

**UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA
FILHO” CAMPUS DE BAURU**

Beatriz Augusto Domingos

**Ecologia e modo de vida do fitopatógeno diazotrófico *Herbaspirillum
rubrisubalbicans* por meio de genômica comparativa**

Bauru
2024

Beatriz Augusto Domingos

**Ecologia e modo de vida do fitopatógeno diazotrófico *Herbaspirillum
rubrisubalbicans* por meio de genômica comparativa**

Trabalho de Conclusão de Curso, apresentado à Faculdade de Ciências, Campus de Bauru, Universidade Estadual Paulista “Júlio de Mesquita Filho”, como parte das exigências para obtenção do título de Bacharelado em Ciências Biológicas.

Orientador: Prof^o. Dr^o. Filipe Pereira Matteoli

Bauru
2024

D671e Domingos, Beatriz Augusto
Ecologia e modo de vida do fitopatógeno diazotrófico
Herbaspirillum rubrisubalbicans por meio de genômica comparativa /
Beatriz Augusto Domingos. -- Bauru, 2024
39 p.

Trabalho de conclusão de curso (Bacharelado - Ciências
Biológicas) - Universidade Estadual Paulista (UNESP), Faculdade de
Ciências, Bauru
Orientador: Filipe Pereira Matteoli

1. Microbiologia. 2. Bioinformática. 3. Bactéria. 4. Fitopatógeno. 5.
Diazotrófico. I. Título.

BEATRIZ AUGUSTO DOMINGOS

**Ecologia e modo de vida do fitopatógeno diazotrófico *Herbaspirillum*
rubrisubalbicans por meio de genômica comparativa**

Trabalho de Conclusão de Curso, apresentado à Faculdade de Ciências, Campus de Bauru, Universidade Estadual Paulista “Júlio de Mesquita Filho”, como parte das exigências para obtenção do título de Bacharelado em Ciências Biológicas.

Orientador: Profº. Drº. Filipe Pereira Matteoli

BANCA EXAMINADORA

Profº. Drº. Filipe Pereira Matteoli

Drº. Francisnei Pedrosa Silva

Profº. Drº. Ricardo Utsunomia

Bauru, 02 de dezembro de 2024.

AGRADECIMENTOS

Gostaria de expressar minha profunda gratidão a todos que de alguma forma contribuíram para a realização deste trabalho. Agradeço primeiramente à minha família, pelo apoio e amor durante toda a trajetória. Aos meus amigos, pela companhia, compreensão e incentivo, especialmente nos momentos mais desafiadores.

Agradeço à política de cotas e à Unesp pelo acesso ao ensino superior gratuito e de qualidade, assim como pela oportunidade de crescimento acadêmico e pessoal, e à FAPESP, pelo apoio financeiro que possibilitou a realização deste projeto. Ao Cursinho Ferradura, pela contribuição à permanência na universidade e apoio incondicional.

Um agradecimento especial ao meu orientador, Filipe, pela orientação valiosa, pelas inúmeras contribuições que enriqueceram este trabalho e a minha visão sobre a carreira acadêmica.

Aos meus professores.

RESUMO

Herbaspirillum rubrisubalbicans é uma bactéria Gram-negativa, pertencente à ordem *Burkholderiales*, que possui os genes *nif* presentes de forma conservada em todos os genomas conhecidos o que justifica a caracterização do fenótipo diazotrófico para a espécie. Entretanto, em culturas de sorgo e cana-de-açúcar foi reportado que *H. rubrisubalbicans* estabelece uma relação fitopatogênica endofítica, caracterizada pelo surgimento de mosqueado clorótico. Assim, *H. rubrisubalbicans* apresenta fenótipos considerados excludentes. Por isso, propõe-se no presente estudo a caracterização da ecologia e o modo de vida de *Herbaspirillum rubrisubalbicans*, com ênfase na dualidade diazotrófico *versus* fitopatogênico por meio de genômica comparativa. Os dados genômicos foram obtidos a partir do RefSeq (NCBI) totalizando um *dataset* com nove *assemblies*; destes, quatro estão completos e o restante apresentam-se em *drafts*. A avaliação da qualidade dos *assemblies* revelou que a média do tamanho dos genomas foi de 5.777.246,67 pb $\pm$ 199.422,0672. Dentre os *drafts*, o *assemblie* de *H. rubrisubalbicans* Os45 é o mais fragmentado do *dataset* com 144 *contigs*, N50 de 248.240 pb e L50 de 7 *contigs*. A anotação dos genes foi realizada com o *software* Bakta, após o *benchmark* contra os *softwares* PROKKA e PGAP, considerando o número de proteínas hipotéticas preditas, CDS, tRNAs e a velocidade do anotador. A qualidade biológica da anotação foi aferida pelo *software* BUSCO, que demonstrou média de 98,74% de presença de ortólogos completos, com desvio padrão de 0,014, uma clara indicação da qualidade do *dataset*. A análise de ANI identificou que todos os genomas possuem 95% ou mais de identidade de nucleotídeos e há formação de dois grupos similares entre si. A partir do desenvolvimento do pangenoma da espécie, identificou-se a presença de 7.881 genes, nos quais 3.580 estão presentes no *hard core*, 2.687 compõem o grupo *shell* e 1514 são integrantes do grupo *cloud*; portanto, tem-se uma alta homogeneidade gênica. Foram identificados 1.572 genes benéficos associados a plantas disponíveis no PLaBase e 498 ocorrências de genes presentes em 20 espécies bacterianas reconhecidas como fitopatogênicas através do PHI-base.

ABSTRACT

Herbaspirillum rubrisubalbicans is a Gram-negative bacterium that belongs to the order *Burkholderiales*, and harbors *nif* genes conserved across all known genomes, which justifies the characterization of a diazotrophic phenotype for the species. However, in sorghum and sugarcane cultures, it has been reported that *H. rubrisubalbicans* establishes an endophytic phytopathogenic relationship, characterized by the appearance of chlorotic streaking. *H. rubrisubalbicans* presents phenotypes that are considered mutually exclusive. Therefore, this study proposes the characterization of the ecology and lifestyle of *Herbaspirillum rubrisubalbicans*, with an emphasis on the duality of diazotrophic versus phytopathogenic, through comparative genomics. Genomic data were obtained from RefSeq (NCBI), resulting in a dataset of nine assemblies; of these, four are complete, and the rest are drafts. The assembly quality evaluation revealed that the average genome size was 5.777.246.67 bp $\pm$ 199.422.0672. Among the drafts, the assembly of *H. rubrisubalbicans* Os45 is the most fragmented in the dataset with 144 contigs, an N50 of 248,240 bp, and an L50 of 7 contigs. The gene annotation was performed using the software Bakta, after benchmarking against the software PROKKA and PGAP, considering the number of predicted hypothetical proteins, CDS, tRNAs, and the annotator's speed. The biological quality of the annotation was assessed with the software BUSCO, which showed an average of 98.74% of complete ortholog presence, with a standard deviation of 0.014, a clear indication of the quality of the dataset. The ANI analysis identified that all genomes more than 95% nucleotide identity. From the development of the species pangenome, the presence of 7881 genes was identified, of which 3580 are part of the hard core, 2687 make up the shell group, and 1514 are members of the cloud group; thus, there is a high genetic homogeneity. A total of 1572 plant-associated genes were identified from PLaBase, and 498 occurrences of genes present in 20 bacterial species recognized as phytopathogenic were identified through PHI-base.

LISTA DE FIGURAS

Figura 1. Sintomas de mosqueado clorótico vermelho em folha de sorgo	18
Figura 2- Fluxograma geral da metodologia aplicada	20
Figura 3 - Fluxograma de análise de grau de completude de genes ortólogos de cópia única nos genomas	21
Figura 4 - Controle de Qualidade - Métricas Matemáticas	24
Figura 5- Resultado do <i>benchmark</i> entre os anotadores	25
Figura 6 - Controle de Qualidade - Genes Ortólogos de Cópia Única	26
Figura 7 - Controle de Qualidade - Análise de ANI	27
Figura 8 - Anotação de genes	28
Figura 9 - Genes PGPR pertencentes às três categorias disponíveis no PLaBase	29
Figura 10- Distribuição dos <i>hits</i> de genes patogênicos	30

LISTA DE TABELAS

Tabela 1 - Genomas de <i>H. rubrisubalbicans</i> disponíveis no Genbank	23
Tabela 2 - Identificação de genes relacionados à interação bactéria-planta em <i>H. rubrisubalbicans</i> .	29
Tabela 3 - Número de hits de genes de <i>H. rubrisubalbicans</i> contra o PHI-base.	30
Tabela 4 - Número de genes distintos por espécies fitopatogênicas em <i>H. rubrisubalbicans</i> .	31

LISTA DE ABREVIATURAS E SIGLAS

ANI	Identidade média de nucleotídeos
bGWAS	Bacterial genome-wide association studies
CDS	Sequências codificadoras
FBN	Fixação biológica de nitrogênio
NCBI	National Center for Biotechnology International
PGPR	Plant Growth Promoting Bacteria

SUMÁRIO

1. INTRODUÇÃO	12
1.1 Genômica comparativa e o estudo de bactérias causadoras de fitopatologias	12
1.2 <i>Gênero Herbaspirillum</i>	15
1.3 <i>Herbaspirillum rubrisubalbicans</i>	16
2. JUSTIFICATIVA	18
3. OBJETIVO GERAL	19
3.1 Objetivos específicos	19
4. MATERIAIS E MÉTODOS	19
4.1 Obtenção dos genomas e controle de qualidade	20
4.2 Anotação de genes	21
4.3 Pangenoma de <i>H. rubrisubalbicans</i>	22
5. RESULTADOS	23
5.1 Qualidade dos assemblies e Benchmark de anotadores	23
5.2 Anotação de genes	27
5.3 Pangenoma de <i>H. rubrisubalbicans</i>	28
5.4 Identificação de genes PGPR e quantificação de genes fitopatogênicos	28
6. DISCUSSÃO	31
7. CONCLUSÃO	33
REFERÊNCIAS BIBLIOGRÁFICAS	33

1. INTRODUÇÃO

1.1 Genômica comparativa e o estudo de bactérias causadoras de fitopatologias

Os microrganismos existem desde a origem da vida na Terra, há pelo menos 3,8 bilhões de anos (Cavicchioli et. al, 2019), eles foram fundamentais para o surgimento da primeira célula eucariótica, da qual todas as plantas e animais, incluindo os seres humanos, foram derivados (Rappuoli et. al, 2023) e permanecem desempenhando papéis fundamentais para o planeta, envolvendo o ciclo do carbono, a disponibilização de nutrientes, a saúde de animais e plantas, impactando atividades humanas essenciais como a agricultura e a cadeia trófica (Cavicchioli et. al, 2019).

Dentro deste contexto, a associação entre plantas e microrganismos, incluindo bactérias, remonta a mais de 400 milhões de anos. Embora existam registros fósseis de interações planta-fungo, como as micorrizas arbusculares, não há registros fósseis disponíveis do início da associação planta-bactéria devido à natureza frágil da célula bacteriana (Lemanceau *et al.*, 2017). No entanto, como os microrganismos existiam antes das plantas e colonizaram todos os ecossistemas, é provável que as bactérias tenham formado associações simbióticas, comensais e parasitárias com plantas em um período muito mais antigo do que se imagina. Essas interações envolvem comunicação química e física, permitindo que bactérias se estabeleçam e colonizem partes das plantas, como raízes, caules e folhas (Santoyo, 2022).

As interações entre plantas e bactérias apresentam importância econômica, como é o caso da bactéria Gram negativa, aeróbia *Pseudomonas syringae* que causa necroses em folhas, frutos e caules, e pode sobreviver em diversos habitats (Monteil et al., 2014). A espécie fitopatogênica *P. syringae* dispõe de alta variabilidade genética, com pelo menos 50 patovares que infectam diferentes tipos de plantas, incluindo monocotiledôneas e dicotiledôneas, e possuem diversos fatores de virulência (Green et al., 2010). Assim como *P. syringae*, o complexo de espécies *Ralstonia solanacearum* apresenta relevância econômica, pois compreende linhagens geneticamente distintas de bactérias aeróbias e Gram-negativas, sendo considerado o complexo fitobacteriano mais impactante no mundo (Mansfield et al., 2012). Espécies deste complexo infectam principalmente as raízes de cerca de 250 espécies de plantas dentro de 54 famílias diferentes, invadindo e se multiplicando nos vasos do xilema, causando murcha generalizada e morte das plantas infectadas (Prior et al., 2016).

As bactérias possuem genomas compactos, com altas taxas de mutações, sujeitos a diferentes mecanismos de transferência horizontal de genes (*horizontal gene transfers* –

HGTs), estes fenômenos são importantes forças evolutivas nos procariotos (Ochman, et al. 2000). Os genes transferidos horizontalmente são classificados como xenólogos (Patterson, 1988), um tipo de homologia, assim como a paralogia e a ortologia (Fitch, 2000). Além desses elementos, as bactérias realizam reprodução clonal, caracterizada por ser uma reprodução por fissão binária (Mishra; Forche; Anderson, 2021). Este conjunto de fatores permite assumir que a correlação genótipo x fenótipo é altamente fidedigna e constitui a base teórica dos estudos de associação genômica ampla (bacterial genome-wide association studies – bGWAS) (Power; Parkhill; De Oliveira, 2017). Apesar disso, uma espécie de bactéria pode diferir em até 5% da sua identidade média de nucleotídeos (ANI) e há alto grau de variabilidade gênica na mesma espécie; por isso a generalização de fenótipos para uma espécie não é uma realidade plausível.

Diante da complexidade genômica bacteriana, o avanço das tecnologias de sequenciamento massivo paralelo que incluem plataformas como Illumina, Ion Torrent e Pacific Biosciences, representam um avanço significativo na análise genômica. Ao sequenciar simultaneamente um grande número de fragmentos de DNA ou RNA, essas metodologias proporcionam resultados mais rápidos e econômicos em comparação ao disponível antes do advento do sequenciamento por amplificação em ponte (Illumina). Essa capacidade de analisar genomas em larga escala tem impulsionado a pesquisa em diversas áreas da biologia molecular e da genética (Alonso et al., 2018).

A popularização do sequenciamento massivo paralelo propiciou o desenvolvimento de estudos de associação genômica ampla (GWAS) generalizados e aqueles voltados para bactérias, que surgem como ferramentas para compreender as relações entre genótipo e fenótipo em populações bacterianas. Os bGWAS constituem um conjunto de técnicas de genômica comparativa destinadas a identificar variantes genéticas em genomas que se relacionam com características fenotípicas dentro de uma população, sua implementação se tornou viável com o aumento do número de isolados sequenciados em genomas completos (Coll *et al.*, 2022). Estes estudos têm sido utilizados para investigar a base genética de diversas características bacterianas, como resistência a antibióticos (Chewapreecha *et al.*, 2014; Salipante *et al.*, 2015; Earle *et al.*, 2016; Coll *et al.*, 2018), vulnerabilidade a desinfetantes (Pidot *et al.*, 2018), especificidade para hospedeiros (Sheppard & Didelot, 2013; Richardson, et al., 2018), entre outros (Coll *et al.*, 2022). A expansão do uso de bGWAS é facilitada pelo desenvolvimento de ferramentas especializadas que se adequam às particularidades dos genomas bacterianos, originalmente inspiradas em ferramentas criadas para GWAS em humanos (Lees, et al., 2018).

Uma das abordagens mais contemporâneas para estudar populações de cepas da mesma espécie de bactérias é por meio de detecção de genes ou alelos associados a fenótipos em pangenomas. O termo pangenoma foi cunhado por Tettelin *et al.* (2005) em pesquisas sobre infecções neonatais em humanos por diferentes cepas de *Streptococcus agalactiae* para descrever coleções de genes disponíveis em toda uma população de um grupo taxonômico. É possível definir categorias para classificar as proteínas dentro de um pangenoma, com base em parâmetros que estabelecem suas respectivas porcentagens de presença nos genomas analisados (Beier, 2022). A categoria *hard core* refere-se às proteínas essenciais para a estrutura do táxon, que devem estar presentes em 99% (considerando ausências de erros de montagem) dos genomas. Já o *soft core* engloba proteínas que, embora não sejam vitais, aparecem em 95% dos genomas. O grupo *shell* é composto por proteínas que ocorrem em 15% dos genomas, portanto, proteínas específicas que podem estar relacionadas a características que distinguem linhagens ou grupos funcionais. Por fim, as proteínas classificadas como *cloud* são raras, presentes em menos de 15% dos genomas (Beier, 2022).

Adicionalmente, Tettelin *et al.* (2008) propuseram uma dicotomia para os pangenomas, classificando-os como abertos ou fechados. Um pangenoma aberto caracteriza-se por sua natureza dinâmica e em constante expansão, portanto, à medida que novos genomas são adicionados à análise, o repertório genético total - medido em número de genes - aumenta indefinidamente, revelando um fluxo constante de genes inéditos (Santos, 2024). Um exemplo de pangenoma aberto foi demonstrado por meio da análise de 128 genomas de *Bacillus subtilis*, espécie Gram-positiva encontrada no solo e em associação com plantas (WU; WANG; GAO, 2021). Em contraste, um pangenoma fechado apresenta um limite definido: a inclusão de novos genomas não altera significativamente o conjunto de genes já conhecidos, indicando que a diversidade genética dentro daquele grupo está contida nos genomas previamente sequenciados (Santos, 2024). A bactéria fitopatogênica, Gram-negativa *Xanthomonas campestris*, responsável por acometer plantas hospedeiras de importância econômica como a banana e o pimentão, é um exemplo de espécie que apresenta um pangenoma fechado de acordo com análise pangenômica estrutural realizada em uma população de 14 genomas da espécie (Guimarães, 2020).

No âmbito das pesquisas que buscam compreender as relações entre bactérias que causam doenças em plantas e seus hospedeiros, os estudos de associação genômica ampla (GWAS) têm se mostrado uma ferramenta eficaz (Holtappels, 2024). Essas análises permitem avaliar a qualidade dos genomas, estudar o conjunto completo de genes de uma população bacteriana (pangenoma) e construir árvores filogenéticas. Um estudo sobre

Pseudomonas syringae (Holtappels, 2024), por exemplo, utilizou ferramentas de bGWAS para classificar as cepas por meio da identificação do grupo *hard core* e propor estratégias para controlar a infecção deste procarioto em plantas, demonstrando a importância desse conhecimento para a agricultura.

1.2 Gênero *Herbaspirillum*

O gênero de bactérias Gram-negativas *Herbaspirillum* foi descrito nos anos 80 por Baldani e colaboradores (1986). A espécie tipo do gênero foi denominada *Herbaspirillum seropedicae*, em alusão ao nome do município no qual se localiza a Universidade Federal Rural do Rio de Janeiro, instituição na qual a pesquisa fora desenvolvida. As 15 espécies conhecidas de *Herbaspirillum* apresentam ampla distribuição, colonizando ambientes aquáticos, solo, sedimentos e tecidos vegetais (Tomazini, 2022). Taxonomicamente, o gênero pertence ao filo Pseudomonadota (anteriormente Proteobacteria), classe Betaproteobacteria, ordem Burkholderiales, família Oxalobacteraceae.

Nos anos seguintes, isolados de *Pseudomonas rubrisubalbicans*, um fitopatógeno foliar brando de cana-de-açúcar (*Saccharum barberi* e *Saccharum officinarum*), identificado pela primeira vez em 1927 no estado da Louisiana, nos Estados Unidos (Hale & Wilkie, 1972), foram reclassificados para o gênero *Herbaspirillum* (Baldani *et al.*, 1996; Gillis *et al.*, 1991). Interessantemente, tanto a espécie reconhecidamente diazotrófica *H. seropedicae*, quanto a fitopatogênica *H. rubrisubalbicans*, demonstraram serem capazes de realizar fixação do nitrogênio atmosférico (Pimentel *et al.*, 1991), uma etapa crucial do ciclo deste elemento químico.

Estudos ecológicos envolvendo *H. seropedicae*, *H. rubrisubalbicans* e *H. frisingense* demonstraram a capacidade destas espécies de se estabelecerem como endófitos em tecidos vegetais (James; Olivares 1998; Rothballer *et al.*, 2008). Apesar das similaridades, *H. seropedicae* se destacou em múltiplos estudos que apontaram maior espectro de plantas hospedeiras e de ocorrência geográfica nesta espécie do que em *H. rubrisubalbicans* e *H. frisingense*, com relatos em diferentes cultivares de cana-de-açúcar (incluindo híbridos de *S. officinarum*, *S. barberi*, *S. spontaneum*), milho (*Z. mays*) (Matteoli *et al.*, 2020), sorgo (*S. bicolor*), arroz (*O. sativa*) (Olivares *et al.*, 1996; Matteoli *et al.*, 2020) e gramíneas como *Pennisetum purpureum*, *Digitaria decumbens*, *Brachiaria decumbens* (Matteoli *et al.*, 2020). Além destas espécies de Poaceae, *H. seropedicae* também se mostrou capaz de colonizar espécies da família Leguminosae e espécies frutíferas tropicais (Matteoli *et al.*, 2020). Em contrapartida, *H. frisingense* e *H. rubrisubalbicans* apresentaram estreito espectro de plantas

hospedeiras, além de resultados que indicaram a presença desta última em folhas com sintomas de mosqueado clorótico em cana-de-açúcar nos EUA, Austrália e China (Tan *et al.*, 2010; Quiñones-Valdez *et al.*, 2022) e mosqueado clorótico vermelho em sorgo (*Sorghum bicolor*) (Tuleski *et al.*, 2019a).

Indubitavelmente, o fenótipo diazotrófico foi por muitos anos a principal motivação para os estudos neste gênero. O sequenciamento genômico da cepa *H. seropedicae* SmR1 (Pedrosa *et al.*, 2011) se tornou um dos pilares da genômica comparativa de bactérias promotoras do crescimento vegetal (*plant growth promoting bacteria* - PGPB). Entretanto, isolados clínicos de *H. seropedicae* diminuíram as possibilidades dessa espécie ser amplamente aceita como constituinte de inoculantes comerciais (Suwantararat *et al.*, 2015).

1.3 *Herbaspirillum rubrisubalbicans*

Os dados disponíveis na literatura apontam que *H. rubrisubalbicans* apresenta uma série de características específicas que podem justificar seu comportamento fitopatogênico. No que tange ao processo de colonização dos hospedeiros, destaca-se que, assim como *H. seropedicae*, *H. rubrisubalbicans* coloniza a planta através de superfícies descontínuas da raiz e ferimentos (Monteiro *et al.*, 2012). De maneira passiva, a bactéria penetra e se dissemina pelos espaços intercelulares e feixes vasculares das partes aéreas, resultando na colonização e fixação no xilema (Monteiro *et al.*, 2012; Olivares *et al.*, 1996).

O processo de interação bactéria-planta sofre interferência de aspectos relacionados à motilidade e presença de pilus, bem como as propriedades genômicas que formam estas estruturas. A motilidade bacteriana é essencial para a colonização, ao entrar em contato com o tecido vegetal, as bactérias podem se deslocar dentro da planta e invadir os tecidos do hospedeiro (Trivedi *et al.*, 2020). Tomazini (2022) estudou as principais formas de motilidade em *H. rubrisubalbicans* M1, isolada da cana-de-açúcar (*Saccharum officinarum*) descrevendo a presença de motilidade flagelar natatória, em ambientes líquidos, motilidade deslizante (*swarming*), que envolve flagelos e o pilus tipo IV e deslocamento sobre superfícies (Trivedi *et al.*, 2020), além da motilidade espasmódica (*twitching*) que decorre da extensão, ligação e tração do pilus do tipo VI que confere agilidade à fimbria, possibilitando a movimentação através de impulso por contato em substratos orgânicos, importante para a colonização da superfície (Charles-Orszag *et al.*, 2024).

Os pili, por sua vez, são filamentos proteicos, compostos por várias subunidades, que se localizam na superfície das células bacterianas (Ramirez *et al.*, 2020). Eles desempenham um papel crucial como iniciadores na interação patógeno-hospedeiro (Zhou *et al.*, 2021),

assim como para a motilidade bacteriana, como mencionado. Em relação ao pili de *H. rubrisubalbicans* M1, Tomazini (2022) aponta a presença das proteínas Pile, PilU, PilV, PilX e PilW, presentes em *Pseudomonas aeruginosa* e PilH, PilI, PilJ, PilK e PilL (Craig *et al.*, 2019), dispostas em *N. gonorrhoeae*. A disponibilidade de uma vasta gama de proteínas responsáveis pela formação do pili, confere à *H. rubrisubalbicans* M1 ampla variedade estrutural, o que sugere uma grande capacidade de adaptação ao ambiente e aos hospedeiros (Tomazini, 2022).

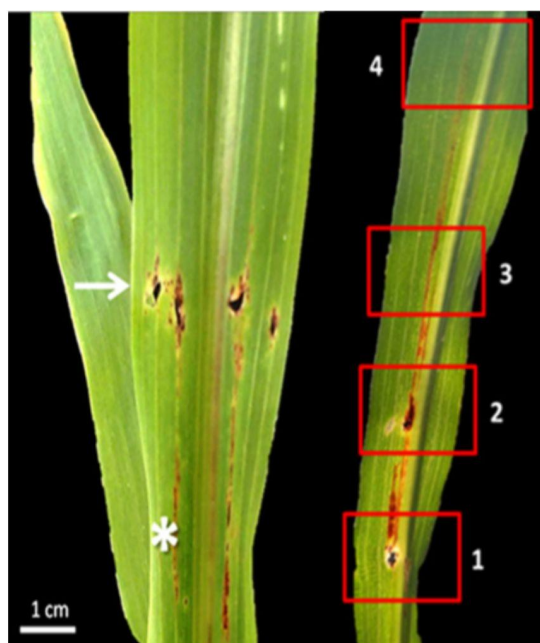
Assim como as características relacionadas à motilidade e à estrutura dos pili, a formação de biofilme afeta a interação entre bactérias e plantas, uma vez que desempenha um papel primordial na colonização eficaz do hospedeiro (Li *et al.*, 2013). Salienta-se que o biofilme bacteriano consiste em uma comunidade estruturada de microrganismos, que podem ser da mesma espécie ou de espécies diversas, envolvidos em uma matriz de exopolissacarídeos. No ambiente, a formação do biofilme está associada à proteção contra tensões de cisalhamento, desidratação, substâncias tóxicas e ataques de protozoários (Sauer *et al.*, 2022). Em contrapartida aos efeitos benéficos, a formação de biofilme bacteriano pode gerar diminuição da recepção de nutrientes para os microrganismos, em decorrência da estrutura ser uma barreira ao ambiente, ocasionando a desaceleração do metabolismo da comunidade (Branda, 2005). Em *H. rubrisubalbicans* M1, há ocorrência do *cluster ws* responsável pela codificação de uma via de biossíntese de celulose bacteriana, envolvido na formação de biofilme e no processo de desenvolvimento da doença do mosqueado vermelho em *Sorghum bicolor* (Tomazini, 2022; Tuleski *et al.*, 2018).

Outros fatores presentes em *H. rubrisubalbicans* M1 cooperam para que a espécie seja classificada como um patógeno, como é o caso da presença de sistema de secreção tipo III, que atua na secreção de toxinas (Schmidt *et al.*, 2012; Tomazini, 2022), a diferenciação dos lipopolissacarídeos, associados ao processo de colonização de raízes (Balsanelli *et al.*, 2010) e a diminuição no tamanho do *cluster II* responsável pelo sistema de secreção tipo VI, homólogo a *Burkholderia spp.*, que pode estar relacionado a interação bactéria-planta (Tomazini, 2022). Apesar disso, todas as cepas de *H. rubrisubalbicans* possuem genes amplamente reconhecidos como benéficos às plantas, como o *cluster nif*, responsável pela fixação de nitrogênio atmosférico (Pedrolo *et al.*, 2022).

A fim de analisar o desenvolvimento do mosqueado clorótico vermelho ou *red stripe disease* no sorgo, Tuleski e colaboradores (2019) inocularam a cepa *Herbaspirillum rubrisubalbicans* M1 no caule de sorgo BTx623 com uma agulha hipodérmica, a aproximadamente 1 cm acima do solo. Dessa forma foi possível observar os sintomas da

doença (Figura 1), como as áreas necróticas, sinalizadas pelas setas e a listra vermelha ao longo da folha identificada pelo asterisco.

Figura 1. Sintomas de mosqueado clorótico vermelho em folha de sorgo.



Legenda. A sinalização de 4 pontos na folha indica diferentes sintomas ao longo da planta: em 1 tem-se o ponto sintomático inicial, 2 apresenta um ponto necrótico, 3 contém uma listra vermelha e 4 é uma área que não apresenta sintomas, entretanto, está a 1 cm acima da última área de sintomas.

Poucos estudos relacionados à população de *H. rubrisubalbicans* foram desenvolvidos, evidência disto é a presença de informações genômicas e transcriptômicas disponíveis apenas relacionadas ao isolado *H. rubrisubalbicans* M1 como exposto anteriormente. De igual modo, a disponibilidade de somente nove genomas sequenciados no banco de dados públicos Genbank (NCBI, 2024), indicam a escassez de dados moleculares associados à espécie.

2. JUSTIFICATIVA

Herbaspirillum rubrisubalbicans é um fitopatógeno endofítico diazotrófico, capaz de realizar a fixação de nitrogênio e de causar doença em plantas; dois fenótipos considerados excludentes no estudo das interações bactéria-planta. Além disso, os estudos apontam a conservação dos genes *nif* em todos os genomas da espécie, o que é intrigante devido à baixa prevalência deste *cluster* gênico em outras espécies fixadoras de nitrogênio. A escassez de

estudos genômicos detalhados de *H. rubrisubalbicans* impossibilita o entendimento dessa dualidade em relação à ecologia da população.

Para melhor compreender a ecologia desta espécie, o presente estudo propõe o uso de ferramentas computacionais de análise genômica para a realização de análises que possibilitem estudar a conservação e a distribuição de genes associados à fitopatogênese e a efeitos benéficos em plantas nos nove genomas públicos de *H. rubrisubalbicans* disponíveis no NCBI.

3. OBJETIVO GERAL

Caracterizar a ecologia e o modo de vida de *H. rubrisubalbicans* por meio de análises genômicas, com ênfase na dualidade diazotrófico *versus* fitopatogênico.

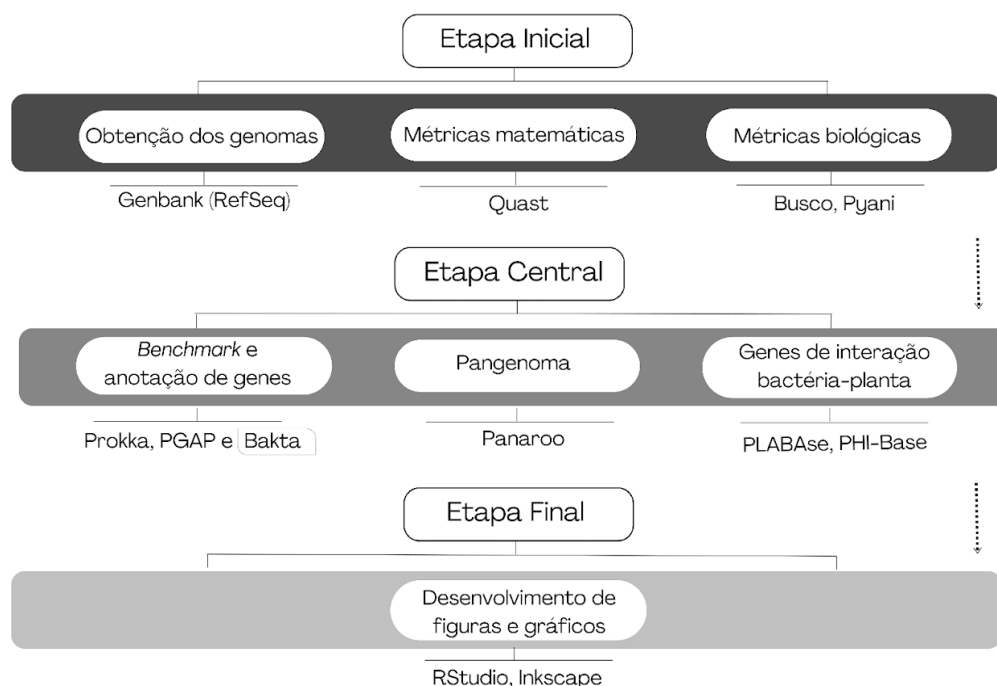
3.1 Objetivos específicos

- Analisar a qualidade dos genomas disponíveis de *H. rubrisubalbicans*, predizer seus genes e estudar a genômica da espécie;
- Quantificar genes associados à promoção do crescimento vegetal em *H. rubrisubalbicans*;
- Quantificar genes associados à fitopatogenia em *H. rubrisubalbicans*;
- Construir o pangenoma de *H. rubrisubalbicans*, correlacionar os genes de relação benéfica e fitopatogênica com o pangenoma;

4. MATERIAIS E MÉTODOS

As análises foram realizadas em ambiente Unix, dotado de processamento compartimentalizado via *secure shell* (ssh). Os algoritmos de análise ou conversão de *outputs* foram escritos em python 3 ou R. Todos os *softwares* utilizados são gratuitos e houve a preferência aos algoritmos já consolidados. Para *download* e armazenamento dos *softwares* utilizou-se o administrador de pacotes Miniconda versão 24.4.0. A metodologia, bem como os principais *softwares* utilizados para a realização do presente trabalho está disposta na Figura 2 e será detalhada nesta seção.

Figura 2. Fluxograma geral da metodologia aplicada.



Legenda. O projeto consistiu de três etapas principais. Etapa inicial: destinou-se a averiguar a qualidade matemática e biológica dos dados. Etapa central: buscou-se elucidar a significância biológica dos dados através do conhecimento dos genes, sua distribuição no pangenoma, assim como identificação e quantificação dos genes associados às plantas. Etapa final: projetada para processamento de dados para representação gráfica.

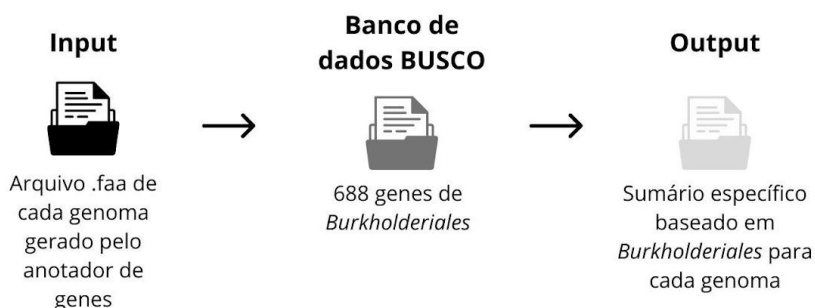
4.1 Obtenção dos genomas e controle de qualidade

A obtenção dos *assemblies* deu-se por meio da ferramenta *datasets* (NCBI, 2020). Durante o download dos genomas, utilizou-se a *flag* para baixar apenas os genomas depositados no banco de dados Refseq do National Center for Biotechnology International (Wenjun *et. al*, 2021). Aferiu-se o tamanho total e o conteúdo GC de todos os genomas em busca de identificar possíveis discrepâncias entre o conjunto de dados, enquanto as variáveis N50, L50 e número de contigs foram analisadas apenas nos cinco genomas disponíveis como *drafts*, a fim de analisar a fragmentação.

Para compor os padrões de qualidade biológica do conjunto de dados utilizou-se o grau de completude de genes ortólogos de cópia única nos genomas, aferido por meio do *software* BUSCO versão 5.2.2 (Manni *et al.*, 2021). Os genes ortólogos são homólogos que evoluíram a partir do mesmo gene em um ancestral comum (Pierce, 2016). O *software* BUSCO avaliou a presença de 688 destes genes compartilhados pela ordem *Burkholderiales* nos genomas de *H. rubrisubalbicans*. A escolha da ordem *Burkholderiales* se deu em razão desta ser o nível taxonômico mais baixo disponível em relação ao gênero *Herbaspirillum* dentre os bancos de dados do BUSCO. O conjunto dessas etapas pode ser visualizado na Figura 2. Para a realização da análise de genes ortólogos de cópia única, o arquivo de entrada

do BUSCO deve estar em formato fasta de aminoácidos (.faa), em decorrência dessa necessidade, as análises de qualidade biológica por meio de ortólogos de cópia única e a anotação de genes (detalhada na próxima sessão) foram realizadas de maneira concomitante.

Figura 3 - Fluxograma de análise de grau de completude de genes ortólogos de cópia única nos genomas.



Legenda: Fluxograma de análise de grau de completude de genes ortólogos de cópia única nos genomas.

A análise de identidade média dos nucleotídeos (ANI), realizada por meio da ferramenta Pyani (Pritchard *et al.*, 2015), foi utilizada para confirmar se todos os genomas são correspondentes a espécie *Herbaspirillum rubrisubalbicans*, evitando dados contaminados em posteriores análises, portanto, também para aferir a qualidade biológica dos *assemblies*. O arquivo de entrada para a análise de ANI consiste nos próprios genomas em formato fasta, contendo as sequências de nucleotídeos.

4.2 Anotação de genes

Para aferir o conteúdo gênico total codificado pelos genomas de *H. rubrisubalbicans*, realizou-se o *benchmark* entre os anotadores Prokka (Seemann, 2014), NCBI-PGAP (Tatusova *et al.*, 2016) e Bakta (Schewengers *et al.*, 2021) considerando sequências codificadoras (CDS), proteínas hipotéticas, tRNAs e tempo para a predição gênica de um genoma. A escolha dos três anotadores deu-se por razões distintas: o anotador Prokka é caracterizado pela sua consolidada atuação e difusão na área da bioinformática; o *software* PGAP foi desenvolvido para uso próprio do NCBI e, posteriormente, disponibilizado para a comunidade (Tatusova *et al.*, 2016), conferindo confiabilidade a ferramenta. Por fim, o *software* Bakta apresenta-se como uma ferramenta mais atual, com a proposta de anotação em tempo reduzido sem a utilização de alinhamento (Schewengers *et al.*, 2021). Para o desenvolvimento da análise, foram escolhidos apenas os 4 genomas completos, com o

objetivo de otimizar o tempo e desenvolver comparações mais equivalentes.

Após avaliações de tempo e qualidade a ferramenta Bakta se mostrou a melhor opção para este estudo, uma vez que combinou um bom desempenho em termos de quantidade de proteínas hipotéticas, tRNAs e CDS com um tempo de execução significativamente menor. Portanto, a predição de genes ocorreu por meio do *software* Bakta versão 1.9.3, através do banco de dados na versão *light*. Foi realizado um *script* em *loop* para que todos os genomas fossem analisados por meio de um único comando.

4.3 Pangenoma de *H. rubrisubalbicans*

Utilizou-se o *software* Panaroo (Tonkin-Hill, 2020) para desenvolvimento do pangenoma, permitindo estudos detalhados dos genomas central, acessório e único, assim como a inferência na distribuição dos fenótipos de interesse na espécie. Para esta análise, foram fornecidos todos os arquivos da anotação de genes no formato .gff3 provenientes do *output* do anotador de genes.

A análise permitiu o estudo do grupo *hard core* (genoma central), ou seja, aqueles presentes em todas as nove cepas analisadas, a partir da configuração a opção -a core. Para otimizar o tempo de processamento, foram alocados 10 *threads*, conforme indicado pela opção -t na linha de comando. O modo de limpeza dos dados foi definido como "*strict*", garantindo critérios rigorosos para a inclusão de genes na análise. Os genes foram clusterizados em famílias gênicas a partir de *threshold* de 95% de identidade.

4.4 Identificação de genes benéficos às plantas e de genes fitopatogênicos

A identificação dos genes associados aos fenótipos PGPR e fitopatogênico foi realizada por meio de alinhamentos locais entre as proteínas de *H. rubrisubalbicans* e dois bancos de dados. Para a identificação dos genes associados ao fenótipo PGPR, foi utilizado o banco de dados PGPT-Pred do PLaBase (Patz *et al.*, 2021). Através das categorias disponibilizadas pelo PLaBase, os genes encontrados foram categorizados em três grupos: efeitos diretos, indiretos e putativos. Para a identificação dos genes associados à fitopatogenicidade foi realizada por meio do banco de dados PHI-base (Urban *et al.*, 2022). Em ambos os casos, o blastp foi realizado pelo software Diamond (Buchfink *et al.*, 2015).

5. RESULTADOS

5.1 Qualidade dos *assemblies* e *Benchmark* de anotadores

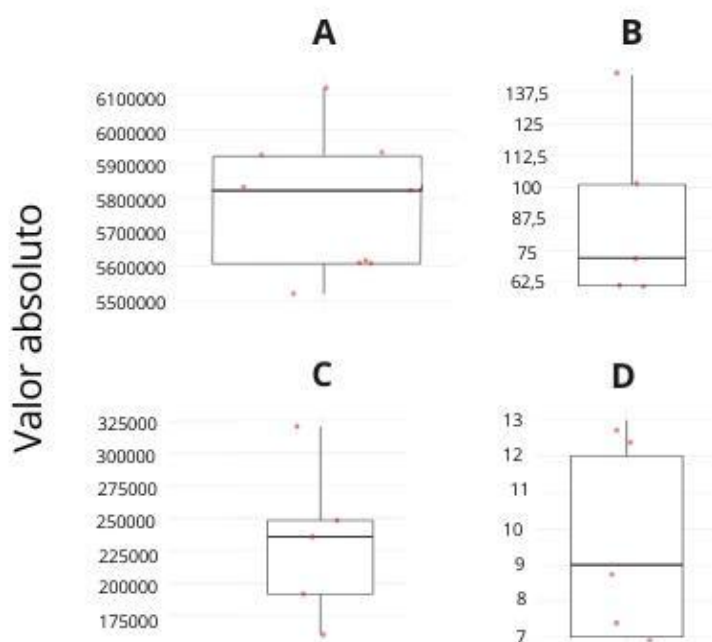
A partir dos dados disponíveis no banco de dados públicos do NCBI, construiu-se o conjunto de dados do presente estudo (Tabela 1), sendo este composto por nove genomas de *H. rubrisubalbicans*, dos quais cinco estão montados em *contigs* caracterizando-os como *drafts*, e quatro estão dispostos em apenas um *contig*, classificando-os como completos.

Tabela 1. Genomas de *H. rubrisubalbicans* disponíveis no Genbank.

Cepa	Nível de montagem	Contigs	Sequenciamento	Local de isolamento
M1	Completo	1	454, Illumina	Ilhas Maurício
DSM	Completo	1	PacBio	Estados Unidos
Os34	Completo	1	PacBio	China
SLBN7	Completo	1	PacBio	Estados Unidos
NBRC	<i>Drafts</i>	101	Illumina HiSeq	Estados Unidos
Os49	<i>Drafts</i>	61	Illumina HiSeq	China
Os44	<i>Drafts</i>	61	Illumina HiSeq	China
Os38	<i>Drafts</i>	72	Illumina HiSeq	China
Os45	<i>Drafts</i>	144	Illumina HiSeq	China

Legenda: Caracterização inicial dos genomas de *H. rubrisubalbicans* disponíveis pelo NCBI.

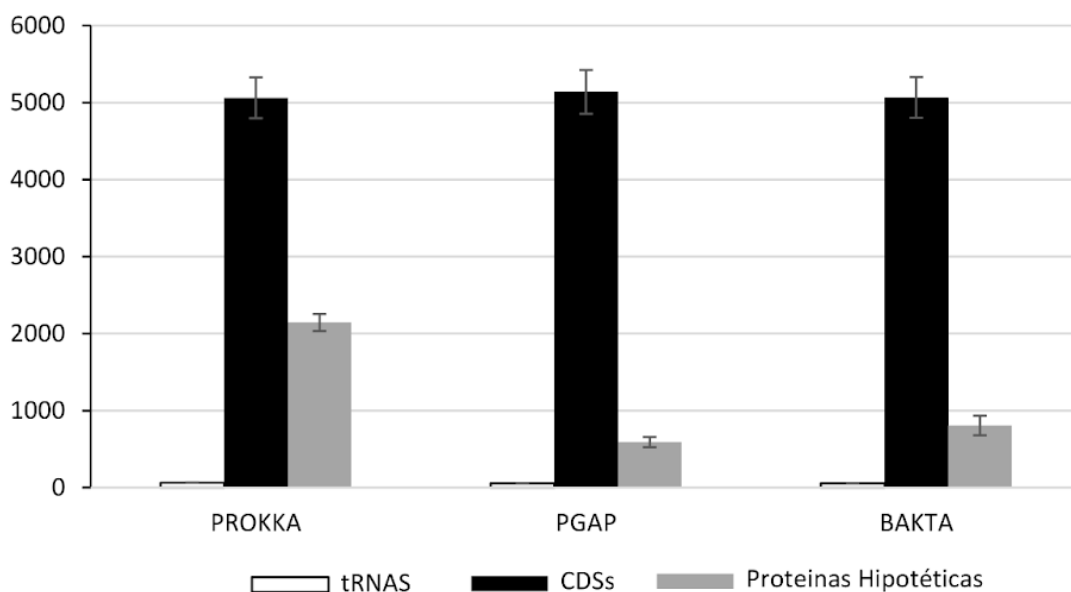
A partir da análise das métricas matemáticas por meio da ferramenta Quast, obteve-se que o maior genoma possui 6.122.717 pb (Os34) e o menor, possui 5.520.192 pb (NBRC), portanto, uma amplitude de 602.525 pb. A média do tamanho dos genomas foi de 5.777.246,67 pb $\pm$ 199.422,0672 (Figura 3A). Dentre os *drafts*, o *assembly* de *H. rubrisubalbicans* Os45 é o mais fragmentado do *dataset* com 144 contigs, N50 de 248.240 pb e L50 de 7 contigs (Figuras 4B, C e D). Todos os *assemblies* apresentaram conteúdo GC de 61%, com variação apenas em escala decimal.

Figura 4 - Controle de Qualidade - Métricas Matemáticas.

Legenda: Boxplot realizado a partir do controle de qualidade matemática por meio do *software* Quast. Tem-se em (A) o tamanho dos nove genomas em pares de base, (B) o número de *contigs* dos *drafts* do *dataset*, (C) o N50 dos *drafts* do *dataset* e (D) o N50 dos *drafts* do *dataset*.

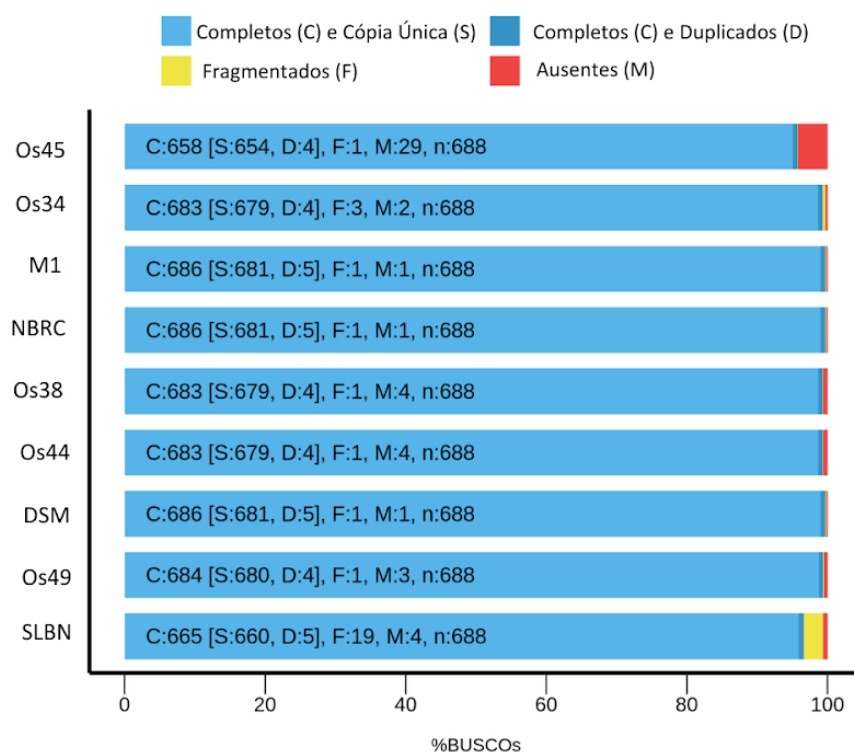
A partir do *benchmark* desenvolvido em busca do melhor anotador, constatou-se que o *software* Prokka demonstrou um desempenho abaixo do nível dos demais anotadores, visto que apresentou o maior número de proteínas hipotéticas (2144), enquanto Bakta (590) e PGAP (806) apresentaram valores significativamente menores. No fator quantidade de tRNAs anotados, as ferramentas Prokka, Bakta e PGAP apresentaram desempenho similar, revelando a existência de 62, 59 e 58 tRNAs, respectivamente (Figura 4). O parâmetro quantidade de sequências codificantes (CDS) também se apresentou com resultados próximos entre os três anotadores, sendo em ordem crescente: Prokka (5062), Bakta (5067) e PGAP (5139). A partir dos resultados analisados até então, o *software* Prokka foi desclassificado em decorrência do seu alto número de proteínas hipotéticas que poderiam prejudicar análises posteriores. Dessa forma, a variável tempo mostrou-se decisiva para a escolha do anotador, uma vez que o *software* PGAP levou cerca de 100 minutos para finalizar a anotação de um genoma, enquanto a ferramenta Bakta necessitou de apenas sete minutos para completar a anotação de cada um dos nove genomas presentes no conjunto de dados.

Figura 5 - Resultado do *benchmark* entre os anotadores.



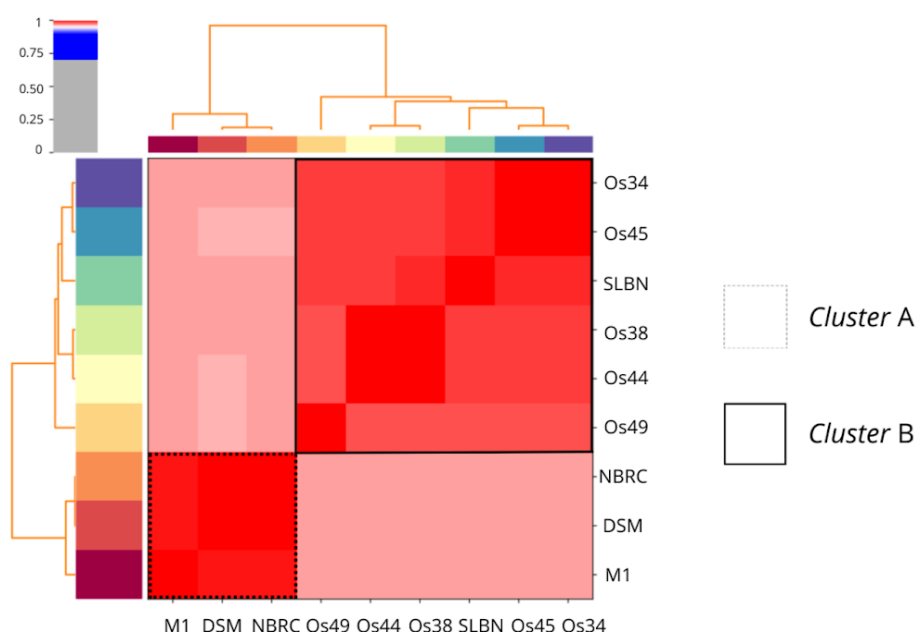
Legenda: *Benchmark* considerando os valores absolutos de tRNAs, CDSs e proteínas hipotéticas encontrados por cada um dos anotadores no conjunto de dados de 4 genomas completos de *H. rubrisubalbicans*.

Por meio da análise do *software* BUSCO, constatou-se que as cepas *H. rubrisubalbicans* M1, NBRC e DSM apresentaram a maior porcentagem de completude de ortólogos de cópia única do conjunto de dados (99,7%), indicando a presença de 688 genes compartilhados pela ordem *Burkholderiales*. O genoma *H. rubrisubalbicans* Os45, detentor da maior fragmentação no *dataset*, apresentou a menor porcentagem de presença de ortólogos completos (95,7%) com 658 ortólogos dos 688 genes totais encontrados no banco de dados (Figura 6). A análise de ortólogos de cópia única demonstrou uma média de 98,74% de presença de ortólogos completos, com desvio padrão de 0,014.

Figura 6 - Controle de Qualidade - Genes Ortólogos de Cópia Única.

Legenda: Análise gerada pelo software BUSCO baseada em 688 genes ortólogos da ordem *Burkholderiales*.

A análise da porcentagem de identidade de nucleotídeos revelou que todos os *assemblies* de *H. rubrisubalbicans* possuem identidade superior a 95%. Além desse principal aspecto demonstrado, é possível identificar a clusterização dos genomas em dois grupos distintos (Figura 7): o cluster A contendo as cepas *H. rubrisubalbicans* M1, *H. rubrisubalbicans* DSM e *H. rubrisubalbicans* NBRC e o grupo B, integrando as cepas *H. rubrisubalbicans* Os49, *H. rubrisubalbicans* Os44, *H. rubrisubalbicans* Os38, *H. rubrisubalbicans* SLBN, *H. rubrisubalbicans* Os 45 e *H. rubrisubalbicans* Os34.

Figura 7 - Controle de Qualidade - Análise de ANI.

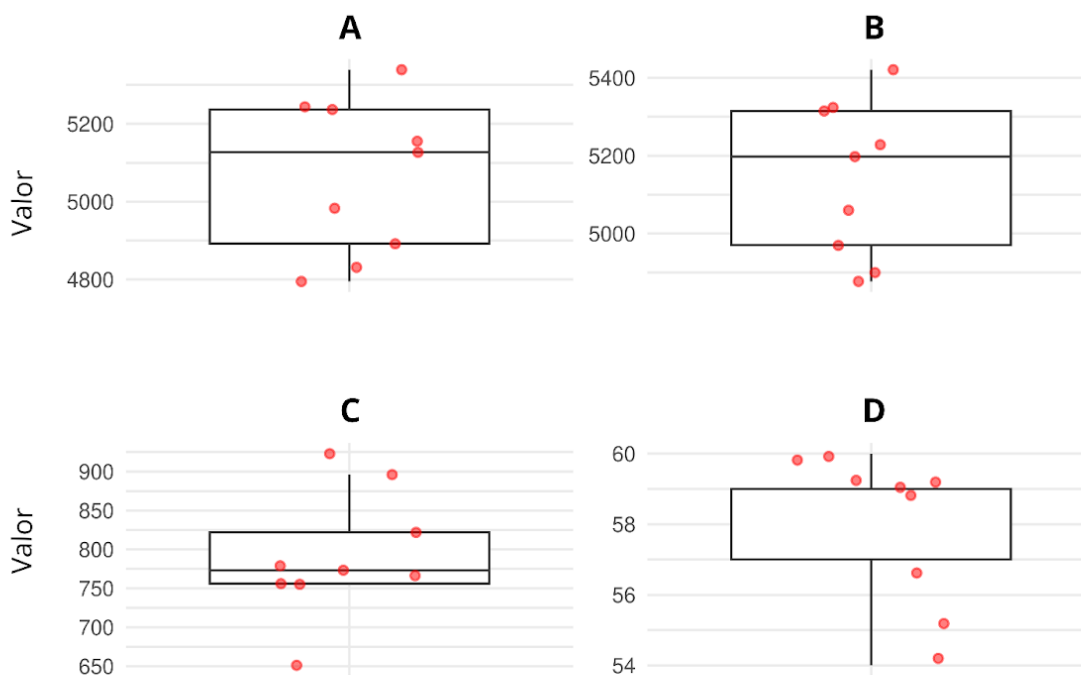
Legenda: Identidade média de nucleotídeos dos nove genomas de *H. rubrisubalbicans* através do Pyani.

5.2 Anotação de genes

A anotação dos genes (Figura 8) evidenciou que os genomas apresentam, aproximadamente 5.200 genes por genoma. A cepa *H. rubrisubalbicans* Os34 apresentou o maior número de genes (5.421), enquanto a linhagem *H. rubrisubalbicans* M1 apresentou o menor (4.877). Além dos genes, a anotação também identificou a quantidade de CDS, proteínas hipotéticas e tRNAs.

A variável tRNAs manteve-se, assim como no processo de *benchmark*, como o fator com menor amplitude de valor, sendo a maior quantidade encontrada nas cepas Os45 e Os38 (60) e a menor em *H. rubrisubalbicans* Os34 (54).

Figura 8 - Anotação de genes.



Legenda: Boxplot realizado a partir da anotação de genes gerada pelo Bakta. Tem-se (A) a quantidade de CDS, (B) genes, (C) proteínas hipotéticas e (D) tRNAs do *dataset*.

5.3 Pangenoma de *H. rubrisubalbicans*

O pangenoma de *H. rubrisubalbicans* revelou a presença de 7881 genes em todo o conjunto de dados, destes, 3580 compõem o *hard core*, portanto estão presentes em 99% ou mais dos genomas. O grupo *shell* é composto por 2687 genes e o grupo *cloud*, por 1514 genes.

5.4 Identificação de genes PGPR e quantificação de genes fitopatogênicos

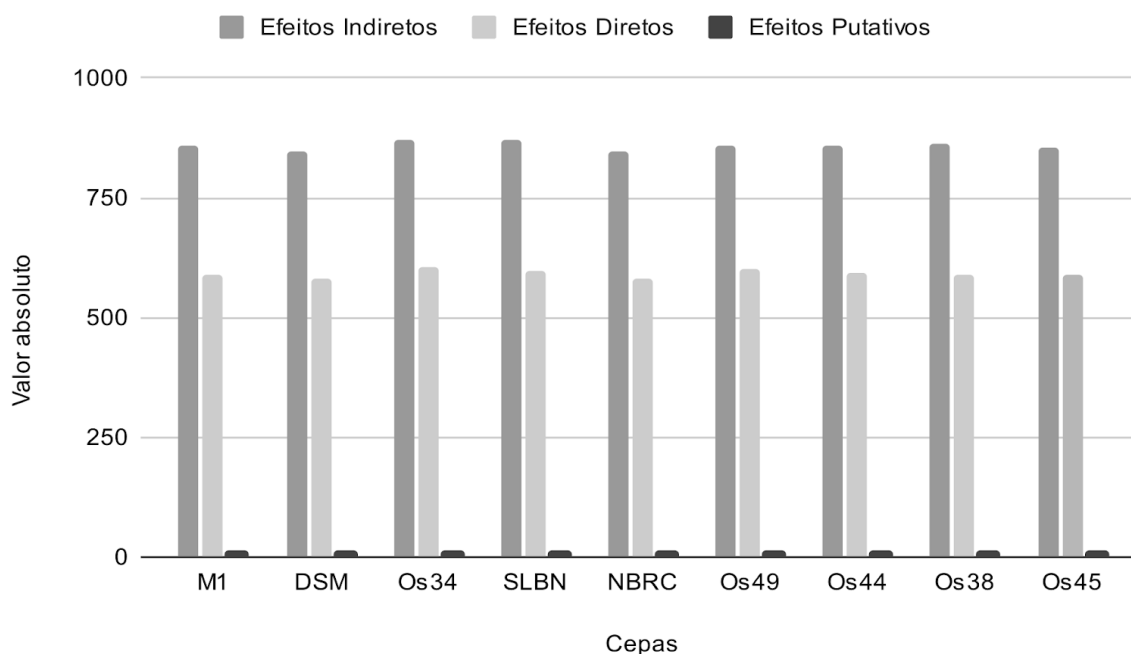
A análise *in silico* de genes associados às plantas a partir do banco de dados PGPT-Pred do PLaBase, nomeados como PGPT's, identificou a presença de 1572 genes em todo o conjunto de dados do estudo (Tabela 2). A categorização funcional geral destes genes, baseada na classificação do PLaBase, revelou a presença de 638 genes envolvidos em efeitos diretos, 922 indiretos, além de 12 genes com funções putativas.

Tabela 2. Identificação de genes relacionados à interação bactéria-plantas em *H. rubrisubalbicans*.

Categorias	Total	M1	DSM	Os34	SLBN	NBRC	Os49	Os44	Os38	Os45
Efeitos Indiretos	922	858	847	871	870	848	860	857	863	854
Efeitos Diretos	638	590	583	606	597	582	601	593	591	589
Efeitos Putativos	12	11	11	11	11	11	10	10	11	10
Total	1572	1459	1441	1488	1478	1441	1471	1460	1465	1453

Legenda: As categorias foram consideradas de acordo com a classificação disponível pelo PLaBase.

Apesar dos genomas se apresentarem em diferentes estados de qualidade, não houve diferenças significativas entre os genomas quanto ao número de genes pertencentes às três categorias encontrados (Figura 9).

Figura 9. Genes PGPT pertencentes às três categorias disponíveis no PLaBase.

Legenda: No eixo x, as cepas analisadas constituem as nove cepas de *H. rubrisubalbicans* e no eixo y tem-se o valor absoluto de genes.

A análise proveniente do PHI-Base indicou a presença de 1594 *hits*, ou seja, ocorrências de genes vinculados a espécies patogênicas presentes no conjunto de dados

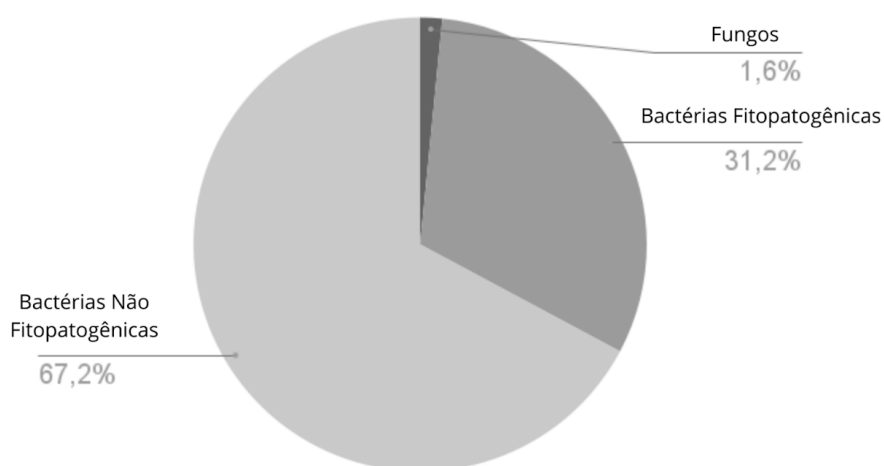
estudado. A distribuição das ocorrências ao longo de cada genoma pode ser visualizada na Tabela 3.

Tabela 3. Número de hits de genes de *H. rubrisubalbicans* contra o PHI-base.

PHIBase	Total	M1	DSM	Os34	SLBN	NBRC	Os49	Os44	Os38	Os45
<i>Hits</i>	1594	179	175	178	190	174	177	175	178	168

A partir dos hits gerados pelo PHI-Base, identificou-se a presença de genes provenientes de 40 espécies patogênicas distintas. De acordo com a ocorrência e o vínculo do gene a espécie patogênica, classificou-se a ocorrência dos hits em três grupos: bactérias fitopatogênicas, bactérias não fitopatogênicas e fungos fitopatogênicos (Figura 10).

Figura 10 – Distribuição dos *hits* de genes patogênicos.



Legenda: Distribuição das porcentagens das ocorrências de genes patogênicos encontrados na população de *H. rubrisubalbicans*.

Na categoria de bactérias fitopatogênicas que possuem genes presentes no conjunto de dados estudado, identificou-se a presença de 15 espécies distintas. As espécies encontradas, assim como a ocorrência de genes verificados pelo PHI-Base podem ser visualizados por meio da Tabela 4.

Tabela 4. Número de genes distintos por espécies fitopatogênicas em *H. rubrisubalbicans*.

Espécie	Ocorrência de genes
<i>Agrobacterium viti</i>	1
<i>Pectobacterium carotovorum</i>	1
<i>Xanthomonas albilineans</i>	1
<i>Xylella fastidiosa</i>	1
<i>Pseudomonas cichorii</i>	8
<i>Pantoea ananatis</i>	9
<i>Dickeya dadantii</i>	18
<i>Paracidovorax citrulli</i>	18
<i>Erwinia amylovora</i>	23
<i>Xanthomonas citri</i>	26
<i>Burkholderia glumae</i>	55
<i>Pseudomonas syringae</i>	58
<i>Xanthomonas campestris</i>	65
<i>Ralstonia solanacearum</i>	107
<i>Xanthomonas oryzae</i>	107

Legenda: Todas as bactérias presentes são descritas na literatura como fitopatogênicas.

6. DISCUSSÃO

O conjunto de dados de *H. rubrisubalbicans* do presente trabalho, composto por nove genomas, nos quais, cinco são *drafts*, demonstra que essa espécie ainda é pouco estudada pela comunidade científica. Trabalhos de genômica comparativa envolvendo bactérias patogênicas desenvolvem a investigação a partir de *datasets* maiores, como é o caso do estudo de Bertucceli e colaboradores (2023) que utilizaram 46 genomas de *Pseudomonas aeruginosa*. Além disso, outras bactérias fitopatogênicas possuem grande quantidade de genomas disponíveis em bancos de dados públicos, como por exemplo, *Xanthomonas campestris* e *Pseudomonas syringae* que possuem, respectivamente, 886 e 884 genomas disponíveis no

NCBI. Ao analisar um *dataset* pequeno é necessário cautela com as inferências visto que os dados podem não refletir a complexidade total da população, portanto, o pequeno número de genomas gera limitações ao estudo.

Os resultados das métricas matemáticas para controle de qualidade do conjunto de dados fornecidas pelo Quast são outro aspecto a ser considerado sobre o conjunto de dados. A fragmentação de 55% dos genomas, poderia indicar uma qualidade baixa para realização das próximas análises, no entanto, o N50 de 248.240 pb do genoma mais fragmentado (Os45), indica que até mesmo a maior fragmentação não possui um baixo valor de N50, característica que possibilita a inferência de uma boa qualidade do genoma, apesar do formato em *draft*.

A análise realizada a partir do BUSCO comprovou a qualidade biológica dos dados, em decorrência da alta média (98,74%) de genes ortólogos de cópia única encontrados em todos os genomas. Além disso, os genomas compartilham 95% ou mais de identidade de nucleotídeos, o que confirma que são da mesma espécie. Uma possível inferência sobre a formação de subgrupos na população de *H. rubrisubalbicans* pode ser vislumbrada a partir da análise de ANI que possibilitou a identificação de dois grupos mais similares entre si. A formação de dois grupos distintos, pode indicar um caminho para novas análises nesta espécie tão homogênea genomicamente.

A homogeneidade genômica fica evidente nos dados do pangenoma, que revelou que cerca de 3580 genes estão presentes em todos os genomas, fazendo parte do grupo de genes denominado *hard core*. Assumindo que os genomas de *H. rubrisubalbicans* possuem cerca de 5200 genes, 68% deste total é conteúdo comum a toda espécie, e apenas 32% do conteúdo genômico restante é acessório (*soft core* ou *shell*) ou único (*cloud*). Curiosamente, esse quadro repete algo descrito para outro fitopatógeno, a enterobactéria *Dickeya solani*, que apresenta 84% dos genes do pangenoma comuns a todos os 22 genomas estudados (Motyka-Pomagruk et al., 2020). Um core genomas grande em relação ao tamanho total do genoma é indicativo de alto grau de especialização ecológica de uma espécie, uma vez que a perda ou ganho de genes ocorre de forma aleatória na população, e a retenção destes genótipos decorre por ganho de fitness.

A presença de 1572 genes envolvidos em efeitos diretos, indiretos e putativos demonstrados pelo PlaBase comprova o fenótipo de associação de *H. rubrisubalbicans* às plantas, entretanto, a presença de 31,2% de genes de bactérias fitopatogênicas na população de *H. rubrisubalbicans* reforça a ideia de fitopatogenicidade da espécie. Dessa forma, há dificuldade em enquadrar esta bactéria em um fenótipo bem definido, o que enfatiza que são

necessários estudos comparativos mais aprofundados para determinar as diferenças entre os genomas e aspectos evolutivos destes genes.

7. CONCLUSÃO

Os resultados aqui apresentados permitem concluir que *H. rubrisubalbicans* é uma espécie de bactéria com alto teor de homogeneidade genômica, conforme revelado pelo seu pangenoma, que apresenta alto número de genes relacionados tanto ao fenótipo benéfico às plantas quanto ao fenótipo fitopatogênico. Esses genes se apresentam conservados, não permitindo inferir diferenças de estilo de vida intra-espécie.

REFERÊNCIAS BIBLIOGRÁFICAS

- ALONSO A, BARRIO P.A. et al. Current state-of-art of STR sequencing in forensic genetics. *Electrophoresis*, v. 39, n. 21, p.2655- 2668, 2018.
- ASHRAFI, S. et al. Two New Rhizobiales Species Isolated from Root Nodules of Common Sainfoin (*Onobrychis viciifolia*) Show Different Plant Colonization Strategies . *Microbiology Spectrum*, v. 10, n. 5, 2022.
- BALDANI, J. I. et al. Characterization of *Herbaspirillum seropedicae* gen. nov., sp. nov., a root-associated nitrogen-fixing bacterium. *International Journal of Systematic Bacteriology*, v. 36, n. 1, p. 86–93, 1986.
- BALDANI, J. I. et al. Emended Description of *Herbaspirillum*; Inclusion of [*Pseudomonas*] *rubrisubalbicans*, a Mild Plant Pathogen, as *Herbaspirillum rubrisubalbicans* comb. nov.; and Classification of a Group of Clinical Isolates (EF Group 1) as *Herbaspirillum* Species 3. *International Journal of Systematic Bacteriology*, v. 46, n. 3, p. 802–810, 1996.
- BALSANELLI, E. et al. *Herbaspirillum seropedicae* *rfbB* and *rfbC* genes are required for maize colonization. *Environmental Microbiology*, v. 12, n. 8, p. 2233–2244, 2010.
- BEIER, S., THOMSON, N.R. Panakeia - a universal tool for bacterial pangenome analysis. *BMC Genomics* v. 23, n. 265, 2022.
- BENSON, D. A. *et al.* GenBank. *Nucleic Acids Research*, v. 46, n. D1, p. D41–D47. 2018.
- BRANDA, S. S.; VIK, A. S.; FRIEDMAN, L. & KOLTER, R. Biofilms: the matrix revisited. *Trends in microbiology*. v. 13, n. 1, p. 20–26, 2005.
- BRYNILDSRUD, O. et al. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology*, v. 17, n. 1, p. 25, 2016.

BUCHFINK, B., XIE, C. & HUSON, D. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, v. 12, p. 59–60, 2015.

CAVICCHIOLI, R., RIPPLE, W.J., TIMMIS, K.N. et al. Scientists' warning to humanity: microorganisms and climate change. *Nat Rev Microbiol*, v. 17, p. 569–586, 2019.

CHARLES-ORSZAG, A *et al.* Adhesion pilus retraction powers twitching motility in the thermoacidophilic crenarchaeon *Sulfolobus acidocaldarius*. *Nature Commun* v. 15, n. 5051, 2024.

COLLINS, C. & DIDELOT, X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *Commun Biol*, v. 5, n. 266, 2022.

COLL, F., GOULIOURIS, T., BRUCHMANN, S. *et al.* PowerBacGWAS: a computational pipeline to perform power calculations for bacterial genome-wide association studies. *Commun Bio*, v.5, n. 1, 2022.

CRAIG, L., FOREST, K. T., & MAIER, B. Type IV pili: dynamics, biophysics and functional consequences. *Nature Reviews Microbiology*, v. 17, p. 429–440, 2019.

TOMAZINI, E. Determinação de genes envolvidos na interação de *Herbaspirillum rubrisubalbicans* com *Sorghum bicolor*. Tese—Curitiba: Universidade Federal do Paraná, 2022.

FARHAT, M. R. et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nature Genetics*. v. 45, n. 10, p. 1183–1189, 2013.

FITCH, W. Homology: a personal view on some of the problems. *Trands in genetics*, v. 16, n. 5, 2000.

GILLIS, M. et al. Taxonomic relationships between [*Pseudomonas*] *rubrisubalbicans*, some clinical isolates (EF group 1), *Herbaspirillum seropedicae* and [*Aquaspirillum*] *Autotrophicum*. Belgium: *Developments in Plant and Soil Sciences*, v. 48, 1991.

GREEN, S. (2010). Comparative genome analysis provides insights into the evolution and adaptation of *Pseudomonas syringae* pv. *aesculi* on *Aesculus hippocastanum*. *PLoS ONE*, v.5, n. 4, p. 1–14, 2010.

GUIMARÃES, A. Abordagens bioinformáticas no estudo pangenômico de *Xanthomonas campestris*. Dissertação - Universidade Federal de Pelotas, 2020.

GUREVICH, A. et al. QUASt: Quality assessment tool for genome assemblies. *Bioinformatics*, v. 29, n. 8, p. 1072–1075, 2013.

HALE, C. N., & PAULA WILKIE, J. A comparative study of *Pseudomonas* species pathogenic to sorghum. *New Zealand Journal of Agricultural Research*, v. 15, n. 3, p. 448–456, 1972

HARSHEY, R. M. Bees aren't the only ones: swarming in gram-negative bacteria. *Mol. Microbiol.* v.13, p. 389-94. 1994.

HOLTAPPELS, D., ABELSON, S. A., NOUTH, S. C., RICKUS, G. E. J., AMARE, S. Z., GILLER, J. P., JIAN, D. Z., & KOSKELLA, B. Genomic characterization of *Pseudomonas syringae* pv. *syringae* from Callery pear and the efficiency of associated phages in disease protection. *Microbiology Spectrum*, v. 12, n. 3, 2024.

JAMES, E. K.; OLIVARES, F. L. Infection and Colonization of Sugar Cane and Other Graminaceous Plants by Endophytic Diazotrophs. *Critical Reviews in Plant Sciences*, v. 17, n. 1, p. 77–119, 1998.

KIM, Y. et al. Current status of pan-genome analysis for pathogenic bacteria. *Current Opinion in Biotechnology* Elsevier Ltd, , 1 jun. 2020.

LEES, J. A., GALARDINI, M., Bentley, S. D., Weiser, J. N. & Corander, J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, v. 34, p. 4310–4312, 2018.

LEES, J. A. et al. Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions. *American Society for Microbiology*, n. 2161– 2129, 2020.

LEMANCEAU P, *et al.* Plant Communication With Associated Microbiota in the Spermosphere, Rhizosphere and Phyllosphere, Elsevier, v. 82, 2017.

MANNI, M. *et al.* BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols*, v. 1, n. 12, 1 dez. 2021.

MANSFIELD, J. *et al.* Top 10 plant pathogenic bacteria in molecular plant pathology. *Molecular Plant Pathology*, v. 13, n. 6, p. 614-629, 2012.

MATTEOLI, F. P. et al. *Herbaspirillum. Em: Beneficial Microbes in Agro-Ecology: Bacteria and Fungi.* [s.l.] Elsevier, 2020. p. 493–508.

MATTICK, J. S. Type IV pili and twitching motility. *Annu. Rev. Microbiol.* v. 56, p. 289-314. 2002.

MISHRA, A.; FORCHE, A.; ANDERSON, M. Z. Parasexuality of *Candida* Species. *Frontiers in cellular and infection microbiology*, v. 11, 2021.

MONTEIL, C. L. Features of air masses associated with the deposition of *Pseudomonas syringae* and *Botrytis cinerea* by rain and snowfall. *ISME Journal*, v. 8, n.11, p. 2290–2304, 2014.

MONTEIRO, R. A., BALSANELLI, E.; TULESKI, T. R.; FAORO, H.; CRUZ, L. M.; WASSEM, R.; BAURA, V. A.; TADRA-SFEIR, M. Z.; WEISS, V.; DAROCHA, W. D.; MULLER-SANTOS, M.; CHUBATSU, L. S.; HUERGO, L. F.; PEDROSA, F. O., SOUZA, E. M. Genomic comparison of the endophyte *Herbaspirillum seropedicae* SmR1 and the phytopathogen *Herbaspirillum rubrisubalbicans* M1 by Suppressive Subtractive Hybridization and partial genome sequencing. *FEMS Microbiology Ecology*, v. 80, n. 2, p. 441-51, 2012.

MONTYKA-POMAGRUK, A., ZOLEDOWSK, S., MISZTAK, A.E. *et al.* Comparative genomics and pangenome-oriented studies reveal high homogeneity of the agronomically relevant enterobacterial plant pathogen *Dickeya solani*. *BMC Genomics*, v. 21, n. 44, 2020.

OCHMAN, H, LAWRENCE, J. G., GROISMAN, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature*, v. 405, n. 6784, p. 128-142, 2000.

OLIVARES, F. L.; BALDANI, V. L. D.; REIS, V. M.; BALDANI, J. I.; DOBEREINER, J. Occurrence of the endophytic diazotrophs *Herbaspirillum* spp in roots, stems, and leaves, predominantly of Gramineae *Biol. Fertil. Soils*, v. 21, n. 3, p. 197– 200, 1996.

PATTERSON, C. Homology in classical and molecular biology. *Molecular and Evolution*, v. 5, n. 6, p. 603-625, 1988.

PATZ, S. *et al.* PLaBAsE: A comprehensive web resource for analyzing the plant growth-promoting potential of plant-associated bacteria. [s.d.].

PEDROLO, A., MATEOLLI, F., SOARES, C., ARISI, A. Comparative Genomics Reveal the High Conservation and Scarce Distribution of Nitrogen Fixation *nif* Genes in the Plant-Associated Genus *Herbaspirillum*. *Microbial Ecology*, v 86, 2022.

PEDROSA, F. O. *et al.* Genome of *Herbaspirillum seropedicae* strain SmR1, a specialized diazotrophic endophyte of tropical grasses. *PLoS Genetics*, v. 7, n. 5, maio 2011.

PIDOT, S. J. *et al.* Increasing tolerance of hospital *Enterococcus faecium* to handwash alcohols. *Science translational medicine*, v. 10, p. 452, 2018.

PIERCE, Benjamin A. *Genética - Um Enfoque Conceitual*, 5ª edição. Rio de Janeiro: Guanabara Koogan, p.532, 2016.

PIMENTEL, J. P. *et al.* Dinitrogen fixation and infection of grass leaves by *Pseudomonas rubrisubalbicans* and *Herbaspirillum seropedicae*. *Plant Soil*, v. 137, p. 61–65, 1991.

POWER, R. A.; PARKHILL, J.; DE OLIVEIRA, T. Microbial genome-wide association studies: lessons from human GWAS. *Nature Reviews Genetics* Nature Publishing Group, v. 3, p. 41-50, 2017.

PRIOR, P., *et al.* Genomic and proteomic evidence supporting the division of the plant pathogen *Ralstonia solanacearum* into three species. *BMC Genomics*, v.17, n. 1, p 1–11, 2016.

QUIÑONES-VALDEZ, R. *et al.* Spatial pattern of mottled stripe disease. *APS Publications*, v. 94, n. 0191–2917, p. 126–134, 2022.

RAMIREZ, N.A., DAS A., TON-THAT H. New paradigms of pilus assembly mechanisms in gram-positive actinobacteria. *Trends Microbiol.* v. 28, n. 12, p. 999–1009, 2020.

RAPPUOLI, R., YOUNG, P., RON, E. *et al.* Save the microbes to save the planet. A call to action of the International Union of the Microbiological Societies (IUMS). *One Health Outlook* v. 5, n. 5, 2023.

RICHARDSON, E. J. *et al.* Gene exchange drives the ecological success of a multi host bacterial pathogen. *National Ecology & Evolution*, v. 2, p. 1468–1478, 2018.

SANTOYO, G. How plants recruit their microbiome? New insights into beneficial interactions, *Journal of Advanced Research*, v. 40, p. 45-58, 2022.

ROTHBALLER, M. *et al.* Endophytic root colonization of gramineous plants by *Herbaspirillum frisingense*. *FEMS Microbiology Ecology*, v. 66, n. 1, p. 85–95, 2008.

SANTOS, T. M. Explorando a diversidade genética e taxonômica no grupo extended Asfarviridae: Um estudo de genômica comparativa. Tese - Universidade Federal de Minas Gerais, 2024.

SAUER, K.; STOODLEY, P.; GOERES, D. M.; HALL-STOODLEY, L.; BURMØLLE, M.; STEWART, P. S.; BJARNSHOLT, T. The biofilm life cycle: expanding the conceptual model of biofilm formation. *Nature Reviews Microbiology*, v. 20, p. 608–620, 2022.

SCHMIDT, M. A. *et al.* The type III secretion system is necessary for the development of a pathogenic and endophytic interaction between *Herbaspirillum rubrisubalbicans* and Poaceae. *BMC Microbiology*, v. 12, 2012.

SCHWENGERS, O., JELONEK, L., DIECKMANN, M. A., BEYVERS, S., BLOM, J., & GOESMANN, A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial genomics*, v. 7, n. 11, 2021.

SHEPPARD, S. & DIDELOT, X. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Biological Science*, v. 110, n. 29, p. 11923–11927, 2013.

SUWANTARAT, N. et al. Fatal case of *Herbaspirillum seropedicae* bacteremia secondary to pneumonia in an end-stage renal disease patient with multiple myeloma. *Diagnostic Microbiology and Infectious Disease*, v. 82, n. 4, p. 331–333, 2015.

TATUSOVA, T. et al. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, v. 44, n. 14, p. 6614–6624, 2016.

TETTELIN, H., et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial “pan-genome”. *Proceedings of the National Academy of Sciences*, v. 102, n. 39, 2005.

TETTELIN, H., RILEY, D., CATATTUTO, C., & MEDINI, D. Comparative genomics: The bacterial pan-genome. *Current Opinion in Microbiology*, v. 11, n. 5, p. 472–477, 2008.

TIMMUSK S., GRANTCHAROVA N, WAGNER, E. G. H. *Paenibacillus polymyxa* Invades Plant Roots and Forms Biofilms. *Appl Environ Microbiol*, v. 71, n. 11, p. 7292 - 7300, 2005.

TOLKIN-HILL, G., MACALASDAIR, N., RUIS, C. et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol*, v. 21, n. 180, 2020.

TORSTEN, S. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, v. 30, n. 14, p. 2068–2069, 2014.

TRIVEDI, P., LEACH, J.E., TRINGE, S.G. et al. Plant–microbiome interactions: from community assembly to plant health. *Nat Rev Microbiol*, vol. 18, p. 607–621, 2020.

TULESKI, T. R. et al. *Herbaspirillum rubrisubalbicans* as a phytopathogenic model to study the immune system of *Sorghum bicolor*. *Molecular Plant-Microbe Interactions*, v. 33, n. 2, p. 235–246, 2019a.

TULESKI, T. R. et al. Cellulose production increases sorghum colonization and the pathogenic potential of *Herbaspirillum rubrisubalbicans* M1. *Scientific Reports*, v. 9, n. 1, 1 dez. 2019b.

URBAN, M. et al. PHI-base: The pathogen-host interactions database. *Nucleic Acids Research*, v. 48, n. D1, p. D613–D620, 2020.

WENJUN, L. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Research*, v. 49, n. D1, p. D1020–D1028, 2021.

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

ZHOU X., LI Y., PENG X., REN B., XIAO L., LI M., GUO Q. From Healthy Microflora to Disease. *Atlas of Oral Microbiology: Basic Biology of Oral Microbes*; pp. 1–24, 2021