte do International Nucleotide Sequence Database Collaboration, que compreende: DNA DataBank of Japan (DDBJ), European Nucleotide Archive (ENA) e o próprio GenBank (NCBI). Essas três organizações compartilham dados diariamente (Cochrane et al., 2016).

A informatização foi impulsionada com os crescentes bancos de dados de informação biológica, com a necessidade de desenvolvimento de novas abordagens para análise e apresentação dos dados e complexidade das investigações. A bioinformática consolida-se como nova área de conhecimento científico, com aplicações como reconhecimento de sequências gênicas, predição de configuração tridimensional de proteínas, identificação de inibidores e promotores de regiões codificantes, organização e relação de funções biológicas, análises de expressão gênica, estudos filogenéticos, entre outros (Verli, 2014). Ferramentas computacionais evoluíram para acompanhar essas aplicações atreladas as quantidades progressivas de dados biológicos, com maior capacidade de armazenamento, maior velocidade de processamento e menores custos.

A união da biotecnologia e informática é parte essencial nas análises funcionais, estruturais e interativas dos estudos genômicos, proteômicos e transcriptômicos. Gradativamente a bioinformática tem sua ênfase principal migrando do acúmulo de dados para a interpretação deles, como a criação de ferramentas promissoras que conduzem a descrição e análise de genomas.

2.5. Montagem de genomas

O processo de montagem do genoma consiste num conjunto de procedimentos para organizar (mapear) o grande número de fragmentos (*reads*) de DNA gerados no sequenciamento. Esse processo ocorre a partir do agrupamento dos fragmentos em sequências contíguas (*contigs*) e ordenadas (*scaffolds*). Os *contigs* são resultados do alinhamento das múltiplas sequências de *reads* em um consenso, enquanto os *scaffolds*, também chamados de *supercontigs* ou *metacontigs*, definem a ordem e orientação dos contigs no cromossomo (Miller, et al. 2010).

Existem duas abordagens diferentes na montagem de genomas, por referência e montagem *de novo*. A montagem por referência ocorre por mapeamento dos fragmentos de leitura contra uma sequência genômica conhecida como referência, preferencialmente do mesmo grupo taxonômico. Na montagem *de novo* (latim: desde o princípio), vários métodos são utilizados para agrupar as leituras com base no alinhamento entre elas. A abordagem por “Overlap-layout-consensus” (OLC) baseia-se em grafo de sobreposição, funcionando em três etapas. Inicialmente encontra-se a sobreposição entre as *reads*, cria-se um layout e, por fim, permite a inferência de contigs por alinhamento múltiplo de sequências (MSA) (He et al. 2013). Newbler (Margulies et al., 2005), CAP3 (Huang & Madan, 1999) e Mira (Chevreux, 1999) estão entre os montadores baseados no método OLC.

A estratégia De Bruijn Graph (DBG) há uma quebra dos *reads*, em fragmentos ainda menores (*k-mers*) e então, por sobreposição de k-1 os grafos que reconstroem as sequências contidas no genoma são determinados. Dispensa a sobreposição entre todos, bem como o armazenamento de k-mers individuais e suas informações e, portanto, esta abordagem requer menor esforço computacional (He et al. 2013). Para o DBG, os montadores Velvet (Zerbino e Birney, 2008), SPAdes (Bankevich et al., 2012), SOAPdenovo (Luo et al., 2012) e AbySS (Simpson et al., 2009) estão entre os mais usados. Na verdade, uma diversidade de montadores tem capacidades diferentes quanto aos tamanhos de fragmentos, formatos de arquivos e complexidade do genoma. Além disso, existem também alguns montadores comerciais como o CLC Workbench (www.clcbio.com) e SeqMan (www.dnastar.com) (Miller et al., 2010).

Após a montagem, alguns parâmetros irão determinar a qualidade da montagem como cobertura, conjunto de contigs produzidos, valor de N50, maior e menor contig, total de contigs e total de bases obtidas após montagem. A cobertura genômica refere-se ao número de vezes que uma determinada região do genoma é alinhada por segmentos de leitura, que contribui para aumentar a acurácia de identificação da sequência de DNA na região considerada. E o valor de N50 é baseado na metade do genoma representada em contigs maiores que esse valor (Ekblom & Wolf, 2014). Ou seja, para calcular o N50, ordenam-se os contigs de forma decrescente, somando a partir do maior até que se atinja no mínimo a metade do tamanho total da montagem. O processo de scaffolding (ordenação dos contigs) pode aumentar bastante a média do tamanho dos contigs e, consequentemente, o N50.

Uma vez que diferentes abordagens podem gerar um padrão diferente nos resultados, a mistura de mais de um tipo de montador pode contribuir para finalizar melhor um genoma. É possível realizar abordagens híbridas para a montagem de genomas, além da geração de novo dos contigs, e uma ordenação e orientação com uso de um genoma referência (Miller et al., 2010).

2.6. Análises comparativas de genomas

Os notáveis avanços na tecnologia de sequenciamento, com redução de tempo e custos, e consequentemente, uma massiva quantidade de dados genômicos, dão uma maior oportunidade para explorar análises genômicas *in silico*. Variadas abordagens e ferramentas tem se tornado cada vez mais comuns para essas análises e desafiam a classificação filogenética tradicional. A identidade média de nucleotídeos (Average Nucleotide Identity - ANI) aliada ao cálculo de distância entre genomas (genome to genome distance calculator - ggdc) ou hibridização DNA-DNA digital (dDDH) ou *in silico* podem representar poderosas abordagens para determinar identidade filogenômicas. O ANI representa uma média de valores de identidade/similaridade entre regiões homólogas do genoma, ou *core* do genoma (Kim, 2014). De outra forma, o dDDH inclui a análise da parte variável do genoma (Chaudhry, 2016). O *cutoff* indicado em análises de dDDH é de 70% para delimitação de espécie, e de 95% (+-0,5%) para valores de ANI, indicando opções de classificação taxonômica de genomas que tiverem a sequência disponível (Arahal, 2014).

O Pangenoma representa todo o repertório genômico de um grupo filogeneticamente definido e sua análise pode investigar heterogeneidade e diversificação do genoma de uma dada espécie bacteriana (Basharat, 2016). O pangenoma pode ser classificado como aberto ou fechado de acordo com a lei de Heaps, onde $\alpha \leq 1$ representa um pan genoma aberto e $\alpha > 1$ indica um pangenoma fechado (Tettelin et al. 2008). Ou seja, quando o número de genes aumenta com adição de genomas na análise, temos um pangenoma aberto; enquanto que, se o tamanho do pangenoma (número de genes) se mantem constante mesmo com mais genomas considerados, trata-se de um pangenoma fechado.

A variabilidade de genômica pode ser originada a partir da presença de elementos genéticos móveis, como ilhas genômicas (GIs). GIs são entidades genéticas de provável origem horizontal, normalmente com tamanho médio > 8 kb em genomas de Bacteria e Archeae (Langille, 2010), que contribui para rápida evolução e vantagens de sobrevivência (Lahiri, 2014). Essas podem direcionar uma estratégia na busca por particularidades de cepas de um mesmo grupo com comportamentos distintos.

A mineração e comparação de conteúdos genômicos entre linhagens associadas, podem representar também um direcionamento na compreensão comportamental de diferentes organismos. A elucidação de mecanismos associados à interação planta-microrganismo a partir de análises genômicas já tem sido observada em trabalhos com endofíticos bacterianos como nos gêneros *Burkholderia*, *Bacillus, Pseudomonas, Serratia, Staphylococcus* (Reinhold-Hurek, 2011; Mitter et al., 2013; Liu et al., 2017; Khan et al., 2017; Chaudhry et al., 2017).

3. Objetivos

3.1. Objetivo geral

Estudos genômicos das bactérias endofíticas *Bacillus velezensis* 629 e *Serratia marcescens* 1274, visando a caracterização genômica e identificação de genes relacionados ao processo endofítico em bactérias.

3.2. Objetivos específicos

- Sequenciamento, montagem e anotação genômica da bactéria endofítica *Bacillus velezensis* 629;
- Estudos genômicos do *Bacillus velezensis* 629, incluindo análises de taxonomia do genoma, caracterização quanto à capacidade endofítica e análises comparativas do genoma;
- Sequenciamento, montagem e anotação genômica da bactéria endofítica *Serratia marcescens* 1274;
- Estudos genômicos do *Serratia marcescens* 1274, incluindo análises de taxonomia do genoma, caracterização quanto à capacidade endofítica e análises comparativas do genoma.

Manuscrito 1

High-quality draft genome sequence of *Bacillus amyloliquefaciens* strain 629, an endophyte from *Theobroma cacao*

Brena M. M. SantAnna[a], Phellippe P. A. Marbach[b], Marcelo Rojas-Herrera[c], Jorge T. De Souza[d], Milton R. A. Roque[a], Artur T. L. Queiroz[e] *

Universidade Federal da Bahia, UFBA – Brazil[a]; Universidade Federal do Recôncavo da Bahia, UFRB – Brazil[b]; Centro de Genómica y Bioinformática, Universidade Mayor – Chile[c]; Universidade Federal de Lavras, UFLA – Brazil[d]; CPqGM-Fiocruz/BA – Brazil[e]

* Address correspondence to Artur T. L. Queiroz, artur.queiroz@bahia.fiocruz.br

Abstract

Bacillus amyloliquefaciens strain 629 is an endophyte isolated from *Theobroma cacao* L. Here we report the draft genome sequence (3.9Mb) of *B. amyloliquefaciens* strain 629 containing 16 contigs (*3,903,367* bp), 3,912 coding sequences, and an average 46.5% GC content.

Keywords: Bacilli; endophytic bacterium; genome sequencing

Genome announcement

Bacilli are frequently isolated as endophytes and are common components of the microbiota of several plant species (1, 2). Strain 629 was isolated from a healthy *Theobroma cacao* tree and was initially identified as *Bacillus subtilis* (3), but further analysis based on *gyr*A and *rec*A sequences revealed that its true identity is *Bacillus amyloliquefaciens* (4). This isolate colonizes different host and plant tissues under both sterile and non-sterile conditions and promotes plant growth (3, 4). Strain 629 produces the lipopeptides iturin, fengicin, and surfactin and volatile organic compounds that may be active in the biocontrol of several fungal plant pathogens (Unpublished data) and pathogenic bacteria, including *Curtobacterium flaccumfaciens* pv. *flaccumfaciens* (5). Furthermore, *B. amyloliquefaciens* 629 is currently being used as model to study endophytic colonization (4). This strain is deposited in the Biological Institute Culture Collection of Phytopathogenic Bacteria - IBSBF (Campinas, São Paulo, Brazil) under accession number IBSBF-3106. This collection is registered with the World Data Centre for Microorganisms collection under number WDCM-110.

Genomic DNA from isolate 629 was extracted and sequenced using Ion Torrent

PGM platform (Life Technologies) 318 chip. A total of 7,567,586 reads with an average length of 330 pb were obtained. All reads were assembled to an initial draft genome of 3,866,991 nucleotides at 443-fold coverage using the SPAdes Genome Assembler version 3.5.0, generating 129 unoriented contigs, with a GC content of 46.5%, (N50: 285,363 bp).

Contigs were ordered using CONTIGuator 2.3 (*http://contiguator.sourceforge.net/*) (6) with the *B. amyloliquefaciens* CC178 genome, the closest available, as a reference (GenBank accession CP006845.1). Subsequently, 34 contigs with 3,8Mb were aligned with the reference genome to order the contigs. A total of 95 contigs (only 9 >600bp) corresponding to 29,876 nucleotides were not mapped to the reference genome. These sequences were identified as redundant contigs, according to BLAST results, and were removed from the assembly. To solve the repetitive sequences and the remaining gaps the *MapRepeat pipeline* (7) was used, resulting in the final high-quality draft genome sequence with 16 contigs, containing 3,903,367 bp.

Genome annotation was performed with RAST version 2.0 server (8). The genome of strain 629 is composed of 4,013 predicted genes, including 3,912 protein-coding sequences, 82 tRNAs, and 19 copies of the genes for 5S, 16S, and 23S rRNA. The genome of strain 629 is closely related to that of *B. amyloliquefaciens* CC178 with an identity of 99% (97% coverage) and also has a similar numbers of predicted genes (9).

Subsequent analysis of the genome content of *B. amyloliquefaciens* 629 and its comparison with phylogenetically-related strains will help to determine key aspects of its interaction with the environment, plants, and other microorganisms.

Nucleotide sequence accession number. The *Bacillus amyloliquefaciens* strain 629 whole genome shotgun (WGS) project has been deposited at DDBJ/EMBL/GenBank under the accession no. **LGYP00000000**. The version described in this paper is the first version, **LGYP01000000**, and consists of sequences **LGYP01000001-LGYP01000016**.

Acknowledgements

The authors thank CAPES and CNPq for financial support through the PVE project n° 61/2011, Zayda M. Moreira, Artur Silva, Rommel T. J. Ramos and Luis

Pacheco for DNA extraction, sequencing and assembly assistance. JTS acknowledges his productivity scholarship.

References

1. White, J. F., Torres, M. S., Sullivan, R. F., Jabbour, R. E., Chen, Q., Tadych, M., Irizarry, I., Bergen, M. S., Havkin-Frenkel, D.; Belanger, F.C., 2014. Occurrence of *Bacillus amyloliquefaciens* as a systemic endophyte of vanilla orchids. Microsc. Res. Tech., 77: 874–885. doi: 10.1002/jemt.22410.

2. Wang, X.; Liang G., 2014. Control Efficacy of an Endophytic Bacillus amyloliquefaciens Strain BZ6-1 against Peanut Bacterial Wilt, Ralstonia solanacearum. BioMed Research Internationa, v. 2014, Article ID 465435. DOI: 10.1155/2014/465435.

3. Leite, H.A., Silva, A.B., Gomes, F.P., Gramacho, K.P., Faria, J.C., De Souza, J.T., Loguercio, L.L., 2013. *Bacillus subtilis* and *Enterobacter cloacae* endophytes from healthy *Theobroma cacao* L. trees can systemically colonize seedlings and promote growth. Appl. Microbiol. Biotechnol., 97(6):2639-2651. doi: 10.1007/s00253-012-4574-2.

4. Moreira, Z.M., Duarte, E.A.A., Oliveira, T.A.S., Monteiro, F.P., Loguercio, L.L., De Souza, J.T., 2015. Host and tissue preferences of *Enterobacter cloacae* and *Bacillus amyloliquefaciens* for endophytic colonization. African Journal of Microbiology Research, v. 9, p.1352-1356.

5. Martins, S.J., Medeiros, F.H.V., Souza, R.M., Resende, M.L.V., Ribeiro Jr., P.M., 2013. Biological control of bacterial wilt of common bean by plant growth-promoting rhizobacteria. Biological Control, 66:65-71. doi:10.1016/j.biocontrol.2013.03.009.

6. Galardini, M.; Biondi, E.G.; Bazzicalupo, M.; Mengoni, A., 2011. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. Source Code Biol Med., 6: 11. doi: 10.1186/1751-0473-6-11.

7. Mariano, D.C.B.; Pereira, F.L.; Ghosh, P.; Barh, D.; Figueiredo, H.C.P.; Silva, A.; Ramos, R.T.J.; Azevedo, V.A.C., 2015. MapRepeat: an approach for effective assembly of repetitive regions in prokaryotic genomes. Bioinformation 11(6): 276-279.

8. Overbeek R.; Olson, R.; Pusch, G.D.; Olsen, G.J.; Davis, J.J.; Disz, T., Edwards, R.A.; Gerdes, S.; Parrello, B.; Shukla, M.; Vonstein, V.; Wattam, A.R.; Xia, F.; Stevens, R., 2013. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). 42:D206–D214. 10.1093/nar/gkt1226.

9. Kim B.Y., Lee S.Y., Ahn J.H., Song J., Kim W.G., Weon H.Y., 2015. Complete genome sequence of *Bacillus amyloliquefaciens* subsp. *plantarum* CC178, a phyllosphere bacterium antagonistic to plant pathogenic fungi. Genome Announc 3(1):e01368-14. doi:10.1128/genomeA.01368-14.

Manuscrito 2

Novel insights on bacterial endophytic behavior from comparative genomic analysis of *Bacillus velezensis* strain 629

Brena Sant'Anna[a] , Artur Queiroz[b], Luis Pacheco[a], Jorge Souza[c], Ícaro Lopes[a], Felipe Rangel[a], Milton Roque[a] *

[a]Institute of Health Sciences, Federal University of Bahia, Salvador-BA, Brazil; [b]Gonçalo Moniz Institute - Oswaldo Cruz Foundation, Salvador-BA, Brazil; [c]Department of Phytopathology, Federal University of Lavras, Lavras-MG, Brazil.

*Corresponding author: Milton Roque, milton.roque@ufba.br - Tel: +55 71 32838893.

Abstract

Endophytic bacteria have a complex interaction with hosts and molecular basis of these interactions can be studied using comparative genomic. We have generated a high quality 3.9Mbp genomic sequence of *Bacillus velezensis* 629, taxonomy and comparative genomic analysis was used to better understand the pathways involved in the endophytic process and the *Bacillus* classification. Forty-eight complete genomes of *B. amyloliquefaciens* and *B. velezensis* were used in comparative analyses. Phylogenetic analysis based on five conserved genes (16S rRNA, *rpo*D, *gyr*B, *rec*A and *rpo*B) grouped the isolate 629 within plant-associated strains belonging to *B. amyloliquefaciens* or *B. velezensis* group. DNA-DNA hybridization (dDDH) and average nucleotide identity (ANI) values confirm identity between the 629 genomic sequence and *B. velezensis* strains. Therefore, we suggest a reclassification of isolate 629 from *B. amyloliquefaciens* to *B. velezensis*. We predicted that the *B. velezensis* pan-genome contains 8,428 genes, 2,479 into core-genome, 3,244 accessory genes and 1,997 singletons, indicates that the pangenome is open. When compared genomes from 15 different endophytic species, 121 genes were shared, furthermore, 8 genomic islands (GIs) were predicted with 148 genes. Distributed in the endophytic genome 629, many genes fit to lifestyles sharing with plant behavior like phytohormone pathways, nitrate reduction and phosphate solubilization.

Keywords: *Endophytic bacteria; Bacillus velezensis; Genome taxonomy; Pan-genome; Comparative genome analysis; Endophytic behavior.*

Acknowledgements: This work was supported by grants from the Brazilian Funding Agencie: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

1. Introduction

Endophytic bacteria are able to colonize the internal tissues of plants without causing morphological damage or disease. Associations of endophytic microorganisms with their hosts are diverse and complex and the mechanisms are not completely elucidated to date (Santoyo et al. 2016). Bacteria are involved in a plethora of plant-microbe interactions and their abilities to efficiently colonize plants is an essential feature for their uses as biological control agents and plant growth promoters. However, the lack of knowledge about the genetic mechanisms involved in endophytic colonization is one of the reasons why only a few endophytic microorganisms become commercial products for field use. The ability of bacteria to metabolize plant compounds may be the prerequisite for successful endophytic establishment (Malfanova et al. 2013).

Bacterial species of the "operational group *Bacillus amyloliquefaciens*", that was proposed to include the species *B. amyloliquefaciens*, *B. velezensis*, and *B. siamensis* (Fan et al. 2017), are common plant-associated bacteria and, due to their positive effects on resistance to plant diseases, especially those caused by fungi, they have gained considerable attention (Chen et al. 2009; Zouari et al. 2016). There are some taxonomic complexities concerning these species, particularly for the classification of the subspecies *plantarum* of *Bacillus amyloliquefaciens*. According to Dunlap and collaborators (2016), a reclassification based on phylogenomics proposed *Bacillus amyloliquefaciens* subsp. *plantarum, Bacillus methylotrophicus* and *Bacillus oryzicola* as later heterotypic synonyms of *Bacillus velezensis*. *Bacillus velezensis* is an environmental bacterium with multiple biological functions, especially related to plants, such as plant growth promotion (Chen et al. 2014), biocontrol through antimicrobial activity (Cai et al. 2017; Gao et al. 2016) and improvement of plant resistance against pest insects (Rashid et al. 2017).

The *B. amyloliquefaciens* strain 629, initially identified by *16S rDNA*, *gyrA* and *recA* gene sequencing, was isolated from *Theobroma cacao* leaves, and has been shown to promote cacao growth and to control bacterial and fungal infections (Falcäo et al. 2014; Martins et al. 2015). The ability of the 629 strain to produce different lipopeptides when in contact with plant-derived materials has also been demonstrated (Monteiro et al. 2016) and a high quality 3.9Mbp genomic sequence was recently made available by our group (SantAnna et al. 2015). Now, we provide genomic taxonomy data that support the reclassification of the *B. amyloliquefaciens* strain 629 as belonging to the species *Bacillus velezensis*. We also performed a comparative genomic analysis of various plant-associated *B. amyloliquefaciens* and *B. velezensis* strains, what gives novel insights on the mechanisms accounting for the endophytic behavior of strain 629.

2. Materials and Method

2.1. Genome Sequences

The complete genome sequences of 48 strains, 20 deposited as *Bacillus amyloliquefaciens* and 28 as *Bacillus velezensis,* were retrieved from the NCBI GenBank database (http://www.ncbi.nlm.nih.gov/genbank/) prior to April 2017), to compare with the genomic sequence of endophytic strain 629 (GenBank ID: LGYP00000000.1) (Supplementary Table S1). The selected genomes included 41 plant-associated strains (*B. velezensis* and *B. amyloliquefaciens*) and 7 *B. amyloliquefaciens* strains not associated with plants: DSM7, TA208, XH7, LL3, RD7-7, MBE1283 and KHG19.

2.2. Phylogenetic Analysis and Genomic Taxonomy

Phylogenetic analysis was performed based on concatenated sequences from five genes: *16S rRNA*, *rpoD*, *gyrB*, *recA* and *rpoB*, all retrieved from the 48 genome sequences of *B. amyloliquefaciens* and *B. velezensis* and the endophytic strain 629. Phylogenetic trees were built using PhyML (Guindon and Gascuel 2003) with a distance-based ML method, with 1,000 bootstrap iterations.

The taxonomic reclassification of the endophytic strain 629 was performed using a genomic taxonomy approach between strain 629 and the reference genomes. *In silico* or digital DNA-DNA hybridization (dDDH) was combined with Average Nucleotide Identity (ANI) by BLAST to construct an ANI-Matrix estimating all-vs-all distances in a collection of genomes to build similarity clustering (available at http://enve-omics.ce.gatech.edu/ani/index); dDDH was done using GGDC 2.1 server, Genome-to-Genome Distance Calculator (available at http://ggdc.dsmz.de/ggdc.php#) (Meier-Kolthoff et al. 2013).

2.3. Comparative Genome Analysis

2.3.1. Pan-Genome Analysis

The *Bacillus velezensis* Pan-Genome (including core genes, accessory genes and singletons) was analyzed PanWeb (Pantoja et al. 2017). The PanWeb tool (http://www.computationalbiology.ufpa.br/analysis.php) is based on the PGAP pipeline (Zhao et al. 2012). The parameters used were as follows: 0.8 for identity; 0.8 for coverage; E-value cutoff = <1E-5; and Multiparanoid (MP) method.

For the PanWeb analysis, the 44 *B. velezensis* genomic sequences retrieved from GenBank were re-annotated using a RAST (Rapid Annotation using Subsystem Technology) server (version 2.0) to standardize genomic annotations.

2.3.2 Genomic Islands Prediction

Genomic island (GI) distribution among the *B. velezensis* strains was predicted with IslandViewer 4 (Bertelli et al. 2017). This method predicts GIs based on two sequence composition methods, IslandPath-DIMOB and SIGI-HMM, and a comparative method called IslandPick (Zhang et al. 2015).

2.3.3 Secondary metabolite prediction

The gene clusters required for secondary metabolite biosynthesis were predicted by antiSMASH (Medema et al. 2011). Biological functions of the identified genes were predicted using Gene Ontology annotations.

2.3.4 Prediction of genes involved in endophytic behavior

The complete gene content of fifteen genomic sequences of endophytic bacterial species (Supplementary Table S2) was compared with the gene content of the 629 strain using EDGAR Server 2.1 (Blom et al., 2016) to identify essential biological functions that could potentially contribute to endophytic behavior. In addition, a thorough analysis of the genome of strain 629 was performed to identify the genes involved in the endophytic process as previously reported in literature.

3. Results and Discussion

3.1. Taxonomic reclassification of the endophytic strain 629

The molecular phylogenetic tree, based on five conserved concatenated genes - *16S rRNA*, *rpoD, gyrB, recA* and *rpoB,* revealed that strain 629 clustered with *B. velezensis*

FZB42, TrigoCor1448, 9D-6 and plant-associated *B. amyloliquefaciens* CC178 (Figure 1). All these lineages are associated with plants, and all were formerly considered as subspecies *plantarum,* but are currently classified as *B. velezensis.* Corroborating with the ANI values and dDDH analysis by ggdc, which has been replacing conventional DDH analysis. The calculated values of ANI (>95%) and ggdc (formula 2 >70%) (Meier-Kolthoff et al. 2013) indicate a greater identity within the genomes of *B. velezensis* and the plant-related genomes of *B. amyloliquefaciens* (Figure 2 and Table 1). Our genomic taxonomy analysis suggests the reclassification of the endophytic 629 strain from *B. amyloliquefaciens subsp. plantarum* to *Bacillus velezensis*. In fact, the former taxonomic classification is considered a later heterotypic synonym of *B. velezensis* according to Dunlap et al. (2016).

Bacterial species can be better delineated when placed together with a tetranucleotide (Tetra) frequency correlation coefficient (Scortichini et al. 2013) species boundary threshold >0.989 (Fan et al. 2017). Tetra analysis corroborated the results obtained by ANI and ggdc (or dDDH) with values higher than 0.998 for *B. velezensis* as well as most of the *B. amyloliquefaciens* isolates (Table 1).

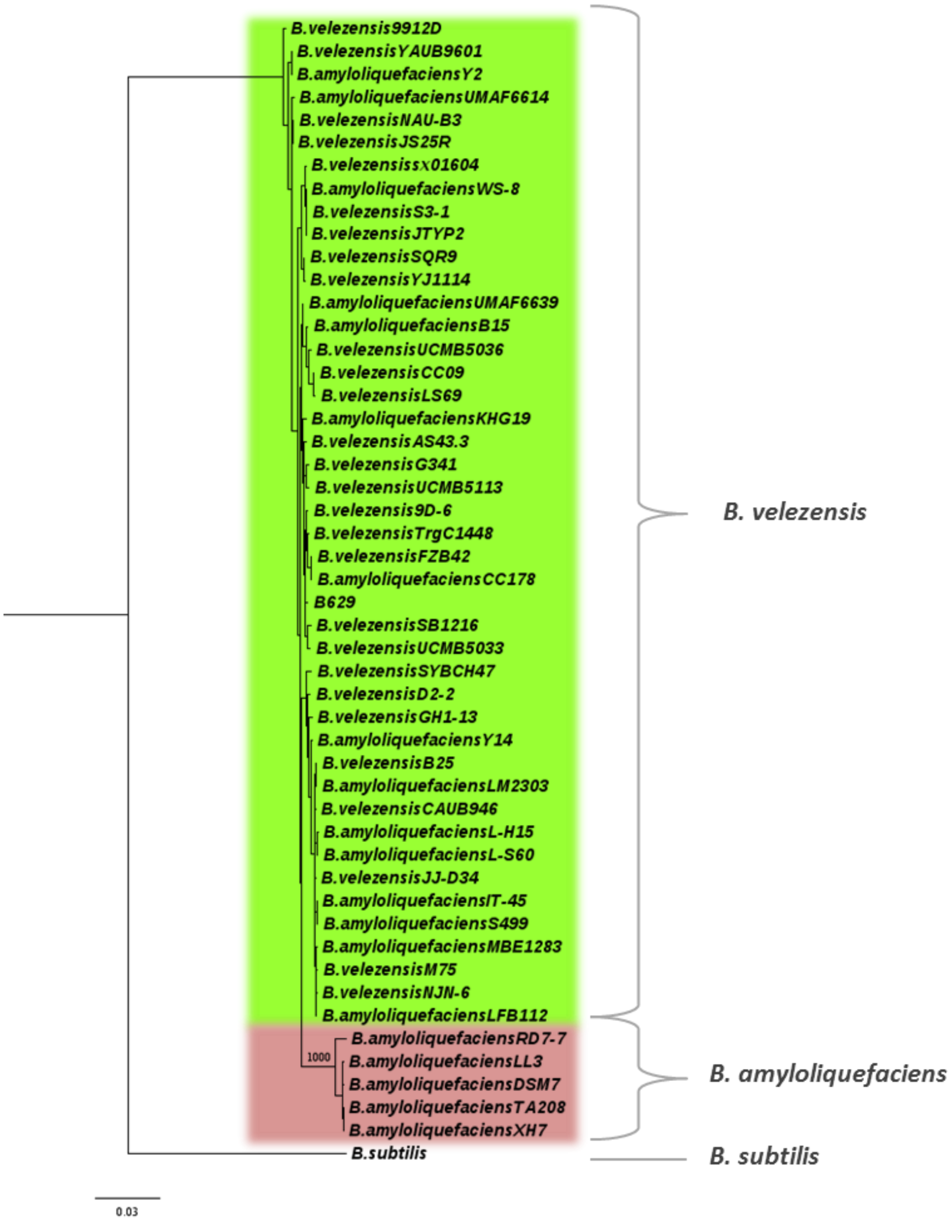


Fig. 1 Phylogenetic tree based on concatenated sequences from five genes, *16S rRNA*, *rpoD*, *gyrB*, *recA* and *rpoB*, showing the phylogenetic relationship between members of the *B. amyloliquefaciens* and *B. Velezensis*. This tree was generated with PhyML (bootstrap 1,000) under the model GTR selected by jModelTest.

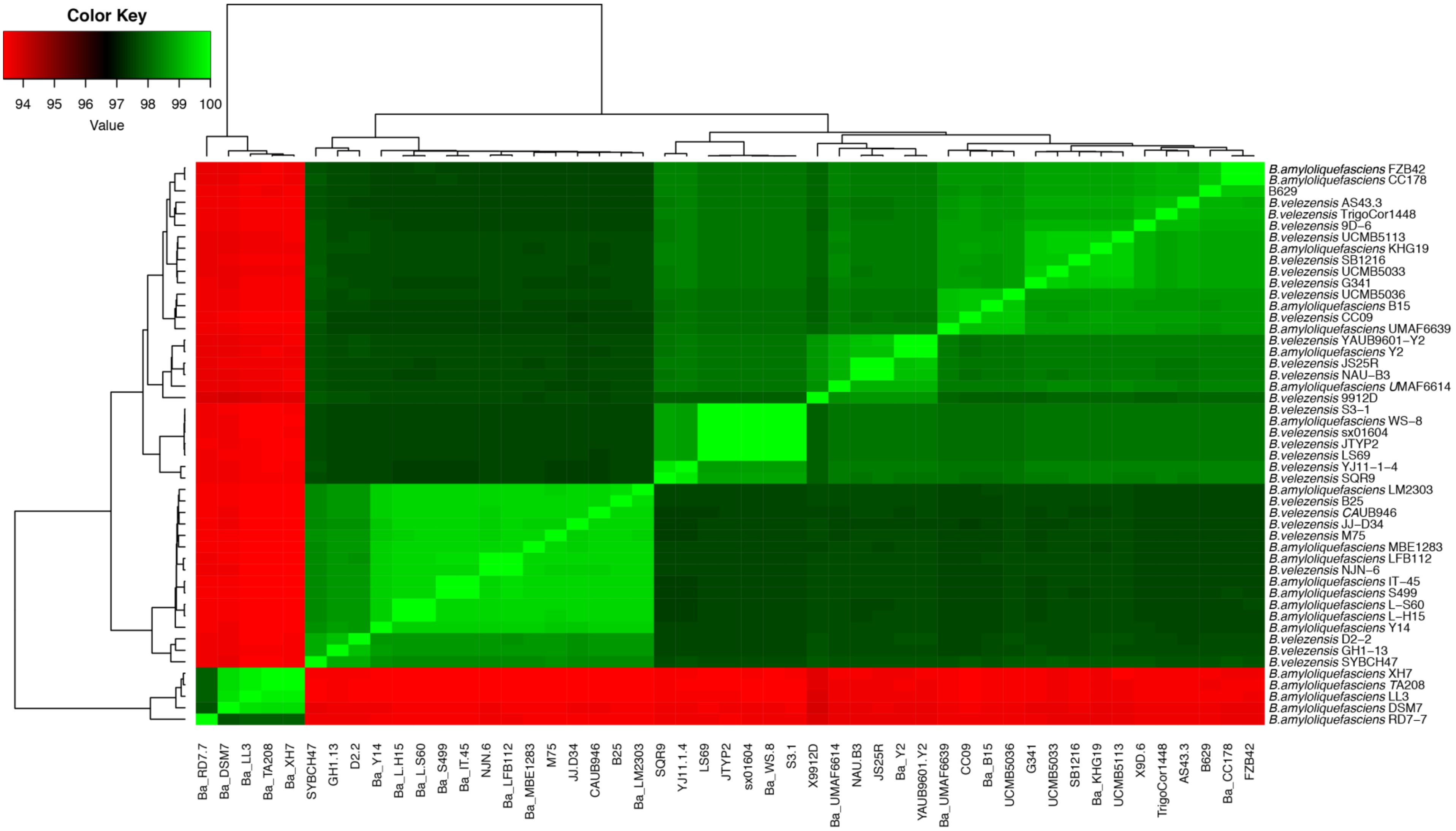


Fig. 2 Heat-map of Average Nucleotide Identity (ANI) values amongst different strains of *B. amyloliquefaciens* and *B. velezensis*. Strain 629 is B629 and five strains of *B. amyloliquefaciens* are grouped separately (Red) by having very low ANI values. Names and features of strains are in Supplementary Table 1.

Table 1
Comparative genomic analysis by ANI, TETRA and dDDH values (GGDC) to *B. velezensis* and *B. amyloliquefaciens* with strain 629 as query.

Query: *Bacillus* strain 629 (GC: 46.6%)				
Genomes	**%GC**	**GGDC**	**ANI**	**Tetra**
B. velezensis FZB42	46.5	94.20	99.23	0.99976
B. amyloliquefaciens CC178	46.5	94.20	99.24	0.99976
B. velezensis TrigoCor1448	46.5	91.50	99.00	0.99979
B. velezensis 9D-6	46.4	91.10	98.92	0.99956
B. velezensis AS43.3	46.6	91.00	98.94	0.99982
B. velezensis UCMB5113	46.7	90.30	98.81	0.99977
B. velezensis UCMB5033	46.2	90.20	98.80	0.99934
B. amyloliquefaciens KHG19	46.6	90.20	98.79	0.99977
B. velezensis G341	46.5	90.10	98.80	0.99963
B. velezensis SB1216	46.8	89.90	98.80	0.99949
B. velezensis UCMB5036	46.6	88.70	98.59	0.99977
B. velezensis CC09	46.1	88.20	98.60	0.99893
B. amyloliquefaciens UMAF6639	46.3	87.80	98.58	0.99970
B. amyloliquefaciens B15	46.5	87.80	98.57	0.99960
B. velezensis SQR9	46.1	86.80	98.36	0.99897
B. velezensis YJ11-1-4	46.4	86.70	98.34	0.99955
B. amyloliquefaciens UMAF6614	46.5	85.90	98.34	0.99963
B. velezensis YAUB9601-Y2	45.9	85.60	98.27	0.99802
B. amyloliquefaciens Y2	45.9	85.50	98.28	0.99805
B. velezensis NAU-B3	46.0	85.30	98.25	0.99818
B. velezensis JS25R	46.4	85.30	98.25	0.99946
B. amyloliquefaciens WS-8	46.5	84.80	98.19	0.99973
B. velezensis sx01604	46.5	84.80	98.19	0.99973
B. velezensis S3-1	46.5	84.80	98.19	0.99973
B. velezensis JTYP2	46.5	84.80	98.19	0.99973
B. velezensis LS69	46.5	84.70	98.18	0.99968
B. velezensis 9912D	46.0	82.90	98.00	0.99813
B. velezensis SYBCH47	46.4	81.40	97.76	0.99960
B. velezensis GH1-13	46.2	80.40	97.62	0.99911
B. velezensis D2-2	46.7	80.40	97.65	0.99956
B. velezensis M75	46.6	80.20	97.57	0.99959
B. amyloliquefaciens S499	46.6	80.20	97.59	0.99956
B. amyloliquefaciens IT-45	46.6	80.20	97.60	0.99957
B. velezensis NJN-6	46.6	80.10	97.57	0.99945
B. amyloliquefaciens LFB112	46.7	80.00	97.56	0.99960
B. amyloliquefaciens Y14	46.4	79.90	97.59	0.99951
B. velezensis CAUB946	46.5	79.80	97.57	0.99945
B. amyloliquefaciens LM2303	46.7	79.80	97.58	0.99934
B. velezensis B25	46.7	79.80	97.55	0.99959
B. velezensis JJ-D34	46.2	79.60	97.56	0.99878
B. amyloliquefaciens MBE1283	46.5	79.60	97.55	0.99957
B. amyloliquefaciens L-S60	46.7	79.50	97.56	0.99965
B. amyloliquefaciens L-H15	46.7	79.50	97.56	0.99965
B. amyloliquefaciens RD7-7	46.3	56.30	93.61	0.99867
B. amyloliquefaciens DSM7	46.1	55.70	93.62	0.99754
B. amyloliquefaciens XH7	45.8	55.40	93.50	0.99801
B. amyloliquefaciens TA208	45.8	55.30	93.55	0.99770
B. amyloliquefaciens LL3	45.7	55.30	93.54	0.99747

Tetra values were below the cutoff only when compared to the non-plant-related strains *B. amyloliquefaciens*, DSM7, TA208 and LL3, which also presented low levels of ANI and dDDH per ggdc (Table 1). The RD7-7 and XH7 strains also presented values below the cutoff. A study by Rückert et al. (2011) presented genomic differences between plant-associated and non-plant-associated *B. amyloliquefaciens* strains, proposing a taxonomic distinction between FZB42 (plant-related) and DSM7 (not plant-related) strains.

In this study, only the non-plant associated *B. amyloliquefaciens* strains DSM7, TA208, LL3, RD7-7 e XH7 remain with the same classification, the others are *B. velezensis* (Figure 1). However, *B. amyloliquefaciens* and *B. velezensis* are not possible to distinguish only based on the relation with plants, since the strains KHG19 and MBE1283 are as *B. velezensis* into phylogenetic analysis (Figure 1) and genomic taxonomy (Figure 2 and Table 1) and are not known as plant-associated. To 629 strain the ANI and dDDH values are high for species classification, respectively, 98.79% e 90.20% for KHG19 and 97.55% e 79.6% for MBE1283.

According to the values presented in Table 1, there is a pending process in the reclassification of some *B. amyloliquefaciens* strains to *B. velezensis*, especially from the *plantarum* subspecies. Almost half of the complete genomes deposited in NCBI as *B. velezensis* were *B. amyloliquefaciens* that were reclassified, strains: FZB42, CC09, AS43.3, TrigoCor1448, UCMB5113, UCMB5033, UCMB5036, 9D-6, CAU B946, YAU B9601-Y2, NAU-B3, SQR9, JS25R, NJN-6 and SYBC H47.

3.2. Pan-Genome of *B. velezensis*

The pan-genome analysis was performed in order to better understand the genome repertoire of *B. velezensis* 629 and to identify the genes potentially involved in the endophytic behavior of this species. Figure 3 and Table 2 presents the predictions of core genes, accessory genes and strain-specific genes (singletons) of 44 *B. velezensis*, according to the previous genome taxonomy results, using two different approaches.

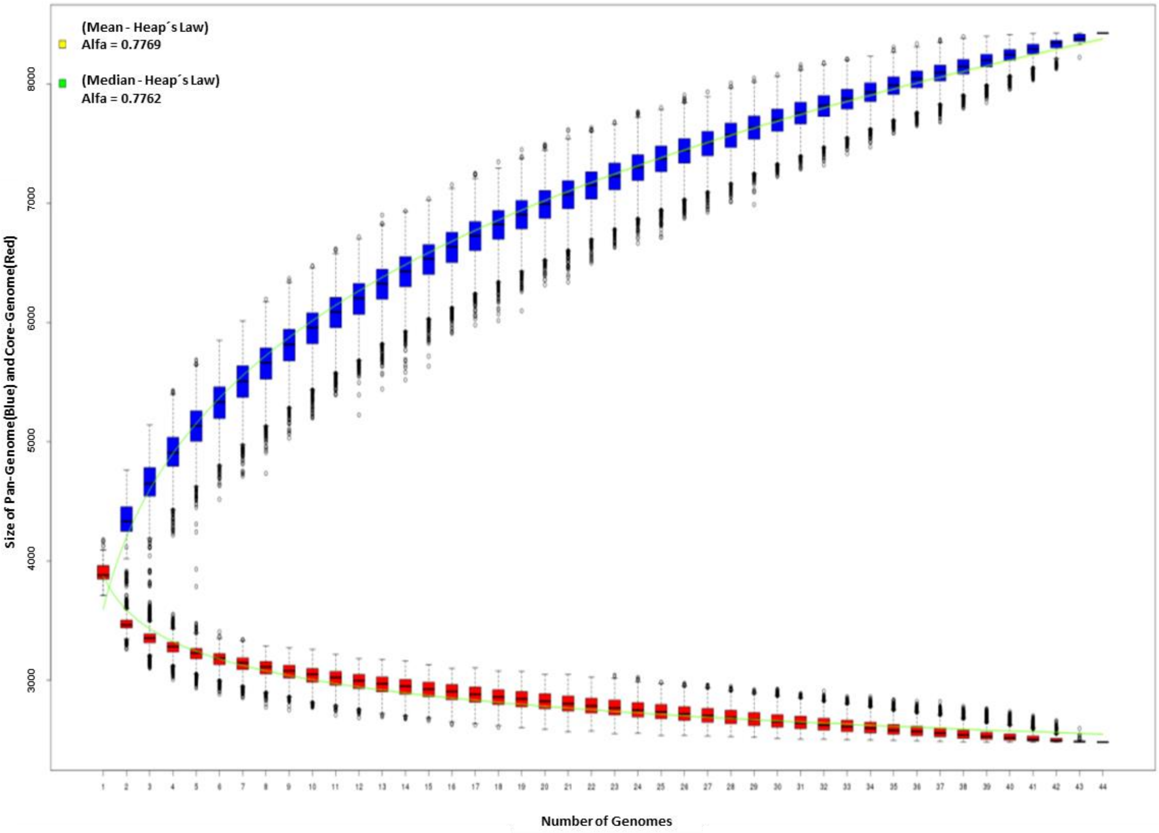


Fig. 3 Pan-genome and core-genome of *B. velezensis*. Graph representing the pan-genome (top raw) and core-genome (bottom row) of the 44 analyzed genomes. Also shows the α coefficient value of Heap’s Law, mean or median, less than 1, which mean an open pan-genome.

Table 2
Bacillus velezensis pan-genome analysis performed by PanWeb server based on 44 genomes of this species.

B. velezensis	**PanWeb (PGAP)** Number of genes
Pan-genome	8,428
Core-genome	2,479
Accessories	3,952
Singletons	1,997

Figure 3 shows the pan-genome and core-genome of *B. velezensis*, as well as the α coefficient value defined by curve fitting based on Heap's Law using the medians and means of the distributions. According to Heaps’ Law, α ≤ 1 represents an open pan-genome and α > 1 indicates a closed pan-genome (Tettelin et al. 2008). Thus, the pan-genome of *B. velezensis,* based on 44 genomes, is open.

The *B.velezensis* pan-genome prediction by PanWeb showed a higher number of total genes (8,428) and detailed results are provided in Supplementary Figure S1. This analysis indicated that most genes are dispensable (~46%), which evidences high variability within the group.

3.3. Presence of Genomic Islands in the Genome of *B. velezensis* 629

Genomic islands (GIs) are genetic entities of probable horizontal origin that are usually > 8 kbp in size in bacterial and archaeal genomes and contribute to rapid evolution and survival advantages (Langille et al. 2010). The horizontal acquisition of GIs in different isolates of *B. velezensis* may be a contributing factor for the open structure of the pan-genome and can be a source of acquisition of biological functions related to the endophytic behavior. Our analysis predicted 8 GIs ranging between 4 and 68 kbp in the *B. velezensis* 629 genome, supported by at least one method (SIGI-HMM, IslandPath-DIMOB, IslandPick) (Figure 4). These GIs include 148 genes, of which 47 (31%) code for hypothetical proteins. All the genes on the GIs were displayed in the Supplementary Table S3. Six genes present in GIs are shared with endophytic *B. velezensis* CC09: beta-lactamase A, histidine kinase, *cwlD*, *msrB*, *yncF* and *yokF* (Supplementary Table S4). Importantly, genes encoding transcription regulatory proteins were identified in 4 GIs from *B.velezensis* 629, which are involved in transcriptional networks related to quorum sensing, response to stress and sporulation (Supplementary Figure S2).

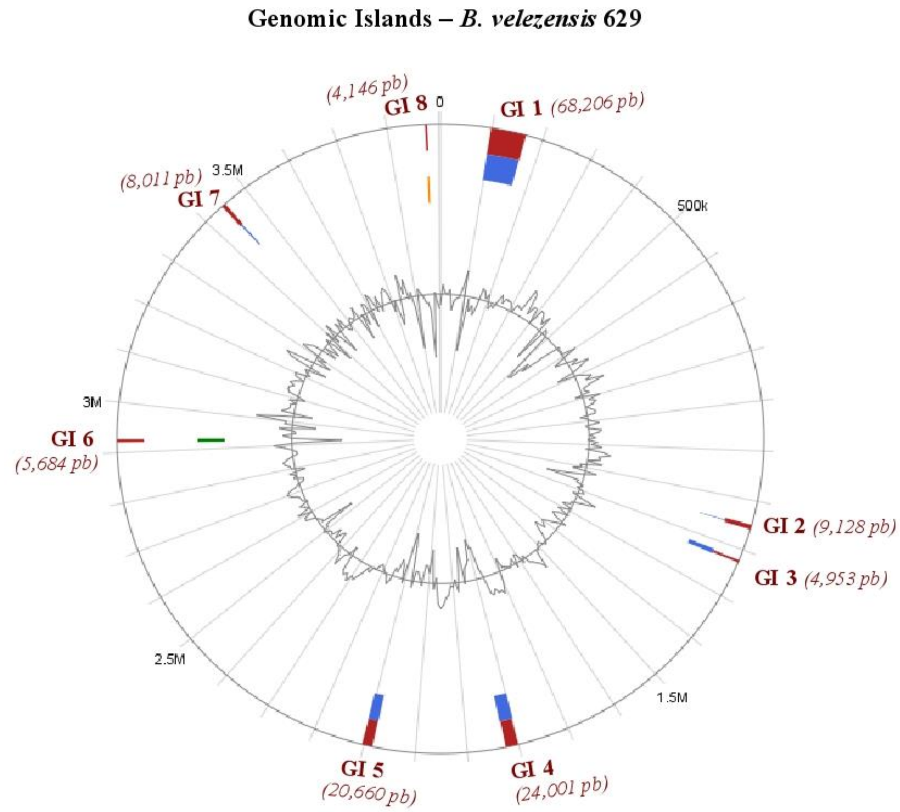


Fig. 4 Circular visualization of Genomic Islands (GIs) prediction of *B. velezensis* 629 by IslandViewer 4. More than 140kb distributed in 8 GIs were predicted. GIs prediction by IslandPath-DIMOB, SIGI-HMM, and IslandPick approaches.

3.4. Secondary metabolites biosynthesis

Twelve gene clusters with secondary metabolites biosynthesis (about 19% of the total genome) were identified in the *B. velezensis* 629 genome. These encode Lantipeptides, Nrps, Terpenes, Transatpks, Transatpks-Nrps, T3pks, Bacteriocin-Nrps and others (Table 3). There are also gene clusters responsible for the synthesis of surfactin, bacillaene, macrolactin, butyrosin, difficidin, bacillibactin, fengycins, and bacilysin (Table 3). Five of these clusters are shared with endophytic strain *B. velezensis* CC09 (2, 4, 8, 9 and 10), 5 with *B. subtilis* BsN5 (2, 6, 7, 11 and 12) and only 1 with *B. megaterium* Q3 (2) among the endophytic genomes studied. These are important manifestations of a plant defense mechanism and can cope with competing microorganisms and inhibit the growth of phytopathogenic fungi or bacteria, enhancing the potential biological control agent.

Table 3
Secondary metabolite clusters identified in the genome of *B. velezensis* 629 by AntiSmash 4.0 and shared among endophytes.

Cluster	**Type Synthetase**	**Metabolites**	**Size (pb)**	**MIBiG BGC-ID (%)**	**Shared to**
1	Lantipeptide	-	23041	-	-
2	Nrps	Surfactin	65408	BGC0000433_c1 (91%)	*B. velezensis* CC09*, *B. subtilis* BsN5 (78%), *B. megaterium* Q3
3	Otherks	Butirosin	41244	BGC0000693_c1 (7%)	-
4	Terpene	-	20740	-	B. velezensis CC09
5	Transatpks	Macrolactin	85900	BGC0000181_c1 (100%)	-
6	Transatpks-Nrps	Bacillaene	102454	BGC0001089_c1 (100%)	*B. subtilis* BsN5
7	Transatpks-Nrps	Fengycin	137831	BGC0001095_c1 (100%)	*B. subtilis* BsN5
8	Terpene	-	21883	-	B. velezensis CC09*
9	T3pks	-	41244	-	B. velezensis CC09*
10	Transatpks	Difficidin	94235	BGC0000176_c1 (100%)	B. velezensis CC09
11	Bacteriocin-Nrps	Bacillibactin	66787	BGC0000309_c1 (100%)	*B. subtilis* BsN5
12	Other	Bacilysin	41418	BGC0001184_c1 (100%)	*B. subtilis* BsN5

*Clusters shared with others strains of *B. velezensis* (Cai et al., 2017).

It has been recently shown that the rice-associated *B. velezensis* strain LS69 exhibits activity against a diverse spectrum of pathogenic bacteria and has 10 secondary metabolites clusters involved in nonribosomal synthesis of polyketides (macrolactin, bacillaene and difficidin), lipopeptides (surfactin, fengycin, bacilysin and iturin A) and bacteriocins (amylolysin and amylocyclicin) (Liu et al. 2017). Besides, the plant-associated *B. velezensis* NJN6 has bacillomycin D and macrolactin with significant antagonistic effects against *Fusarium oxysporum* and *Ralstonia solanacearum*, respectively (Yuan et al. 2012).

3.5. Prediction of genes involved in endophytic behavior

3.5.1. Comparative analysis with endophytic strains

Comparative genomic analysis of endophytic bacteria can contribute to identification of a specific set of genes related to the plant niche adaptation. In addition, comparative analyzes of the genome may determine gene signatures or gene clusters by endophytic bacteria. The comparative genome analysis of strain 629 together with 15 other genomes of endophytic bacteria shown 121 shared genes (Supplementary Table S5). According to Gene Ontology annotation, 60% of the orthologs shared by endophytic species were related to metabolic processes (Figure 5); 22.1% of these genes are related to nitrogen compound metabolic processes (Figure 5).

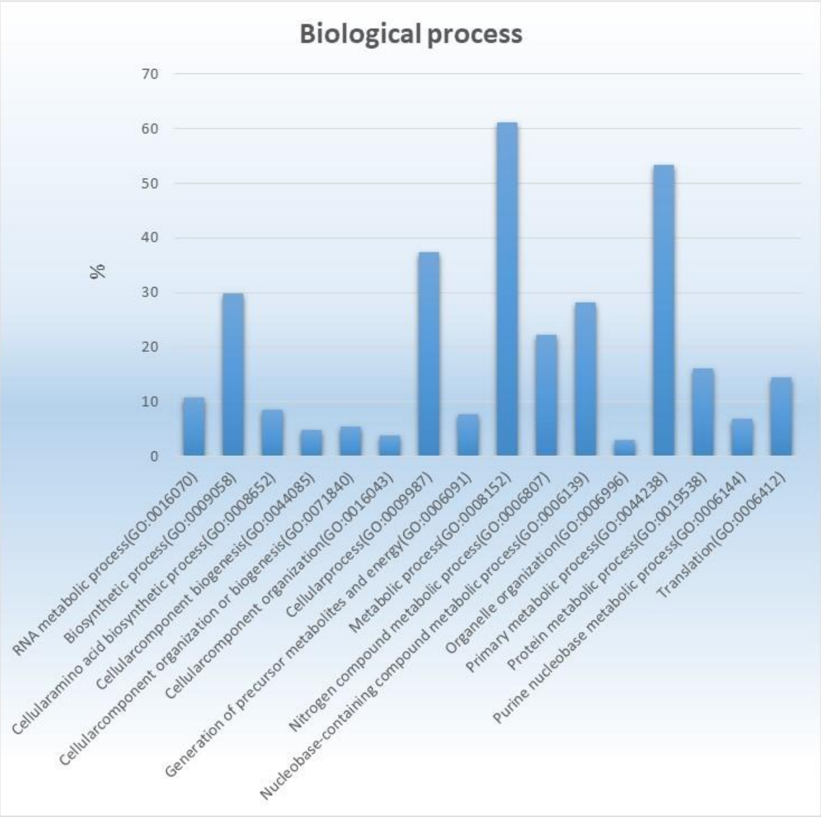


Fig. 5 Distribution of biological processes of 121 genes shared among strain 629 and remaining 15 endophytic genome strains.

We found glutamate synthase (gltA), involved in the nitrogen fixation process by assimilation of ammonia (van den Heuvel et al. 2004); pABA, involved in tryptophan biosynthesis; seven genes associated with stress response, as chaperone proteins DnaJ, DnaK and GroEL, which actively participate in the response to hyperosmotic and heat shock by preventing the aggregation of stress-denatured proteins and by disaggregating proteins; Elongation factor Tu (EF-Tu), associated with tolerance to heat and elicitation of plant basal defense (Fu and Prasad 2012); thiC protein, which acts on thiamine biosynthesis, may be involved in the activation of plant defense reactions, plant growth and as a cofactor in several reactions such as IAA synthesis (Palacios and Bashan 2014).

3.5.2. Genes related to endophytic capacity (colonization and survival) of *B. velezensis* 629

Several genes related to the process of plant / microorganism interaction and putatively responsible for endophytic behavior are present in the genome of strain 629 and shared with some of the endophytes studied, such as: numerous transcriptional regulators: LysR family, involved in bacterial virulence, metabolism, quorum sensing and motility (Maddocks and Oyston 2008); AraC that can regulate functions in carbon metabolism, response to stress and virulence strategy (Santos et al. 2009); LrgB family, with hydrolases controlling activity, avoiding invasive breakage of the plant cell wall and, consequently, reducing the defense response.

In chemotaxis, the role played seems crucial in the adaptation to the host, including in flagella and pili biosynthesis, as is the case of the genes present: *cheB, cheA, cheW* and *cheY*; as well as the flagellar components (such as FlgC, Flil, FliP, FliQ, FlhA), followed by MotA and MotB (Motility).

For effective colonization of plant tissue, endophytic bacteria need to provide enzymes capable of neutralizing the oxidative process of the plant in response to abiotic stress or colonizing microorganisms. In the genome of *B.velezensis* 629, some of these enzymes were found as catalases (*katA*, *katE* and *katX*), superoxide dismutases (*sodA*, *sodB* and *sodC*) and glutathione peroxidase indicating their resilience under conditions of oxidative stress.

Inducible defense responses in plants may occur by recognition of PAMPs / MAMPS (pathogen or microbe associated molecular pattern), but the relationship with endophytic responses is varied and compatible interactions and / or a specific cascade of signals may allow colonization depending on the genotype of both (Reinhold-Hurek and Hurek

2011). Although flagellins are recognized as MAMPs, the components of the flagellar apparatus are present in the genome of endophytic bacteria 629.

3.5.3. Genomic traits related to plant metabolism

Additionally, we found enzymes involved in trehalose metabolism, cell osmoprotection and carbon routes to microorganisms and plants (Iturriaga et al. 2009), ABC transporter for siderophore and siderophore biosynthesis, which allows survival in the low iron environment and can efficiently compete for this element with other microorganisms, including phytopathogens (Malfanova et al. 2013); Glycine betaine ABC transport system, known to capacity for abiotic stress tolerance in plants (Giri 2011). Genes involved in phosphate solubilization, such as *ywqF* (Khan et al. 2017), improving host plant growth and decrease the use of soluble phosphate fertilizers. The presence and expression of these genes must occur in response to the conditions encountered by an endophytic when it comes into contact with the interior of the host.

In the *B. velezensis* strains 629 and CC09 the genes to fix nitrogen (nitrogenase complex, such as *nifH*, *nifK and nifD*) are absent, but have systems of regulation involved in nitrogen metabolism, glutamine synthetase and glutamate synthase (shared by all 15 endophytic genomes already mentioned). In addition, also lacked the region encoding ACC deaminase (*acds*) for plant growth promotion, although presenting gene cluster related to IAA biosynthesis from Trp biosynthetic pathway and/or Trp-independent IAA biosynthetic pathway (e.g. by indole-3-glycerol phosphate synthase) (Ouyang et al. 2000). The endophytic bacterium *S. marcescens* RSC-14 (Khan et al. 2017), also shows plant growth promotion, and does not contain the ethylene-regulating enzyme (acds). As natural regulators of plant growth, plant hormones directly influence physiological processes and interaction with microorganisms.

3.5.4. Endophytic behavior X Endophytic bacteria

There is no standard or signature common to all endophytes related to any host plant. Each interaction will be specific so as not to trigger a strong defense response sufficient to prevent colonization of the endophytic. In contrast, all endophytes have a beneficial interaction with their host, which seems to be related to the reduction in this response, and the host may be able to distinguish colonization by endophytic or pathogenic microorganisms (Cavalcante et al. 2007). Endophytic bacteria and rhizobacteria share varied lifestyles mechanisms associated with plants such as: plant growth promotion, biological control of phytopathogens, nitrogen fixation, phosphate solubilization, etc. Strain 629 presents all of these traits in their genome, by similar metabolic pathways of other endophytes. Based on genomic analysis the capacity to penetrate and survive inside plant tissues are multifactorial, including characteristics such as motility, quorum sensing, resistance to stress, ability to adapt to the environment and to survive to defense mechanisms of plants. Santoyo et al. (2016) by bioinformatics approach distinguish endophytic from rhizospheric PGPB summarized in ~ 40 genes procuring transporter proteins, secretion and delivery systems, plant polymer degradation or modification, transcriptional regulation, detoxification, redox potential maintenance, and unkown functions.

4. Conclusion

Our data supports taxonomic reclassification of the *Theobroma cacao*-related *Bacillus* strain 629, from the species *Bacillus amyloliquefaciens* to the species *Bacillus velezensis*. Pan-genomic analysis demonstrated that the species *B. velezensis* possess an open pan-genome, with the majority of genes belonging to the categories of accessory and strain-specific genes (singletons). This associated to a high number of genes present in probable genomic islands indicates that the species acquires many biological functions by horizontal transfer. Furthermore, the detailed study of genes participating in shared metabolic processes between the endophytic bacterial strains analyzed, which are present in genomic islands, contributes new species to understanding essential characteristics related to endophytic colonization.

Conflict of interest

The authors declare no conflict of interest.

Supplementary Material (Apêndice D)

The online version of this article contains supplementary material, which is available to authorized users.

Supplementary Material 1: Supplementary Figure S1.pdf (146Kb)
Supplementary Material 2: Supplementary Figure S2.pdf (860Kb)
Supplementary Material 3: Supplementary Table S1.doc (50Kb)
Supplementary Material 4: Supplementary Table S2.doc (30Kb)
Supplementary Material 5: Supplementary Table S3.doc (114Kb)
Supplementary Material 6: Supplementary Table S4.doc (17Kb)
Supplementary Material 7: Supplementary Table S5.doc (92Kb)

References

Bertelli C, Laird MR, Williams KP, Simon Fraser University Research Computing Group, Lau BY, Hoad G, Winsor GL, Brinkman FS (2017). IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Research 3:45(W1):W30-W35. doi: 10.1093/nar/gkx343

Blom J, Kreis J, Spanig S, Juhre T, Bertelli C, Ernst C, Goesmann A (2016) EDGAR 2.0: an enhanced software platform for comparative gene content analyses. Nucleic Acids Research, 44(W1):W22-8. doi: 10.1093/nar/gkw255

Cai XC, Liu CH, Wang BT, Xue YR (2017) Genomic and metabolic traits endow Bacillus velezensis CC09 with a potential biocontrol agent in control of wheat powdery mildew disease. Microbiological Research 196:89–94. http://dx.doi.org/10.1016/j.micres.2016.12.007

Cavalcante JJ, Vargas C, Nogueira EM, Vinagre F, Schwarcz K, Baldani JI, Ferreira PC, Hemerly AS (2007). Members of the ethylene signalling pathway are regulated in sugarcane during the association with nitrogen-fixing endophytic bacteria. Journal of Experimental Botany, 58(3):673-686

Chen XH, Koumoutsi A, Scholz R, Eisenreich A, Schneider K, Heinemeyer I, Morgenstern B, Voss B, Hess WR, Reva O, Junge H, Voigt B, Jungblut PR, Vater J, Süssmuth R, Liesegang H, Strittmatter A, Gottschalk G, Borriss R (2007) Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium Bacillus amyloliquefaciens FZB42. Nat Biotechnol 25(9):1007–1014

Dunlap CA, Kim SJ, Kwon SW, Rooney AP (2016) *Bacillus velezensis* is not a later heterotypic synonym of *Bacillus amyloliquefaciens*; *Bacillus methylotrophicus, Bacillus amyloliquefaciens subsp. plantarum* and '*Bacillus oryzicola*' are later heterotypic synonyms of *Bacillus velezensis* based on phylogenomics. Int. J. Syst. Evol. Microbiol 66(3):1212–1217. Doi: 10.1099/ijsem.0.000858

Falcão LL, Silva-Werneck JO, Vilarinho BR, Silva JP, Pomella AWV, Marcellino LH (2014) Antimicrobial and plant growth-promoting properties of the cacao endophyte *Bacillus subtilis* ALB629. J Appl Microbiol 116(6):1584–1592. Doi:10.1111/jam.12485

Fan B, Blom J, Klenk HP, Borriss R (2017) *Bacillus amyloliquefaciens*, *Bacillus velezensis*, and *Bacillus Siamensis* Form an "Operational Group *B. amyloliquefaciens*" within the *B. Subtilis* Species Complex. Front Microbiol 8:22. doi: 10.3389/fmicb.2017.00022

Fu J, Momcilovic I, Prasad PVV (2012) Roles of Protein Synthesis Elongation Factor EF-Tu in Heat Tolerance in Plants. Journal of Botany, 2012, Article ID 835836, 8 pages. Doi:10.1155/2012/835836

Gao Z, Zhang B, Liu H, Han J, Zhang Y (2016) Identification of endophytic *Bacillus velezensis* ZSY-1 strain and antifungal activity of its volatile compounds against *Alternaria solani* and *Botrytis cinerea*. Biological Control 105:27-39. http://dx.doi.org/10.1016/j.biocontrol.2016.11.007

Giri J (2011) Glycinebetaine and abiotic stress tolerance in plants. Plant Signaling & Behavior 6(11):1746-1751. Doi: 10.4161/psb.6.11.17801

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52(5):696-704. Doi: 10.1080/10635150390235520

Iturriaga G, Suarez R, Nova-Franco B (2009) Trehalose metabolism: from osmoprotection to signaling. Int. J. Mol. Sci 10(9):3793-3810. Doi:10.3390/ijms10093793

Khan AR, Park G-S, Asaf S, Hong S-J, Jung BK, Shin J-H (2017) Complete genome analysis of *Serratia marcescens* RSC-14: A plant growth-promoting bacterium that alleviates cadmium stress in host plants. PLoS ONE 12(2):e0171534. doi:10.1371/journal.pone.0171534

Langille MGI, Hsiao WWL, Brinkman FSL (2010) Detecting genomic islands using bioinformatics approaches. Nature Rev Microbiology 8(5):373-82. doi:10.1038/nrmicro2350

Liu G, Kong Y, Fan Y, Geng C, Peng D, Sun M (2017) Whole-genome sequencing of *Bacillus velezensis* LS69, a strain with a broad inhibitory spectrum against pathogenic bacteria. J Biotechnol 249:20-24. Doi: 10.1016/j.jbiotec.2017.03.018

Maddocks SE, Oyston PC (2008) Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. Microbiology, 154:3609–3623. doi: 10.1099/mic.0.2008/022772-0

Malfanova N, Lugtenberg BJJ, Berg G (2013) Bacterial Endophytes: Who and Where, and What Are They Doing There? In: Molecular Microbial Ecology of the Rhizosphere: Volume 1 & 2 (ed F. J. de Bruijn), John Wiley & Sons, Inc., Hoboken, NJ, USA. Doi: 10.1002/9781118297674.ch36

Martins SJ, de Medeiros FHV, de Souza RM, de Faria AF, Cancellier EL, Silveira HRO, de Rezende MLV, Guilherme LRG (2015) Common bean growth and health promoted by rhizobacteria and the contribution of magnesium to the observed responses. Appl Soil Ecol 87:49–55. https://doi.org/10.1016/j.apsoil.2014.11.005

Medema MH, Kai B, Cimermancic P, Jager VD, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 39 (8), 339–346. doi:10.1093/nar/gkr466

Meier-Kolthoff JP, Auch AF, Klenk HP, GöKer M (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 14:60. doi:10.1186/1471-2105-14-60

Monteiro FP, Medeiros FHV, Ongena M, Franzil L, Souza PE, Souza JT (2016) Effect of temperature, pH and substrate composition on production of lipopeptides by Bacillus amyloliquefaciens 629. Afr J Microbiol Res 10(36):1506-1512. Doi: 10.5897/AJMR2016.8222

Ouyang J, Shao X, Li J (2000) Indole-3-glycerol phosphate, a branchpoint of indole-3-acetic acid biosynthesis from the tryptophan biosynthetic pathway in *Arabidopsis thaliana*. The Plant Journal, 24(3), 327-333.

Palacios OA, Bashan Y, de-Bashan LE (2014) Proven and potential involvement of vitamins in interactions of plants with plant growth-promoting bacteria - an overview. Biol Fertil Soils 50(3), 415-432. Doi: 10.1007/s00374-013-0894-3

Pantoja Y, Pinheiro K, Veras A, Araújo F, Lopes de Sousa A, Guimarães LC, Silva A. Rammos RTJ (2017) PanWeb: A web interface for pan-genomic analysis. PLoS ONE 12(5): 1-9. https://doi. org/10.1371/journal.pone.0178154

Rashid MH, Khan A, Hossain MT and Chung YR (2017) Induction of Systemic Resistance against Aphids by Endophytic *Bacillus velezensis* YC7010 via Expressing *PHYTOALEXIN DEFICIENT4* in Arabidopsis. Front. Plant Sci 8:211. doi: 10.3389/fpls.2017.00211

Reinhold-Hurek B, Hurek T (2011) Living inside plants: bacterial endophytes. Curr. Opin. Plant Biol. 14(4):435-443. Doi: 10.1016/j.pbi.2011.04.004

Rückert C, Blom J, Chen XH, Reva O, Borriss R (2011) Genome sequence of *B. amyloliquefaciens* type strain DSM7T reveals differences to plant-associated *B. amyloliquefaciens* FZB42. Journal of Biotechnology, 155(1):78-85. doi:10.1016/j.jbiotec.2011.01.006.

SantAnna BM, Marbach PP, Rojas-Herrera M, De Souza JT, Roque MR, Queiroz AT. (2015) High-Quality Draft Genome Sequence of *Bacillus amyloliquefaciens* strain 629, an Endophyte from *Theobroma cacao*. Genome Announcements, 3 (6), e01325-15. Doi:10.1128/genomeA.01325-15.

Santos CL, Tavares F, Thioulouse J, Normand P (2009) A phylogenomic analysis of bacterial helix-turn-helix transcription factors. FEMS Microbiol Rev 33(2):411–429. Doi:10.1111/j.1574-6976.2008.00154.x

Santoyo G, Moreno-Hagelsieb G, Orozco-Mosqueda MC, Glick BR (2016) Plant growth-promoting bacterial endophytes. Microbiol Res., 183:92-99. Doi: 10.1016/j.micres.2015.11.008.

Scortichini M, Marcelletti S, Ferrante P, Firrao G (2013) A genomic redefinition of *Pseudomonas avellanae* species. Plos One, 8(9):1-16. Doi: 10.1371/journal.pone.0075794.

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Curr. Opinion in Microbiol., 11(5)472–477. Doi: 10.1016/j.mib.2008.09.006

van den Heuvel RHH, Curti B, Vanoni MA, Mattevi A (2004) Glutamate synthase: a fascinating pathway from L-glutamine to L-glutamate. Cell. Mol. Life Sci 61:669– 681. Doi: 10.1007/s00018-003-3316-0

Yuan J, Li B, Zhang N, Waseem R, Shen Q, Huang Q (2012) Production of Bacillomycin- and Macrolactin-Type Antibiotics by *Bacillus amyloliquefaciens* NJN-6 for Suppressing Soilborne Plant Pathogens. Journal of agricultural and food chemistry. 60(12):2976-81. Doi: 10.1021/jf204868z.

Zhang X, Peng C, Zhang G, Gao F (2015) Comparative analysis of essential genes in prokaryotic genomic islands. Scientific reports, 5:12561. Doi: 10.1038/srep12561

Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J (2012) PGAP: pan-genomes analysis pipeline. Bioinformatics 28(3):416–418. Doi: 10.1093/bioinformatics/btr655

Zouari I, Jlaiel L, Tounsi S, Trigui M (2016). Biocontrol activity of the endophytic *Bacillus amyloliquefaciens* strain CEIZ-11 against *Pythium aphanidermatum* and purification of its bioactive compounds. Biological Control 100:54-62. Doi: http://dx.doi.org/10.1016/j.biocontrol.2016.05.012

Manuscrito 3

Genome sequence of *Serratia marcescens* strain 1274, an endophytic bacterium isolated from *Agave sisalana*

Brena M. M. Sant'Anna[1], Jorge T. De Souza[2], Phellippe P. A. Marbach[3], Vasco Azevedo[4], Artur Silva[5], Rommel T. J. Ramos[5], Luis G. C. Pacheco[1,] Artur T. L. Queiroz[6], Milton R. A. Roque[1*]

1. Universidade Federal da Bahia, UFBA – Brazil; 2. Universidade Federal de Lavras, UFLA – Brazil; 3. Universidade Federal do Recôncavo da Bahia, UFRB – Brazil; 4. Institute of Biological Sciences, Federal University of Minas Gerais (UFMG), Belo Horizonte, MG, Brazil; 5. Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil; 6. CPqGM-Fiocruz/BA – Brazil

*Corresponding author: milton.roque@ufba.br (M. Roque)

Abstract

Serratia marcescens strain 1274 is a Gram-negative and non-pigmented endophytic bacterium isolated from the leaves of *Agave sisalana*. The genome of this strain has a predicted size of 5,214,868 bp and contains 4,861 protein coding sequences, 104 RNA sequences and an average 59.8% GC content. Seven gene clusters that code for secondary metabolites with antibacterial activity and products that contribute to plant-microbe interactions were identified. The comparative analysis of 22 complete *S. marcescens* genome sequences revealed a pan-genome with 10,450 non-redundant coding sequences and a core genome of 2,855 (27%) genes. Further analyses of this genome sequence may provide novel insights into the possible molecular mechanisms of endophytic behavior.

Keywords: *Serratia marcescens*; Genome; Endophytic bacteria

Genome announcement

Serratia marcescens is a Gram-negative rod-shaped bacterium, member of the *Enterobacteriaceae* family, with strains that show diverse life styles. They may associate with plants as both endophytes (Khan et al. 2017; Li et al. 2015) or as phytopathogens (Wang et al. 2015). Endophytic microorganisms can colonize the internal tissues of plants without negative effects or symptoms of disease. These endophytes may provide beneficial effects to their hosts, including plant growth promotion (Mitter et al. 2013), resistance to stresses (Sziderics et al. 2007) and biological control of phytopathogens (Krishnan et al. 2015). These bacteria are potential sources of natural molecules to be employed in medical sciences and other areas such as agriculture and industry (Doley and Jha 2015; Mitter et al. 2013).

The non-pigmented endophytic bacterium *Serratia marcescens* 1274 was isolated from healthy leaves of sisal (*Agave sisalana*) in the Brazilian semi-arid region of Bahia State, Brazil. This strain was deposited in the Collection of Bacteria from the Environment and Health (CBAS - WFCC) at Fundação Oswaldo Cruz (Fiocruz) under accession number CBAS 643.

Genomic DNA from strain 1274 was extracted and sequenced using the Ion Torrent PGM platform (Life Technologies) in a 318 chip. A total of 5,466,729 single reads with an average length of 236bp were obtained. Read quality assessment was performed with the FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc); filtering and trimming were performed the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), using a Phred score cut-off of 20.

De novo genome assembly was performed using SPAdes Genome Assembler (version 3.5.0) (Nurk et al. 2013); curation with the Lasergene 11 Core Suite (DNASTAR) was done to remove redundant contigs. Scaffolding was done using Mauve (Darling et al. 2004), with *S. marcescens* strain RSC14 (GenBank: CP012639.1) as the reference genome (Fig1). GapBlaster (de Sá et al. 2016) was employed to close gaps using data from another assembly performed with Mira 4.0 (Chevreux et al. 1999). Manual curation was performed through CLC Genomics Workbench 8 (Qiagen, USA) and Artemis 16.0.0 software (Rutherford et al. 2000).

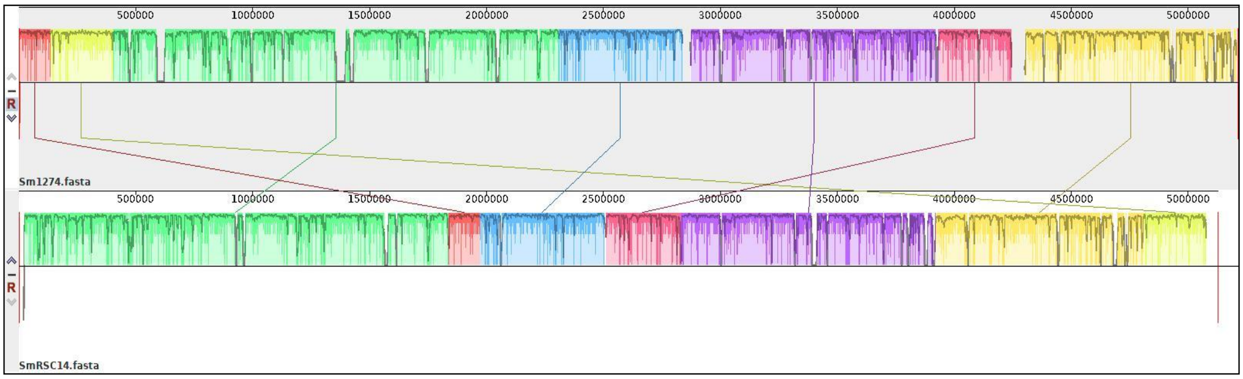


Fig 1. Synteny analysis of *Serratia marcescens* 1274 genome and the reference *S. marcescens* strain RSC14 with Mauve - multiple genome alignment.

Genome annotation was achieved with Rapid Annotations using Subsystem Technology (RAST) version 2.0 server (Overbeek et al. 2014). tRNAs and rRNAs predictions were confirmed using the software tools tRNAScan-SE 1.21 (Lowe and Eddy 1997) and RNAmmer 1.2 (Lagesen et al. 2007), respectively.

The genome assembly resulted in 6 scaffolds with a total size of 5,214,868bp and an average 59.8% GC content (Fig2) (N50: 1,503,108pb; largest contig: 2,174,531). The genome of strain 1274 is composed of 4,965 predicted genes, including 4,861 protein-coding sequences, 86 tRNAs and 6 copies of the ribosomal operon containing the genes 5S, 16S and 23S rRNA, similar to other *S. marcescens* strains deposited in the NCBI database. A gap in the ~3.8Mb region coincides with a breakdown of GC content (Fig2), which may represent regions of high complexity, such as repetitive sequences or transposons.

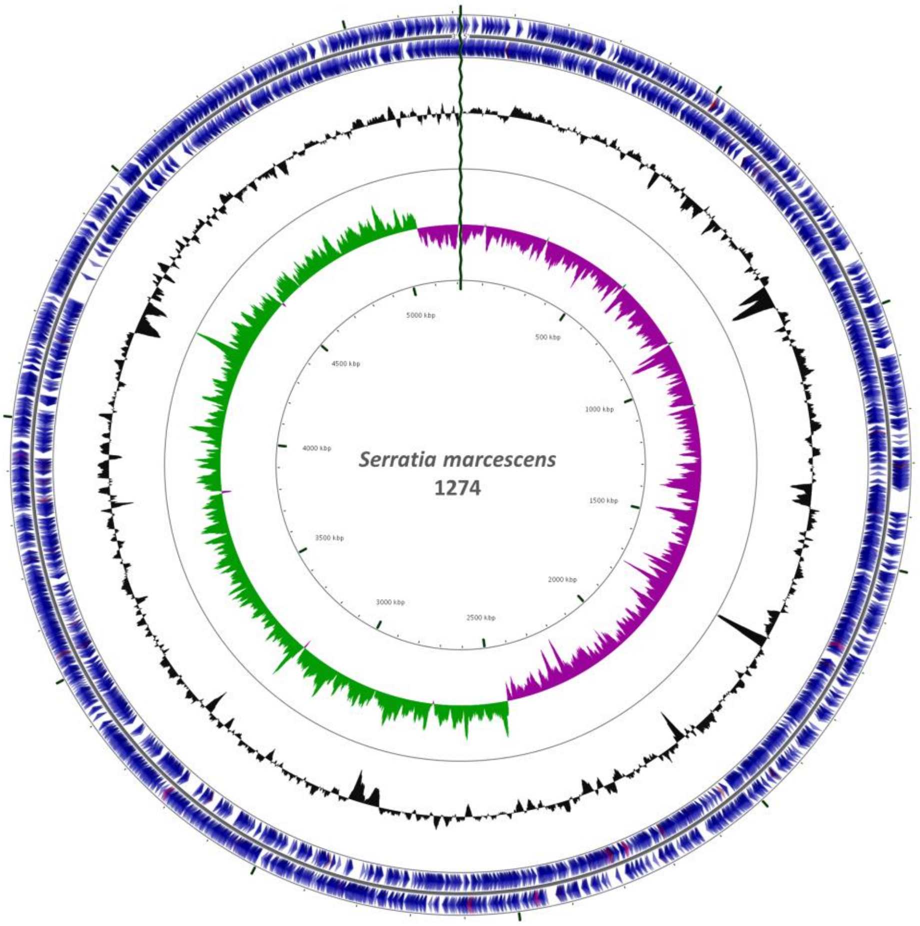
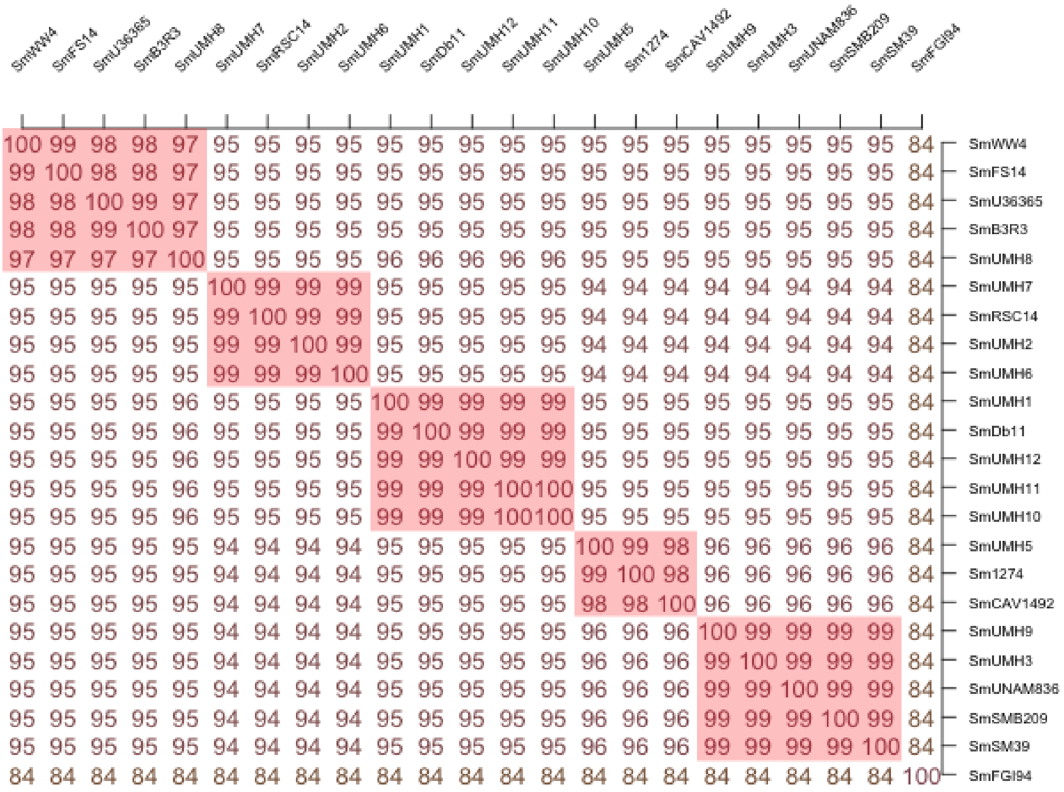


Fig 2. Circular map of the genome of *Serratia marcescesns* strain 1274. From outside to the center: rings 1 and 2 show protein-coding genes oriented in the forward and reverse directions (with RNA sequences in red); ring 3 shows G + C% content plot, and the inner-most ring shows GC skew, purple indicating negative values and green, positive values.

Analysis by antiSMASH (Weber et al. 2015) suggests seven putative secondary metabolite gene clusters for thiopeptide, hserlactone, siderophore and non-ribosomal peptide synthetase (Nrps), such as lysobactin, enterobactin and turnerbactin. These latter clusters, also present in endophytic strains *S. marcescens* RSC-14 and FS14, are involved in iron uptake and nitrogen fixation. Nearly 6.3% of the genome of strain 1274 is dedicated to secondary metabolites production.

For comparative genome analysis, a new annotation was created on the RAST server (version 2.0) for 22 complete genome sequences retrieved from GenBank, in order to homogenize the genome annotations. The *S. marcescens* genome sequences included: strain RSC-14 (CP012639.1); strain FS14 (CP005927); strain B3R3 (CP013046.1); strain DB11 (HG326223.1); strain WW4 (CP003959.1); strain CAV1492 (CP011642.1); strain SM39 (AP013063.1); strain U36365 (CP016032.1); strain UNAM836 (CP012685.1); strain SMB2099 (HG738868.1); strain UMH1 (CP018915.1); strain UMH2 (CP018924.1); strain UMH3 (CP018925.1); strain UMH5 (CP018917.1); strain UMH6 (CP018926.1); strain UMH7 (CP018919.1); strain UMH8 (CP018927.1); strain UMH9 (CP018923.1); strain UMH10 (CP018928.1); strain UMH11 (CP018929.1); strain UMH12 (CP018930.1); and *Serratia* sp. strain FGI94 (CP003942.1).

As a robust genome based criteria to determine phylogenomic identity, Average Nucleotide Identity (ANI) was performed using ANI-Matrix estimating all-vs-all distances in a collection of genomes to build similarity clustering (available at http://enve-omics.ce.gatech.edu/ani/index) (Fig3). Although some strains present low ANI values, into cutoff ≥95% (Arahal 2014), a digital DNA-DNA Hybridization (dDDH), was performed and presented values above limit of 70% (Meier-Kolthoff et al. 2013), being members of the same species (Supplementary Table S1). The strain FGI94 show a lower values in both analyzes, besides a high number of strain-specific genes (1,786) and was removed from the next studies because it was considered outside the group. Though strain 1274 is an environmental (endophytic) isolate, the closest strains were the clinical isolates, as UMH5 and CAV1492.



	SmWW4	SmFS14	SmU36365	SmB3R3	SmUMH8	SmUMH7	SmRSC14	SmUMH2	SmUMH6	SmUMH1	SmDb11	SmUMH12	SmUMH11	SmUMH10	SmUMH5	Sm1274	SmCAV1492	SmUMH9	SmUMH3	SmUNAM836	SmSMB209	SmSM39	SmFGI94
SmWW4	100	99	98	98	97	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	84
SmFS14	99	100	98	98	97	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	84
SmU36365	98	98	100	99	97	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	84
SmB3R3	98	98	99	100	97	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	84
SmUMH8	97	97	97	97	100	95	95	95	95	96	96	96	96	96	95	95	95	95	95	95	95	95	84
SmUMH7	95	95	95	95	95	100	99	99	99	95	95	95	95	95	94	94	94	94	94	94	94	94	84
SmRSC14	95	95	95	95	95	99	100	99	99	95	95	95	95	95	94	94	94	94	94	94	94	94	84
SmUMH2	95	95	95	95	95	99	99	100	99	95	95	95	95	95	94	94	94	94	94	94	94	94	84
SmUMH6	95	95	95	95	95	99	99	99	100	95	95	95	95	95	94	94	94	94	94	94	94	94	84
SmUMH1	95	95	95	95	96	95	95	95	95	100	99	99	99	99	95	95	95	95	95	95	95	95	84
SmDb11	95	95	95	95	96	95	95	95	95	99	100	99	99	99	95	95	95	95	95	95	95	95	84
SmUMH12	95	95	95	95	96	95	95	95	95	99	99	100	99	99	95	95	95	95	95	95	95	95	84
SmUMH11	95	95	95	95	96	95	95	95	95	99	99	99	100	100	95	95	95	95	95	95	95	95	84
SmUMH10	95	95	95	95	96	95	95	95	95	99	99	99	100	100	95	95	95	95	95	95	95	95	84
SmUMH5	95	95	95	95	95	94	94	94	94	95	95	95	95	95	100	99	98	96	96	96	96	96	84
Sm1274	95	95	95	95	95	94	94	94	94	95	95	95	95	95	99	100	98	96	96	96	96	96	84
SmCAV1492	95	95	95	95	95	94	94	94	94	95	95	95	95	95	98	98	100	96	96	96	96	96	84
SmUMH9	95	95	95	95	95	94	94	94	94	95	95	95	95	95	96	96	96	100	99	99	99	99	84
SmUMH3	95	95	95	95	95	94	94	94	94	95	95	95	95	95	96	96	96	99	100	99	99	99	84
SmUNAM836	95	95	95	95	95	94	94	94	94	95	95	95	95	95	96	96	96	99	99	100	99	99	84
SmSMB209	95	95	95	95	95	94	94	94	94	95	95	95	95	95	96	96	96	99	99	99	100	99	84
SmSM39	95	95	95	95	95	94	94	94	94	95	95	95	95	95	96	96	96	99	99	99	99	100	84
SmFGI94	84	84	84	84	84	84	84	84	84	84	84	84	84	84	84	84	84	84	84	84	84	84	100

Fig3 Average Nucleotide Identity (ANI) values between 23 different strains of *S. marcescens* (cutoff ≥95%).

The PanWeb tool (http://www.computationalbiology.ufpa.br/analysis.php) (Pantoja et al. 2017) was used to calculate the core genome (subset of genes shared by all the selected strains), the pan-genome (core genome plus the accessory genes of all analyzed strains) and singletons (strain-specific) genes of *S. marcescens*. The parameters used were as follows: 0.8 for identity; 0.8 for coverage; 1e-5 cutoff for E-value. The genome sequence of *S. marcescens* strain 1274 expands the pan-genome of the species to 10,450 genes (with 215 strain-specific additional genes), 2,855 (27%) genes belonging to the core genome and 3,862 accessory genes (37%). The low ANI value for some *S. marcescens* strains, but appropriate dDDH values and the considerable amount of core genes may reflect the genomic diversity and phenotypic plasticity of this species.

This genome (CP019927) will provide novel insights into the possible molecular mechanisms of plant-interaction of this endophytic bacterium.

Accession numbers

The genome sequencing data were deposited in the NCBI Sequence Read Archive under the BioProject number PRJNA371353. The assembled scaffolds of *S. marcescens* strain 1274 were deposited in the NCBI Whole Genome Shotgun database under the accession number of CP019927.2.

Conflict of interest

The authors declare that they have no conflict of interest in the publication.

References

Arahal DR (2014) Whole-Genome Analyses: Average Nucleotide Identity. Methods in Microbiology - New Approaches to Prokaryotic Systematics, Vol 41, p103-122. DOI: 10.1016/bs.mim.2014.07.002

Chevreux B, Wetter T, Suhai S (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99, pp. 45-56.

Darling ACE, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. 14:1394–1403.10.1101/gr.2289704.

de Sá PHCG, Miranda F, Veras A, Melo DM, Soares S, Pinheiro K, Guimarães L, Azevedo V, Silva A, Ramos RTJ (2016) GapBlaster — A Graphical Gap Filler for Prokaryote Genomes. PLOS ONE, 11(5): e0155327. DOI:10.1371/journal.pone.0155327

Doley P, Jha DK (2015) Antimicrobial Activity of Bacterial Endophytes from Medicinal Endemic Plant *Garcinia lancifolia* Roxb. Annals of Plant Sciences 4.12: 1243-1247

Khan AR, Ullah I, Khan AL, Park GS, Waqas M, Hong SJ, Jung BK, Kwak Y, Lee IJ, Shin JH (2015) Improvement in phytoremediation potential of *Solanum nigrum* under cadmium contamination through endophytic-assisted *Serratia* sp. RSC-14 inoculation. Environ Sci Pollut Res, 22:14032–14042. DOI 10.1007/s11356-015-4647-8

Krishnan N, Gandhi K, Peeran MF, Kuppusami P, Thiruvengadam R (2015) Molecular characterization and in vitro evaluation of endophytic bacteria against major pathogens of rice. African Journal of Microbiology Research, 9(11), 800-813.

Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. 35:3100–3108. 10.1093/nar/gkm160.

Li P, Kwok AHY, Jiang J, Ran T, Xu D, Wang W, Leung FC (2015) Comparative Genome Analyses of *Serratia marcescens* FS14 Reveals Its High Antagonistic Potential.

PLoS ONE 10(4): e0123061. doi:10.1371/journal.pone.0123061.

Meier-Kolthoff JP, Auch AF, Klenk HP, GöKer M (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 14:60. doi:10.1186/1471-2105-14-60

Mitter B, Petric A, Chain PS, Trognitz F, Nowak J, Compant S, Sessitsch A (2013) Genome Analysis, Ecology, and Plant Growth Promotion of the Endophyte *Burkholderia phytofirmans* Strain PsJN. Molecular Microbial Ecology of the Rhizosphere, Volume 1, Chapter 81, First Edition. Edited by Frans J. de Bruijn.

Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, Prjibelski AD, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, McLean JS, Lasken R, Clingenpeel SR, Woyke T, Tesler G, Alekseyev MA, Pevzner PA (2013) Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. *In*: Deng M, Jiang R, Sun F, Zhang X (ed), Research in Computational Molecular Biology, vol 7821. Springer, Berlin, Heidelberg.

Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R (2014) The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). 42:D206–D214. 10.1093/nar/gkt1226.

Pantoja Y, Pinheiro K, Veras A, Araújo F, Lopes de Sousa A, Guimarães LC, Silva A. Rammos RTJ (2017) PanWeb: A web interface for pan-genomic analysis. PLoS ONE 12(5): 1-9. https://doi. org/10.1371/journal.pone.0178154

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B (2000) Artemis: sequence visualization and annotation. 16 (10): 944–945. DOI: 10.1093/bioinformatics/16.10.944

Sziderics AH, Rasche F, Trognitz F, Sessitsch A, Wilhelm E (2007) Bacterial endophytes contribute to abiotic stress adaptation in pepper plants (*Capsicum annuum* L.). Canadian Journal of Microbiology, Vol. 53, No. 11: pp. 1195-1202. 2007. doi: 10.1139/W07-082

Wang XQ, Bi T, Li XD, Zhang LQ, Lu SE (2015) First Report of Corn Whorl Rot Caused by *Serratia marcescens* in China. Journal of Phytopathology, Short Communication. Doi: 10.1111/jph.12366

Weber T, Blin K, Duddela S, Krug D, Kim Hyun Uk, Bruccoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, Breitling R, Takano E, Medema MH (2015) antiSMASH 3.0 - a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Research, Volume 43, Issue W1, 1 July 2015, Pages W237–W243, https://doi.org/10.1093/nar/gkv437.

Manuscrito 4

Endophytic Behavior in *Serratia marcescens*: New Insights from Comparative Genomics of Strain 1274

Brena M. M. Sant'Anna[1], Artur T. L. Queiroz[2], Icaro Lopes¹, Luis G. C. Pacheco[1,] , Milton R. A. Roque[1*]

1. Universidade Federal da Bahia, UFBA – Brazil; 2. CPqGM-Fiocruz/BA – Brazil;

*Correspondencing authors: milton.roque@ufba.br (M. Roque)

Abstract

Serratia marcescens strain 1274 is a plant-associated bacterial endophyte that was isolated from healthy the leaves of *Agave sisalana*. Endophytic microorganisms colonize the internal tissues of plants without causing any negative effects and can bring several benefits. Interactions between microorganism and the host are diverse and complex, the molecular basis of these interactions is not yet fully understood. Genomic analysis can aid to predict the set of gene-related functions that are involved in plant-endophyte interactions. The genome of 1274 was explored and a comparison was carried out from five representative genomes (endophytics, plant pathogenic and clinical) of *Serratia marcescens.* Were predicted 38 genomic islands (GIs) in strain 1274 with 460 genes, and only seven are exclusive of endophytic genomes, such as accessory components of type VI secretion system (T6SS) and regulation genes of Type 1 fimbriae protein. Genomic analysis showed the presence of genes that encode function involved in endophytic behavior in 1274 genome and may represent advantageous differences in interaction with the host plant.

1. Introduction

Serratia marcescens is an opportunistic, potentially human-pathogenic, Gram-negative rod-shaped bacterium, which belongs to family Enterobacteriaceae. Furthermore, there are isolates that can be found naturally in the environment in association with plants as endophytes and as phytopathogens (Khan et al., 2017; Wang et al., 2015a). Of the *S. marcescens* complete genome sequences available in public databases until recently, there are diverse lifestyles, as endophytes, phytopatogens, not plant-related and clinicals. Two correspond to the genome of an endothytic lineage, strain RSC-14 (Khan et al., 2015) and strain FS14 (Li et al., 2015).

By definition, endophytic microorganisms can colonize the internal tissues of plants without causing any negative effects or symptoms of disease. The bacteria are found inside the roots, stems and leaves of plants and can live inside cells, in the intercellular spaces or in the vascular system (Ryan et al., 2008). Endophytic bacteria have many beneficial effects to their hosts, including plant growth promotion (Mitter et al., 2013), resistance to stress (Sziderics et al., 2007) and biological control of pathogens (Krishnan et al., 2015). In addition, these bacteria provide sources of natural molecules to be employed in medicine and other areas such as agriculture and industry (Doley and Jha, 2015; Mitter et al., 2013).

Bacterial endophytes and rhizosphere-associated bacteria have similar beneficial effects for the host plant, nevertheless, endophytic bacteria might interact more closely within the host, this condition provide low competition to carbon sources and protected to environmental changes than rhizosphere soil bacteria (Reinhold-Hurek and Hurek, 1998). Interactions between microorganism and the host are diverse and complex, and molecular basis of these interactions not well described (Ryan et al., 2008). The genome analysis can provide new and useful information on genetic traits from endophytic bacteria that might be essential to interaction with plants (Mitter et al., 2013). Genetic engineering of endophytic bacteria can be easier than plants genetic engineering (Newman and Reynolds, 2005).

Comparative genome analysis is a powerful tool that can predict the possible gene-related functions that are involved in switching lifestyles adapted by bacteria at different conditions (Ali, 2014). A genomic study of endophytic bacterial strain, *Serratia marcescens* RSC-14 (isolated from *Solanum nigrum*), revealed the presence of genes that explain Cd-tolerant and plant growth promoting (Khan, 2017). The *S. marcescens* genome is highly dynamic, which reflects the diversity of environmental niches that the bacterium occupies. The numerous accessory genes emphasize diversity of the species and the dynamic nature of its genome, as when resistance level are correlated by the presence of multidrug-resistance genes (Moradigaravand et al., 2016).

The remarkable advancement of DNA sequencing technology and powerful genomic data means an effort to explore genomic processes *in silico*. To better understand genomic attributes that can contribute to endophytism in *Serratia* spp., we performed a comparative genome analysis of the endophytic strain *S. marcescens* 1274, including comparisons to recently sequenced genomes of other plant-associated *S. marcescens* isolates as well as to clinical isolates. Ours effort carried out to heterogeneity analysis of this species, to understand divergence in *S. marcescens* endophytic isolates seeking to fill in some of the gaps that involve understanding through genomic analyzes, of endophytic capacity in relation to other lifestyles, such as phytopathogens.

2. Material and methods

***Serratia marcescens* genomes**

The following genomes were used for comparative genome analysis between *Serratia marcescens* 1274 and others strains of the *S. marcescens* species with different lifestyles; the GenBank (NCBI) access numbers and feature of the strain are given in brackets: strain 1274 [CP019927.2, endophytic] (SantAnna, 2018); strain RSC-14 [CP012639.1, endophytic] (Khan, 2017); strain FS14 [CP005927, endophytic] (Li, 2015); strain B3R3 [CP013046.1, Phytopathogenic] (Wang, 2015a); strain CAV1492 [CP011642.1, clinical] (Tatusova, 2014); strain SM39 [AP013063.1, clinical] (Iguchi et al. 2014).

Comparative genome analysis

In order to understand molecular strategies to endophytic behavior and adaptation to the host, the endophytic genome of strain 1274 were compared with others *S. marcescens* strains analyzing their genomic architectures and targeted content. The antiSMASH database (Weber et al., 2015) was used to searching secondary metabolite biosynthesis gene clusters and T346Hunter web tool to predict bacterial secretion systems (Martínez-García et al., 2015).

The IslandViewer 4 tool (Bertelli et al. 2017) has identified genomic islands (GIs) in the current study based on two sequence composition methods, IslandPath-DIMOB and SIGI-HMM, and comparative methods named IslandPick. A comparison of genes in GIs was analyzed between strains of *S. marcescens* that contrast in lifestyle by OrthoVenn (Wang et al., 2015b).

The comparative analyzes focused on the sharing of characteristics between the endophytic genomes and the contrast with phytopathogenic and clinical strains. The genome mining of endophytic 1274 was lead from the presence and comparative analysis of genes that encode function involved in endophytic behavior. Functional annotation and search for domains to confirm gene sequences in results of comparative analysis were performed by Uniprot/Swissprot and pHMMER (EMBL-EBI) respectively.

3. Results and discussion

The bacterial behavior is very similar into rhizobacteria and phytopathogenic strains compared with plant host interactions. A couple of processes came out to be essential to endophytic capacity of a bacterium, such as motility and adhesion, detoxification of reactive oxygen species (ROS), plant polymer degradation, presence of secretion systems and iron and nitrogen uptake (Santoyo et al., 2016; Hong et al. 2016).

Table 1 shows the features of strains of 1274, RSC14, FS14 (endophytic), B3R3 (phytopathogenic) and CAV1492, SM39 (clinical), although smaller genomes, endophytic strains conserved the presence of secondary metabolites, genomic islands, and Type VI secretion system components.

Table 1. Architecture and features of the *Serratia marcescens* genomes.

Features	***Serratia marcescens strains***					
	1274	**RSC14**	**FS14**	**B3R3**	**CAV1492**	**SM39**
GenBank (NCBI)	CP019927.2	CP012639.1	CP005927	CP013046.1	CP011642.1	AP013063.1
Genome size (Mb)	5,21	5.12	5.24	5.59	5.82	5.32
Genes	4,979	4,849	4,918	5,356	5,649	5,091
Genomic Islands (GIs)	38	31	40	48	62	35
Secondary Metabolite (Clusters)	7	8	6	7	7	8
T6SS clusters (n^o components)	2 (14;18)	2 (11;15)	2 (11;15)	2 (15;11)	2 (12;15)	1 (15)
T4SS	Type G	-	-	Type G	Type P	-
Source	Endophytic	Endophytic	Endophytic	Pyhtopatogen	Clinical	Clinical

*T6SS and T4SS: Type VI and Type IV Secretion system;

Antibiotics and secondary metabolites

Gene clusters potentially related to biosynthesis of antibiotics and secondary metabolites, was done by searching against the antiSMASH database, was found Thiopeptide, Hserlactone, Siderophore and non-ribosomal peptides sintetase (Nrps) (Table 2).

A cluster of Nrps including turnerbactin biosynthetic genes, no found to clinical strain

CAV1492, are involved with catecholate siderophore expression and contribute to iron and nitrogen fixation in the host (Liu et al., 2017). Also implicated in plant benefits, the clusters of Siderophore and Nrps/Enterobactin are involved in the acquisition and transport of iron and Hserlactone (homoserine lactone) as quorum-sensing signals. Antibacterial activity is present by Lysobactin and Thiopeptide, especially due to its profile against Gram-positive bacteria (Just-Baringo et al., 2014).

Table 2
Gene clusters potentially involved in the synthesis of secondary metabolites and antibiotics in the genome of *Serratia marcescens* 1274 by AntiSmash 4.0 and shared among others *S.marcescens* strains.

Cluster	Type Synthetase	Metabolites	Size (pb)	MIBiG*	Shared to strains
1	Thiopeptide	O-antigen	26,442	BGC0000781_c1	RSC-14; FS14; B3R3; CAV1492; SM39
2	Hserlactone	-	20,692	-	RSC-14
3	Nrps	Enterobactin	70,150	BGC0000343_c1	RSC-14; FS14; B3R3; SM39
4	Nrps	Lysobactin	54,628	BGC0000385_c1	-
5	Siderophore	-	11,859	-	CAV1492
6	Nrps	-	86,129	-	-
7	Nrps	Turnerbactin	57,430	BGC0000451_c1	RSC-14; FS14; B3R3; SM39

*Minimal Information about a Biosynthetic Gene cluster (MIBiG)

Comparative genomics to endophytic analyzes

In comparative genomic analysis, the genes encoded by the endophytic *S. marcescens* strains were subtracted from the clinicals and phytopatogenic strains. The six *S. marcescens* strains (shown in Table 1) share 3,802 genes while the three endophytes present 4,042 core genes. A little difference was observed for the plant-associated process and for this analysis EDGAR Server 2.1 (Blom et al., 2016) was used (Supplementary Figure S1).

The genome mining of the strain 1274 (Fig. 1) shows protein-coding genes involved in important processes of endophytic behavior. Furthermore, Genomic Islands (GIs) were predicted and their locations and contents analyzed.

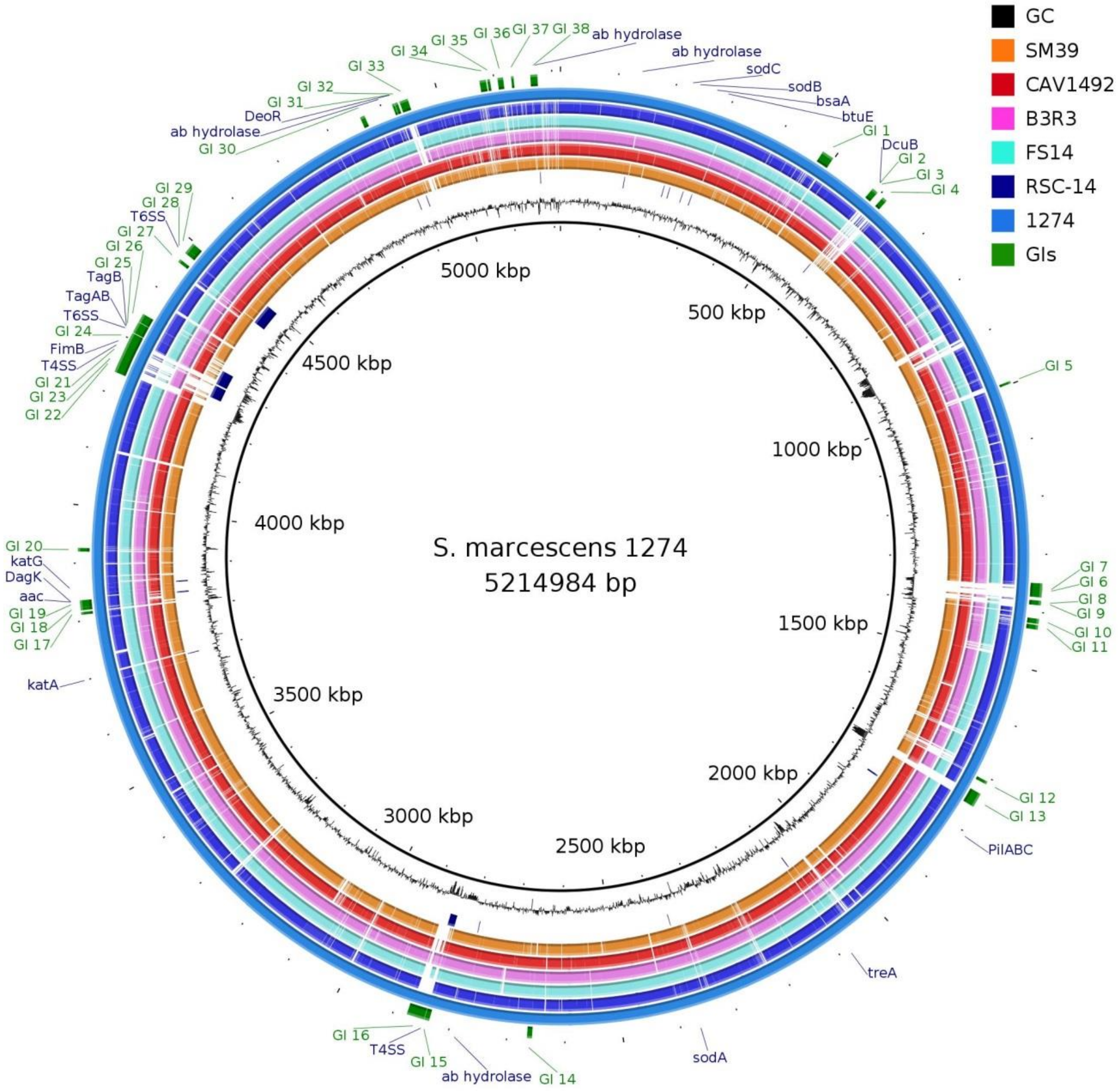
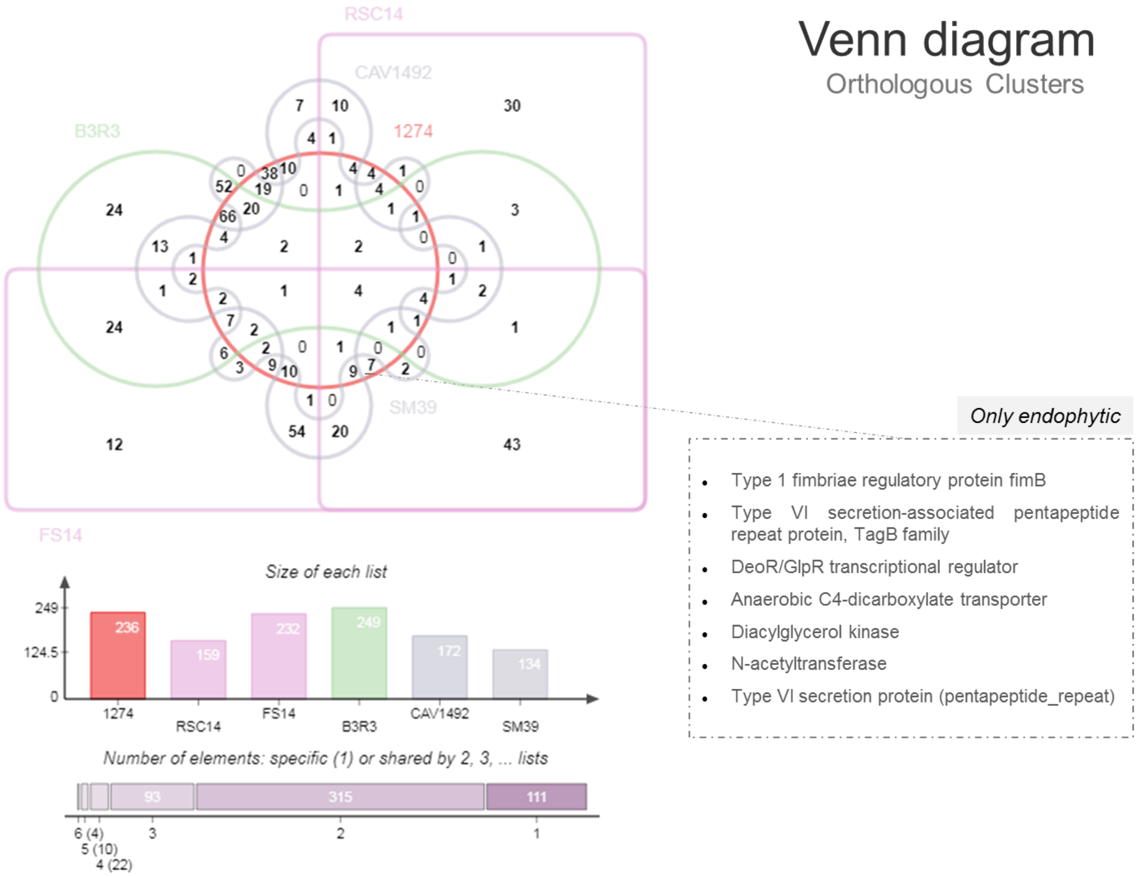


Fig.1 Genome alignment of endophytics *S. marcescens* 1274, RSC-14 and FS14, phytopatogenic B3R3 and clinicals CAV1492 and SM39. The outermost circle highlights some shared genes related with plant interaction in blue and Genomic Islands (GIs) in strain 1274 in green.

The predicted Genomic Island (GI) were compared and there is no evidence of sharing between strains and only four genes, from different islands present similarity, components of Type VI secretion systems *tssM, tssB* and *tssA* and pilus assembly *fimC* protein (Fig.2). The genomic islands characterized by the horizontal transfer of DNA contain genes involved in adaptation strategies. In our analyses was founded a higher difference to phytopathogenic and clinical isolates (B3R3, CAV1492 and SM39) that including several sequences referring to phages and toxins in their GIs, absent in the endophytic strains.



Only endophytic

- Type 1 fimbriae regulatory protein fimB
- Type VI secretion-associated pentapeptide repeat protein, TagB family
- DeoR/GlpR transcriptional regulator
- Anaerobic C4-dicarboxylate transporter
- Diacylglycerol kinase
- N-acetyltransferase
- Type VI secretion protein (pentapeptide_repeat)

Fig.2 Comparative analysis in Venn diagram showing the shared genes in predicted Genomic Islands of six representative *S. marcescens* 1274 (endophytic), RSC-14 (endophytic), FS14 (endophytic), B3R3 (phytopathogenic), CAV1492 (clinical) and SM39 (clinical). The cluster number in each component is displayed in the bar-plots and protein-coding genes shared only three endophytic bacteria are presented in the featured frame.

Through the comparative analysis of known features associated with GIs (Fig.2), 7 CDS are exclusively shared by endophytic strains, including regulatory proteins to fimbrial complex (Type 1 pili) and T6SS and transcriptional regulators DeoR Family with several copies in the genome envolved into negatively control genes to carbohydrate metabolism (Elgrably-Weiss et al., 2006). Besides the "Anaerobic C4-dicarboxylate transporter", member of the C4-Dicarboxylate Uptake (Dcu) family, involved in anaerobic growth of bacteria (Unden, 2016); and "Diacylglycerol kinase", engaged in recycling of diacylglycerol produced during the turnover of membrane phospholipid, increases the response to bacterial cell stress (Horn & Sanders, 2011). Considered essential in bacterial growth in challenging environments, along with the "Dcu" protein family, represent an increase in the adaptive capacity of endophytic bacteria. In addition, N-acetyltransferases (NATs) are enzymes wich broad specificity for aromatic amines and can catalize the transfer of acetyl groups. These enzymes can mediate to the adaptation of bacteria to their various niches by biotransformation of a variety of potential toxic aromatic compounds, including antibiotics (Kubiak et al., 2017; Martins et al, 2008).

Type VI secretion system (T6SS)

Commonly present in gram negative bacteria, the bacterial Type VI secretion system (T6SS) is a molecular machine used to carry effectors to prokaryotic or eukaryotic cells and represent a significant fitness advantage interbacterial competition (Bernal, 2018). The colonization by T6SS-active bacteria provides benefits in the interaction with plant host and can be identified in numerous bacterial endophytes (Ali et al., 2014).

T6SS is composed of a gene cluster varied with core components, called *tss* genes, and T6SS-associated genes (*tag* genes) (Shalom et al., 2007). Recently, Li et al. (2015) demonstrated the genetic organization the T6SS clusters found in *Serratia* sp. and separated four group families. The genome 1274 harbors two clusters of T6SS, including structural components VgrG, ClpV and core essential components Imp A-M (TssA-M homologs) (Fig. 3).

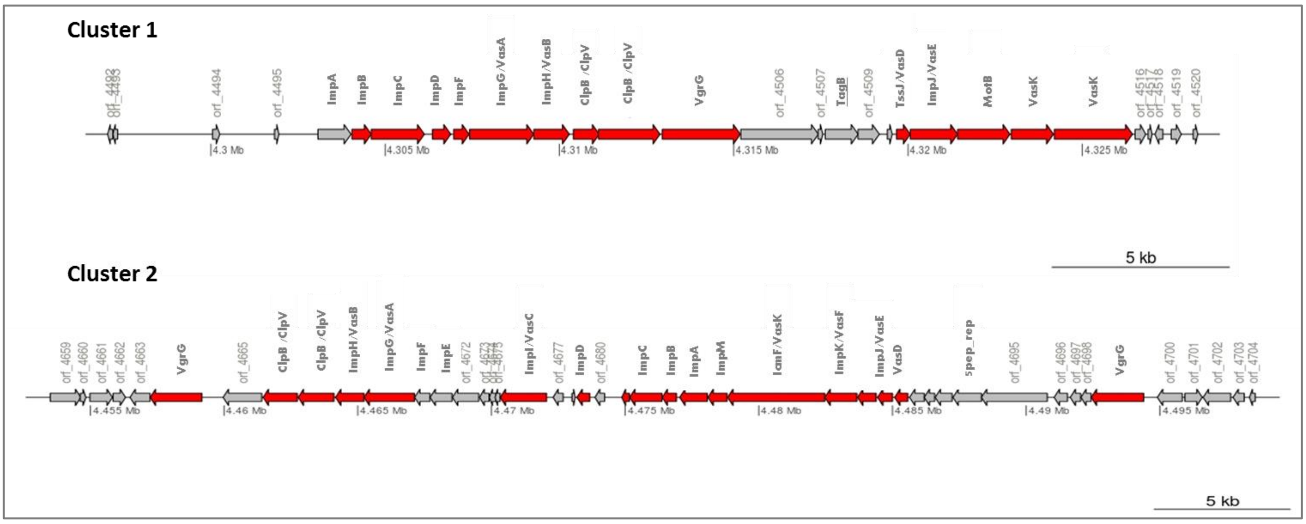


Fig. 3 Genetic architecture of T6SS clusters from *S. marcescens* 1274. The conserved core gene components of the T6SS are indicated in red and the *TagB*, associated pentapeptide repeat protein, is only in the first cluster.

Three type VI secretion proteins (*tssM, tssB and tssA*) were shared from all GIs of the *S. marcescens* strains (Fig.2), although all have clusters that encode the type VI secretion system (T6SS). Numerous components were found in several GIs, however arranged on different architectures. A Type VI secretion associated pentapeptide repeat protein, *TagB* family, that is accessory component that contribute to regulation of T6SS (Bernal, 2018), exclusively presented in GIs of endophytic strains.

Type IV secretion system (T4SS)

During plant colonization, bacterial process of adhesion are required to effective establishment in the endosphere, we found SfmA and SfmH (Fimbria-like adhesin) predicted to GIs and the Type IV secretory systems (T4SS) components were shared with pathogenic strain B3R3. The T4SS participate of macromolecules transporting like as proteins and protein-DNA complexes (Rego *et al.*, 2010). Moreover, some components are required on pilus biogenesis, as surface filaments or protein adhesin (Christie & Cascales, 2014). Hardoim (2015) detected T4SS more prominently among endophytes than among rhizobacteria, with protein-encoding genes involved in adhesion to host also more prominently than in nodulating symbionts.

Adhesion by Type I fimbriae

The first step of the colonization process in plant surface is the attachment of bacterial cells, which can be mediated by bacterial structures like as fimbriae. Among the seven coding genes indicated as exclusive of endophytic strains, the *FimB* (type 1 fimbriae regulatory protein), upstream of the *fimA*, is involved in control the expression of type 1 fimbriae (Klemm, 1986) (Fig.2). The type 1 fimbrial proteins are encoded by the *fim* gene cluster (*FimA-I-C-D-F-G*) and components *fimA, fimF* and *fimH* are essential in their assembly (Zeiner et al., 2012).

The B3R3 shows a *FimB* out of Genomic islands regulatong adhesion system in the host cell (Fig.4). The occurrence of *FimB* (type 1 fimbriae regulation protein) in endophytic strains 1274, RSC14 and FS14 resulting by horizontal transfer and absent in clinical strains CAV1492 and SM39, representing a potential advantage presents in endophytic bacteria.

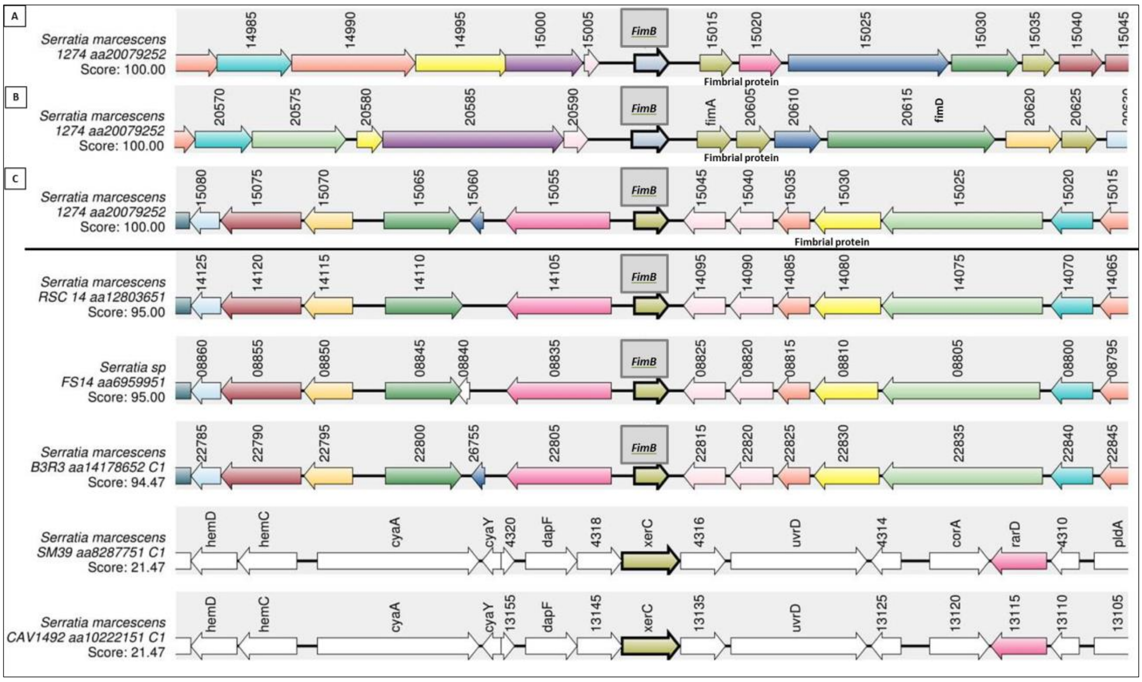


Fig.4 The synteny analysis across gene *FimB* in *Serratia marcescens* strain 1274 and together with other strains RSC14, FSC14, B3R3, CAV1492 and SM39 (as depicted by the program SyntTax). A, B and C represent the occurrence of that architecture in the endophytic genome 1274 and just below the other genomes. Highlighting the *FimB* gene present in *S. marcescens* plant-associated, strains RSC14, FSC14 and B3R3.

Adhesion and Twitching motility

The Type IV pili are multifunctional involving in twitching motility (Type IVa pili) and adhesion (Type IVb pili) and required *pilT* (Berne et al., 2015). As specific genes to endophytic behaviors, Type IV pili *PilA* and *PilT* (twitching motility) (Reinhold-Hurek *et al.*, 2015) are present in the plant pathogenic strain and *Serratia marcescens* endophytics (1274, RSC-14, FS14). Twitching motility refers a movement form flagella-independent of bacteria, widely distributed in endophytic strains (Mitter, 2013). The *PilT* are responsible to retractile force of the movement and is necessary for invasion of and establishment inside the plant (Böhm et al., 2006).

Plant colonization and plant polymer degradation

Successful plant tissue colonization involves the ability to interact and respond to obstacles naturally encountered in the host plant. The adaptation include antioxidant defense strategies, which can be mediated by first line defense antioxidants as superoxide dismutase (SOD), catalase (CAT) and glutathione peroxidase (GPX) (Ighodaro, 2017). Shared between endophytic strains 1274, RSC14 and FS14, were found glutathione synthetase (GSS), glutathione S-transferase (GSTs), glutathione peroxidase, superoxide dismutases (sodA, sodB and sodC) reactive oxygen species (ROS) allow to colonize plant tissue (Sessitsch *et al.*, 2012; Vicente 2016). The Glycoside hydrolases family was found in endophytes strains, as “alpha/beta hydrolase fold family protein”, “trehalose-6-phosphate hydrolase” and an endoglucanase, related to plant polymer degradation and play a key role in success of colonization (Vicente, 2016; Perez-Donoso, 2010). This hydrolases, involved to carbohydrates metabolism, allow to endophytic strains colonize plants (Ali, 2014).

Plant growth promotion

Various mechanisms may be appropriate in plant growth promoting bacteria, either by acquisition of resources, as nitrogen, phosphorous and iron, or modulating plant hormones, auxin, cytokinin or ethylene (Santoyo, 2016). In modulation of phytohormones, the genome of the strain 1274 evidenced an “Indole-3-glycerol phosphate synthase” (*trpc*), a precursor in tryptophan biosynthetic pathway involved to plant hormones indole-3-acetic acid (IAA) biosynthesis, one of the two major pathways for IAA in plants (Mano & Nemoto, 2012). Was also stablished in 1274 genome siderophores clusters, including non-ribosomal peptides (Table 2), and enzymes associated with the assimilation and regulation of nitrogen. In a plant growth-promoting experiment, the nitrogen fixation characteristics were relevant from Turnerbactin cluster in Pseudomonas psychrotolerans (Liu et al., 2017).

Plant defense response

Plant defense mechanisms trigger responses by recognition of invading microorganisms and may represent a major obstacle in the colonization of beneficial microorganisms. This recognition systems include microbe-associated molecular patterns (MAMPS), represented by flagellum, glycoproteins, lipopolysaccharides (Reinhold-Hurek et al., 2015). Although not well established role to flagellar proteins and Type III secretion system (T3SS) in endophytic competence or immune response of the host (Reinhold-Hurek and Hurek, 2011; Piromyou *et al.*, 2015; Vicente *et al.*, 2016), we found flagellar apparatus in all endophytic genomes (1274, RSC-14 and FS14) and the Negative regulator of flagellin synthesis (FlgM). The lack of flagella may improve endophytic colonization when the defense response is reduced (Iniguez *et al.*, 2005).

Despite of the presence of T3SS and T4SS are widespread in pathogens, flagella can be important to efficient endophytic colonization, found in 1274 strain can represents some relantionship whit host that contribute to colonization and where flagellins do not appear to act as PAMPS-eliciting defense responses (Buschart et al., 2012).

4. Conclusions

This study reveals the endophytic *S. marcescens* 1274 and strains isolated from diverse environments, with different adaptability profiles and behavior but with a high core genome. Since this species is able to live in diverse environments, sharing high number of genes, extending of genes will be continuous. The *S. marcescens* genomes studied no shows a higher variation but it has a survival capacity in several niche.

In this study we indicate the advantageous performance of the type VI secretion system and adhesion by type 1 fimbriae in the process of endophytic colonization. Furthermore, other features are essential as the balance in the interaction with the host, strategies for locomotion and adherence, quorum-sensing and benefits for the plant. The success as plant colonizing depends of bacterial ecology and the regulation of their responses, for establishment to competitive endophytic.

References

Ali S., Duan J., Charles T.C., Glick B.R. 2014. A bioinformatics approach to the determination of genes involved in endophytic behavior in *Burkholderia* spp. Journal of Theoretical Biology 343, 193–198. http://dx.doi.org/10.1016/j.jtbi.2013.10.007

Bernal P., Llamas M.A., Filloux A. 2018. Type VI secretion systems in plant-associated bacteria. Environmental Microbiology, 20(1), 1–15.

Berne C., Ducret A., Hardy G.G., Brun Y.V. 2015. Adhesins involved in attachment to abiotic surfaces by Gram-negative bacteria. Microbiol Spectr. 3(4). doi:10.1128/microbiolspec.MB-0018-2015.

Bertelli, C., Laird, M. R., Williams, K. P., Simon Fraser University Research Computing Group, Lau, B. Y., Hoad, G., Brinkman, F. S. 2017. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Research*, *45*(Web Server issue), W30–W35. http://doi.org/10.1093/nar/gkx343

Blom J, Kreis J, Spanig S, Juhre T, Bertelli C, Ernst C, Goesmann A. 2016. EDGAR 2.0: an enhanced software platform for comparative gene content analyses. Nucleic Acids Research, 44(W1):W22-8. doi: 10.1093/nar/gkw255

Bohm M., Hurek T., Reinhold-Hurek B. 2007. Twitching Motility Is Essential for Endophytic Rice Colonization by the N2-Fixing Endophyte *Azoarcus* sp. Strain BH72. Mol PlantMicrobe Interact 20(5):526–533. doi:10.1094/MPMI-20-5-0526

Buschart, A., Sachs, S., Chen, X., Herglotz, J., Krause, A., and Reinhold-Hurek, B. 2012. Flagella mediate endophytic competence rather than act as MAMPS in rice - Azoarcus sp. strain BH72 interac- tions. Mol. Plant Microbe Interact. 25, 191–199.

Christie P.J., Cascales E. 2005. Structural and dynamic properties of bacterial Type IV secretion systems (Review). Mol Membr Biol., 22(0): 51–61.

Doley, P., Jha, D.K. 2015. Antimicrobial Activity of Bacterial Endophytes from Medicinal Endemic Plant Garcinia lancifolia Roxb. Annals of Plant Sciences 4.12: 1243-1247

Elgrably-Weiss, M., Schlosser-Silverman, E., Rosenshine, I., Altuvia, S., 2006. DeoT, a DeoR-type transcriptional regulator of multiple target genes. FEMS Microbiol. Lett. 254, 141–148

Hardoim PR, van Overbeek LS, Berg G, Pirttilä AM, Compant S, Campisano A, Döring M, Sessitsch A. 2015. The hidden world within plants: ecological and evolutionary considerations for defining functioning of microbial endophytes. Microbiol Mol Biol Rev, 79(3). doi:10.1128/MMBR.00050-14.

Hong CE, Jeong H, Jo SH, Jeong JC, Kwon SY, An D, Park JM. 2016. A leaf-inhabiting endophytic bacterium, *Rhodococcus* sp. KB6, enhances sweet potato resistance to black rot disease caused by *Ceratocystis fimbriata*. J Microbiol Biotechnol 26:488–492

Horn W.D.V., Sanders C.R. 2013. Prokaryotic Diacylglycerol Kinase and Undecaprenol Kinase. Annu Rev Biophys. 41: 81–101. doi:10.1146/annurev-biophys-050511-102330.

Ighodaro O.M., Akinloye O.A. 2017. First line defence antioxidants-superoxide dismutase (SOD), catalase (CAT) and glutathione peroxidase (GPX): Their fundamental role in the entire antioxidant defence grid. Alexandria Journal of Medicine, 2017. https://doi.org/10.1016/j.ajme.2017.09.001

Iguchi, A., Nagaya, Y., Pradel, E., Ooka, T., Ogura, Y., Katsura, K., Kurokawa, K., Oshima, K., Hattori, M., Parkhill, J., Sebaihia, M., Coulthurst, S.J., Gotoh, N., Thomson, N.R., Ewbank, J.J., Hayashi, T. 2014. Genome evolution and plasticity of *Serratia marcescens*, an important multidrug-resistant nosocomial pathogen. Genome Biol. Evol. 6(8):2096–2110. doi:10.1093/gbe/evu160

Iniguez, A.L., Dong, Y., Carter, H.D., Ahmer, B.M., Stone, J.M., Triplett, E.W. 2005. Regulation of enteric endophytic bacterial colonization by plant defenses. Mol. Plant-Microbe Interact. 18:169–78.

Just-Baringo X, Albericio F, Álvarez M. Thiopeptide Antibiotics: Retrospective and Recent Advances. Marine Drugs. 2014;12(1):317-351. doi:10.3390/md12010317.

Khan, A. R., Ullah, I., Khan, A. L., Park, G. S., Waqas, M., Hong, S. J., Jung, B. K., Kwak, Y., Lee, I. J., Shin, J. H. 2015. Improvement in phytoremediation potential of *Solanum nigrum* under cadmium contamination through endophytic-assisted *Serratia* sp. RSC-14 inoculation. Environ Sci Pollut Res , 22:14032–14042. DOI 10.1007/s11356-015-4647-8

Khan AR, Park G-S, Asaf S, Hong S-J, Jung BK, Shin J-H 2017. Complete genome analysis of *Serratia marcescens* RSC-14: A plant growth-promoting bacterium that alleviates cadmium stress in host plants. PLoS ONE 12(2): e0171534. doi:10.1371/journal.pone.0171534

Klemm P. 1986. Two regulatory *fim* genes, *fimB* and *fimE*, control the phase variation of type 1 fimbriae in *Escherichia coli.* EMBO Journal vol.5 no.6 pp.1389-1393.

Krishnan, N., Gandhi, K., Peeran, M. F., Kuppusami, P., & Thiruvengadam, R. 2015. Molecular characterization and in vitro evaluation of endophytic bacteria against major pathogens of rice. African Journal of Microbiology Research, 9(11), 800-813.

Kubiak X., Duval R., Pluvinage B., Chaffotte A.F., Dupret J. M., Rodrigues-Lima F. 2016. Xenobiotic-metabolizing enzymes in Bacillus anthracis: molecular and functional analysis of a truncated arylamine N-acetyltransferase isozyme. British Journal of Pharmacology, 174 2174–2182. DOI:10.1111/bph.13647

Li, P., Kwok, A.H.Y., Jiang, J., Ran, T., Xu, D., Wang, W., Leung, F.C. 2015. Comparative Genome Analyses of *Serratia marcescens* FS14 Reveals Its High Antagonistic Potential. PLoS ONE 10(4): e0123061. doi:10.1371/journal.pone.0123061.

Liu R, Zhang Y., Chen P., Lin H., Ye G., Wang Z., Ge C., Zhu B., Ren D. 2017. Genomic and phenotypic analyses of *Pseudomonas psychrotolerans* PRS08-11306 reveal a turnerbactin biosynthesis gene cluster that contributes to nitrogen fixation. Journal of Biotechnology, Vol 253, p 10-13. https://doi.org/10.1016/j.jbiotec.2017.05.012

Mano Y., Nemoto K. 2012. The pathway of auxin biosynthesis in plants. Journal of Experimental Botany, Vol. 63, No. 8, pp. 2853–2872. doi:10.1093/jxb/ers091

Martínez-García PM, Ramos C, Rodríguez-Palenzuela P (2015) T346Hunter: A Novel Web-Based Tool for the Prediction of Type III, Type IV and Type VI Secretion Systems in Bacterial Genomes. PLoS ONE 10(4): e0119317. doi:10.1371/journal. pone.0119317

Martins M, Pluvinage B, Li de la Sierra-Gallay I, Barbault F, Dairou J,Dupret J-Met al.(2008). Functional and structural characterization ofthe arylamine N-acetyltransferase from the opportunistic pathogenNocardia farcinica.JMolBiol383:549–560

Mitter, B., Petric, A., Chain, P.S., Trognitz, F., Nowak, J., Compant, S., Sessitsch, A. 2013. Genome Analysis, Ecology, and Plant Growth Promotion of the Endophyte *Burkholderia phytofirmans* Strain PsJN. Molecular Microbial Ecology of the Rhizosphere, Volume 1, Chapter 81, First Edition. Edited by Frans J. de Bruijn.

Moradigaravand D., Boinett C.J., Martin V., Peacock S. J., Parkhill J. 2016. Recent independent emergence of multiple multidrug-resistant *Serratia marcescens* clones within the United Kingdom and Ireland. Genome Research 26:1101–1109. doi: 10.1101/gr.205245.116.

Newman, L.A., Reynolds, C.M. 2005. Bacteria and phytoremediation: new uses for endophytic bacteria in plants. TRENDS in Biotechnology, Vol.23 No.1.

Piromyou, P., Songwattana, P., Greetatorn, T., Okubo, T., Kakizaki, K. C., Prakamhang, J., … Minamisawa, K. 2015. The Type III Secretion System (T3SS) is a Determinant for Rice-Endophyte Colonization by Non-Photosynthetic *Bradyrhizobium*. *Microbes and Environments*, *30*(4), 291–300. http://doi.org/10.1264/jsme2.ME15080

Pérez-Donoso A. G., Sun Q., Roper M. C., Greve L. C., Kirkpatrick B., Labavitch J. M. 2010. Cell wall-degrading enzymes enlarge the pore size of intervessel pit membranes in healthy and *Xylella fastidiosa*-infected grapevines. *Plant Physiol.* 152 1748–1759

Rego, A.T., Chandran, V., Waksman, G. 2010. Two-step and one-step secretion mechanisms in Gram-negative bacteria: contrasting the type IV secretion system and the chaperone-usher pathway of pilus biogenesis. Biochem. J. 425, 475–488. doi:10.1042/BJ20091518

Reinhold-Hurek, B., Hurek, T. 1998. Life in grasses: diazotrophic endophytes. Trends Microbiol. 6:139–144.

Reinhold-Hurek, B., Hurek, T. 2011. Living inside plants: bacterial endophytes. Curr. Opin. Plant Biol. 14:435–43

Reinhold-Hurek, B., Bunger, W., Burbano, C.S., Sabale, M., Hurek, T. 2015. Roots shaping their microbiome: global hotspots for microbial activity. Annu. Rev. Phytopathol., 53:403–24. doi: 10.1146/annurev-phyto-082712-102342

Ryan, R.P., Germaine, K., Franks, A., Ryan, D.J., Dowling, D.N. 2008. Bacterial endophytes: recent developments and applications. FEMS Microbiol Lett 278, 1-9. DOI:10.1111/j.1574-6968.2007.00918.x

SantAnna BM, Marbach PP, Rojas-Herrera M, De Souza JT, Roque MR, Queiroz AT. 2015. High-Quality Draft Genome Sequence of *Bacillus amyloliquefaciens* strain 629, an Endophyte from *Theobroma cacao*. Genome Announcements, 3 (6), e01325-15. Doi:10.1128/genomeA.01325-15.

Santoyo G, Moreno-Hagelsieb G, Orozco-Mosqueda MC, Glick BR 2016. Plant growth-promoting bacterial endophytes. Microbiol Res., 183:92-99. Doi: 10.1016/j.micres.2015.11.008.

Sessitsch, A., Hardoim, P., Döring, J., Weilharter, A., Krause, A., Woyke, T., Mitter, B., Hauberg-Lotte, L., Friedrich, F., Rahalkar, M., Hurek, T., Sarkar, A., Bodrossy, L., van Overbeek, L., Brar, D., van Elsas, J.D., Reinhold-Hurek, B. 2012. Functional Characteristics of an Endophyte Community Colonizing Rice Roots as Revealed by Metagenomic Analysis. Molecular Plant-Microbe Interactions, Vol. 25, No. 1, 29.

Shalom, G., Shaw, J.G., Thomas, M.S. 2007. In vivo expression technology identifies a type VI secretion system locus in *Burkholderia pseudomallei* that is induced upon invasion of macrophages. Microbiology 153, 2689e2699.

Sziderics, A.H., Rasche, F., Trognitz, F., Sessitsch, A., Wilhelm, E. 2007. Bacterial endophytes contribute to abiotic stress adaptation in pepper plants (Capsicum annuum L.). Canadian Journal of Microbiology, Vol. 53, No. 11: pp. 1195-1202. 2007. (doi: 10.1139/W07-082)

Tatusova T., Ciufo S., Fedorov B., O'Neill K., Tolstoy I. 2014. RefSeq microbial genomes database: new representation and annotation strategy. Nuclei Acids Res 42:D553–D559

Vicente, C., Nascimento, F., Barbosa, P., Ke, H-M., Tsai, I.J. Hirao, T., Cock, P.J.A., Kikuchi, T., Hasegawa, K., Mota, M. 2016. Evidence for an opportunistic and endophytic lifestyle of the *Bursaphelenchus xylophilus*-associated bacteria *Serratia marcescens* PWN146 isolated from wilting *Pinus pinaster*. Microbial Ecology, Vol. 72, nº 3, p669-681. DOI: 10.1007/s00248-016-0820-y

Wang, X.Q., Bi, T., Li, X.D., Zhang, L.Q., Lu, S.E. 2015a. First Report of Corn Whorl Rot Caused by *Serratia marcescens* in China. Journal of Phytopathology, Short Communication. Doi: 10.1111/jph.12366

Wang, Y., Coleman-Derr, D., Chen, G., & Gu, Y. Q. 2015b. OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Research*, *43*(Web Server issue), W78–W84. http://doi.org/10.1093/nar/gkv487

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., Medema, M. H. (2015). antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research*, *43*(Web Server issue), W237–W243. http://doi.org/10.1093/nar/gkv437

Zeiner S.A., Dwyer B.E., Clegg S. 2012. FimA, FimF, and FimH Are Necessary for Assembly of Type 1 Fimbriae on *Salmonella enterica* Serovar Typhimurium. Infection and Immunity, 80(9), p. 3289 –3296. doi: 10.1128/IAI.00331-12

5. Considerações finais

- Os genomas de *Bacillus velezensis* 629 e *Serratia marcescens* 1274 apresentam conteúdos gênicos que representam características benéficas na interação com plantas e que facilitam o processo de colonização no tecido vegetal;
- Novas ferramentas aplicadas para análises de taxonomia do genoma evidenciaram uma inconstância nas classificações bacterianas, especialmente no gênero *Bacillus*, que podem representar atualizações nos depósitos já existentes;
- A *Serratia marcescens* 1274, como bactéria gram-negativa, apresenta vantagens no processo de colonização do hospedeiro por possuir sistemas de secreção tipo VI (T6SS);
- Reguladores do T6SS e do sistema de adesão por fimbrias tipo I, encontrados em ilhas genômicas, representam um diferencial em isolados endofíticos de *S. marcescens*;
- Processos metabólicos comuns em endofíticos como capacidade de resistir aos mecanismos de defesa das plantas, capacidade de colonização e interações benéficas como a mediação de fitormônios, produção de antimicrobianos e assimilação de nutrientes podem ser mediados por diferentes vias e as estratégias de colonização serão específicas da relação hospedeiro-endofítico;
- Não é possível destacar uma assinatura molecular entre endofíticos já que as interações endofítico-hospedeiro possuem características específicas.

6. Referências

Ali S., Duan J., Charles T.C., Glick B.R. 2014. A bioinformatics approach to the determination of genes involved in endophytic behavior in *Burkholderia spp.* Journal of Theoretical Biology 343, 193–198. http://dx.doi.org/10.1016/j.jtbi.2013.10.007

Arahal D.R. 2014. Whole-Genome Analyses: Average Nucleotide Identity. Methods in Microbiology - New Approaches to Prokaryotic Systematics, Vol 41, p103-122. DOI: 10.1016/bs.mim.2014.07.002

Bankevich A., Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology, 19(5):455-77. DOI: 10.1089/cmb.2012.0021.

Basharat, Z., Yasmin, A. 2016. Pan-genome Analysis of the Genus *Serratia*. Manuscript, arXiv preprint arXiv:1610.04160.

Berg, G.; Krechel, A.; Ditz, M.; Sikora, R.A.; Ulrich, A.; Hallmann, J. Endophytic and ectophytic potato-associated bacterial communities differ in structure and antagonistic function against plant pathogenic fungi. FEMS Microbiol. Ecol. v.51, p.215-229, 2005

Bilofsky, H. S.; Burks, C: The GenBank genetic sequence data bank. Nucleic Acids Res. v. 16, 1988, p. 1861-1863.

Bleidorn, C. 2015. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. Systematics and Biodiversity, 1-8. DOI: 10.1080/14772000.2015.1099575

Borriss R. (2011). Use of plant-associated *Bacillus* strains as biofertilizers and biocontrol agents in agriculture, in Bacteria in Agrobiology: Plant Growth Responses, ed Maheshwari D. K. (ed). Springer-Verlag Berlin Heidelberg, 41–76. DOI 10.1007/978-3-642-20332-9_3

Buschart, A., Sachs, S., Chen, X., Herglotz, J., Krause, A., and Reinhold-Hurek, B. 2012. Flagella mediate endophytic competence rather than act as MAMPS in rice - *Azoarcus sp*. strain BH72 interactions. Mol. Plant Microbe Interact. 25, 191–199.

Cai XC, Liu CH, Wang BT, Xue YR (2017) Genomic and metabolic traits endow *Bacillus velezensis* CC09 with a potential biocontrol agent in control of wheat powdery mildew disease. Microbiological Research 196:89–94. http://dx.doi.org/10.1016/j.micres.2016.12.007

Chaudhry, V., Patil, P. 2016. Genomic investigation reveals evolution and lifestyle adaptation of endophytic *Staphylococcus epidermidis*.Scientific Reports, 6:19263. DOI: 10.1038/srep19263.

Chevreux, B., Wetter, T., Suhai, S. 1999. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99, pp. 45-56.

Chin-A-Woeng, T.F.C.; Bloemberg, G.V.; Mulders, I.H.M.; Dekkers, L.C.; Lugtenberg, B.J.J. Root colonization by phenazine-1-carboxamide- producing bacterium *Pseudomonas chlororaphis* PCL1391 is essential for biocontrol of tomato foot and root rot. Mol. Microbe-Plant Interact., v.12, p.1340-1345, 2000.

Cochrane, G. *et al.* The International Nucleotide Sequence Database Collaboration.

Nucleic Acids Research, 2016, Vol. 44, Database issue, D48–D50. doi: 10.1093/nar/gkv1323.

Doley, P., Jha, D.K. 2015. Antimicrobial Activity of Bacterial Endophytes from Medicinal Endemic Plant Garcinia lancifolia Roxb. Annals of Plant Sciences 4.12: 1243-1247

Dunlap CA, Kim SJ, Kwon SW, Rooney AP (2016) *Bacillus velezensis* is not a later heterotypic synonym of *Bacillus amyloliquefaciens*; *Bacillus methylotrophicus*, *Bacillus amyloliquefaciens subsp. plantarum* and '*Bacillus oryzicola*' are later heterotypic synonyms of *Bacillus velezensis* based on phylogenomics. Int. J. Syst. Evol. Microbiol 66(3):1212–1217. Doi: 10.1099/ijsem.0.000858

Ekblom R., Wolf J.B.W. 2014. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl., 7(9): 1026–1042.

Fan B, Blom J, Klenk HP, Borriss R (2017) *Bacillus amyloliquefaciens*, *Bacillus velezensis*, and *Bacillus Siamensis* Form an "Operational Group *B. amyloliquefaciens*" within the *B. Subtilis* Species Complex. Front Microbiol 8:22. doi: 10.3389/fmicb.2017.00022

Fleischmann, R. D.; Adams, M. D.; White, O.; Clayton, R. A.; Kirkness, E. F.; Kerlavage, A. R.; Bult, C. J.; Tomb, J. F.; Dougherty, B. A.; Merrick, J. M.; *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science, v. 269, p. 496-512, 1995.

Fraser, C. M.; Gocayne, J. D.; White, O.; Adams, M. D.; Clayton, R. A.; Fleischmann, R. D.; Bult, C. J.; Kerlavage, A. R.; Sutton, G.; Kelley, J. M.; Fritchman, R. D.; Weidman, J. F.; Small, K. V.; Sandusky, M.; Fuhrmann, J.; Nguyen, D.; Utterback, T. R.; Saudek, D. M.; Phillips, C. A.; Merrick, J. M.; Tomb, J. F.; Dougherty, B. A.; Bott, K. F.; Hu, P. C.; Lucier, T. S.; Peterson, S. N.; Smith, H. O.; Hutchison, C. A., 3rd; Venter, J. C. The minimal gene complement of Mycoplasma genitalium. Science, v. 270, p. 397-403, 1995.

Guo, B.; Wang, Y.; Sun, X.; Tang, K. Bioactive natural products from endophytes: a review. Appl. Biochem. Microbiol., v.44, p.136-142, 2008.

Grada A, Weinbrecht K. 2013. Next-generation sequencing: methodology and application. J Invest Dermatol. 133:e11. doi:10.1038/jid.2013.248

He Y., Zhang Z., Peng X., Wu F., Wang J. 2013. De Novo Assembly Methods for Next Generation Sequencing Data. Tsinghua Science and Technology, 18(5): 500-514.

Huang, X., Madan, A. 1999. CAP3: A DNA sequence assembly program. Genome Research, 9, 868-877. doi:10.1101/gr.9.9.868

Hutchison III, C. A. 2007. DNA sequencing: bench to bedside and beyond. Nucleic Acids Res. v. 35, p. 6227-6237.

International Human Genome Sequencing Consortium (IHGSC) 2001. Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Khan AR, Park G-S, Asaf S, Hong S-J, Jung BK, Shin J-H (2017) Complete genome analysis of *Serratia marcescens* RSC-14: A plant growth-promoting bacterium that alleviates cadmium stress in host plants. PLoS ONE 12(2):e0171534. doi:10.1371/journal.pone.0171534

Kim, M.; Oh, H. S.; Park, S. C.; Chun, J. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. International Journal of Systematic and Evolutionary Microbiology, 64, 346–351. DOI 10.1099/ijs.0.059774-0.

Krishnan, N., Gandhi, K., Peeran, M. F., Kuppusami, P., & Thiruvengadam, R. 2015. Molecular characterization and in vitro evaluation of endophytic bacteria against major pathogens of rice. African Journal of Microbiology Research, 9(11), 800-813.

Lahiri, A., Sanchini, A., Semmler, T., Schafer, H., Lewin, A. 2014. Identification and comparative analysis of a genomic island in Mycobacterium avium subsp. hominissuis. FEBS Lett., 588, pp. 3906–3911.

Li, P., Kwok, A.H.Y., Jiang, J., Ran, T., Xu, D., Wang, W., Leung, F.C. 2015. Comparative Genome Analyses of *Serratia marcescens* FS14 Reveals Its High Antagonistic Potential. PLoS ONE 10(4): e0123061. doi:10.1371/journal.pone.0123061.

Liu G, Kong Y, Fan Y, Geng C, Peng D, Sun M (2017) Whole-genome sequencing of *Bacillus velezensis* LS69, a strain with a broad inhibitory spectrum against pathogenic bacteria. J Biotechnol 249:20-24. Doi: 10.1016/j.jbiotec.2017.03.018

Langille, M.G.I., Hsiao, W.W.L., Brinkman, F.S.L. 2010. Detecting genomic islands using bioinformatics approaches. Nature Reviews - Microbiology, vol 8:373-382.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience, 1(1):18. doi: 10.1186/2047-217X-1-18

Margulies M., Egholm M., Altman W. E., *et al.* 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.

Malfanova, N.; Lugtenberg, B. J. J.; Berg, G. Bacterial Endophytes: Who and Where, and What Are They Doing There? Molecular Microbial Ecology of the Rhizosphere, Volume 1, First Edition. Edited by Frans J. de Bruijn, Chapter 36, 2013.

Mendes, R.; Azevedo, J.L. Valor biotecnológico de fungos endofíticos isolados de plantas de interesse econômico. In: Maia, L.C.; Malosso, E.; Yano-Melo A.M. (Ed.). Micologia: avanços no conhecimento. Ed. Universitária da UFPE, Recife, pp.129-140, 2007.

Miller JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. Genomics 95: 315–327.

Mitter, B., Petric, A., Chain, P.S., Trognitz, F., Nowak, J., Compant, S., Sessitsch, A. 2013. Genome Analysis, Ecology, and Plant Growth Promotion of the Endophyte Burkholderia phytofirmans Strain PsJN. Molecular Microbial Ecology of the Rhizosphere, Volume 1, Chapter 81, First Edition. Edited by Frans J. de Bruijn.

Mukherjee, S.; Stamatis, D.; Bertsch, J.; Ovchinnikova, G.; Verezemska, O.; Isbandi, M.; Thomas, A. D.; Ali, R; Sharma, K; Kyrpides, N. C.; Reddy, T. B. K. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. Nucl. Acids Res. (2016); doi: 10.1093/nar/gkw992

Newman, L.A., Reynolds, C.M. 2005. Bacteria and phytoremediation: new uses for endophytic bacteria in plants. TRENDS in Biotechnology, Vol.23 No.1.

Reinhold-Hurek, B., Hurek, T. 1998. Life in grasses: diazotrophic endophytes. Trends Microbiol. 6:139–144.

Reinhold-Hurek, B., Hurek, T. 2011. Living inside plants: bacterial endophytes. Curr. Opin. Plant Biol. 14:435–43

Reinhold-Hurek, B., Bunger, W., Burbano, C.S., Sabale, M., Hurek, T. 2015. Roots shaping their microbiome: global hotspots for microbial activity. Annu. Rev. Phytopathol., 53:403–24. doi: 10.1146/annurev-phyto-082712-102342

Ruiz-García, C., Béjar, V., Martı´nez-Checa, F., Llamas, I., Quesada, E. 2005. Bacillus velezensis sp. nov., a surfactant producing bacterium isolated from the river Vélez in Málaga, southern Spain.

Ryan, R.P., Germaine, K., Franks, A., Ryan, D.J., Dowling, D.N. 2008. Bacterial endophytes: recent developments and applications. FEMS Microbiol Lett 278, 1-9. DOI:10.1111/j.1574-6968.2007.00918.x

Sanger, F.; Nicklen, S.; Coulson, A. R. DNA sequencing with chain terminating inhibitors. Proc Natl Acad Sci USA, v. 74, p. 5463-5467, 1977.

Sessitsch, A., Hardoim, P., Döring, J., Weilharter, A., Krause, A., Woyke, T., Mitter, B., Hauberg-Lotte, L., Friedrich, F., Rahalkar, M., Hurek, T., Sarkar, A., Bodrossy, L., van Overbeek, L., Brar, D., van Elsas, J.D., Reinhold-Hurek, B. 2012. Functional Characteristics of an Endophyte Community Colonizing Rice Roots as Revealed by Metagenomic Analysis. Molecular Plant-Microbe Interactions, Vol. 25, No. 1, 29.

Simpson J.T., Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res., 19, 1117–1123.

Staden, R.; Beal, K. F.; Bonfield, J. K. The Staden package, 1998. Methods Mol Biol. v. 132, 2000, p. 115-130.

Sziderics, A.H., Rasche, F., Trognitz, F., Sessitsch, A., Wilhelm, E. 2007. Bacterial endophytes contribute to abiotic stress adaptation in pepper plants (Capsicum annuum L.). Canadian Journal of Microbiology, Vol. 53, No. 11: pp. 1195-1202. 2007. (doi: 10.1139/W07-082)

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pangenome. Curr. Opinion in Microbiol., 11(5)472–477. Doi: 10.1016/j.mib.2008.09.006

van Dijk, E. L.; Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet 2014; 30:418-26.

Venter J.C., Adams, M.D., Myers E. W. et al. 2001. The sequence of the human genome. Science 291, 1304-1351.

Verli, H. 2014. Bioinformática da Biologia à flexibilidade molecular. B615, cap. 1, Sociedade Brasileira de Bioquímica e Biologia Molecular, Porto Alegre.

Wang, X.Q., Bi, T., Li, X.D., Zhang, L.Q., Lu, S.E. 2015. First Report of Corn Whorl Rot Caused by Serratia marcescens in China. Journal of Phytopathology, Short Communication. Doi: 10.1111/jph.12366

Zerbino,D.R., Birney, E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res., 18, 821–829.

APÊNDICES

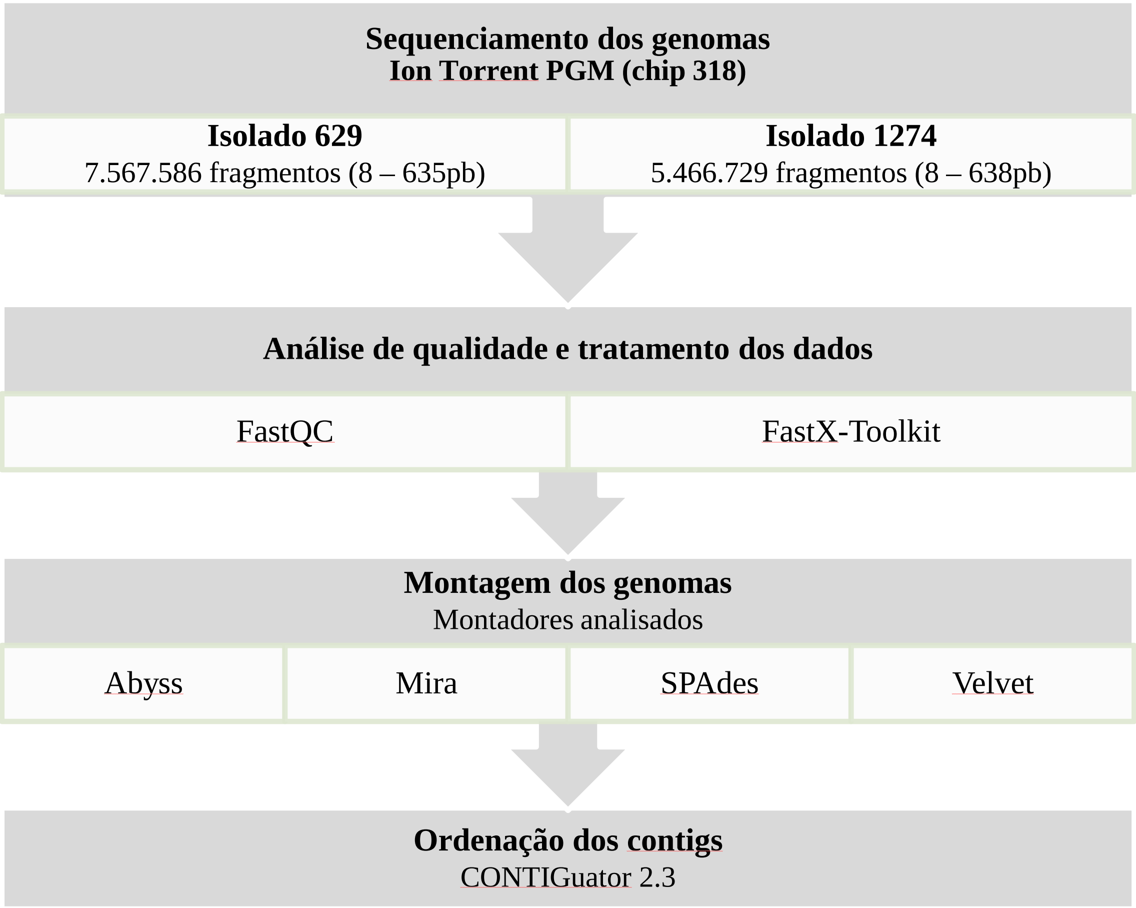
APÊNDICE A

Tratamento dos dados de sequenciamento e definição de montadores e parâmetros na montagem de genomas dos isolados 629 e 1274

Resumo

O processo de montagem de genomas varia de acordo com a plataforma de sequenciamento que gera os dados e o tipo de fragmentos. Variados montadores são constantemente desenvolvidos no intuito de aperfeiçoar os resultados de montagem de genomas. Diferentes parâmetros no tratamento de dados brutos e no processo de montagem podem variar e se adequar melhor a cada situação. Os genomas deste estudo, dos isolados 629 e 1274, apresentaram melhor qualidade dos dados quando trimados (t20/l50) e filtrados (q20/p80) antes da montagem. No processo de montagem *de novo*, o SPAdes gerou melhores resultados entre os montadores testados, utilizando valores de K de 33 (629) e 127 (1274) e *cutoff* de cobertura automático em ambas montagens. Os isolados 629 e 1274 obtiveram 129 e 116 fragmentos, respectivamente, no final da montagem *de novo*.

Resumo gráfico



Sequenciamento dos genomas
Ion Torrent PGM (chip 318)

Isolado 629
7.567.586 fragmentos (8 – 635pb)

Isolado 1274
5.466.729 fragmentos (8 – 638pb)

Análise de qualidade e tratamento dos dados

FastQC

FastX-Toolkit

Montagem dos genomas
Montadores analisados

Abyss

Mira

SPAdes

Velvet

Ordenação dos contigs
CONTIGuator 2.3

1. Introdução

Dois genomas bacterianos, previamente identificados por análises do 16S como *Bacillus amyloliquefaciens* 629 e *Serratia marcescens* 1274, foram sequenciados em plataforma Ion Torrent PGM (chip 318). Para avaliar melhor estratégia de montagem, os dados foram tratados e alguns montadores testados com diferentes parâmetros.

As análises foram realizadas com acesso a um servidor com capacidade para processamento e armazenamento dos dados, hospedado na Fiocruz-BA. A maioria das ferramentas utilizadas foram realizadas por linha de comando e algumas noções básicas para desenvolver as análises e utilizar as ferramentas incluem alguns comandos primordiais listados no Box 1.

Box 1 - Lista de comandos

- **cat**: Mostra o conteúdo de um arquivo e é muito usado também para concatenar arquivos, como por exemplo fazendo cat a.txt b.txt > c.txt" para juntar o arquivo a.txt e b.txt num único de nome "c.txt".
- **cd**: Mudar de diretório atual, como por exemplo: cd diretório, cd .., cd /.
- **cp**: Copiar arquivos.
- **grep**: Procura um arquivo por um padrão, sendo um filtro muito útil e usado, por exemplo: um cat a.txt | grep ola irá mostrar-nos apenas as linhas do arquivo a.txt que contenham a palavra "ola".
- **less**: Paginação de arquivos, funciona como o "more", para visualizar conteúdo do arquivo.
- **ls**: Lista o conteúdo de uma diretório.
- **ls -lh**: Lista o conteúdo do diretório com detalhes.
- **man**: Manual muito completo, pesquisa informação acerca de todos os comandos que necessitemos de saber, como por exemplo man find.
- **mkdir**: Criar uma diretório, vem de "make directory".
- **more**: Mostra o conteúdo de um arquivo, mas apenas um ecrã de cada vez, ou mesmo output de outros comandos, como por exemplo ls | more.
- **mv**: Move ou renomeia arquivos ou diretórios.
- **pwd**: Mostra-nos o caminho por inteiro da diretório em que nos encontramos em dado momento.
- **tar**: Cria ou extrai arquivos, muito usado como programa de backup ou compressão de arquivos.
- **rm**: Apaga arquivos, vem de remove. É preciso ter cuidado com o comando "rm*" pois apaga tudo sem confirmação por defeito.
- **wc**: Conta linhas, palavras e mesmo caracteres num arquivo.

2. Materiais e Métodos

2.1. Análises da qualidade dos dados

O sequenciamento realizado por Ion Torrent PGM (Chip 318) gerou dados de 7.567.586 fragmentos (reads) para o genoma 629 e 5.466.729 reads do genoma 1274. A qualidade dos fragmentos foi avaliada utilizando FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), considerando aceitáveis valores Phred iguais ou superiores a 20.

Para trimar e filtrar os dados foi utilizado o pacote de ferramentas FastX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), especificamente o *FASTQ Quality Trimmer* e o *FASTQ Quality Filter*. Foram consideradas bases com valor Phred igual ou maior que 20 para 80% das leituras. Os comandos utilizados para trimagem e filtro estão representados nos quadros 1 e 2, respectivamente:

Quadro 1 - Trimagem dos dados brutos utilizando *FASTQ Quality Trimmer*

```
fastq_quality_trimmer -Q 33 -t 20 -l 50 -i arquivo.fastq -o
arquivo_t20_l50.trimming.fastq
```

*onde,

- -Q 33, indica a referência para dados oriundos de plataforma Ion Torrent;
- -t 20, representa limite do valor Phred = 20 para as bases que serão removidas na extremidade dos *reads*;
- -l 50, remove *reads* menores que 50 pb

Quadro 2 - Filtro aplicado nos dados brutos utilizando *FASTQ Quality Filter*

```
fastq_quality_filter -Q 33 -q 20 -p 80 -i arquivo_t20_l50.trimming.fastq -o
arquivo_t20_p80_filter.fastq
```

*onde,

- -q 20, representa limite de Phred = 20 para as bases
- -p 80, determina a porcentagem mínima de 80% dos *reads* apresentarem o valor de qualidade Phred determinado em “-q”

2.2. Montadores e parâmetros

Foram analisados quatro montadores: Mira (Chevreux et al., 1999) (OLC), Abyss (Simpson et al., 2009), SPAdes (Bankevich et al., 2012) e Velvet (Zerbino & Birney, 2008) (DBG). Na montagem com o Mira, um arquivo chamado “*manifestfile*” é criado com os parâmetros da montagem, como descrito no Quadro 3.

Quadro 3 - Modelo de arquivo “manifestfile” utilizado na montagem com Mira com parâmetros padrões, foi utilizado por apresentar melhor resultado.

“*manifestfile*”

```
project = NomedoProjeto
job = genome,denovo,accurate
parameters = -GE:not=5 -NW:cac=no

readgroup = Iontor_NomedoGrupo
data = arquivo.fastq
technology = iontor
```

As montagens realizadas por SPAdes, Abyss e Velvet utilizam abordagem DBG, que exige determinação de alguns parâmetros como o valor de k-mer. Em todos os casos, apenas os melhores resultados são apresentados. Para o SPAdes, o comando seguiu uma linha comum aos dois genomas como representada no Quadro 4, com variadas opções de valores de k-mer e corte de cobertura como automático.

Quadro 4 - Linha de comando para montagem por SPAdes

Comando:

```
python spades.py -k 21,33,55,77,99,127 --cov-cutoff auto --iontorrent -s arquivo.fastq -
o directory_out
```

Para montagem com o Abyss, diferentes valores de K foram testados, mas o k=33 foi a melhor opção em ambos os genomas. O comando com parâmetros padrões de média de cobertura estão no Quadro 5.

Quadro 5 - Linha de comando para montagem dos genomas utilizando Abyss

Comando:

```
abyss -k 33 -c 150 -e 100 arquivo.fastq -o arquivo_saída.fa
```

A montagem por Velvet envolve dois passos, o primeiro comando por velveth, os arquivos são preparados. E no segundo comando, utilizando velvetg, é realizada a montagem por Grafo De Bruijn. O Quadro 6 apresenta os dois comandos.

Quadro 6 - Comandos para montagem utilizando montador Velvet.

1º comando:

```
velveth output_directory 31 -fastq -short arquivo.fastq
```

2º comando:

```
velvetg output_directory -max_coverage 150
```

O QUAST (Quality Assessment Tool for Genome Assemblies) foi utilizado para apresentar as estatísticas da montagem, como valores de N50, número de contigs e tamanho do genoma montado.

Ordenação dos contigs

A ordenação dos contigs gerados foi realizado utilizando um genoma de referência filogeneticamente próximo aos genomas montados, a partir da ferramenta CONTIGuator 2.3 (*http://contiguator.sourceforge.net/*) (GALARDINI, 2011).

Resultados

Análises da qualidade dos dados

Os dados que foram trimados e filtrados utilizando o FastX-Toolkit estão apresentados na Tabela 1. Antes e após o tratamento dos dados, a análise de qualidade dos dados foi observada no FastQC, que demonstrou resultados satisfatórios após a trimagem e filtragem dos dados.

Tabela 1 – Resultados dos dados tratados dos genomas 629 e 1274.

Bacillus Amyloliquefaciens 629					
Dados brutos			Dados tratados (t20/l50, q20/p80)		
% GC	Reads (Tamanho)	Total de Sequências	% GC	Reads (Tamanho)	Total de Sequências
46%	8 – 635	7.567.586	46	50 – 617	6.262.510
Serratia marcescens 1274					
Dados brutos			Dados tratados (t20/l50, q20/p80)		
% GC	Reads (Tamanho)	Total de Sequências	% GC	Reads (Tamanho)	Total de Sequências
59	8 – 638	5.466.729	59	50 – 634	4.267.824

Análises dos montadores e parâmetros

Nas análises de montadores, foram utilizados dados brutos por apresentarem qualidade aceitável. A Tabela 2 resume os resultados para os genomas do *B. amyloliquefaciens* 629 e *S. marcescens* 1274. Os genomas de referência foram selecionados de acordo com a identidade da sequência do gene 16S rRNA de cada um.

Tabela 2 – Resultados das montagens, por diferentes montadores, dos isolados *B. amyloliquefaciens* 629 e *S. marcescens* 1274.

<table>
<tr><td colspan="8"><b><i>Genoma – Bacillus amyloliquefaciens</i> 629 (</b>length: 3,9Mb)
<i>Ref.: B. amyloliquefaciens LH15 (length: 3.9Mb; GC: 46,7%)</i></td></tr>
<tr><td><b>Montador</b></td><td><b>Contigs</b></td><td><b>Total pb</b></td><td><b>N50</b></td><td><b>L50</b></td><td><b>%GC</b></td><td><b>Kmer</b></td><td><b>Contigs após Ordenação</b></td></tr>
<tr><td><b>Mira</b></td><td>179</td><td>4.027.772</td><td>46.957</td><td>21</td><td>46,49%</td><td>-</td><td><b>146</b></td></tr>
<tr><td><b>SPAdes</b></td><td>129</td><td>3.866.991</td><td>285.363</td><td>4</td><td>46,49%</td><td>33</td><td><b>33</b></td></tr>
<tr><td><b>Abyss</b></td><td>5.165</td><td>3.663.692</td><td>2.812</td><td>410</td><td>46,49%</td><td>33</td><td><b>1.105</b></td></tr>
<tr><td><b>Velvet</b></td><td>48.211</td><td>7.404.007</td><td>582</td><td>2</td><td>46%</td><td>31</td><td><b>Nenhum contig mapeado</b></td></tr>
<tr><td colspan="8"></td></tr>
<tr><td colspan="8"><b>Genoma – <i>Serratia marcescens</i> 1274 (</b>length: 5,2Mb)
<i>Ref.: S. marcescens WW4 (length: 5,2Mb; GC: 59,5%)</i></td></tr>
<tr><td><b>Assembler</b></td><td><b>Contigs</b></td><td><b>Total length</b></td><td><b>N50</b></td><td><b>L50</b></td><td><b>GC%</b></td><td><b>Kmer</b></td><td><b>Contigs após Ordenação</b></td></tr>
<tr><td><b>Mira</b></td><td>135</td><td>5.288.248</td><td>81.550</td><td>19</td><td>59,77%</td><td>-</td><td><b>96</b></td></tr>
<tr><td><b>SPAdes</b></td><td>116</td><td>5.215.196</td><td>589.570</td><td>3</td><td>59,86%</td><td>127</td><td><b>16</b></td></tr>
<tr><td><b>Abyss</b></td><td>7.575</td><td>4.108.694</td><td>1.471</td><td>886</td><td>59,18%</td><td>33</td><td><b>1.186</b></td></tr>
<tr><td><b>Velvet</b></td><td>38.894</td><td>5.822.081</td><td>552</td><td>2</td><td>55,06%</td><td>31</td><td><b>Nenhum contig mapeado</b></td></tr>
</table>

A montagem com SPAdes possibilita o uso de diferentes parâmetros na montagem *de novo* e diferentes valores de -cov_cutoff e k-mer foram testados (Tabela 3). Os resultados da qualidade da montagem são avaliados, principalmente, quanto ao menor número de contigs, maior valor de N50 e tamanho máximo dos contigs e número de pares de base (pb) gerados de acordo com o esperado. Magoc (2013) também obteve melhores resultados utilizando o SPAdes quando comparado aos montadores Mira, Abyss e Velvet, na montagem de 12 genomas bacterianos.

Tabela 3 – Análise de parâmetros na montagem *de novo* utilizando SPAdes 3.5.0 com os genomas do *B. amyloliquefaciens* 629 e *S. marcescens* 1274.

Genoma – *Bacillus amyloliquefaciens* 629					
SPAdes [options]	**K-mer**	Resultados			
		N50	Max	Contigs	Total pb
-cov_cutoff auto, 5 e 3	127	338.968	536.517	240	3.922.394
	99	169.629	535.971	203	3.889.321
	33	285.363	772.303	129	3.866.991
Genoma – *Serratia marcescens* 1274					
SPAdes [options]	**K-mer**	Resultados			
		N50	Max	Contigs	Total pb
-cov_cutoff auto	127	548.516	1.671.598	116	5.215.196
	99	270.913	724.907	260	5.216.607
	33	69.090	230.425	548	5.191.550
-cov_cutoff 5	127	381.708	947.590	130	5.222.143
-cov_cutoff 3	127	423.636	1.562.393	122	5.213.912

O valor de K=127 (K-mer) apresentou melhores resultados para a montagem do isolado 1274. Os resultados com o parâmetro de montagem "-cov_cutoff auto" ativado apresentaram melhores resultados na montagem do genoma da 1274, contudo não significou diferença na montagem da 629.

Considerações finais

O processo de montagem de genomas envolve diversos pacotes e programas que são aplicados na análise de qualidade dos dados, tratamento dos dados, montagem e visualização dos resultados. Além disso, exige o domínio do uso das ferramentas e na análise dos resultados. A compreensão da importância dos tratamentos dos dados (*reads* e bases) antes de se iniciar uma montagem de genoma e a escolha de parâmetros/opções no processo de montagem são crucial para obter resultados mais satisfatórios.

5. Referências

Bankevich, A. *et al.* 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology, Vol. 19, n 5. DOI: 10.1089/cmb.2012.0021.

Chevreux B, Wetter T, Suhai S (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99, pp. 45-56.

Galardini, M.; Biondi, E.G.; Bazzicalupo, M.; Mengoni, A. 2011. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. Source Code Biol Med., 6: 11. Doi: 10.1186/1751-0473-6-11.

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q. 2013. GAGE-B: An evaluation of genome assemblers for bacterial organisms. Bioinformatics, 29(14): 1718-1725. doi:10.1093/bioinformatics/btt273.

Simpson, J.T. et al. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res., 19, 1117–1123.

Zerbino, D.R., & Birney, E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res., 18, 821–829.

APÊNDICE B

Manuscrito (co-autoria):

GATOOL: a fast and user-friendly Genome Assembly web TOOL for Ion Torrent data

BMC Bioinformatics

GATOOL: a fast and user-friendly Genome Assembly web TOOL for Ion Torrent data.

--Manuscript Draft--

Manuscript Number:	
Full Title:	GATOOL: a fast and user-friendly Genome Assembly web TOOL for Ion Torrent data.
Article Type:	Software
Section/Category:	Sequence analysis (applications)
Funding Information:	
Abstract:	Background: The advances of Next Generation Sequencing (NGS) technologies exponentially increased the production of sequence data, mainly associated with the reduction of cost per sequenced-base and contributing to the expansion of genome sequencing projects. However, there are many sequencing platforms which require different strategies and software to perform an appropriate genome assembly. The installation of multiple programs, and dealing with their specificities and requirements, are also needed to perform data analysis and finishing the assembly. Additionally, advanced computational skills are required for performing most of the genome analysis protocols involving usage of command-line.. Moreover, the plethora of commands and parameters at the user's disposal increases the risk of operational errors. Aiming to facilitate the assembly process and reduce the operational errors risk we developed GATOOL, an user-friendly pipeline interface bacterial genome assembly using Ion Torrent data. Furthermore, we applied the pipeline using raw and SRA data to evaluate the performance of GATOOL. Results: Our tool provides a friendly and intuitive interface to perform analysis without the need of advanced computational skills. The user is guided through the genome assembly steps in an easy and fast way, without page changes. The workflow has two modules: (i) analysis/preprocessing: allows evaluation of of read qualities and also preprocessing such as trimming and quality filter; (ii) assembly/orientation: allows the performing of de novo assembly, choosing SPAdes or Velvet software. The tool also evaluate the assembly quality with QUAST. After this process, the contig orientation can be performed using a reference genome by the CONTIGuator. In the end of the process, scaffolds and contigs are made available in the user's folder. We also performed a comparative study with seven different Sequence Read Archive samples to validated the tool. All SRA samples were obtained from NCBI. GATOOL outperformed all the previously performed assemblies. Moreover, the N50 statistics, the number of contigs and scaffolds were better compared to other analysis. Conclusions: GATOOL is a complete tool to preprocessing analysis, genome assembly and contigs orientation of bacterial genomes. The tool can be used in a personal computer or installed in a server, as a web tool. Both interfaces are identical. GATOOL is made open-source and is available at: https://sourceforge.net/projects/gatool-beta/.
Corresponding Author:	Matheus Brito Oliveira, Esp. Instituto Federal de Educacao Ciencia e Tecnologia da Bahia BRAZIL
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Instituto Federal de Educacao Ciencia e Tecnologia da Bahia
Corresponding Author's Secondary Institution:	
First Author:	Matheus Brito Oliveira
First Author Secondary Information:	

Order of Authors:	Matheus Brito Oliveira
	Brena M. SantAnna
	Pablo Ivan Pereira Ramos
	Artur Trancoso Lopo de Queiróz
Order of Authors Secondary Information:	
Author Comments:	

Title

GATOOL: a fast and user-friendly Genome Assembly web TOOL for ion Torrent data.

Matheus Brito de Oliveira*[1,2], Brena M. SantAnna, Pablo Ivan Pereira Ramos[3] and Artur Trancoso Lopo de Queiróz[2,3,4].

Author's affiliation

[1] Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA), Salvador Bahia, Brazil.

[2] Post-graduation Program in Applied Computation, Universidade Estadual de Feira de Santana, Feira de Santana, Bahia, Brazil.

[3] Instituto Gonçalo Moniz, Fundação Oswaldo Cruz (FIOCRUZ), Salvador, Bahia, Brazil.

[4] Post-graduation Program in Biotechnology in Health and Investigative Medicine, FIOCRUZ, Salvador, Bahia, Brazil.

*** Corresponding author**

Email address: matheusbrito_si@hotmail.com

Abstract

Background:

The advances of Next Generation Sequencing (NGS) technologies exponentially increased the production of sequence data, mainly associated with the reduction of cost per sequenced-base and contributing to the expansion of genome sequencing projects. However, there are many sequencing platforms which require different strategies and software to perform an appropriate genome assembly. The installation of multiple programs, and dealing with their specificities and requirements, are also needed to perform data analysis and finishing the assembly. Additionally, advanced computational skills are required for performing most of the genome analysis protocols involving usage of command-line.. Moreover, the plethora of commands and parameters at the user's disposal increases the risk of operational errors. Aiming to facilitate the assembly process and reduce the operational errors risk we developed GATOOL, an user-friendly pipeline interface bacterial genome assembly using Ion Torrent data. Furthermore, we applied the pipeline using raw and SRA data to evaluate the performance of GATOOL.

Results:

Our tool provides a friendly and intuitive interface to perform analysis without the need of advanced computational skills. The user is guided through the genome assembly steps in an easy and fast way, without page changes. The workflow has two modules: (i) analysis/preprocessing: allows evaluation of of read qualities and also preprocessing such as trimming and quality filter; (ii) assembly/orientation: allows the performing of *de novo* assembly, choosing SPAdes or Velvet software. The tool also evaluate the assembly quality with QUAST. After this process, the contig orientation can be performed using a reference genome by the CONTIGuator. In the end of the process, scaffolds and contigs are made available in the user's folder. We also performed a comparative study with seven different Sequence Read Archive samples to validated the tool. All SRA samples were obtained from NCBI. GATOOL outperformed all the previously performed assemblies. Moreover, the N50 statistics, the number of contigs and scaffolds were better compared to other analysis.

Conclusions:

GATOOL is a complete tool to preprocessing analysis, genome assembly and contigs orientation of bacterial genomes. The tool can be used in a personal computer or installed in a server, as a web tool. Both interfaces are identical. GATOOL is made open-source and is available at: https://sourceforge.net/projects/gatool-beta/.

Keywords:

NGS, web-tool, Genome assembly, pipeline, de novo, Ion torrent, Bioinformatics.

Background:

Application of next-generation sequencing (NGS) techniques has profoundly impacted the fields of clinical microbiology and molecular epidemiology, also playing an important role in infectious diseases outbreaks management [1]. The complete genome of a pathogen can now be rapidly attained using NGS techniques, and this data can be further utilized for investigating the presence of resistance and virulence genes, as well as in outbreak investigation in nosocomial settings [2].

There are several methodologies for NGS with particular characteristics, such as variable read lengths, number of produced reads and chemistry, with a great range of applications [3]. Although data generation is fast, the analysis process requires a number of software packages, specific programming or scripting skills, and time. Many of the tools are difficult to install, with complex dependencies requirements, multiple parameters and command-line based. Among the reasons for these are that many tools are written by scientists and their focus rely more on the analysis problem itself than in usability, while most user-friendly tools are commercial, limiting their usage.

Recently, several tools were developed to simplify the genome assembly process, such as Orione [4], CLC Workbench (Qiagen, USA), Lasergene Suite [5], IonGAP [6] and SIMBA [7]. The Orione is a free web platform allowing genome analysis and assembly. This tool still requires too many configurations steps for both processes. Both CLC Workbench and Lasergene are easy to install and present good usability, but are only available commercially. Another free web tool with good usability is the IonGap, which performs steps such as read quality check, with a downside being that the user can not preprocess the fragments, for example could not filter the fragments that would meet a quality of sequencing of 80%. The open-source web tool SIMBA provides several functionalities and assembly alternatives. However, the assemblers using the Bruijn graph do not allow the user to change the values of k-*mer* nor to choose multiple values, as in the case of the Minia [8] genome assembler. Genome assemblies using the Bruijn graph, the value of k is crucial for good assembly [9].

To provide an alternative that allows circumventing these limitations in bacterial genome assembly, we developed GATOOL (Genome Assembly Tool). GATOOL provides a friendly interface capable of guiding the user through simple steps: from the easy installation, well defined analysis, preprocessing, assembly to the final scaffolding. Through these processes user interaction is required only in few steps. GATOOL can be installed on a personal computer for use or on a Linux server, being available through the web and can be used by any operating system through a browser.. The major tool objective is to help non-technical users and scientists to perform genome assembly through a friendly user interface.

Methods

The GATOOL processing pipeline is structured in two major modules: 1) analysis/preprocessing and 2) assembly/scaffolding. The interface was developed using PHP (http://secure.php.net) and Bootstrap front-end framework (http://getbootstrap.com).

Analysis and Preprocessing

The first module has implemented was FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). These implementations together allows the quality analysis and read processing. Also, the multiple reports generated during this step help the user in choosing the best filtering parameters, offering a dynamic intervention in preprocessing step while also permitting the comparison of different preprocessing strategies.

Genome Assembly and Scaffolding

For assembly/scaffolding process GATOOL has two softwares implemented: SPAdes 3.7.1 [10] and Velvet 1.2.10 [11]. After preprocessing step, the assembly setup is started. The user can choose between preprocessed and not preprocessed FASTQ files. Next, for SPAdes assembly there are two options: fast assembler and MDA Single cell. Furthermore, for the Velvet assembly the user can choose the minimum number of contigs. For both assemblers, the GATOOL interface requires the k-*mer* number set up by user. Finally, the assembly quality evaluation is performed by QUAST software [administered against 12] after assembly finishing.

The settings for SPAdes assembly in the "Fast Assembly" and "Single cell-MDA" options are 21,33,55,77 k-mers (-k 21,33,55,77), "--iontorrent" and 21,33,55 k-mers (-k 21,33,55), "--iontorrent", respectively. As we determined, the GATOOL is primarily tuned to analyze ion torrent data. However, if the user choose use the basic SPAdes parameters, the values are automated set up as 21, 33, 55, 77, 99 and 127 k-mers. For Velvet assembly the user must provide multiple k-*mer* values.

The scaffolding process and *in silico* gap closing are performed with a reference genome. The user must choose a reference using a genome search interface implemented. This interface searches the NCBI reference genome database and automatic retrieves the sequence data. All procedures, including reference genome search is performed in the GATOOL interface without leaving the tool. CONTIGuator

[13] is used for the scaffolding process. However, all data for each procedure is available to the user and can be exported. This allows users to analyze their data with other softwares alongside GATOOL.

Validation using samples from the SRA

The SRA samples were obtained through the NCBI repository (https://www.ncbi.nlm.nih.gov/sra). We collected seven datasets that respected the following criteria: 1) data sequenced using the Ion Torrent platform; 2) that these datasets have publications with genome drafts or complete genomes sequences associated. (Table 1) summarizes information on the selected SRA datasets, as well as the reference genome used for each sample during assembly.

Table 1: Information about SRA samples

Species name	SRA Accession Id	Total sequenced Bases	%GC	References NCBI accession
Clostridium autoethanogenum	SRR1748018 [14]	99,5 Mbp	31%	NC_022592.1
Corynebacterium pseudotuberculosis	SRR3312980 [15]	388,9 Mbp	52%	NC_017301.1
Mycobacterium ulcerans	ERR732677 [6]	422,0 Mbp	63%	CP000325.1
Staphylococcus aureus	ERR493460 [6]	328,5 Mbp	32%	NZ_CP009828.1
Escherichia coli	SRR3707448 [16]	65,7 Mbp	46%	NC_000913.3
Escherichia coli O104:H21	SRR927598 [17]	142,4 Mbp	50%	NC_018658.1

Pedobacter sp. NL19	SRR1769012 [18]	1,1 Gbp	39%	NZ_CP012996.1

Results and Discussion:

Tool development and interface

We developed a simple and powerful tool for performing genome assembly with preprocessing, genome assembly and scaffolding/gap-closing. The protocol embedded different approaches and programs to maximize the quality of the finished genome (Fig. 1). All steps are performed through a friendly and fluid interface. All processes can be performed by users, without requiring advanced computational skills.

(insert Fig. 1 here)

Fig. 1: Workflow representing the assembling process and each step for the generation of a consensus sequence along with their respective software / methods

Validation using case study datasets

After the tool development, we performed a validation using public ion torrent data available in SRA repository. After platform filtering, using bacterial genome and ion torrent data, we retrieved seven SRA samples. Thus, we performed the analysis with GATOOL in a 2.50 GHz 64-bit Dual-Core CPU (Intel Core™ i5-7300HQ) and 8GB of RAM, running Ubuntu 16.04 LTS.

Most of the samples (6 samples) were sequenced by the Ion Torrent (PGM) sequencing system with 200 bp reading chemistry. A sample was originally sequenced using the Illumina platform and later sequenced using Ion Torrent. The original assembly statistics were retrieved in each SRA-related publication or genome project. These results will be compared to the GATOOL statistics. We applied the GATOOL to the SRA data and the reference genomes were retrieved from NCBI. The GATOOL sets outweigh all previous analyzes, providing better N50 numbers, and genome lengths were similar to those reported previously. In addition, our most time-consuming analysis (2 hours and 5 minutes) was 1 hour and 27 minutes faster than the previous best fitting (3 hours and 32 minutes).

For the SRR1748018, SRR927598 and SRR1769012 samples we performed quality filtering (trimming reads with below phred 20). For the other samples no preprocessing were performed. The increase of the error rate, makes it necessary to change the length of k-mer, for example selection of shorter ones [10]. For preprocessed samples the assembly were performed with the fast option. The other samples were assembled with standard settings. Both approaches reach in excellent results (Table 2). The samples SRR1748018, SRR3312980, ERR493460 and SRR927598 showed assemblies with quality of draft genomes (40, 7, 21 and 74 oriented contigs, respectively). Even without the gap-closing steps the tool provided good and accurate results. Only the SRR1748018 sample showed better assembly result than compared to GATOOL, with 100 contigs against 321. Nonetheless, the GATOOL provided better N50 (245.313 versus 115.901 from Newbler). Despite higher contig number in assembly process, our tool performed the scaffolding in this sample. In the end of process GATOOL showed 40 contigs in the final assembly and outperforming the previous assembly [14].

(insert Table 2 here)

We also compared the GATOOL performance with proprietary tools such as CLC Workbench. Thus, our tool presented better results compared to CLC Workbench different versions [16,17,18]. For instance, the SRR1769012 analysis resulted in 78 contigs against 201 generated by the CLC Workbench. Further, the N50 was 5-fold higher (260.508 against only 57.428 from CLC Workbench). In general, the GATOOL result were better than the previous reported (Table 2).

Conclusions:

Herein, we presented GATOOL, for bacterial assembly using ion Torrent data. Our tool offers a simple and straightforward genome finishing protocol and it can be used in a personal computer or configured as web tool. Moreover, The user identifies and retrieve a reference genome with an implemented search interface without leave the tool environment. The simpler and friendly-user interface guide the user through all steps, allowing parameter changes and reanalysis from quality analysis, through genome assembly and to scaffolding.

Furthermore, GATOOL showed impressive genome assembly performance. the contig number and the N50 statistics were better in all SRA samples tested. Our tool performed the assembly faster than previous assemblies reported. However, this can be associated to hardware used and we do not retrieve this information from other studies. Also, the genome length were similar to the reference genome. These observations are consistent with the reference genomes, indicating the GATOOL accuracy.

Availability and requirements

- **Project name:** GATOOL - Genome Assembly Tool
- **Project home page:** https://sourceforge.net/projects/gatool-beta/
- **Operating system(s):** Linux 64bit (Server), Platform independent (Client)
- **Programming languages:** PHP

- **Other requirements:** NCBI-BLAST+, Biopython library, Apache Server
- **License:** GPL v3
- **Any restrictions to use by non-academics:** None

The documentation can be obtained in the project home page, inside the "MANUAL" folder.

Abbreviations

NGS: Next Generation Sequencing

SRA: Sequence Read Archive

Declarations

Author's contributions

MBO: wrote the manuscript and developed the source code of the software; BMS, ATLQ, PIPR gave insights about the manuscript; BMS helped with the software methodology; All author's read and approved the final manuscript.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

Not applicable.

Funding

Not applicable.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

References

[1] Deurenberg HR, Bathoorn E, Chelbowicz AM, Couto N, Ferdous M, García-Coobs S, Kooistra-Smida AMD, Raangs EC, Rosema S, Veloo ACM, Zhou K, Friedrich AW, Rossen, AWJ. Application of next generation sequencing in clinical microbiology and infection prevention. Journal of Biotechnology. 2017;243:16-24.

[2] Tang P, Croxen MA, Hasan MR, Hsiao WWL, Hoang LM. Infection control in the new age of genomic epidemiology. American Journal of Infection Control. 2017; doi:10.1016/j.ajic.2016.05.015.

[3] Metzker ML. Sequencing technologies — the next generation. Nature Reviews Genetics. 2009; doi:10.1038/nrg2626.

[4] Cuccuru G, Orsini M, Pinna A, Sbardellati A, Soranzo N, Travaglione A, Paolo U, Gianluigi Z, Fotia G. Orione, a web-based framework for NGS analysis in microbiology. Bioinformatics. 2014; doi:10.1093/bioinformatics/btu135.

[5] Burland TG. DNASTAR's Lasergene sequence analysis software. Methods Mol Biol Clifton NJ. 2000;132:71–91.

[6] Baez-Ortega A, Lorenzo-Diaz F, Hernandez M, Gonzalez-Vila CI, Roda-Garcia JL, Colebrook M, Flores C. IonGAP: integrative bacterial genome analysis for Ion Torrent sequence data. Bioinformatics. 2015; doi:10.1093/bioinformatics/btv283.

[7] Mariano DCB, Pereira FL, Aguiar EL, Oliveira LC, Benevides L, Guimarães LC, Folador EL, Sousa TJ, Ghosh P, Barh D, Figueiredo HCP, Silva A, Ramos RTJ, Azevedo VAC. SIMBA: a web tool for managing bacterial genome assembly generated by Ion PGM sequencing technology. BMC Bioinformatics. 2016; doi:10.1186/s12859-016-1344-7.

[8] Chikhi R, Rizk G. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. Algorithms Mol Biol AMB. 2013;8:22.

[9] Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 2014;doi:10.1093/bioinformatics/btt310

[10] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a New genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77.

[11] Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research.2008;18:821-829.

[12] Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013; doi:10.1093/bioinformatics/btt086.

[13] Galardini M, Biondi EG, Bazzicalupo M, Mengoni A. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. Source Code Biol Med. 2011;6:11

[14] Brown SD, Nagaraju S, Utturkar S, De Tissera S, Segovia S, Mitchell W, Land ML, Dassanayake A, Köpke M. Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of Clostridium autoethanogenum and analysis of CRISPR systems in industrial relevant Clostridia. Biotechnology for Biofuels. 2014; doi:10.1186/1754-6834-7-40.

[15] Araújo CLDA, Dias LM, Veras AAO, Alves JTC, Cavalcante ALQ, Dowson CG, Azevedo V, Ramos RTJ, Silva A, Carneiro, AR. Whole-Genome Sequence of Corynebacterium pseudotuberculosis 262 Biovar equi Isolated from Cow Milk. Genome Announcements. 2016; doi:10.1128/genomeA.00176-16.

[16] Valat C, Goldstone RJ, Hirchaud E, Haenni M, Smith DGE, Madec Jean-Yves. Draft Genome Sequences of Enterohemorrhagic Escherichia coli Encoding Extended-Spectrum Beta-Lactamases. Genome Announcements. 2016; doi:10.1128/genomeA.01633-15.

[17] Gonzalez-Escalona N, McFarland MA, Rump LV, Payne J, Andrzejewski D, Brown EW, Evans PS, Croley TR. Draft Genome Sequences of Two O104:H21 Escherichia coli Isolates Causing Hemorrhagic Colitis during a 1994 Montana Outbreak Provide Insight into Their Pathogenicity. Genome Announcements. 2013; doi:10.1128/genomeA.00805-13.

[18] Santos T, Cruz A. Draft Genome Sequence of Pedobacter sp. Strain NL19, a Producer of Potent Antibacterial Compounds. Genome Announcements. 2016; doi:10.1128/genomeA.00184-15.

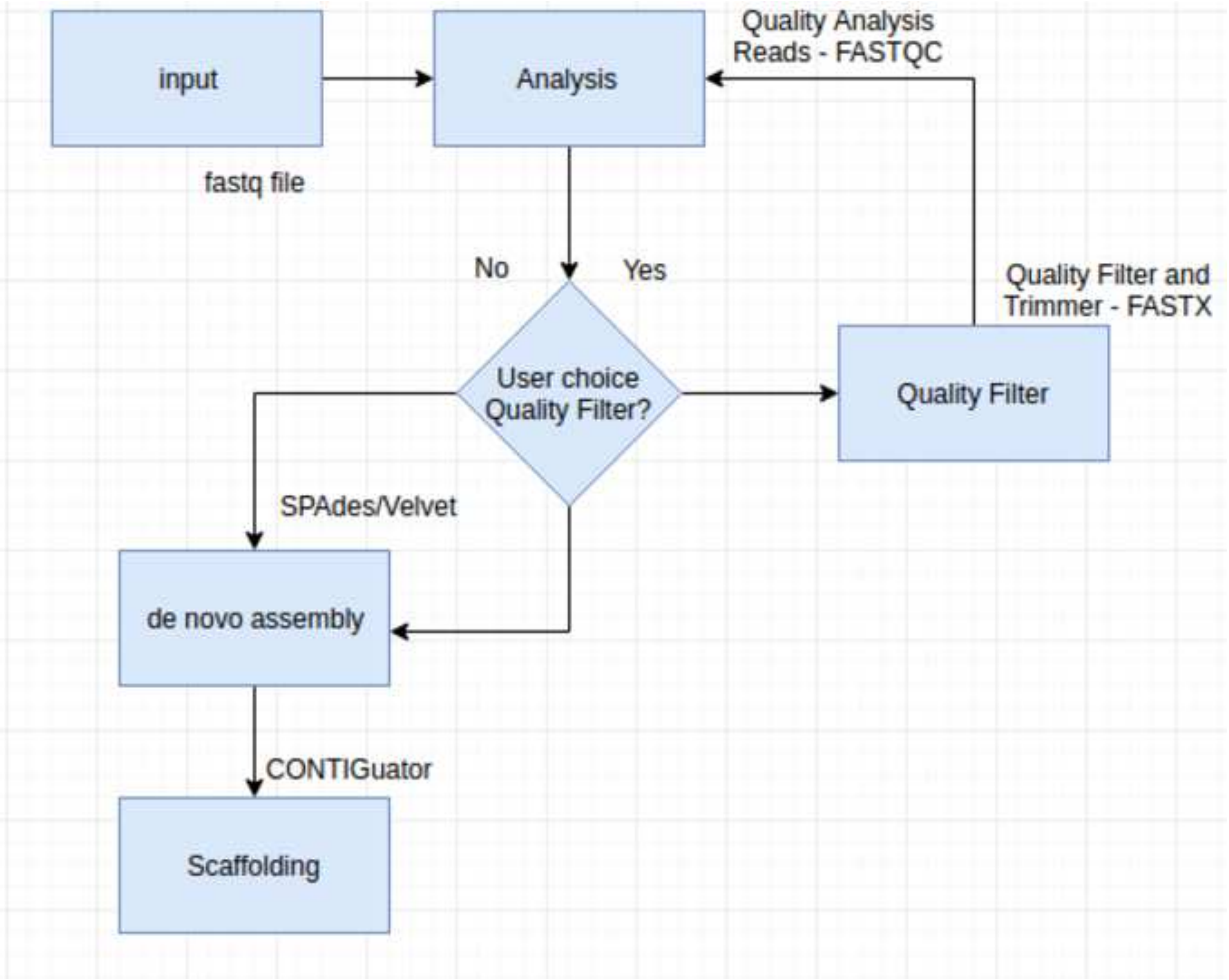
Table 2: Comparison of the results of GATOOL against other authors.

Articles	SRR1748018	SRR3312980	ERR732677	ERR493460	SRR3707448	SRR927598	SRR1769012
Softwares	Newbler 2.6	Mira 4.0.2	ionGAP		CLC Workbench 6.5.2	CLC Workbench 5.5.1	CLC Workbench 7.0.3
Articles results	contigs: **100** N50: 115.901 Size:4.320.000 Time: -	contigs: 29 N50: 333.604 Size:2.342.591 Time: -	contigs:1.352 N50: 7.674 Size: - Time: ≥20h	contigs: 94 N50: 218.499 Size: - Time:3h32min	contigs: 818 N50: 85kbp Size: 5.868Mb Time: -	contigs: 769 N50: - Size:4.929.288 Time: -	contigs: 201 N50: 57.428 Size:5.988.703 Time: -
GATOOL results	contigs: 321 N50: **245.313** Size: 4.524.980 Time: 20 min Scaffold: **40**	contigs: **9** N50: **628.361** Size:2.342.591 Time: 57 min Scaffold: **7**	contigs: **968** N50: **8.384** Size:5.128.282 Time:1h07min Scaffold: **791**	contigs: **32** N50: **314.833** Size:2.766.025 Time:42min Scaffold: **21**	contigs: **9** N50: **134.786** Size: 266.701 Time: 13 min Scaffold: -	contigs: **111** N50: **113.884** Size:4,955.474 Time:25min Scaffold: **74**	contigs: **78** N50: **260.508** Size:5.995.690 Time:2h05min Scaffold: 14

The best results are marked in bold on the table;
The scaffold value equal to 14 in SRA **-** SRR1769012 was not considered a good value due to the high number of base pairs that were excluded during the scaffolding process;
contigs column are based on the default value of the minimum number of contigs for QUAST, which is 500.
-some orientation process fails for an unknown reason and some authors did not report this data.

Figure

Click here to download Figure fluxogram gatool.png



APÊNDICE C

Artigo publicado – Manuscrito 1

High-Quality Draft Genome Sequence of *Bacillus amyloliquefaciens* Strain 629, an Endophyte from *Theobroma cacao*

genomeA
Journals.ASM.org

High-Quality Draft Genome Sequence of *Bacillus amyloliquefaciens* Strain 629, an Endophyte from *Theobroma cacao*

Brena M. M. SantAnna,[a] Phellippe P. A. Marbach,[b] Marcelo Rojas-Herrera,[c] Jorge T. De Souza,[d] Milton R. A. Roque,[a] Artur T. L. Queiroz[e]

Universidade Federal da Bahia (UFBA), Salvador, Brazil[a]; Universidade Federal do Recôncavo da Bahia (UFRB), Cruz das Almas, Brazil[b]; Centro de Genómica y Bioinformática, Universidade Mayor, Santiago, Chile[c]; Universidade Federal de Lavras (UFLA), Lavras, Brazil[d]; Centro de Pesquisas Gonçalo Moniz (CPqGM)-FIOCRUZ, Salvador, Brazil[e]

***Bacillus amyloliquefaciens* strain 629 is an endophyte isolated from *Theobroma cacao* L. Here, we report the draft genome sequence (3.9 Mb) of *B. amyloliquefaciens* strain 629 containing 16 contigs (3,903,367 bp), 3,912 coding sequences, and an average 46.5% G+C content.**

Received 29 September 2015 **Accepted** 1 October 2015 **Published** 19 November 2015

Citation SantAnna BMM, Marbach PPA, Rojas-Herrera M, De Souza JT, Roque MRA, Queiroz ATL. 2015. High-quality draft genome sequence of *Bacillus amyloliquefaciens* strain 629, an endophyte from *Theobroma cacao*. Genome Announc 3(6):e01325-15. doi:10.1128/genomeA.01325-15.

Address correspondence to Artur T. L. Queiroz, artur.queiroz@bahia.fiocruz.br.

Bacilli are frequently isolated as endophytes and are common components of the microbiota of several plant species (1, 2). Strain 629 was isolated from a healthy *Theobroma cacao* tree and was initially identified as *Bacillus subtilis* (3), but further analysis based on *gyr*B and *rec*A sequences revealed that its true identity is *Bacillus amyloliquefaciens* (4). This isolate colonizes different host and plant tissues under both sterile and nonsterile conditions and promotes plant growth (3, 4). Strain 629 produces the lipopeptides iturin, fengicin, and surfactin and volatile organic compounds that may be active in the biocontrol of several fungal plant pathogens (unpublished data) and pathogenic bacteria, including *Curtobacterium flaccumfaciens* pv. flaccumfaciens (5). Furthermore, *B. amyloliquefaciens* 629 is currently being used as a model to study endophytic colonization (4). This strain is deposited in the Biological Institute Culture Collection of Phytopathogenic Bacteria (IBSBF) (Campinas, São Paulo, Brazil) under accession no. IBSBF-3106. This collection is registered with the World Data Centre for Microorganisms collection under no. WDCM-110.

Genomic DNA from isolate 629 was extracted and sequenced using the Ion Torrent PGM platform (Life Technologies) 318 chip. A total of 7,567,586 reads with an average length of 330 pb were obtained. All reads were assembled to an initial draft genome of 3,866,991 nucleotides at 443-fold coverage using the SPAdes Genome Assembler version 3.5.0, generating 129 unoriented contigs, with a G+C content of 46.5%, (N_{50}: 285,363 bp).

Contigs were ordered using CONTIGuator 2.3 (http://contiguator.sourceforge.net/) (6) with the *B. Amyloliquefaciens* CC178 genome, the closest available, as a reference (GenBank accession no. CP006845.1). Subsequently, 34 contigs with 3.8 Mb were aligned with the reference genome to order the contigs. A total of 95 contigs (only 9 > 600 bp) corresponding to 29,876 nucleotides were not mapped to the reference genome. These sequences were identified as redundant contigs, according to BLAST results, and were removed from the assembly. To solve the repetitive sequences and the remaining gaps the MapRepeat pipeline (7) was used, resulting in the final high-quality draft genome sequence with 16 contigs, containing 3,903,367 bp.

Genome annotation was performed with RAST version 2.0 server (8). The genome of strain 629 is composed of 4,013 predicted genes, including 3,912 protein-coding sequences, 82 tRNAs, and 19 copies of the genes for 5S, 16S, and 23S rRNA. The genome of strain 629 is closely related to that of *B. amyloliquefaciens* CC178 with an identity of 99% (97% coverage) and also has a similar numbers of predicted genes (9).

Subsequent analysis of the genome content of *B. amyloliquefaciens* 629 and its comparison with phylogenetically related strains will help to determine key aspects of its interaction with the environment, plants, and other microorganisms.

Nucleotide sequence accession numbers. The *Bacillus amyloliquefaciens* strain 629 whole-genome shotgun (WGS) project has been deposited at DDBJ/EMBL/GenBank under the accession no. LGYP00000000. The version described in this paper is the first version, LGYP01000000, and consists of sequences LGYP01000001 to LGYP01000016.

ACKNOWLEDGMENTS

We thank CAPES and CNPq for financial support through the PVE project 61/2011, Zayda M. Moreira, Artur Silva, Rommel TJ. Ramos, and Luis Pacheco for DNA extraction, sequencing, and assembly assistance. J.T.S. acknowledges his productivity scholarship.

REFERENCES

1. **White JF, Torres MS, Sullivan RF, Jabbour RE, Chen Q, Tadych M, Irizarry I, Bergen MS, Havkin-Frenkel D, Belanger FC.** 2014. Occurrence of *Bacillus amyloliquefaciens* as a systemic endophyte of vanilla orchids. Microsc Res Tech **77:**874–885. http://dx.doi.org/10.1002/jemt.22410.
2. **Wang X, Liang G.** 2014. Control efficacy of an endophytic *Bacillus amyloliquefaciens* strain BZ6-1 against peanut bacterial Wilt, *Ralstonia solanacearum*. BioMed Res Int **2014:**465435. http://dx.doi.org/10.1155/2014/465435.
3. **Leite HAC, Silva AB, Gomes FP, Gramacho KP, Faria JC, De Souza JT, Loguercio LL.** 2013. *Bacillus subtilis* and *Enterobacter cloacae* endophytes from healthy *Theobroma cacao* L. trees can systemically colonize seedlings

and promote growth. Appl Microbiol Biotechnol **97:**2639–2651. http://dx.doi.org/10.1007/s00253-012-4574-2.

4. **Moreira ZM, Duarte EAA, Oliveira TAS, Monteiro FP, Loguercio LL, De Souza JT.** 2015. Host and tissue preferences of *Enterobacter cloacae* and *Bacillus amyloliquefaciens* for endophytic colonization. Afr J Microbiol Res **9:**1352–1356. http://dx.doi.org/10.5897/AJMR2015.7475.
5. **Martins SJ, de Medeiros FHV, de Souza RM, de Resende MLV, Ribeiro PM, Jr.** 2013. Biological control of bacterial wilt of common bean by plant growth-promoting rhizobacteria. Biol Contr **66:**65–71. http://dx.doi.org/10.1016/j.biocontrol.2013.03.009.
6. **Galardini M, Biondi EG, Bazzicalupo M, Mengoni A.** 2011. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. Source Code Biol Med **6:**11. http://dx.doi.org/10.1186/1751-0473-6-11.
7. **Mariano DC, Pereira FL, Ghosh P, Barh D, Figueiredo HC, Silva A, Ramos RT, Azevedo VA.** 2015. MapRepeat: an approach for effective assembly of repetitive regions in prokaryotic genomes. Bioinformation **11:**276–279. http://dx.doi.org/10.6026/97320630011276.
8. **Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R.** 2013. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucl Acid Res **42:**D206–D214. http://dx.doi.org/10.1093/nar/gkt1226.
9. **Kim BY, Lee SY, Ahn JH, Song J, Kim WG, Weon HY.** 2015. Complete genome sequence of *Bacillus amyloliquefaciens* subsp. *plantarum* CC178, a phyllosphere bacterium antagonistic to plant pathogenic fungi. Genome Announc **3**(1):e01368-14. http://dx.doi.org/10.1128/genomeA.01368-14.

APÊNDICE D

Materiais Suplementares – Manuscrito 2

Supplementary Figure S1

Supplementary Figure S2

Supplementary Table S1

Supplementary Table S2

Supplementary Table S3

Supplementary Table S4

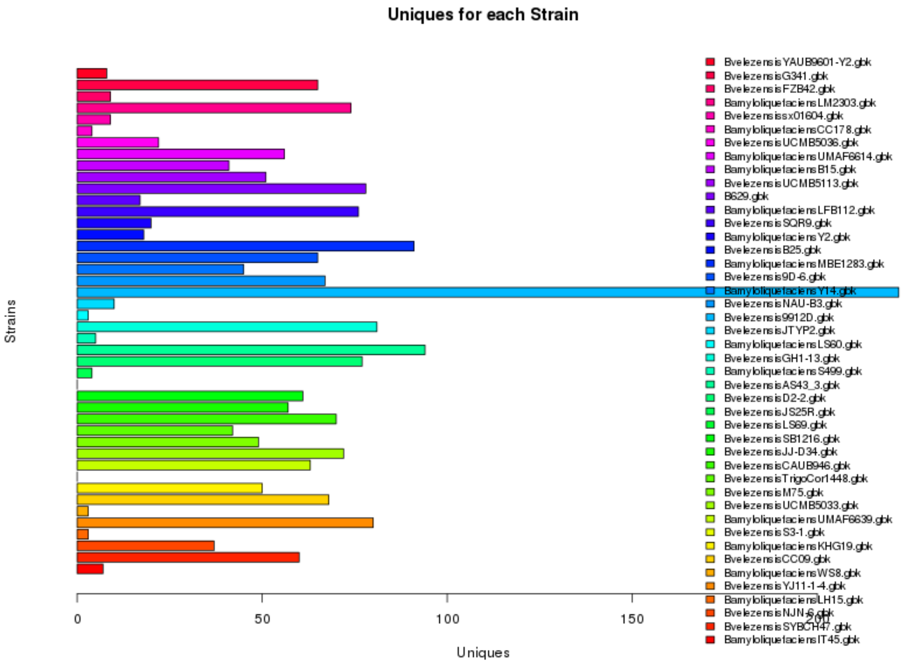
Supplementary Table S5

Novel insights on bacterial endophytic behavior from comparative genomic analysis of *Bacillus velezensis* strain 629

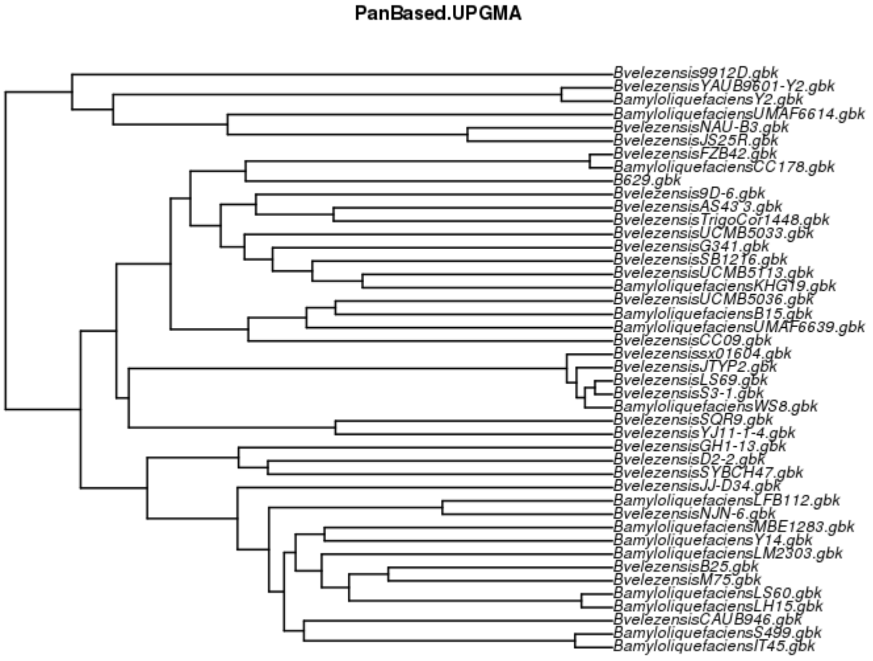
Brena Sant'Anna[a], Artur Queiroz[b], Luis Pacheco[a], Jorge Souza[c], Ícaro Lopes[a], Felipe Rangel[a], Milton Roque[a *]
[a]*Institute of Health Sciences, Federal University of Bahia, Salvador-BA, Brazil;* [b]*Gonçalo Moniz Institute - Oswaldo Cruz Foundation, Salvador-BA, Brazil;* [c]*Department of Phytopathology, Federal University of Lavras, Lavras-MG, Brazil.*
***Corresponding author: Milton Roque, milton.roque@ufba.br - Tel: +55 71 32838893.**

Supplementary Material 1

A



B



Supplementary Figure S1. A) Bar graph showing the number of unique genes found in individual strains of *B. velezensis* displays the number of unique genes for each evaluated strain, allowing the user to determine which strains have larger or smaller numbers of unique genes. B) Phylogenomic tree based on the UPGMA algorithm from *B. velezensis* genomes.

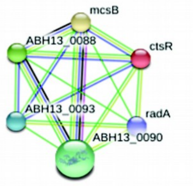
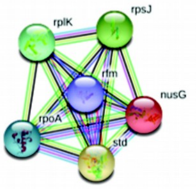
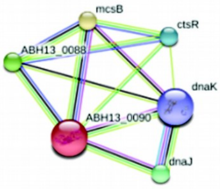
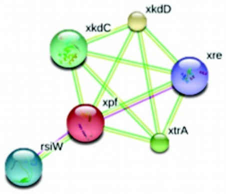
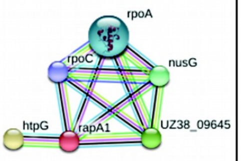
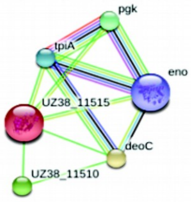
Novel insights on bacterial endophytic behavior from comparative genomic analysis of *Bacillus velezensis* strain 629

Brena Sant'Anna[a] , Artur Queiroz[b], Luis Pacheco[a], Jorge Souza[c], Ícaro Lopes[a], Felipe Rangel[a], Milton Roque[a *]

[a]*Institute of Health Sciences, Federal University of Bahia, Salvador-BA, Brazil;* [b]*Gonçalo Moniz Institute - Oswaldo Cruz Foundation, Salvador-BA, Brazil;* [c]*Department of Phytopathology, Federal University of Lavras, Lavras-MG, Brazil.*

***Corresponding author: Milton Roque, milton.roque@ufba.br - Tel: +55 71 32838893.**

Supplementary Material 2

GI	ID	Bio Process	Function	Associations / Enrichment	String
1	Transcriptional regulator CtsR	- Regulation of transcription; - Response to stresses;	- Transcriptional regulator of stress and heat shock response	mcsB (Stress response system; Involved in the regulation of many critical cellular processes, such as protein homeostasis, motility, competence, and stringent and stress responses). McsA/ABH13_0088 (Activates the phosphorylation activity of the protein-arginine kinase McsB) YacL/ABH13_0093 (Uncharacterized protein) ClpC/ABH13_0090 (Competence gene repressor; required for cell growth at high temperature) radA (plays a role in repairing DNA breaks)	 mcsB ABH13_0088 ctsR ABH13_0093 radA ABH13_0090 http://string.embl.de/cgi/network.pl?taskId=kbyc96YbaaAb
1	Transcription antitermination protein NusG	-Regulation of DNA-templated transcription; -Transcription antitermination	- Participates in transcription elongation, termination and antitermination; - NusG-stimulated pausing is sequence specific.	rpoA (DNA transcription) rplK (ribosomal protein L11; Forms part of the ribosomal stalk which helps the ribosome interact with GTP-bound translation factors) rpsJ (ribosomal protein S10; involved in the binding of tRNA to the ribosomes) rfm (DNA-directed RNA polymerase subunit beta) std (DNA-directed RNA polymerase subunit beta)	 https://string-db.org/cgi/network.pl?taskId=eTdkHHTdplQ5
1	ClpC (ABH13_0090)	- Regulation of transcription; - protein metabolic process	- Competence gene repressor; required for cell growth at high temperature. - Negative regulator of comK expression.	mcsA -ABH13_0088 (Activates the phosphorylation activity of the protein-arginine kinase mcsB) mcsB (Stress response system; Involved in the regulation of many critical cellular processes, such as protein homeostasis, motility, competence, and stringent and stress responses). ctsR (Controls the expression of the cellular protein quality control genes clpC, clpE and clpP, as well as mcsA and mcsB) dnaK (Acts as a chaperone) dnaJ (response to hyperosmotic and heat shock)	 http://string.embl.de/cgi/network.pl?taskId=1cqj1JeKPRXk
2	prophage LambdaBa01, positive control factor Xpf	- Regulation of transcription	- transcription factor activity	risW (Is the anti-sigma factor for extracytoplasmic function sigma factor SigW xkdC (Phage-like element PBSX protein; May function as a transcriptional antiterminator) xkdD (Phage-like element PBSX protein) xre (Repressor of PBSX - Necessary for the maintenance of the lysogenic state) xtrA (Phage-like element PBSX protein)	 http://string.embl.de/cgi/network.pl?taskId=tYAvXkuftBD2
4	Transcriptional regulator DeoR (UZ38_11515)	- Transcription regulation	- DeoR family transcriptional regulator. - involved in sugar catabolism	tpiA (Involved in the gluconeogenesis. pgk (involved in subpathway that synthesizes pyruvate from D-glyceraldehyde 3-phosphate) eno (degradation of carbohydrates via glycolysis) deoC (Catalyzes a reaction to generate 2-deoxy- D-ribose 5-phosphate) YxeH -UZ38_11510 (Hydrolase activity)	 http://string.embl.de/cgi/network.pl?taskId=pEM4miZJw8lt
5	rapA1	- Hydrolase activity	- response regulator aspartate phosphatase - Regulation of sporulation pathways, competence and biofilm formation	RpoA (DNA-dependent RNA polymerase) RpoC (DNA-dependent RNA polymerase) NusG (Participates in transcription elongation, termination and antitermination) DinG -UZ38_09645 (Probable helicase involved in DNA repair and perhaps also replication HtpG (Molecular chaperone. Has ATPase activity) YvtA and YyxA (serine-type endopeptidase activity)	 http://string.embl.de/cgi/network.pl?taskId=1rIhUr9BTKSX

Supplementary Figure S2. Transcription regulatory genes in GIs of *B. velezensis* 629 and their protein-protein interaction networks by String database (https://string-db.org/).

Novel insights on bacterial endophytic behavior from comparative genomic analysis of *Bacillus velezensis* strain 629

Brena Sant'Anna[a] , Artur Queiroz[b], Luis Pacheco[a], Jorge Souza[c], Ícaro Lopes[a], Felipe Rangel[a], Milton Roque[a] *
[a]*Institute of Health Sciences, Federal University of Bahia, Salvador-BA, Brazil;* [b]*Gonçalo Moniz Institute - Oswaldo Cruz Foundation, Salvador-BA, Brazil;* [c]*Department of Phytopathology, Federal University of Lavras, Lavras-MG, Brazil.*
***Corresponding author: Milton Roque, milton.roque@ufba.br - Tel: +55 71 32838893.**

Supplementary Material 3

Supplementary Table S1. List and description of the strains of *B. amyloliquefaciens* and *B. velezensis* used in the study.

Species	Strain	Genome size (Mb)	GenBank	Features
Bacillus amyloliquefaciens	CC178	3,91	CP006845.1	Plant-associated (biocontrol against fungi)
	B15	4,0	CP014783.1	Plant-associated (biocontrol against fungi)
	LFB112	3,94	CP006952.1	Plant-associated (biocontrol against animal pathogens)
	UMAF6614	4,0	CP006960.1	Plant-associated (Biocontrol)
	UMAF6639	4,0	CP006058.1	Plant-associated (Biocontrol)
	Y2	4,2	CP003332.1	Plant-associated, PGPB (Biocontrol)
	IT 45	3,93	CP004065.1	Plant-associated, PGPR
	LH15	3,9	CP010556.1	Plant-associated, PGPR (biocontrol against fungi)
	LS60	3,9	CP011278.1	Plant-associated (Biocontrol)
	S499	3,93	CP014700.1	Plant-associated
	WS-8	3,92	CP018200.1	Plant-associated, PGPB
	Y14	3,95	CP017953.1	Plant-associated, PGPR (isolated from rhizosphere of peanut)
	LM2303	3,98	CP018152.1	Plant-associated, PGPB (Biocontrol) (reclassif.. *B.velezensis*)
	MBE1283	3,97	CP013727.1	Isolated from Korean alcoholic beverage
	KHG19	3,95	CP007242.1	Industrial application; Isolated from Korean traditional soybean paste -alfa amilase
	DSM 7	3,98	FN597644.1	Industrial application; Isolated from soil - amilase e protease
	LL3	4,0	CP002634.1	Industrial application Produção de ac. glutâmico
	TA208	3,93	CP002627.1	Industrial application
	XH7	3,93	CP002927.1	Industrial application Produção de purina nucleosídeo guanosina
	RD7-7	3,68	CP016913.1	Industrial application Isolated from fermented soy sauce; (biocontrol)
Bacillus velezensis	FZB42	3,91	CP000560.1	Plant-associated, PGPB (Biocontrol)
	CAU B946	4,01	HE617159.1	Plant-associated, PGPR (Biocontrol)
	YAU B9601 Y2	4,24	HE774679.1	Plant-associated (Biocontrol)
	AS43.3	3,96	CP003838.1	Plant-associated, (Biocontrol)
	UCMB5036	3,91	HF563562.1	Plant-associated, PGPR
	UCMB-5033	4,07	HG328253.1	Plant-associated, PGPR
	UCMB5113	3,88	HG328254.1	Plant-associated, PGPR (Biocontrol and stress resistence)

	NAU-B3	4,19	HG514499.1	Plant-associated, PGPR
	TrigoCor1448	3,95	CP007244.1	Plant-associated (Biocontrol)
	SQR9	4,11	CP006890.1	Plant-associated, PGPR (Biocontrol)
	JS25R	4,01	CP009679.1	Plant-associated (Biocontrol)
	NJN-6	4,05	CP007165.1	Plant-associated, PGPR (Biocontrol)
	G341	4,0	CP011686.1	Plant-associated (Biocontrole)
	B25	3,86	LN999829.1	Plant-associated, PGPB (health enhancement)
	CC09	4,16	CP015443.1	Plant-associated (Endophytic), PGPB (Biocontrol)
	S3-1	3,92	CP016371.1	Plant-associated, PGPB (Biocontrol)
	LS69	3,91	CP015911.1	Plant-associated, PGPB (Biocontrol)
	9912D	4,24	CP017775.1	Plant-associated (Biocontrol)
	M75	4,0	CP016395.1	Plant-associated (Biocontrol)
	SYBC H47	3,88	CP017747.1	Plant-associated (Biocontrol)
	GH1-13	4,14	CP019040.1	Plant-associated, PGPR
	sx01604	3,92	CP018007.1	Plant-associated, PGPB
	JTYP2	3,92	CP020375.1	Plant-associated (Biocontrol)
	SB1216	3,81	CP015417.1	Isolated from soil
	9D-6	3,96	CP020805.1	Isolated from rhizosphere soil (Biocontrol)
	JJ-D34	4,10	CP011346.1	Isolated from Korean traditional fermented soybean paste *(B. methylotrophicus)*
	YJ11-1-4	4,0	CP011347.1	Isolated from Korean traditional fermented soybean paste *(B. methylotrophicus)*
	D2-2	3,92	CP014990.1	Isolated from Korean traditional fermented soybean paste

* PGPB: Plant Growth Promotion Bacteria; * PGPR: Plant Growth Promoting Rhizobacteria

Novel insights on bacterial endophytic behavior from comparative genomic analysis of *Bacillus velezensis* strain 629

Brena Sant'Anna[a] , Artur Queiroz[b], Luis Pacheco[a], Jorge Souza[c], Ícaro Lopes[a], Felipe Rangel[a], Milton Roque[a *]
[a]*Institute of Health Sciences, Federal University of Bahia, Salvador-BA, Brazil;* [b]*Gonçalo Moniz Institute - Oswaldo Cruz Foundation, Salvador-BA, Brazil;* [c]*Department of Phytopathology, Federal University of Lavras, Lavras-MG, Brazil.*
***Corresponding author: Milton Roque, milton.roque@ufba.br - Tel: +55 71 32838893.**

Supplementary Material 4

Supplementary Table S2. List and description of endophytic genomes used in this study.

	Organism	Host	Benefits for plant	GenBank
1	*Azoarcus* sp. BH72	Rice and other grasses	Nitrogen fixation	AM406670.1
2	*Azospirillum* sp. B510	*Oryza sativa* cv. Nipponbare	Nitrogen fixation, plant growth promotion, increase in seed yield, enhanced disease resistance	AP010946.1
3	*Bacillus megaterium* Q3	Root of tobacco	Degrading quinclorac (herbicide)	CP010586.1
4	*Bacillus subtilis* BsN5	*Amorphophallus konjac*	Antimicrobial activity and biocontrol	CP002468
5	*Bacillus velezensis* CC09	*Cinnamomum camphora*	Biocontrol ability; volatile organic compound synthesis.	CP015443.1
6	*Burkholderia* sp. KJ006 (3 chromosomes)	Surface-sterilized rice root	Antifungal activity, plant growth promotion, degradation of aromatic compounds	CP003514.1 CP003515.1 CP003516.1
7	*Enterobacter cloacae* ENHKU01	*Capsicum annuum*	Biocontrol; antifungical and antimicrobial.	CP003737
8	*Gluconacetobacter diazotrophicus* Pal5	Sugarcane	nitrogen fixation, plant growth promotion, synthesis of auxin (plant growth promoter) and bacteriocins (biocontrol).	AM889285.1
9	*Herbaspirillum seropedicae* SmR1	Rice and sugarcane	Nitrogen fixation and Plant Growth Promoter by ethylene signaling pathway (ACC deaminase), indole acetic acid and auxins	CP002039.1
10	*Klebsiella pneumoniae* 342	Corn	Nitrogen fixation	CP000964.1
11	*Paraburkholderia phytofirmans* PsJN (2 chromosomes)	Onion roots	Plant Growth Promoter (by 1-aminocyclopropane-1-carboxylate (ACC) deaminase)	CP001052.1 CP001053.1
12	*Pseudomonas fluorescens* PICF7	Olive root	Biocontrol of pathogens (fungi)	CP005975.1
13	*Pseudomonas putida* W619	*Populus deltoides*	Promote plant growth	CP000949.1
14	*Serratia marcescens* FS14	*Atractylodes macrocephala* Koidz	Antagonistic action against pathogens (production of prodigiosin and bacteriocins)	CP005927.1
15	*Serratia marcescens* RSC-14	*Solanum nigrum*	Growth-promoting (indole-3-acetic acid (IAA)), tolerância ao cádmio	CP012639.1

Novel insights on bacterial endophytic behavior from comparative genomic analysis of *Bacillus velezensis* strain 629

Brena Sant'Anna[a] , Artur Queiroz[b], Luis Pacheco[a], Jorge Souza[c], Ícaro Lopes[a], Felipe Rangel[a], Milton Roque[a *]
[a]*Institute of Health Sciences, Federal University of Bahia, Salvador-BA, Brazil;* [b]*Gonçalo Moniz Institute - Oswaldo Cruz Foundation, Salvador-BA, Brazil;* [c]*Department of Phytopathology, Federal University of Lavras, Lavras-MG, Brazil.*
***Corresponding author: Milton Roque, milton.roque@ufba.br - Tel: +55 71 32838893.**

Supplementary Material 5

Supplementary Table S3. List of all genes in predicted Genomic Islands (GIs) in *B. velezensis* 629 by IslandViewer4.

GI 1 – Size of region: 68206 bp

Gene start	Gene end	Product
98070	98192	hypothetical protein
109137	109601	Transcriptional regulator CtsR
109615	110172	Nucleotide excision repair protein, with UvrB/UvrC motif
110172	111263	Putative ATP:guanido phosphotransferase YacI (EC 2.7.3.-)
111569	113698	ATP-dependent Clp protease, ATP-binding subunit ClpC / Negative regulator of genetic competence clcC
114242	117283	Lanthionine biosynthesis protein LanM
117332	119050	Lipid A export ATP-binding/permease protein MsbA
119056	120093	extracellular serine protease
121595	122677	DNA integrity scanning protein DisA
122791	123891	Membrane-associated protein containing RNA-binding TRAM domain and ribonuclease PIN-domain, YacL B.s
123904	124602	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (EC 2.7.7.60)
124583	125071	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (EC 4.6.1.12)
125163	126614	Glutamyl-tRNA synthetase (EC 6.1.1.17) @ Glutamyl-tRNA(Gln) synthetase (EC 6.1.1.24)
126926	127579	Serine acetyltransferase (EC 2.3.1.30)
127576	128976	Cysteinyl-tRNA synthetase (EC 6.1.1.16)
128981	129412	COG1939: Ribonuclease III family protein
129396	130145	23S rRNA (guanosine-2'-O-) -methyltransferase rlmB (EC 2.1.1.-)
130152	130664	Hypothetical protein DUF901, similar to C-terminal domain of ribosome protection-type Tc-resistance
130778	131383	RNA polymerase sporulation specific sigma factor SigH
131469	131618	LSU ribosomal protein L33p @ LSU ribosomal protein L33p, zinc-dependent
131651	131830	Preprotein translocase subunit SecE (TC 3.A.5.1.1)

131998	132531	Transcription antitermination protein NusG
132804	133124	LSU ribosomal protein L11p (L12e)
133219	133917	LSU ribosomal protein L1p (L10Ae)
134185	134673	LSU ribosomal protein L10p (P0)
134715	135086	LSU ribosomal protein L7/L12 (P1/P2)
135178	135783	Ribosomal RNA small subunit methyltransferase C (EC 2.1.1.52) ## SSU rRNA m2G1207
136364	139609	DNA-directed RNA polymerase beta subunit (EC 2.7.7.6)
139672	143271	DNA-directed RNA polymerase beta' subunit (EC 2.7.7.6)
143441	143689	Firmicutes ribosomal L7Ae family protein
144261	144731	SSU ribosomal protein S7p (S5e)
144931	146862	Translation elongation factor G
146982	148172	Translation elongation factor Tu
148276	149241	FIG01233798: hypothetical protein
149443	149577	hypothetical protein
149838	150467	LSU ribosomal protein L3p (L3e)
150816	151118	LSU ribosomal protein L4p (L1e)
151118	151405	LSU ribosomal protein L23p (L23Ae)
151437	152270	LSU ribosomal protein L2p (L8e)
153418	153624	SSU ribosomal protein S3p (S3e)
153872	154060	LSU ribosomal protein L16p (L10e)
154050	154250	LSU ribosomal protein L29p (L35e)
154273	154536	SSU ribosomal protein S17p (S11e)
155321	155860	LSU ribosomal protein L5p (L11e)
155883	156068	SSU ribosomal protein S14p (S29e) @ SSU ribosomal protein S14p (S29e), zinc-dependent
156100	156498	SSU ribosomal protein S8p (S15Ae)
156529	157068	LSU ribosomal protein L6p (L9e)
157295	157465	LSU ribosomal protein L18p (L5e)
157484	157990	SSU ribosomal protein S5p (S2e)
158004	158183	LSU ribosomal protein L30p (L7e)
158376	158654	LSU ribosomal protein L15p (L27Ae)
158803	159951	Preprotein translocase secY subunit (TC 3.A.5.1.1)

160003	160656	Adenylate kinase (EC 2.7.4.3)
160653	161399	Methionine aminopeptidase (EC 3.4.11.18)
161724	161942	Translation initiation factor 1
162112	162477	SSU ribosomal protein S13p (S18e)
163070	164014	DNA-directed RNA polymerase alpha subunit (EC 2.7.7.6)
164586	165431	ATPase component of general energizing module of ECF transporters
165407	166276	ATPase component of general energizing module of ECF transporters
166273	167070	Transmembrane component of general energizing module of ECF transporters

GI 2 - 9128 bp

Gene start	Gene end	Product
1147845	1148072	Putative toxin component near putative ESAT-related proteins, repetitive / Repetitive hypothetical p
1149329	1149451	hypothetical protein
1149608	1149730	hypothetical protein
1149976	1150674	hypothetical protein
1151380	1151742	hypothetical protein
1151739	1151855	hypothetical protein
1151875	1152066	hypothetical protein
1152077	1152247	hypothetical protein
1152374	1153147	hypothetical protein
1153243	1153371	hypothetical protein
1153385	1153861	prophage LambdaBa01, positive control factor Xpf
1154030	1154239	hypothetical protein
1154772	1155440	hypothetical protein
1156059	1156973	Beta-lactamase class A

GI 3 - 4953 bp

Gene start	Gene end	Product
1221356	1221934	Phage-like element PBSX protein xkdU
1221931	1222203	FIG01238565: hypothetical protein
1222206	1223837	Uncharacterized protein yqcC
1223850	1224221	FIG01238688: hypothetical protein
1224226	1224423	FIG01236576: hypothetical protein

1225292	1225555	phage related protein
1225569	1225832	holin
1225846	1226724	N-acetylmuramoyl-L-alanine amidase CwlH implicated in mother cell lysis (EC 3.5.1.28)
1226759	1226884	hypothetical protein
GI 4 - 24001 bp		
Gene start	**Gene end**	**Product**
1800971	1803568	Phosphoenolpyruvate synthase (EC 2.7.9.2)
1804975	1805397	hypothetical protein
1805870	1806055	hypothetical protein
1806660	1807574	transcriptional regulator
1807638	1808138	nuclease inhibitor
1808289	1809086	Alcohol dehydrogenase (EC 1.1.1.1)
1810109	1810252	hypothetical protein
1810392	1810514	hypothetical protein
1811001	1812113	Flagellar hook-length control protein FliK
1812475	1812864	YoaW
1813563	1814324	GCN5-related N-acetyltransferase
1814421	1814561	hypothetical protein
1814556	1815098	GCN5-related N-acetyltransferase
1815750	1816229	FIG01240545: hypothetical protein
1816276	1816503	hypothetical protein
1816788	1817408	hypothetical protein
1817633	1818706	luciferase-like monooxygenase
1818776	1818889	hypothetical protein
1818914	1819534	Chitin binding protein
1820053	1820169	hypothetical protein
1820308	1820442	hypothetical protein
1820578	1821114	FIG01232426: hypothetical protein
1821218	1821346	hypothetical protein
1821400	1821573	hypothetical protein
1821742	1824159	Phage neck

1824538	1824972	Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23)

GI 5 - 20660 bp

Gene start	Gene end	Product
2083075	2083500	UDP-N-acetylglucosamine 4,6-dehydratase (EC 4.2.1.-)
2083534	2083710	unknown
2083713	2083886	Phage protein
2084073	2084330	unknown
2085564	2085740	hypothetical protein
2085942	2086067	hypothetical protein
2086645	2087781	response regulator aspartate phosphatase
2088676	2090088	Zn-dependent hydroxyacylglutathione hydrolase / Polysulfide binding protein
2090247	2091446	FIG002984: FAD-dependent pyridine nucleotide-disulphide oxidoreductase
2091625	2092269	two component transcriptional regulator, LuxR family
2092262	2093335	Two component sensor histidine kinase
2093782	2094561	FIG003846: hypothetical protein
2096170	2096730	unknown
2096731	2097756	Putative toxin component near putative ESAT-related proteins, repetitive / Repetitive hypothetical p
2097798	2098214	Putative toxin component near putative ESAT-related proteins, repetitive / Repetitive hypothetical p
2099540	2100073	Uncharacterized protein ynaB
2100237	2101202	Phage-encoded chromosome degrading nuclease YokF
2101419	2103086	site-specific recombinase
2103127	2103735	UDP-N-acetylglucosamine 4,6-dehydratase (EC 4.2.1.-)

GI 6 - 5684 bp

Gene start	Gene end	Product
2922139	2922285	hypothetical protein
2922341	2922472	hypothetical protein
2922462	2923193	response regulator aspartate phosphatase
2924846	2924968	hypothetical protein
2925131	2926558	hypothetical protein
2926685	2927869	FIG01231814: hypothetical protein

GI 7 - 8011 bp

Gene start	Gene end	Product
3442616	3443206	Chromate transport protein
3443203	3443739	Chromate transport protein
3443897	3444637	Endonuclease V (EC 3.1.21.7)
3444654	3445154	Hypothetical protein, CF-21 family
3445346	3445480	hypothetical protein
3446111	3446425	Hypothetical cytosolic protein
3446544	3446975	Hypothetical protein, CF-38 family
3447605	3448297	Mobile element protein
3448361	3448675	hypothetical protein
3448891	3450627	Putative toxin component near putative ESAT-related proteins, repetitive / Repetitive hypothetical p

GI 8 - 4146 bp

Gene start	Gene end	Product
3873509	3873622	hypothetical protein
3874052	3875803	FIG01232221: hypothetical protein
3876207	3877655	unknown

Novel insights on bacterial endophytic behavior from comparative genomic analysis of *Bacillus velezensis* strain 629

Brena Sant'Anna[a] , Artur Queiroz[b], Luis Pacheco[a], Jorge Souza[c], Ícaro Lopes[a], Felipe Rangel[a], Milton Roque[a] *

[a]Institute of Health Sciences, Federal University of Bahia, Salvador-BA, Brazil; [b]Gonçalo Moniz Institute - Oswaldo Cruz Foundation, Salvador-BA, Brazil; [c]Department of Phytopathology, Federal University of Lavras, Lavras-MG, Brazil.

***Corresponding author: Milton Roque, milton.roque@ufba.br - Tel: +55 71 32838893.**

Supplementary Material 6

Supplementary Table S4. Shared genes present in GIs of the endophytic *B. velezensis* 629 and CC09.

Gene Description / ID	**Activity**
Histidine kinase	Cell signaling associated with stress conditions, osmosensing and chemotaxis.
Beta-lactamase A	Resistance to beta-lactam antibiotics
Prophage-derived endonuclease YokF	Catalyzes the hydrolysis of supercoiled double and single strand DNA and RNA. Involved in chromosomal DNA degradation and cell death caused by thermal stress.
N-acetylmuramoyl-L-alanine amidase cwlD	Step in the formation of muramic delta-lactam residues in spore cortex.
Deoxyuridine 5'-triphosphate nucleotidohydrolase / yncF	Enzyme is involved in nucleotide metabolismo, uracil methylation / thymine synthesis process
Peptide-methionine (R)-S-oxide reductase / msrB	Response to oxidative stress

Novel insights on bacterial endophytic behavior from comparative genomic analysis of *Bacillus velezensis* strain 629

Brena Sant'Anna[a] , Artur Queiroz[b], Luis Pacheco[a], Jorge Souza[c], Ícaro Lopes[a], Felipe Rangel[a], Milton Roque[a *]

[a]*Institute of Health Sciences, Federal University of Bahia, Salvador-BA, Brazil;* [b]*Gonçalo Moniz Institute - Oswaldo Cruz Foundation, Salvador-BA, Brazil;* [c]*Department of Phytopathology, Federal University of Lavras, Lavras-MG, Brazil.*

***Corresponding author: Milton Roque, milton.roque@ufba.br - Tel: +55 71 32838893.**

Supplementary Material 7

Supplementary Table S5. Shared genes (121) from comparative genome analysis of *B. velezensis* 629 with 15 other endophytic genomes.

Genome ref.	Product	ID gene
*Bv*629_5	DNA gyrase subunit B (EC 5.99.1.3)	gyrB
*Bv*629_6	DNA gyrase subunit A (EC 5.99.1.3)	gyrA
*Bv*629_13	Inosine-5-monophosphate dehydrogenase (EC 1.1.1.205) / CBS domain	guaB
*Bv*629_17	Seryl-tRNA synthetase (EC 6.1.1.11)	serS
*Bv*629_26	FIG000557: hypothetical protein co-occurring with RecR	yaaK
*Bv*629_27	Recombination protein RecR	recR
*Bv*629_43	Putative deoxyribonuclease YcfH	ycfH
*Bv*629_55	Ribose-phosphate pyrophosphokinase (EC 2.7.6.1)	prs
*Bv*629_72	Cell division protein FtsH (EC 3.4.24.-)	ftsH
*Bv*629_79	Para-aminobenzoate synthase, amidotransferase component (EC 2.6.1.85) @ Anthranilate synthase, amidotransferase component (EC 4.1.3.27)	pabA
*Bv*629_91	ATP-dependent Clp protease, ATP-binding subunit ClpC / Negative regulator of genetic competence clcC/mecB	clpC
*Bv*629_108	Transcription antitermination protein NusG	nusG
*Bv*629_109	LSU ribosomal protein L11p (L12e)	rplK
*Bv*629_110	LSU ribosomal protein L1p (L10Ae)	rplA
*Bv*629_112	LSU ribosomal protein L7/L12 (P1/P2)	rplL
*Bv*629_115	DNA-directed RNA polymerase beta subunit (EC 2.7.7.6)	rpoB
*Bv*629_117	SSU ribosomal protein S7p (S5e)	rpsG
*Bv*629_119	Translation elongation factor Tu	tuf
*Bv*629_122	LSU ribosomal protein L3p (L3e)	rplC
*Bv*629_130	LSU ribosomal protein L5p (L11e)	rplE
*Bv*629_132	SSU ribosomal protein S8p (S15Ae)	rpsH
*Bv*629_133	LSU ribosomal protein L6p (L9e)	rplF
*Bv*629_135	SSU ribosomal protein S5p (S2e)	rpsE
*Bv*629_138	Preprotein translocase secY subunit (TC 3.A.5.1.1)	secY
*Bv*629_139	Adenylate kinase (EC 2.7.4.3)	adk
*Bv*629_142	SSU ribosomal protein S13p (S18e)	rpsM
*Bv*629_143	DNA-directed RNA polymerase alpha subunit (EC 2.7.7.6)	rpoA
*Bv*629_565	Molybdenum cofactor biosynthesis protein MoaC	moaC
*Bv*629_572	Heat shock protein 60 family chaperone GroEL	groL
*Bv*629_600	GMP synthase [glutamine-hydrolyzing], amidotransferase subunit (EC 6.3.5.2) / GMP synthase [glutamine-hydrolyzing], ATP pyrophosphatase subunit (EC 6.3.5.2)	guaA
*Bv*629_608	Phosphoribosylaminoimidazole carboxylase catalytic subunit (EC 4.1.1.21)	purE
*Bv*629_616	Phosphoribosylformylglycinamidine cyclo-ligase (EC 6.3.3.1)	purM

*Bv*629_618	IMP cyclohydrolase (EC 3.5.4.10) / Phosphoribosylaminoimidazolecarboxamide formyltransferase (EC 2.1.2.3)	purH
*Bv*629_619	Phosphoribosylamine--glycine ligase (EC 6.3.4.13)	purD
*Bv*629_644	FIG007491: hypothetical protein YeeN	yeeN
*Bv*629_812	Thiamin biosynthesis protein ThiC	thiC
*Bv*629_1044	3-oxoacyl-[acyl-carrier-protein] synthase, KASII (EC 2.3.1.179)	fabF
*Bv*629_1080	Thiazole biosynthesis protein ThiG	thiG
*Bv*629_1405	GTP-binding protein TypA/BipA	typA
*Bv*629_1430	Phosphopantetheine adenylyltransferase (EC 2.7.7.3)	coaD
*Bv*629_1473	Carbamoyl-phosphate synthase large chain (EC 6.3.5.5)	carB
*Bv*629_1493	Methionyl-tRNA formyltransferase (EC 2.1.2.9)	fmt
*Bv*629_1498	Ribulose-phosphate 3-epimerase (EC 5.1.3.1)	rpe
*Bv*629_1509	Malonyl CoA-acyl carrier protein transacylase (EC 2.3.1.39)	fabD
*Bv*629_1510	3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100)	fabG
*Bv*629_1511	Acyl carrier protein	acpP
*Bv*629_1517	Signal recognition particle, subunit Ffh SRP54 (TC 3.A.5.1.1)	ffh
*Bv*629_1518	SSU ribosomal protein S16p	rpsP
*Bv*629_1522	tRNA (Guanine37-N1) -methyltransferase (EC 2.1.1.31)	trmD
*Bv*629_1527	Succinyl-CoA ligase [ADP-forming] beta chain (EC 6.2.1.5)	sucC
*Bv*629_1528	Succinyl-CoA ligase [ADP-forming] alpha chain (EC 6.2.1.5)	sucD
*Bv*629_1533	ATP-dependent protease HslV (EC 3.4.25.-)	hslV
*Bv*629_1534	ATP-dependent hsl protease ATP-binding subunit HslU	hslU
*Bv*629_1565	SSU ribosomal protein S2p (SAe)	rpsB
*Bv*629_1566	Translation elongation factor Ts	tsf
*Bv*629_1567	Uridine monophosphate kinase (EC 2.7.4.22)	pyrH
*Bv*629_1568	Ribosome recycling factor	frr
*Bv*629_1569	Undecaprenyl diphosphate synthase (EC 2.5.1.31)	uppS
*Bv*629_1571	1-deoxy-D-xylulose 5-phosphate reductoisomerase (EC 1.1.1.267)	dxr
*Bv*629_1582	Polyribonucleotide nucleotidyltransferase (EC 2.7.7.8)	pnp
*Bv*629_1590	4-hydroxy-tetrahydrodipicolinate synthase (EC 4.3.3.7)	dapA
*Bv*629_1606	RecA protein	recA
*Bv*629_1718	Aconitate hydratase (EC 4.2.1.3) @ 2-methylisocitrate dehydratase (EC 4.2.1.99)	citB
*Bv*629_1793	Glutamate synthase [NADPH] large chain (EC 1.4.1.13)	gltA
*Bv*629_1847	Dihydrolipoamide succinyltransferase component (E2) of 2-oxoglutarate dehydrogenase complex (EC 2.3.1.61)	odhB
*Bv*629_1985	Endonuclease III (EC 4.2.99.18)	nth
*Bv*629_1994	3-methyl-2-oxobutanoate hydroxymethyltransferase (EC 2.1.2.11)	panB
*Bv*629_2016	Tryptophan synthase beta chain (EC 4.2.1.20)	trpB
*Bv*629_2025	Nucleoside diphosphate kinase (EC 2.7.4.6)	ndk
*Bv*629_2031	DNA-binding protein HBsu	hupA
*Bv*629_2197	Methylenetetrahydrofolate dehydrogenase (NADP+) (EC 1.5.1.5) / Methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9)	folD
*Bv*629_2263	Manganese superoxide dismutase (EC 1.15.1.1)	sodA
*Bv*629_2304	Chaperone protein DnaJ	dnaJ
*Bv*629_2305	Chaperone protein DnaK	dnaK
*Bv*629_2395	Aspartyl-tRNA synthetase (EC 6.1.1.12)	aspS
*Bv*629_2409	tRNA-guanine transglycosylase (EC 2.4.2.29)	tgt

*Bv*629_2410	S-adenosylmethionine:tRNA ribosyltransferase-isomerase (EC 5.-.-.-)	queA
*Bv*629_2412	Holliday junction DNA helicase RuvB	ruvB
*Bv*629_2426	GTP-binding protein Obg	obg
*Bv*629_2428	LSU ribosomal protein L27p	rpmA
*Bv*629_2430	LSU ribosomal protein L21p	rplU
*Bv*629_2442	Valyl-tRNA synthetase (EC 6.1.1.9)	valS
*Bv*629_2448	Porphobilinogen synthase (EC 4.2.1.24)	hemC
*Bv*629_2454	ATP-dependent protease La (EC 3.4.21.53) Type I	lonA
*Bv*629_2457	ATP-dependent Clp protease ATP-binding subunit ClpX	clpX
*Bv*629_2461	3-isopropylmalate dehydratase small subunit (EC 4.2.1.33)	leuD
*Bv*629_2462	3-isopropylmalate dehydratase large subunit (EC 4.2.1.33)	leuC
*Bv*629_2463	3-isopropylmalate dehydrogenase (EC 1.1.1.85)	leuB
*Bv*629_2467	Acetolactate synthase large subunit (EC 2.2.1.6)	ilvB
*Bv*629_2474	Ribonuclease PH (EC 2.7.7.56)	rph
*Bv*629_2487	Thioredoxin	trxA
*Bv*629_2503	Phenylalanyl-tRNA synthetase alpha chain (EC 6.1.1.20)	pheS
*Bv*629_2524	LSU ribosomal protein L20p	rplT
*Bv*629_2525	Translation initiation factor 3	infC
*Bv*629_2531	Threonyl-tRNA synthetase (EC 6.1.1.3)	thrS
*Bv*629_2561	Acetyl-coenzyme A carboxyl transferase alpha chain (EC 6.4.1.2)	accA
*Bv*629_2562	Acetyl-coenzyme A carboxyl transferase beta chain (EC 6.4.1.2)	accD
*Bv*629_2593	SSU ribosomal protein S4p (S9e)	rpsD
*Bv*629_2683	S-adenosylmethionine synthetase (EC 2.5.1.6)	metK
*Bv*629_2853	Lipoate synthase	lipA
*Bv*629_2899	Glycine cleavage system H protein	gcvH
*Bv*629_2928	Fumarate hydratase class II (EC 4.2.1.2)	fumC
*Bv*629_2998	tmRNA-binding protein SmpB	smpB
*Bv*629_3036	Triosephosphate isomerase (EC 5.3.1.1)	tpiA
*Bv*629_3037	Phosphoglycerate kinase (EC 2.7.2.3)	pgk
*Bv*629_3039	NAD-dependent glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12)	gapA
*Bv*629_3094	ATP-dependent Clp protease proteolytic subunit (EC 3.4.21.92)	clpP
*Bv*629_3141	Excinuclease ABC subunit A	uvrA
*Bv*629_3142	Excinuclease ABC subunit B	uvrB
*Bv*629_3157	Peptide chain release factor 2	prfB
*Bv*629_3158	Protein export cytoplasm protein SecA ATPase RNA helicase (TC 3.A.5.1.1)	secA
*Bv*629_3270	Large-conductance mechanosensitive channel	mscL
*Bv*629_3271	3-hydroxyacyl-[acyl-carrier-protein] dehydratase, FabZ form (EC 4.2.1.59)	fabZ
*Bv*629_3285	Ammonium transporter	amt
*Bv*629_3312	UDP-N-acetylglucosamine 1-carboxyvinyltransferase (EC 2.5.1.7)	murA
*Bv*629_3318	ATP synthase alpha chain (EC 3.6.3.14)	atpA
*Bv*629_3325	Uracil phosphoribosyltransferase (EC 2.4.2.9)	upp
*Bv*629_3338	Peptide chain release factor 1	prfA
*Bv*629_3344	Transcription termination factor Rho	rho
*Bv*629_3702	GTP-binding and nucleic acid-binding protein YchF	ychF
*Bv*629_3711	tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA	gidA

APÊNDICE E

Material Suplementar – Manuscrito 3

Genome sequence of *Serratia marcescens* strain 1274, an endophytic bacterium isolated from *Agave sisalana*

Brena M. M. Sant'Anna[1], Jorge T. De Souza[2], Phellippe P. A. Marbach[3], Vasco Azevedo[4], Artur Silva[5], Rommel T. J. Ramos[5], Luis G. C. Pacheco[1,] Artur T. L. Queiroz[6], Milton R. A. Roque[1]

1. Universidade Federal da Bahia, UFBA – Brazil; 2. Universidade Federal de Lavras, UFLA – Brazil; 3. Universidade Federal do Recôncavo da Bahia, UFRB – Brazil; 4. Institute of Biological Sciences, Federal University of Minas Gerais (UFMG), Belo Horizonte, MG, Brazil; 5. Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil; 6. CPqGM-Fiocruz/BA – Brazil

*Corresponding author: milton.roque@ufba.br (M. Roque)

Supplementary Table S1

Values of digital DNA-DNA Hibridization (dDDH) of *Serratia marcescens* strains (prefix Sm) against strain WW4. dDDH was done using GGDC 2.1 server, Genome-to-Genome Distance Calculator (available at http://ggdc.dsmz.de/ggdc.php#), considering cutoff above 70% for species classification (Meier-Kolthoff et al. 2013).

Query genome	**Reference genome**	**DDH**	**Model C.I.**	**Distance**
SmWW4	Sm1274	**81.9**	[78 - 85.2%]	0.1228
SmWW4	SmU36365	**91**	[87.9 - 93.3%]	0.0756
SmWW4	SmFS14	**90**	[86.7 - 92.5%]	0.0809
SmWW4	SmB3R3	**86.7**	[83.1 - 89.7%]	0.0983
SmWW4	SmSMB209	**81**	[77.1 - 84.4%]	0.1272
SmWW4	SmRSC14	**82.3**	[78.5 - 85.6%]	0.1205
SmWW4	SmDb11	**84.6**	[80.8 - 87.7%]	0.1092
SmWW4	SmSM39	**77.8**	[73.8 - 81.3%]	0.1436
SmWW4	SmCAV1492	**75.4**	[71.5 - 79%]	0.1553
SmWW4	SmUNAM836	**78.1**	[74.2 - 81.6%]	0.1417
SmWW4	SmUMH12	**82.2**	[78.3 - 85.5%]	0.1213
SmWW4	SmUMH11	**83.2**	[79.4 - 86.5%]	116
SmWW4	SmUMH10	**83.2**	[79.4 - 86.5%]	116
SmWW4	SmUMH9	**81.1**	[77.2 - 84.5%]	0.1268
SmWW4	SmUMH8	**86.5**	[82.9 - 89.4%]	0.0994
SmWW4	SmUMH7	**82.1**	[78.2 - 85.4%]	0.1217
SmWW4	SmUMH6	**79.3**	[75.3 - 82.7%]	0.1361
SmWW4	SmUMH5	**81.9**	[78 - 85.2%]	123
SmWW4	SmUMH3	**77.5**	[73.5 - 81%]	0.1449
SmWW4	SmUMH2	**79.7**	[75.8 - 83.2%]	0.1337
SmWW4	SmUMH1	**84.9**	[81.2 - 88%]	0.1077
SmWW4	SmFGI94	**37.9**	[34.6 - 41.4%]	0.4138

APÊNDICE F

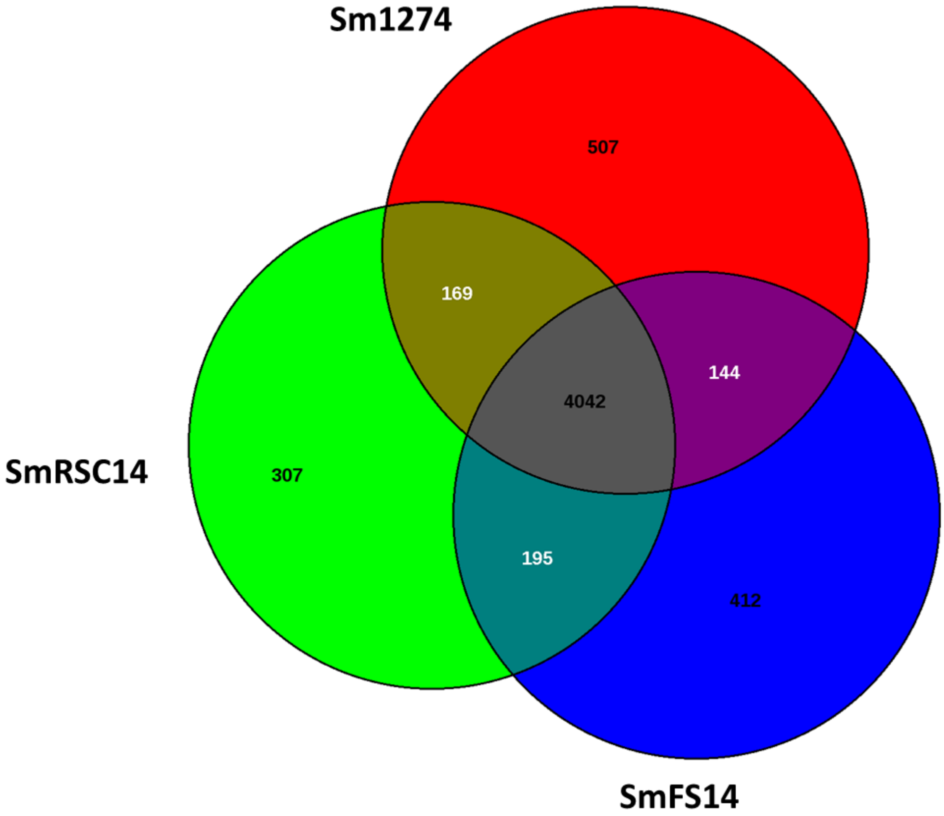
Material Suplementar – Manuscrito 4

Endophytic Behavior in *Serratia marcescens*: New Insights from Comparative Genomics of Strain 1274

Brena M. M. Sant'Anna*1*, Artur T. L. Queiroz*2*, Icaro Lopes[1], Luis G. C. Pacheco*1*, , Milton R. A. Roque*1**
1. Universidade Federal da Bahia, UFBA – Brazil; 2. CPqGM-Fiocruz/BA – Brazil;
*Correspondencing authors: milton.roque@ufba.br (M. Roque)

Supplementary Figure S1

Venn Diagram of comparative analysis between the endophytic strains *Serratia marcescens* 1274, RSC 14 and FS14.


Sm1274
SmRSC14
SmFS14