UNIVERSIDADE FEDERAL DO PARANÁ

RODRIGO LUIS ALVES CARDOSO

ANÁLISE GENÔMICA COMPARATIVA DE BACTÉRIAS DO GÊNERO *Herbaspirillum*

CURITIBA

2015

RODRIGO LUIS ALVES CARDOSO

ANÁLISE GENÔMICA COMPARATIVA DE BACTÉRIAS DO GÊNERO *Herbaspirillum*

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciências-Bioquímica, Setor de Ciências Biológicas, da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de Doutor em Ciências-Bioquímica.

Orientador:
Leonardo Magalhães Cruz, Dr.

CURITIBA

2015

TERMO DE APROVAÇÃO

RODRIGO LUIS ALVES CARDOSO

Análise genômica comparativa de bactérias do gênero *Herbaspirillum*

Tese aprovada como requisito parcial para obtenção do grau de Doutor no curso de Pós-Graduação em Ciências-Bioquímica, Setor de Ciências Biológicas, Universidade Federal do Paraná, pela seguinte banca examinadora:

Prof. Dr. Leonardo Magalhães Cruz – Orientador
Departamento de Bioquímica, UFPR

Prof.ª Dr.ª Rose Adele Monteiro
Departamento de Bioquímica, UFPR

Prof.ª Dr.ª Cyntia Maria Telles F. Picheth
Departamento de Farmácia, UFPR

Dr. Vinícius Almir Weiss
SEPT - UFPR

Dr. Helisson Faoro
Instituto Carlos Chagas, FIOCRUZ-PR

Curitiba, 30 de julho de 2015.

Dedico esse trabalho à minha família por todo
o apoio ao longo desses anos de estudo,
por todos os valores que me foram ensinados
e sem a qual nada eu teria conseguido.

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Leonardo Magalhães Cruz, pela dedicação e ensinamentos ao longo desses anos.

À Prof[a]. Dr[a]. Rose Adele Monteiro, pela dedicação e auxílio nos assuntos de extrema relevância para a execução deste trabalho.

À Prof[a]. Dr[a]. Leda Satie Chubatsu, que sempre foi prestativa e atenciosa.

Ao Prof. Dr. Fábio de Oliveira Pedrosa, pela enorme contribuição neste trabalho, e pela oportunidade de trabalhar em conjunto com o Núcleo de Fixação Biológica de Nitrogênio da UFPR.

Ao Prof. Dr. Emanuel Maltempi de Souza, pelos ensinamentos e pela contribuição neste trabalho através de sugestões e explicações.

Ao Prof. Dr. Roberto Tadeu Raittz, pelas sugestões e idéias a respeito dos problemas que eventualmente surgiram neste trabalho.

Aos demais professores do programa de Pós-Graduação em Ciências-Bioquímica, por toda a atenção.

Ao Prof. Dieval, ao Dr. Helisson, ao Dr. Vinícius, à Dr[a] Giovana, por toda ajuda, pelas informações sempre relevantes, pela prestação de auxílio, e por toda a contribuição neste trabalho.

Ao colega de laboratório Dr. Daniel e ao ex-colega de laboratório Leandro, por terem tornado o ambiente de trabalho sempre alegre, e ao colega de laboratório Robson por ter realizado os testes bioquímicos com as bactérias deste trabalho.

À Dr[a] Michelle, ao Valter e às demais pessoas envolvidas com sequenciamento genômico.

A todo o grupo de Fixação Biológica de Nitrogênio da UFPR, pela participação neste trabalho.

Aos demais colegas de departamento, pela troca de experiências.

A todos os membros da coordenação do curso de Pós-Graduação em Ciências-Bioquímica, pela atenção.

À minha mãe, Catarina, e aos meus irmãos, por todo o apoio.

Ao ex-colega de mestrado Eduardo Tieppo, pela amizade e por ter me ensinado a usar o Photoshop.

À minha amiga Judy, pela amizade e pelos momentos de descontração.

A todos os meus amigos e todas as pessoas que me auxiliaram de alguma forma.

Aos órgãos financiadores: CAPES, CNPq e REUNI.

A Deus, acima de tudo.

RESUMO

Estirpes de *Herbaspirillum* spp. foram isoladas de diversos tipos de ambientes e localizações geográficas, e apresentam grande diversidade metabólica. O gênero *Herbaspirillum* inclui organismos diazotróficos, endofíticos e espécies capazes de degradar compostos fenólicos ou até crescer autotroficamente. Neste trabalho foram sequenciados e anotados os genomas de três espécies de *Herbaspirillum*: *H. autotrophicum* IAM 14942 (bactéria oxidante de hidrogênio), *H. chlorophenolicum* CPW301 (bactéria capaz de degradar 4-clorofenol) e *H. rhizosphaerae* UMS-37 (bactéria isolada da rizosfera de *Allium victorialis*), que foram comparados com os genomas disponíveis das seguintes estirpes de *Herbaspirillum* spp.: *H. seropedicae* SmR1, BR11335, BR11417, AU14040, Os34 and Os45; *H. rubrisubalbicans* M1, BR11504; *H. huttiense* subsp. *putei* IAM 15032; *H. frisingense* GSF30; *H. lusitanum* P6-12; *H. hiltneri* N3; *H. massiliense* JC206; e *Herbaspirillum* sp. GW103, CF444 e YR522. A comparação genômica foi realizada com BLAST todos contra todos, a homologia entre proteínas foi determinada por 50% de cobertura de alinhamento e 50% de identidade. A média de genes dos 19 genomas analisados foi de 4.897, com cerca de 760 genes únicos por genoma (15%). *H. massiliense* parece ser a espécie mais distante das demais, enquanto *H. rubrisubalbicans*, *H. seropedicae*, *H. huttiense* e *H. frisingense* formam um grupo estritamente relacionado (filogrupo 1), e *H. hiltneri*, *H. lusitanum* e *H. rhizosphaerae* formam outro grupo relacionado (filogrupo 2). Essas relações foram confirmadas por análises filogenéticas que utilizaram informações do pangenoma. O *core* e o pangenoma do gênero consistem em aproximadamente 1.400 e 20.000 famílias de genes, respectivamente. A distribuição funcional dos genes do pangenoma, em comparação aos genes do *core*-genoma, mostrou maior número de genes envolvidos no metabolismo de carboidratos e íons inorgânicos, além de genes envolvidos com transcrição. O agrupamento de genes *nif* está presente somente em estirpes de *H. seropedicae* (com exceção da estirpe AU14040), *H. rubrisubalbicans* e *H. frisingense*. Já o agrupamento de genes do sistema de secreção do tipo III é amplamente distribuido dentro do gênero *Herbaspirillum*, e é proposto que esses genes estivessem presentes no ancestral do gênero. Análises taxonômicas baseadas na sequência genômica mostraram que as estirpes de *Herbaspirillum* GW103, CF444 e YR522 podem ser classificadas em novas espécies.

Palavras-chave: *Herbaspirillum*, Comparação Genômica, Pangenoma, *Core*-genoma.

ABSTRACT

Strains of *Herbaspirillum* spp. have been isolated from a multitude of environments and geographic locations, presenting a metabolically diverse genus. It includes diazotrophs, plant endophytes, and also species capable of degrading phenolic compounds or growing autotrophically. We have sequenced and annotated the genome of three species of *Herbaspirillum*, namely: *H. autotrophicum* IAM 14942 (hydrogen-oxidizing bacteria), *H. chlorophenolicum* CPW301 (4-chlorophenol degrading bacteria) and *H. rhizosphaerae* UMS-37 (isolated from rhizosphere of *Allium victorialis*). We compared these genomes with the following available genome sequences of *Herbaspirillum* spp. strains: *H. seropedicae* strains SmR1, BR11335, BR11417, AU14040, Os34 and Os45; *H. rubrisubalbicans* strains M1 and BR11504; *H. huttiense* subsp. *putei* IAM 15032; *H. frisingense* GSF30; *H. lusitanum* P6-12; *H. hiltneri* N3; *H. massiliense* JC206; and *Herbaspirillum* sp. strains GW103, CF444 and YR522. The genome comparison was performed with all-against-all BLAST, the protein homology settled as 50% of align coverage and 50% of protein identity. The average gene content of the 19 analyzed organisms was 4,897, and about 762 unique genes per genome (15%). *H. massiliense* JC206 seems to be the most distant related species, while *H. rubrisubalbicans*, *H. seropedicae* and *H. huttiense* compose a closely related group (phylogroup 1), and *H. hiltneri*, *H. lusitanum* and *H. rhizosphaerae* compose another closely group (phylogroup 2). These relationships were confirmed by phylogenetic analysis using pan-genome tree. The core- and pan-genome of this genus consist of approximately 1,400 and 20,000 groups of genes, respectively. Functional distribution of the pan-genome genes, in comparison to core-genome genes, showed a high number of genes involved in carbohydrate metabolism, inorganic ions metabolism and transcription. The *nif* gene cluster is present only in *H. seropedicae*, *H. rubrisubalbicans* and *H. frisingense* strains. Type III secretion system gene cluster is largely distributed among *Herbaspirillum* genus, and it is hypothesized that this cluster of genes was present in the common ancestor of all *Herbaspirillum* species. Taxonomic analysis, performed with genome sequence, suggest that *Herbaspirillum* spp. strains GW103, CF444 and YR522 are new *Herbaspirillum* species.

Keywords: Herbaspirillum, Genome comparison, Pan-genome, Core-genome.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

4CC - 4-clorocatecol

4CF - 4-clorofenol

5CHMS - 5-cloro-2-hidroximucônico

6-APA - ácido 6-aminopenicilânico

ABC - ATP-*binding cassete*

ACC - 1-aminociclopropano-1-carboxilato

ANI - *Average Nucleotide Identity*

ASiD - *Assembly Assistent for* SOLiD

BBH - *Bidirectional Best Hit*

BLAST - *Basic Local Alignment Search Tool*

BRIG - BLAST *Ring Image Generator*

CDS - *coding sequence* (sequência que codifica para uma proteína)

COG - *Cluster of Orthologous Groups*

DDH - DNA-DNA *hibridization*

EC - *Enzyme Commission*

EMBRAPA - Empresa Brasileira de Pesquisa Agropecuária

EPS - Exopolissacarídeo

GEM - *Genome-Scale metabolic model*

GBDP - *Genome Blast Distance Phylogeny*

GGDC - *Genome-to-Genome Distance Calculator*

GGDH - *Genome-to-Genome Distance Hybridization*

GLIMMER - *Gene Locator and Interpolated Markov Modeler*

HGT - transferência horizontal de genes (do inglês, *horizontal gene transfer*)

IAA - ácido indol-acético (do inglês, *indol acetic acid*)

IS - sequência de inserção (do inglês, *insertion sequence*)

KAAS - KEGG *Automatic annotation server*

KEGG - *Kyoto Encyclopedia of Genes and Genomes*

LPS - lipopolissacarídeo

Mb - Mega base

MEGA - *Molecular Evolutionary Genetics Analysis*

MLSA - *MultiLocus Sequence Analysis*

MUM - *Maximal Unique Match*

MUSCLE - *MUltiple Sequence Comparison by Log-Expectation*

NCBI - *National Center for Biotechnology Information*

NFN - Núcleo de Fixação Biológica de Nitrogênio

NR - não redundante

ORF - *open reading frame* (fase aberta de leitura, possível gene)

PAST - *PAleontological Statistics*

pb - par de base

PEP - fosfoenolpiruvato (do inglês, *phosphoenolpyruvate*)

PFK - fosfofrutoquinase (*phosphofructokinase*)

PGPR - rizobactérias promotoras do crescimento da planta (do inglês, *Plant Growth-Promoting Rhizobacteria*)

PHAST - *PHAge Search Tool*

PTS - sistema fosfotransferase (*phosphotransferase system*)

RAST - *Rapid Annotation using Subsystem Technology*

RPL - RuBisCO-*like protein*

RuBisCO - ribulose 1,5-bifosfato carboxilase/oxigenase

SAET - SOLiD *Accuracy Enhancement Tool*

SOLiD - *Sequencing by Oligonucleotide Ligation and Detection*

SSH - *Suppression Subtactive Hybridization*

T1SS - sistema de secreção do tipo I (do inglês, *type I secretion system*)

T2SS - sistema de secreção do tipo II (do inglês, *type II secretion system*)

T3SS - sistema de secreção do tipo III (do inglês, *type III secretion system*)

T4SS - sistema de secreção do tipo IV (do inglês, *type IV secretion system*)

T5SS - sistema de secreção do tipo V (do inglês, *type V secretion system*)

T6SS - sistema de secreção do tipo VI (do inglês, *type VI secretion system*)

Tat - *twin arginine translocation*

UFPR - Univesidade Federal do Paraná

WGS - *whole-genome shotgun*

SUMÁRIO

1 INTRODUÇÃO

O gênero *Herbaspirillum* é conhecido por abranger bactérias associadas a plantas, capazes de promover o crescimento vegetal através da produção de fitormônios e da fixação biológica de nitrogênio. Bactérias desse gênero foram encontradas em gramíneas e despertaram interesse para seu uso agrícola.

Recentemente, grandes mudanças ocorreram no gênero, que passou a incluir espécies não só associadas a plantas, mas também oriundas de amostras ambientais, como amostras de solo e água. Alguns isolados clínicos dessas bactérias também foram obtidos. Essa diversidade de ambientes explorados por bactérias do gênero *Herbaspirillum* é refletida em uma grande diversidade metabólica descrita dentro desse gênero.

O sequenciamento e análise genômica de estirpes de *Herbaspirillum* spp. associadas a plantas têm auxiliado na descoberta de genes e mecanismos moleculares envolvidos na promoção do crescimento vegetal e na interação planta-bactéria. No entanto, algumas espécies do gênero apresentam características metabólicas interessantes, por exemplo, podem fixar carbono ou degradar o poluente ambiental 4-clorofenol. Assim, o foco das pesquisas nas espécies associadas a plantas tem limitado o estudo da diversidade gênica do gênero *Herbaspirillum*.

Dessa forma, o sequenciamento e a análise genômica de espécies de *Herbaspirillum* com diferentes características metabólicas, bem como a comparação genômica entre as diferentes espécies do gênero, podem ajudar a compreender os mecanismos de interação com a planta, a encontrar o potencial biotecnológico de cada estirpe, a compreender a evolução do grupo e a aprimorar a classificação taxonômica dentro do gênero levando em conta a sequência genômica e seu conteúdo.

2 REVISÃO DE LITERATURA

2.1 Gênero *Herbaspirillum*

2.1.1 Descrição do gênero

As bactérias do gênero *Herbaspirillum* são descritas como células gram-negativas, flageladas e que apresentam forma geralmente vibrióide. O diâmetro dessas células varia de 0,3 a 0,8 µm e o comprimento de 1,4 a 5,0 µm. Essas bactérias realizam metabolismo respiratório típico, não fermentam açúcares e são catalase e oxidase positivas. Elas apresentam preferência por ácidos orgânicos como fontes de carbono e energia, mas muitas espécies podem também utilizar açúcares e alcoóis. As primeiras espécies desse gênero foram descritas como capazes de fixar nitrogênio atmosférico (N_2), sob condições microaeróbicas e crescer na presença de N_2 como única fonte de nitrogênio (bactérias diazotróficas), mas a maioria das espécies do gênero conhecidas atualmente não apresentam essa capacidade metabólica (BALDANI *et al.,* 1986, BALDANI *et al*., 2014). Taxonomicamente o gênero encontra-se na classe *Proteobacteria*, subclasse *Betaproteobacteria*, ordem *Burkholderiales*, família *Oxalobacteraceae* (BALDANI *et al*., 2014).

2.1.2 Taxonomia e filogenia de Herbaspirillum

A espécie tipo do gênero *Herbaspirillum* é *H. seropedicae* (BALDANI *et al*., 1986), que foi encontrada na rizosfera e raízes de milho, sorgo e arroz, além de também ser a primeira bactéria diazotrófica com características endofíticas isolada. Uma década depois, foi descrita a espécie *H. rubrisubalbicans* (BALDANI *et al*., 1996) que, embora seja promotora do crescimento vegetal e diazotrófica, pode causar a doença da estria mosqueada e da estria vermelha em variedades suscetíveis de cana-de-açúcar e sorgo, respectivamente. Posteriormente, foi descrita a espécie *H. frisingense* (KIRCHHOF *et al*., 2001), também diazotrófica, isolada de raízes e folhas de plantas forrageiras tanto na Alemanha quanto no Brasil. A capacidade de essas três espécies fixarem nitrogênio despertou o interesse para

seu uso agrícola e, por consequência, o interesse no estudo desses organismos (BALDANI & BALDANI, 2005). Foi também sugerido o uso dessas bactérias como biofertilizantes (REIS *et al*., 2009; CANELLAS *et al*., 2013).

Com o decorrer dos anos, novas espécies de *Herbaspirillum* foram descritas: *H. lusitanum* (VALVERDE *et al*., 2003), isolada de nódulos de raízes de feijão (*Phaseolus vulgaris*); *H. autotrophicum* (DING & YOKOTA, 2004), encontrada em um lago eutrófico e capaz de fixar carbono (ARAGNO & SCHLEGEL, 1978); *H. huttiense* (DING & YOKOTA, 2004), obtida de amostras de água destilada e de água subterrânea; *H. chlorophenolicum* (IM *et al*., 2004), bactéria que degrada 4-clorofenol, isolada de sedimento em uma região industrial; *H. hiltneri* (ROTHBALLER *et al*., 2006), obtida de raízes de trigo (*Triticum aestivum* var. Naxos); *H. rhizosphaerae* (JUNG *et al*., 2007), encontrada na rizosfera de *Allium victorialis* var. *platyphyllum*; *H. aquaticum*, isolada de água deionizada (DOBRITSA *et. al.*, 2010); e *H. massiliense*, isolada de fezes humanas (LAGIER *et al*., 2012). Muitas dessas novas espécies não foram isoladas de plantas e nenhuma delas apresentou capacidade de fixar nitrogênio. Algumas dessas bactérias podem ser visualizadas na FIGURA 2.1 e características das espécies podem ser visualizadas na TABELA 2.1.

Com a descrição do gênero *Noviherbaspirillum* (LIN *et al*., 2013), representado pela espécie tipo *N. maltae*, as espécies [*H*.] *canariense*, [*H*.] *aurantiacum* e [*H*.] *soli* (CARRO *et al*., 2011), presentes em amostras de solo de uma montanha vulcânica, e [*H*.] *psychrotolerans*, encontrada em uma geleira na Antártica (BAJERSKI *et al*., 2013), foram reclassificadas como *Noviherbaspirillum* (LIN *et al*., 2013).

O agrupamento filogenético das espécies de *Herbaspirillum*, realizado com a sequência do gene que codifica para a subunidade 16S do rRNA, mostra que esses organismos formam dois grupos: 1- *H. seropedicae*, *H. rubrisubalbicans*, *H. frisingense*, *H. huttiense*, *H. aquaticum* e *H. chlorophenolicum*; 2- *H. rhizosphaerae*, *H. lusitanum*, *H. autotrophicum* e *H. hiltneri* (FIGURA 2.2) (MONTEIRO *et al*., 2014).

Isso foi também evidenciado com a descrição do gênero *Paraherbaspirillum* (ANANDHAM *et al*., 2013), representado pela espécie *P. soli*, que está separado de *Herbaspirillum* (MONTEIRO *et al*., 2014), porém posicionado filogeneticamente entre espécies desse gênero dependendo do autor (ANANDHAM *et al*., 2013).

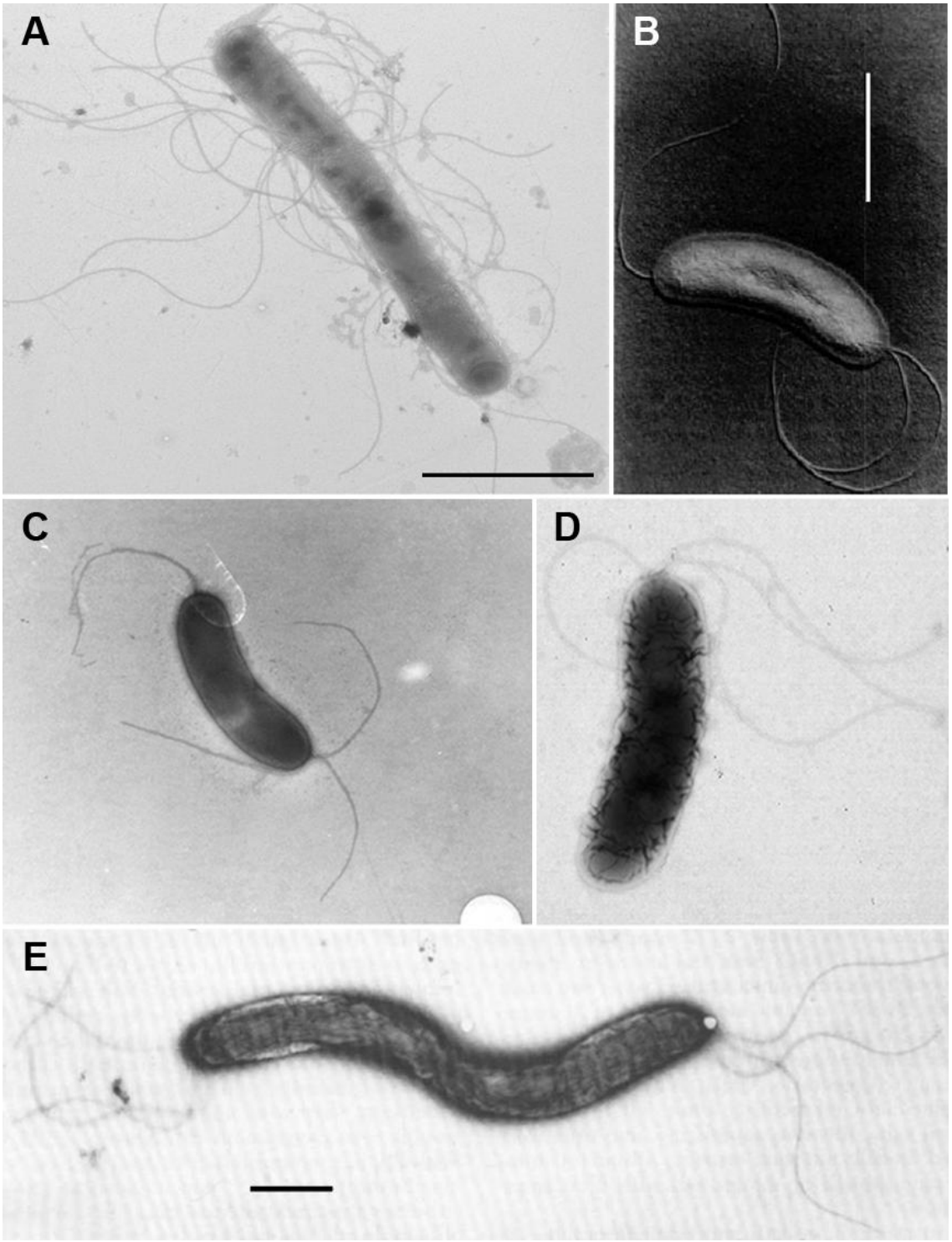


FIGURA 2.1: MORFOLOGIA DE BACTÉRIAS DO GÊNERO *Herbaspirillum*

Em A é representado *H. massiliense* (escala 0,9 µm); em B, *H. seropedicae* (escala 1 µm); em C, uma célula de *Herbaspirillum* spp. isolada de banana (sem escala); em D, *H. aquaticum* (sem escala); em E, *H. autotrophicum* (escala 1 µm)

FONTE: adaptado de LAGIER *et al*. (2012); BALDANI *et al.* (1986); WEBER *et al*. (2001); DOBRITSA *et al*. (2010); ARAGNO & SCHLEGEL (1978)

TABELA 2.1: CARACTERÍSTICAS DAS ESPÉCIES DE *Herbaspirillum*

	H. aquaticum	*H. autotrophicum*	*H. chlorophenolicum*	*H. frisingense*	*H. hiltneri*	*H. huttiense* subsp. *huttiense*	*H. huttiense* subsp. *putei*	*H. lusitanum*	*H. rhizosphaerae*	*H. rubrisubalbicans*	*H. seropedicae*	*H. massiliense*
Morfologia	Bastonetes curvos	Curva a espiral	Bastonetes levemente curvados	Bastonetes curvos, espirais	Bastonetes levemente curvados	Bastonetes levemente curvados	Bastonetes curvos, espirais	Curvada	Bastonetes levemente curvados	Bastonetes levemente curvados	Bastonete, algumas vezes espirais	Bastonete
Tamanho das células (µm)	0,5 x 2	0,6-0,8 x 2-5	0,7 x 2,3	0,5-0,7 x 1,4-1,8	0,5-0,6 x 1,6-2	0,4 x 1,8	0,5-0,7 x 2,1-3,4	0,5 x 1,6	0,3-0,4 x 1,8-2,2	0,6-0,7 x 1,4-1,8	0,6-0,7 x 1,5-5,0	0,44 (diâmetro)
Flagelos	1-4 polares	1-5, tufos bipolares	1, unipolar	1-3, unipolares	2, unipolares	1-3, bipolares	1-4	1-2, polares	bipolar	muitos, unipolares	1-3, bipolares	muitos
Motilidade	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim
Fixação de nitrogênio em meio semi-sólido	não	não	não	sim	não	não	não	sim	não	sim	sim	não
Detecção de *nifD* (ou *nifH*)	não	não	não	sim	não	não	sim*	sim*	não	sim	sim	não
Temperatura para crescimento (^{o}C)	10-35	10-35	30 (ótima)	30-37	26-34	25-37	25-37	20-40	4-34	Até 40	22-38	37 (ótima)
pH para crescimento	5-8	5-8	6-8	6-7 (ótimo)	6-8 (ótimo)	não determinado	6-7 (ótimo)	5-8 (ótimo)	6,5-7,5 (ótimo)	5,7-6,8	5,3-8,0	não determinado
Catalase	+	+	+	+	+	+	+	+	+	+	+	não determinado
Urease	+	+	não determinado	+	não determinado	+	+	+	+	não determinado	+	não determinado
Oxidase	+	+	+	+	+	+	+	+	+	+	+	não determinado
Redução de nitrato	-	-	+	+	não determinado	-	-	-	-	+	+	não determinado

O sinal “+” indica que o teste realizado teve resultado positivo; o sinal “-” indica que o teste teve resultado negativo.

* genes não encontrados no genoma

FONTE: LAGIER *et al*. (2012); BALDANI *et al*. (2014)

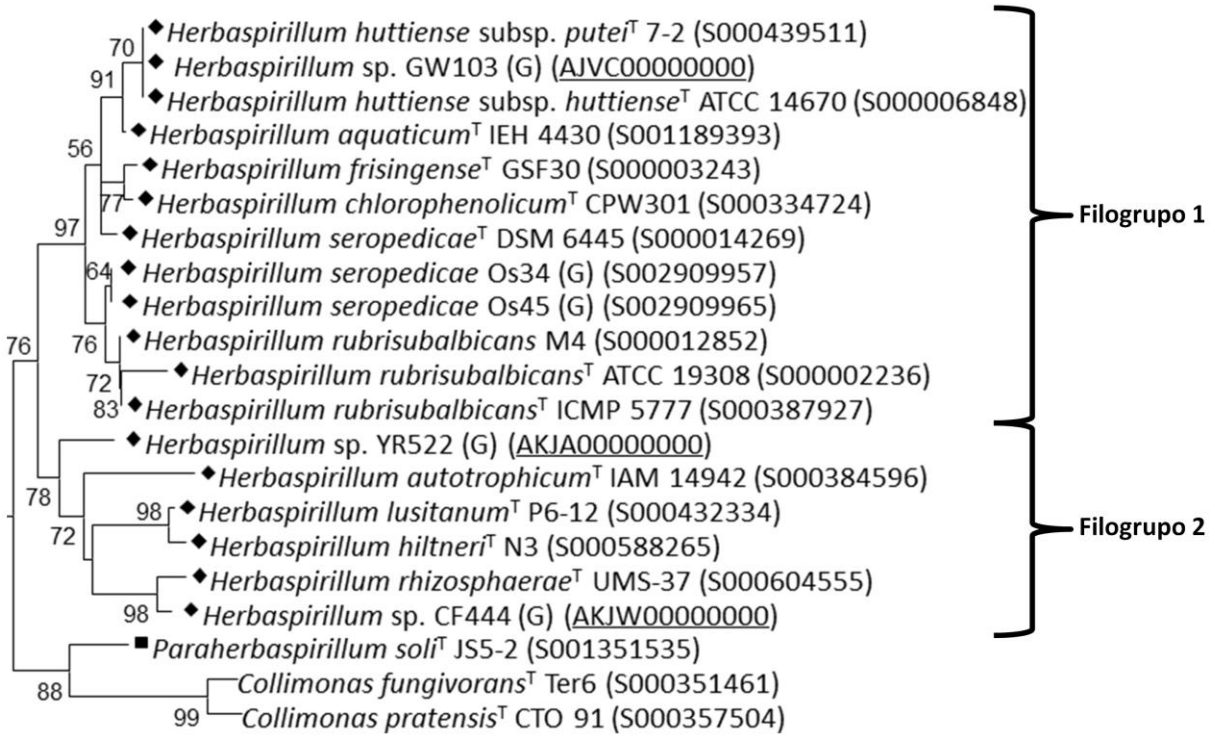


FIGURA 2.2: ÁRVORE FILOGENÉTICA DAS ESPÉCIES DE *Herbaspirillum*
A árvore é baseada na análise do gene 16S rRNA.
FONTE: adaptado de MONTEIRO *et al.,* 2014

Um aspecto notável em determinadas árvores é a presença de espécies de *Collimonas* (entre outros gêneros) agrupadas com *Herbaspirillum* spp. (JUNG *et al*., 2007; CARRO *et al*., 2012). Isso demonstra que a análise de identidade do gene 16S rRNA é limitada para resolver a filogenia de *Herbaspirillum* e apresenta inconsistências a partir de diferentes autores (JUNG *et al*., 2007; CARRO *et al*., 2011; BAJERSKI *et al*., 2013; MONTEIRO *et al.*, 2014).

2.1.3 Diversidade de Herbaspirillum

Relatos na literatura mostraram que estirpes de *Herbaspirillum* spp. foram encontradas nos mais diversos ambientes, como: em arrozais japoneses (ELBELTAGY *et al.*, 2001; ISHII *et. al.* 2009); em depósitos vulcânicos, águas hidrotermais e rochas vulcânicas (LU *et al*., 2008; KAWAICHI *et al*., 2013; KELLY *et al*., 2014); no ovário de *Asobara tabida* (vespa endoparasitóide de *Drosophila*) (ZOUACHE *et al.*, 2009); em água extremamente salina de um lago coberto por gelo (KUHN *et al*., 2014); em canaviais chineses (TAN *et al.*, 2010); em solo contaminado

com fluoranteno (XU *et al.,* 2011); em raízes de chá (*Camellia sinensis*) (GULATI *et al.,* 2011); em bananeiras e abacaxizeiros (CRUZ *et al*., 2001); em arroz silvestre, *Oryza glumaepatula* (JÚNIOR *et al*., 2013). Isso demonstra que esses organismos são mais versáteis do que a descrição que se tem do gênero e apresentam maior diversidade metabólica (JAUREGUI *et al*., 2014).

Além disso, estirpes dessas bactérias foram reportadas em casos clínicos. Elas foram encontradas associadas a celulite (no sentido de infecção) e a bacteremia em um paciente exposto à água doce de um canal (TAN & OEHLER, 2005); em pacientes com fibrose cística (COENYE *et al.,* 2002; CAMPANA *et al.*, 2005; SPILKER *et al.*, 2008); em aneurisma aórtico (SILVA *et al.,* 2006); relacionadas a bacteremia em pacientes com leucemia linfoblástica (ZIGA *et al.*, 2010; CHEN *et al.*, 2011); e também encontradas em pacientes com câncer (CHEMALY *et al*., 2014).

Em um trabalho recente, foi mostrado que diferentes espécies do gênero *Herbaspirillum* são capazes de aderir a células HeLa em baixa densidade, com efeitos citotóxicos discretos, mostrando que *Herbaspirillum* spp. apresentam baixa virulência (MARQUES *et al*., 2015).

2.1.4 Herbaspirillum autotrophicum

Herbaspirillum autotrophicum (do grego *autos*=si mesmo; *trophikos*=nutrição; em latim *autotrophicum*=aquele que nutre a si mesmo), é uma bactéria capaz de crescer autotroficamente, em ambiente aeróbico. Estirpes dessa espécie foram isoladas de um lago eutrófico (lago *Le Loclat*) na Suíça em 1974. Análises microscópicas revelaram células espirais flageladas, com diâmetro entre 0,6 e 0,8 µm e comprimento entre 2,0 e 5,0 µm. Essas células mostraram ser Gram-negativas e continham grânulos citoplasmáticos de poli(3-hidroxibutirato) (ARAGNO & SCHLEGEL, 1978). Características dessa espécie em comparação com outras espécies de *Herbaspirillum* podem ser vistas na TABELA 2.1.

Testes bioquímicos são mostrados na TABELA 2.2 e fontes de carbono e energia que essas bactérias podem utilizar são mostradas na TABELA 2.3. Foi também verificado que essas estirpes fixavam carbono litoautotroficamente utilizando hidrogênio (H_2) como fonte de energia (ARAGNO & SCHLEGEL, 1978).

Devido a semelhanças morfológicas e fisiológicas com bactérias do gênero *Aquaspirillum*, as estirpes foram incluídas inicialmente nesse gênero, porém

classificadas em uma nova espécie chamada [*Aquaspirillum*] *autotrophicum* (ARAGNO & SCHLEGEL, 1978). Posteriormente, com base na filogenia do gene 16S rRNA e na hibridização DNA-DNA, essas estirpes foram reclassificadas como pertencentes ao gênero *Herbaspirillum*. A espécie foi renomeada para *Herbaspirillum autotrophicum* e a estirpe IAM 14942^T (ATCC 29984 = CCUG 12808 = DSM 732 = JCM 21424 = NBRC 15327 = LMG 4326 = VKM B-1394) foi estabelecida como estirpe tipo da espécie (DING & YOKOTA, 2004).

TABELA 2.2: CARACTERÍSTICAS BIOQUÍMICAS DE *H. autotrophicum*

TESTE	RESULTADO[1]	TESTE	RESULTADO[1]
Redução de nitrato	-	Hidrólise de esculina	-
Crescimento anaeróbico com nitrato	-	Produção de pigmento fluorescente solúvel em água	-
Fosfatase	+	Formação de indol	-
Sulfatase	-	H_2S a partir de cisteína	-
Oxidase	+	Ácidos a partir de carboidratos	-
Catalase	+	Crescimento na presença de 1% de sais biliares	+
Urease	+	Crescimento na presença de 1% de glicina	-
Hidrólise de gelatina	-	Crescimento na presença de 3% de NaCl	-
Hidrólise de caseína	-	Redução de selenito	-
Hidrólise de amido	-	Concentração inibitória mínima de penicilina G	60 IU/ml

[1] O sinal '+' indica positivo para o teste, o sinal '-' indica negativo.

FONTE: ARAGNO & SCHLEGEL (1978)

As bactérias como *H. autotrophicum*, que têm a capacidade de crescer autotroficamente utilizando H_2 como doador de elétrons e utilizá-los como fonte de energia para reduzir CO_2, são chamadas de bactérias oxidantes de hidrogênio ou *knallgas*. Elas são encontradas em *taxa* distintos (*Proteobacteria*, *Aquificales*, *Actinobacteria* e *Firmicutes*) e em vários tipos de ambientes (solo, mar, águas termais, entre outros). No entanto, a função ecológica delas é pouco conhecida, embora tenham recebido importância como possíveis produtores de biomassa para fermentação industrial (LEPIDI *et al*., 1990; PUMPHREY *et al*., 2011).

TABELA 2.3: UTILIZAÇÃO DE FONTES DE CARBONO POR *H. autotrophicum*

FONTE DE CARBONO ÚNICA[1]		FONTE DE CARBONO ÚNICA[1]	
Ácidos orgânicos		Aminoácidos (continuação)	
Acetato	+	L-lisina	-
Cis-aconitato	+	L-metionina	-
Antranilato	-	L-ornitina	-
Benzoato	-	L-fenilalanina	+
Butirato	+	L-prolina	+
Caproato	-	L-serina	-
Citrato	+	L-triptofano	+
Fumarato	+	L-tirosina	+
Glicolato	+	L-valina	-
ρ-hidroxibenzoato	+	Açúcares ácidos	
DL-β-hidroxibutirato	+	D-gluconato	+
Poli- β-hidroxibutirato (exógeno)	-	α-cetogluconato	+
DL-isocitrato	+	Açúcares	
α-cetoglutarato	+	L-arabinose	-
DL-lactato	+	Celubiose	-
L-malato	+	D-frutose	-
Malonato	+	D-fucose	-
Mesaconato	+	D-galactose	-
Oxaloacetato	+	D-glucose	-
Propionato	+	Maltose	-
Piruvato	+	D-manose	-
Succinato	+	L-ramnose	-
D-(-)-tartrato	+	Sacarose	-
L-(+)-tartrato	-	Trealose	-
meso-tartrato	-	D-xilose	-
Aminoácidos		Alcoóis	
L-alanina	+	*n*-butanol	-
L-arginina	-	*t*-butanol	-
L-asparagina	+	Etanol	-
L-aspartato	+	Glicerina	-
L-citrulina	-	Manitol	-
L-cisteína	-	Metanol	-
L-glutamato	+	Fenol	-
L-glutamina	+	*n*-propanol	-
Glicina	+	Outras substâncias	
L-histidina	-	Alantoína	-
L-isoleucina	+	Etanolamina	+
L-leucina	-		

[1] O sinal '+' indica positivo para o teste, o sinal '-' indica negativo.

FONTE: ARAGNO & SCHLEGEL (1978)

Essas bactérias são aeróbicas, autotróficas facultativas e realizam o metabolismo mixotrófico, no qual tanto compostos orgânicos quanto inorgânicos podem ser utilizados como fontes de energia. Metabolicamente é sugerido que a maioria delas cresça heterotroficamente em baixa concentração de H_2 e realizem o

metabolismo autotrófico quando há maior disponibilidade desse gás (BOWIEN & SCHLEGEL, 1981; PUMPHREY *et al*., 2011; MATASSA *et al*., 2015).

A enzima chave do metabolismo litoautotrófico é a hidrogenase, uma proteína ligada à membrana da célula que oxida o hidrogênio e promove o bombeamento de prótons, via transferência de elétrons, para quinonas e citocromos. Outras duas enzimas, responsáveis pela fixação de carbono através do ciclo de Calvin, também são consideradas essenciais para esse tipo de metabolismo: fosforibuloquinase (responsável por gerar D-ribulose 1,5-bifosfato) e ribulose 1,5-bifosfato carboxilase/oxigenase (RuBisCO, responsável pela incorporação de uma molécula de dióxido de carbono à D-ribulose 1,5-bifosfato para gerar duas moléculas de 3-fosfoglicerato) (BOWIEN & SCHLEGEL, 1981; MATASSA *et al*., 2015).

2.1.5 Herbaspirillum chlorophenolicum

Herbaspirillum chlorophenolicum (do latim *clorophenolicum*=relacionado a clorofenol) é uma bactéria capaz de utilizar 4-clorofenol como única fonte de carbono e energia. Essa espécie foi isolada de uma amostra retirada do sedimento de um córrego, próximo a uma região industrial em Cheongju (Coréia) e colocada em meio enriquecido com 4-clorofenol. Nesse meio, foi observada a presença de uma bactéria inicialmente descrita como pertencente à espécie *Comamonas testosteroni*, devido a características morfológicas e fisiológicas. Posteriormente, ela foi incluída no gênero *Herbaspirillum* como espécie nova, com base na identidade do gene 16S rRNA e na hibridização DNA-DNA. A estirpe CPW301^{T} (IAM 15024 = JCM 21487 = KCTC 12096 = NBRC 102525) é a estirpe tipo da espécie (IM *et al*., 2004). Características dessa espécie em comparação com outras espécies de *Herbaspirillum* podem ser vistas na TABELA 2.1.

Essa espécie é descrita como apresentando células curvas, flageladas, com 0,7 µm de diâmetro e 2,3 µm de comprimento. Essas células são também Gram-negativas, aeróbicas e móveis, capazes de crescer tanto em fenol quanto 4-clorofenol como única fonte de carbono e energia. Por outro lado, não crescem na presença de açúcares e alcoóis como fontes únicas de carbono e energia (TABELA 2.4). Os genes *nifD* e *nifH* também não foram encontrados (IM *et al*., 2004).

TABELA 2.4: CARACTERÍSTICAS BIOQUÍMICAS E MOLECULARES DE *H. chlorophenolicum*

CARACTERÍSTICA[1]		CARACTERÍSTICA[1]	
Tamanho (µm)	0,7 por 2,3	Crescimento em: (continuação)	
Flagelos	Um, unipolar	**Fenol**	**+**
Crescimento ótimo	30°C	**4-clorofenol**	**+**
pH ótimo	6,0 a 7,0	Conteúdo GC (%)	61,3
nifD	Ausente	Ubiquinona principal	Q-8
nifH	Ausente	Ácidos graxos (%):	
Crescimento em:		$C_{12:0}$	3,7
Meio livre de nitrogênio	-	$C_{14:0}$	0,24
Manose	-	$C_{16:0}$	33,73
N-acetilglucosamina	+	$C_{18:0}$	1,98
Meso-inositol	-	$C_{18:1\omega6c}$	0
Ramnose	-	$C_{18:1\omega7c}$	13,43
Ribose	-	$C_{12:0}$ 2-OH	2,65
Adonitol	-	$C_{10:0}$ 3-OH	1,33
Glicerol	-	$C_{12:0}$ 3-OH	3,27
D-lixose	-	$C_{14:0}$ 2-OH	0,07
Xilitol	-	ciclo-$C_{17:0}$	21,84
D-xilose	-	ciclo-$C_{19:0\ \omega8c}$	1,37
Sorbitol	-	Outros	14,93

[1] o sinal '+' indica que o organismo pode utilizar o composto como única fonte de carbono e energia, o sinal '-' indica que ele não possui essa capacidade

FONTE: IM *et al.* (2004)

As bactérias capazes de degradar clorofenóis são de grande importância ambiental, pois essas substâncias pertencem à classe de compostos organoclorados, considerados de grande risco para o ambiente. Esses compostos podem formar espécies ionizadas conforme a variação de pH do meio, de modo a adquirir diferentes propriedades físico-químicas e diferentes comportamentos químicos nos lugares onde são encontrados. Entre eles, o 4-clorofenol é acumulado no ambiente pelo seu uso em herbicidas, fungicidas e indústrias que fabricam papel. As pessoas podem entrar em contato com essa substância ingerindo alimentos ou água contaminada, ou ainda pelo contato através da pele. É considerado extremamente tóxico, pois possui propriedades carcinogênicas, mutagênicas e citotóxicas (FREIRE *et al*., 2000, CALIZ *et al*., 2011; ARORA & BAE, 2014).

A degradação de 4-clorofenol por bactérias ocorre principalmente pela via do clorocatecol (FIGURA 2.3). Nessa via, o 4-clorofenol (4CF) é convertido a 4-clorocatecol (4CC) pela enzima 2-4CF-monoxigenase (EC 1.14.13.-). O 4CC pode sofrer clivagem *orto* ou *meta*. Na via de clivagem *meta*, o 4CC é convertido a um composto tóxico, chamado semialdeído 5-cloro-2-hidroximucônico (5CHMS) pela

enzima catecol-2, 3-dioxigenase (EC 1.13.11.2). Esse composto muitas vezes é o produto final da via. No entanto, a via pode seguir e esse composto ser degradado a ácido 5-cloro-2-hidrozipenta-2,4-dienóico para, posteriormente, ser degradado até entrar no ciclo do citrato (FIGURA 2.3B) (ARORA & BAE, 2014).

Na via da clivagem *orto*, o 4CC é clivado a 3-cloromuconato pela enzima catecol-1,2-dioxigenase (EC 1.13.11.1); posteriormente o 3-cloromuconato é transformado em *cis*-dienolactona através da liberação de um íon clorido, pela ação da enzima cloromuconato cicloisomerase (EC 5.5.1.7); depois disso, a *cis*-dienolactona é convertida a maleilacetato pela enzima dienolactona hidrolase (EC 3.1.1.45); por fim, o maleilacetato é reduzido a 3-oxoadipato pela enzima maleilacetato redutase (EC 1.3.1.32) (FIGURA 2.3A) (ARORA & BAE, 2014).

Organismos que degradam o 4-clorofenol, como o *H. chlorophenolicum* CPW301, podem ser utilizados para descontaminação ambiental, em um processo chamado de biorremediação. Essa estratégia é bastante promissora, pois apresenta uma boa relação custo-benefício. No entanto, alguns inconvenientes dessa estratégia são: 1- encontrar estirpes especializadas em degradar determinado composto em condições específicas; 2- monitorar e controlar o crescimento e comportamento desses organismos, principalmente frente a compostos sutilmente diferentes daqueles aos quais estão aptos a degradar (FREIRE *et al*., 2000; IM *et al*., 2004; CALIZ *et al*., 2011). Estudos genômicos desses organismos pode ser um passo importante para aprimorar essa estratégia e contornar esses inconvenientes.

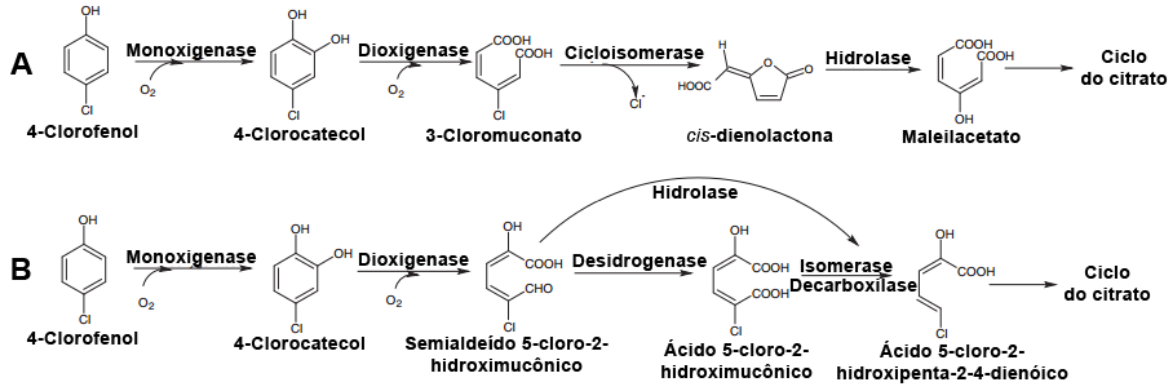


FIGURA 2.3: VIAS DE DEGRADAÇÃO DE 4-CLOFOFENOL

Em ambas as vias o 4-clorofenol é convertido a 4-clorocatecol (4CC). Em A, é mostrada a via de degradação pela clivagem *orto*. Em B, é mostrada a via de degradação pela clivagem *meta*.

FONTE: adaptado de ARORA & BAE (2014)

2.1.6 Herbaspirillum rhizosphaerae

Herbaspirillum rhizosphaerae (do latim *rhizosphaerae*=da rizofera) é uma espécie isolada de uma amostra de solo rizosférico de *Allium victorialis* var. *platyphyllum* (alho vitorino), na ilha de Ulleung, Coréia. É uma bactéria Gram-negativa e possui células curvas, com 0,3 a 0,4 µm de diâmetro e 1,8 a 2,2 µm de comprimento. Apresenta motilidade e as células possuem flagelos bipolares. Sua temperatura ótima para crescimento é entre 25 e 30^{o}C e o pH ótimo para crescimento é entre 6,5 e 7,5. A estirpe UMS-37^{T} (CIP 108917 = KCTC 12558) é a estirpe tipo da espécie (JUNG *et al*., 2007). Características dessa espécie em comparação com outras espécies de *Herbaspirillum* podem ser vistas na TABELA 2.1. Algumas características morfológicas, fisiológicas e bioquímicas estão sumarizadas na TABELA 2.5.

A rizosfera, ambiente de onde *H. rhizosphaerae* foi isolada, corresponde à porção do solo influenciada pela raiz da planta. Devido aos exsudatos ricos em compostos orgânicos (ácidos orgânicos, sideróforos, açúcares, vitaminas, aminoácidos, nucleosídeos e mucilagem) que a planta libera, muitos microrganismos (bactérias, fungos, protozoário, algas, nematóides e microartrópodes) são atraídos pela rizosfera, que estabelece um rico nicho ecológico. Com isso, a rizosfera apresenta uma grande biomassa de bactérias (aproximadamente 10^{10} bactérias por grama de solo) e grande diversidade de *taxa* desses organismos (BERG & SMALLA, 2009; RAAIJMAKERS *et al*., 2009; VACHERON *et al*., 2013).

Algumas bactérias rizosféricas podem promover o crescimento da planta e contribuir para a saúde dessa. Essas bactérias são chamadas de PGPR (do inglês, *plant growth-promoting rhizobacteria* – rizobactérias promotoras do crescimento da planta) e agem através de vários mecanismos, por exemplo, fixando nitrogênio atmosférico, solubilizando fosfato, ou produzindo sideróforos. Além disso, podem produzir fitormônios e inibir o crescimento de fitoparasitas por competição, produzindo antibióticos ou induzindo a resistência sistêmica da planta (FIGURA 2.4) (VACHERON *et al*., 2013; BHARDWAJ *et al*., 2014).

TABELA 2.5: CARACTERÍSTICAS GERAIS DE *H. rhizosphaerae*

CARACTERÍSTICA[1]		CARACTERÍSTICA[1]	
Forma da célula	Levemente curvada	Crescimento em: (continuação)	
Coloração	Branca/leitosa	D-galactose	+
Motilidade	+	Sorbitol	W
Crescimento ótimo (temperatura, em ^{o}C)	25-30	**D-frutose**	**+**
Crescimento ótimo (pH)	6,5-7,5	Outras características:	
Redução de nitrato	-	Fosfatase alcalina	-
Catalese	+	Esterase (C4)	+
nifD	Ausente	Esterase lipase (C8)	+
nifH	Ausente	Valina arilamidase	+
Crescimento em:		Cistina arilamidase	-
D-manose	+	Fosfatase ácida	-
D-ramnose	-	Naftol-AS-BI-fosfohidrolase	+
D-glucose	**+**	β-galactosidase	-
		Conteúdo G+C (%)	59,8 – 60

[1] o sinal '+' indica que o teste apresentou resultado positivo, o sinal '-' indica que o teste apresentou resultado negativo. "W" indica fracamente positivo.

FONTE: JUNG *et al.* (2007)

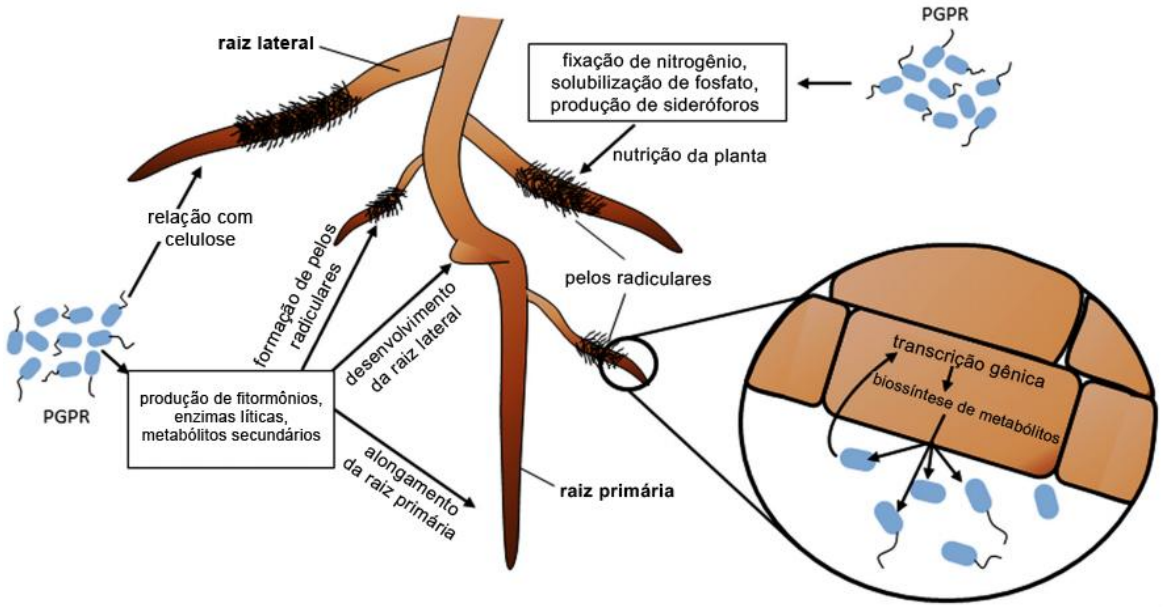


FIGURA 2.4: INTERAÇÃO ENTRE AS RIZOBACTÉRIAS PROMOTORAS DE CRESCIMENTO (PGPR) E A RAIZ DA PLANTA

FONTE: adaptado de VACHERON *et al*., 2013

Muitos antibióticos podem ser isolados de rizobactérias, principalmente aqueles que agem sobre a parede celular, a estrutura da membrana celular e a formação de ribossomos. A produção desses compostos está relacionada com a grande competição entre os organismos que vivem nesse ambiente. Por outro lado, muitas dessas bactérias também apresentam resistência a antibióticos, provavelmente porque a interação entre elas favorece a transferência horizontal de genes (BERG *et al*., 2005, BENEDUZI *et al*., 2012).

2.2 Análise genômica

2.2.1 Montagem e anotação genômica

Um genoma bacteriano geralmente corresponde a um cromossomo circular, que possui uma única origem de replicação, embora alguns genomas possam ser lineares ou apresentar outros *replicons* (GUZMÁN *et al*., 2008). Grande parte desses genomas são sequenciados pela abordagem WGS (do inglês "*whole-genome shotgun*"), que consiste em fragmentar aleatoriamente o DNA genômico total e, posteriormente, sequenciar os fragmentos obtidos. Esses fragmentos sequenciados são chamados de leituras ou, em inglês, *reads*. Dessa forma, o processo de montagem genômica consiste em organizar novamente os *reads* para reconstruir a sequência genômica original (BRODER & VENTER, 2000; NARZISI & MISHRA, 2011). Quando esse processo é realizado sem a utilização de nenhuma sequência genômica prévia, ele é chamado de montagem genômica *de novo* (BERLUNG *et al*., 2011).

Essas montagens *de novo* são realizadas por ferramentas computacionais específicas chamadas de montadores de genomas. Elas partem do princípio básico de que, se dois *reads* possuem uma região sobreposta na sequência das bases, eles ocorrem em uma mesma região do genoma e podem ser usados para estender sua sequência de bases. Baseados nisso, os montadores de genoma geram grafos a partir da sobreposição de *reads*, os quais criam caminhos através desses *reads* para a reconstrução da sequência genômica (FIGURA 2.5A) (NARZISI & MISHRA, 2011; BERLUNG *et al*., 2011).

No entanto, essa dinâmica enfrenta problemas com as regiões de repetição (por exemplo, genes duplicados), nas quais os montadores genômicos encontram

uma série de caminhos alternativos de sobreposições que convergem, divergem e que podem levar à montagem genômicas distintas, muitas delas incorretas (FIGURA 2.5B). Uma alternativa utilizada pelos montadores é mascarar as repetições e montar apenas aquelas regiões que são únicas na sequência genômica, o que acaba inevitavelmente fragmentando as montagens. No entanto, as melhores soluções para resolver as repetições são tanto a utilização de *reads* longos (o suficiente para cobrir essas regiões) quanto a utilização dos chamados *reads* pareados (*reads* sequenciados a partir da ponta de fragmentos que possuem tamanho maior que as repetições), os quais guiam os montadores corretamente frente aos caminhos alternativos proporcionados (FIGURA 2.5B) (CHEN, 2008; MILLER *et al*., 2010; BERLUNG *et al*., 2011).

Por muitas décadas o método de sequenciamento desenvolvido por Sanger e colaboradores (1977), que utiliza eletroforese para separar fragmentos de DNA terminados por didesoxinucleotídeos marcados com fluoróforos, foi utilizado para se obter fragmentos de uma sequência genômica, e era capaz de gerar *reads* de até ~850 pares de bases (MEDINI *et al*., 2008; BERLUNG *et al*., 2011). Porém, avanços tecnológicos produziram novas técnicas de sequenciamento, utilizadas pelos sequenciadores de segunda geração (Roche 454, SOLiD System e Illumina Genome Analyzer) que, embora sejam capazes de gerar maior quantidade de dados e ter baixo custo de sequenciamento, passaram a produzir *reads* de no máximo 500 pares de bases, chamados de *reads* curtos (METZKER, 2010; NARZISI & MISHRA, 2011).

Essa diminuição do tamanho dos *reads* e a produção de grande volume de dados trouxeram consigo alguns problemas adicionais para a reconstrução da sequência genômica original. A maior dificuldade de trabalhar com *reads* curtos, aliada ao grande volume de grafos de sobreposição produzidos durante o processo de montagem, foram responsáveis pela mudança de estratégia e utilização dos chamados grafos de-Brujin. Nessa nova abordagem, os *reads* são fragmentados em subsequências (*mers*) de um tamanho específico (*k*) chamadas *k-mers* (podem também ser chamadas de *n-mers*, *words* ou *seeds*). Assim, os grafos de-Brujin consistem na sobreposição dos *k-mers* e não na sobreposição dos *reads* como um todo (FIGURA 2.5C). No entanto, ao passo que uma das soluções para resolver regiões de repetição era a utilização de *reads* longos, a montagem com *reads* curtos tende a apresentar maior número de repetições não resolvidas, caso não seja usado

o pareamento de *reads* (CHEN, 2008; NARZISI & MISHRA, 2011; BERLUNG *et al*., 2011).

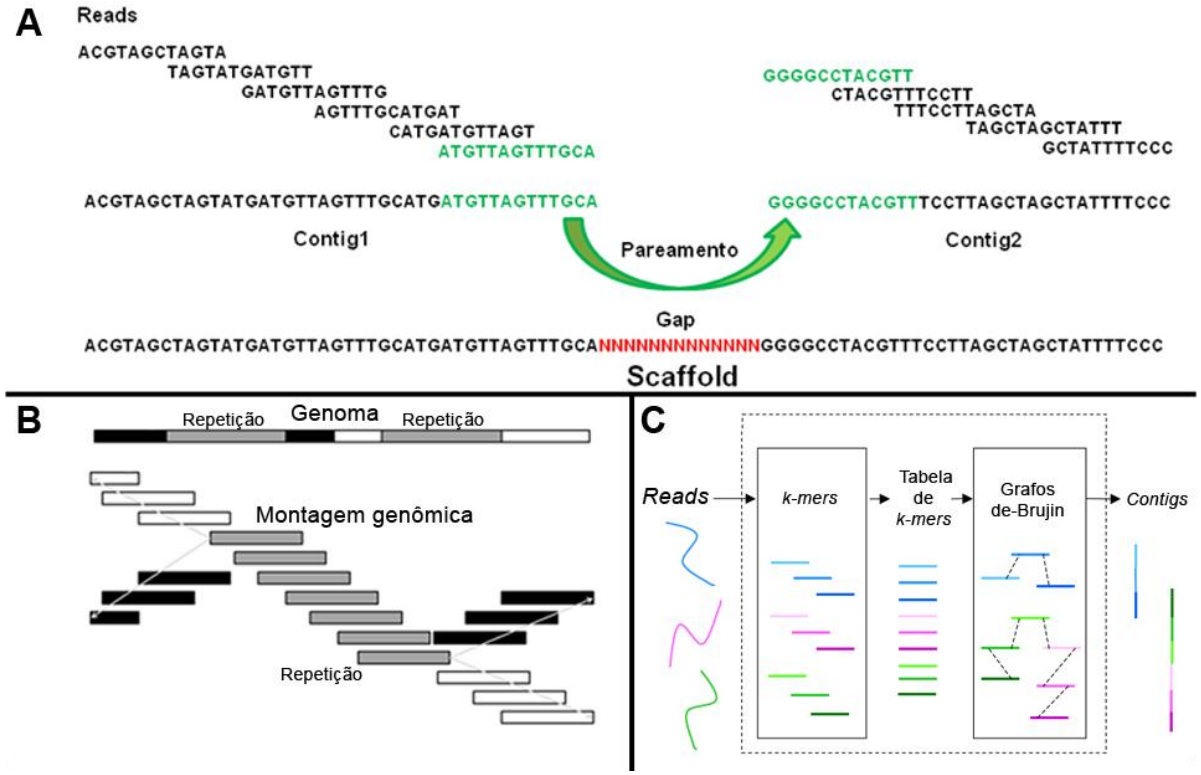


FIGURA 2.5: PROCESSO DE MONTAGEM GENÔMICA

Em A, estão representados grafos de sobreposição de *reads*. Esses grafos são unidos para gerar fragmentos maiores chamados de *contigs* e, se houver informação de pareamento entre os *reads*, os *contigs* podem ser ordenados em *scaffolds*. A informação ausente é chamada de *gap*. Em B, está representado um genoma que apresenta duas cópias de uma mesma região. Durante o processo de montagem essas duas regiões são colabadas e é necessária a informação dos pares (preto/preto, branco/branco) para guiar a montagem pelo caminho correto. Em C, é mostrada a estratégia de montagem com grafos de-Brujin. Nela, os *reads* são fragmentados em *k-mers*, que passam a ser unidos pelos grafos de-Brujin para gerar *contigs*.

FONTE: o autor (2015), adaptado de TIEPPO (2011), CARDOSO (2011) e EDWARDS & HOLT (2013)

Alguns avanços têm sido obtidos com a utilização de *reads* curtos para a montagem genômica, mas esse é um problema desafiador, no qual poucos especialistas conseguem desenhar uma boa estratégia de sequenciamento e montagem de um genoma (SALZBERG *et al*., 2012).

Diante da dificuldade de se obter uma sequência genômica completa, a montagem genômica pode ser arranjada em uma estrutura de dados hierárquica, onde os *reads* são agrupados em sequências contíguas chamadas de *contigs* que, por sua vez, podem ser orientados e ordenados em *scaffolds* (ou *supercontigs*). Através do pareamento de *reads* é possível também estimar o tamanho das lacunas (ou *gaps*, do inglês) que separam os *contigs* que formam um *scaffold* (FIGURA 2.5A)

(MILLER, 2010; NARZISI & MISHRA, 2011). Esse tipo de estrutura resultante de montagens automáticas pode também ser chamado de rascunho (ou *draft*, do inglês) genômico (LAND *et al*., 2015).

A eficiência do processo de montagem é medida pelas características do conjunto de *contigs* e *scaffolds* gerados. Entre elas, podem ser usados como parâmetros de avaliação de montagem: o número de *contigs* obtidos (o quanto a sequência genômica original está fragmentada), o tamanho médio desses *contigs* (qual o tamanho dos fragmentos do *draft* genômico), o tamanho do maior *contig* obtido (o quão perto do genoma total o maior fragmento obtido se aproxima), o tamanho total do conjunto de *contigs* obtidos (o quanto a soma dos fragmentos representa o genoma original) e o tamanho do *contig* N50. O *contig* N50 é entendido como o menor *contig* dentre a soma ordenada dos *contigs* (do maior para o menor) até que se obtenha pelo menos 50% do tamanho de todo o conjunto (MILLER *et al*., 2010); em outras palavras, um *contig* N50 de tamanho N indica que 50% do genoma está representado em *contigs* maiores que N pares de bases.

No entanto, melhorar essas montagens para se obter um *draft* de alta qualidade, ou a sequência genômica completa, requer um processo chamado finalização, o qual é geralmente manual, trabalhoso e demorado. Com a velocidade que as novas tecnologias de sequenciamento genômico produzem dados, há uma defasagem de tempo entre as montagens/finalizações e novos genomas sequenciados. Por esse motivo, essas tecnologias produziram um vasto número de *drafts* genômicos, de modo que 90% dos genomas bacterianos depositados no NCBI (do inglês "*National Center for Biotechnology Information*", http://www.ncbi.nlm.nih.gov) Genbank não estão finalizados (LAND *et al*., 2015).

Depois do processo de montagem, é realizada a anotação genômica, que consiste em analisar e interpretar o genoma sequenciado para extrair dele o seu significado biológico e colocar essa informação dentro do contexto metabólico do organismo. Da mesma forma que a montagem genômica possui uma hierarquia, a anotação genômica pode ser dividida em diferentes etapas de refinamento, que ocorrem nos níveis de nucleotídeos (predição gênica), de proteínas (predição de função gênica) e de metabolismo (construção dos mapas metabólicos do organismo) (STEIN, 2001; STOTHARD & WISHART, 2006; GUZMÁN *et al*., 2008).

Em geral, essas etapas são realizadas pela execução de um *pipeline* (fluxo de execução) de programas em colaboração com a análise de um especialista

(curador). Com isso, programas para predição gênica são utilizados para encontrar sequências codificadoras, que são submetidas a pesquisas de similaridade contra bancos de dados e, por fim, a informação é colocada em um contexto biológico pela relação entre os genes em vias metabólicas. No entanto, a velocidade com que sequências genômicas são geradas pelas técnicas sequenciamento de segunda geração tem também dificultado a revisão das anotações por parte dos curadores (RUST *et al*., 2002; MÉDIGUE & MOSZER, 2007).

Na etapa de predição gênica, o programa comumente utilizado para anotar genomas de procariotos é o GLIMMER (do inglês "*Gene Locator and Interpolated Markov Modeler*" - DELCHER *et al*., 1999), cuja abordagem utiliza um conjunto de sequências de referência para treinar um modelo de predição (que são padrões obtidos através do modelo oculto de Markov) e então esse modelo é aplicado a uma sequência de interesse. Outros programas podem predizer regiões codificantes utilizando diretamente a informação de homologia da sequência, em relação a sequências presentes em bancos de dados (STOTHARD & WISHART, 2006; RICHARDSON & WATSON, 2012).

Entretanto, essa informação é mais utilizada na etapa posterior, que é atribuir funções às proteínas codificadas. Para isso são usadas ferramentas de alinhamento, tal como o BLAST (do inglês "*Basic Local Alignment Search Tool*" - ALTSCHUL *et al*., 1997), que busca proteínas similares à proteína de interesse em bancos de dados, como o UniProt (APWEILER *et al*., 2004). Identificadas as proteínas similares, as informações anotadas são transferidas para a anotação de proteínas homólogas. Além disso, outra estratégia utilizada para predizer a função de uma proteína é a busca por domínios conservados (RUST *et al*., 2002; STOTHARD & WISHART, 2006; RICHARDSON & WATSON, 2012).

Uma vez identificadas as funções das proteínas, as vias metabólicas são reconstruídas por um GEM (do inglês "*Genome-scale metabolic model"*, modelo metabólico de escala genômica), que é gerado com base em dados genômicos conhecidos e que pode ser aplicado a outros genomas (SANTOS *et al*., 2011; LAND *et al*., 2015). Um exemplo é o modelo produzido pelo *Project to Annotate* 1000 *Genomes* (projeto para anotar 1000 genomas), que consiste na anotação de subsistemas (vias metabólicas) realizada por especialistas, bem como na organização e na disponibilização dessas informações (OVERBEEK *et al*., 2014). O banco de dados gerado desse projeto é o SEED (OVERBEEK *et al*., 2014), e uma

ferramenta de anotação que faz uso dele é a plataforma RAST (do inglês "*Rapid Annotation using Subsystem Technology*" – AZIZ *et al*., 2008). Outro projeto desse tipo é o banco de dados integrado KEGG (do inglês "*Kyoto Encyclopedia of Genes and Genomes"* – KANEHISA et al., 2012), utilizado pela ferramenta de anotação KAAS (do inglês "KEGG *Automatic annotation server*" – MORIYA *et al***.**, 2007). Um fluxograma das etapas de anotação e da construção de um GEM pode ser visualizado na FIGURA 2.6.

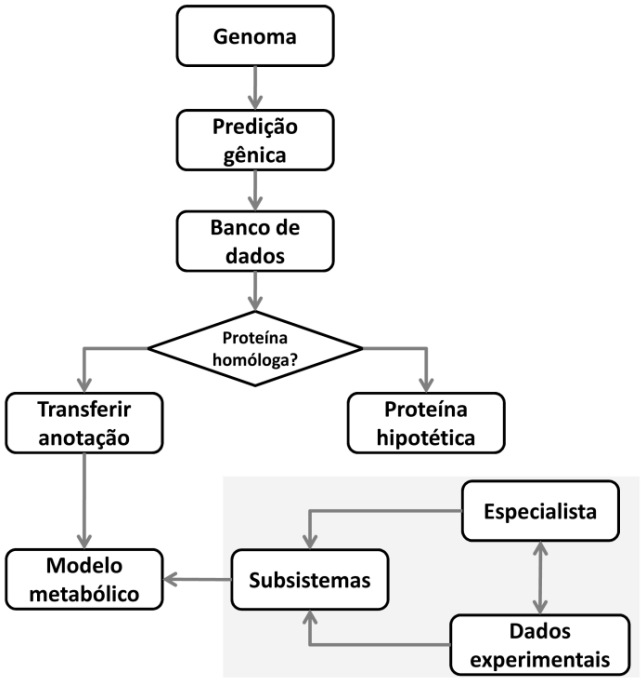


FIGURA 2.6: ETAPAS DE UMA ANOTAÇÃO GENÔMICA DE PROCARIOTO

O quadro em destaque se refere a um GEM, aplicado para posicionar as proteínas anotadas em vias metabólicas.

FONTE: o autor (2015), baseado em RICHARDSON & WATSON (2012) e SANTOS *et al*. (2011)

Enquanto esse fluxo de processos se atenta a predizer principalmente as regiões codificadoras de proteínas, programas de predição específica são utilizados para identificar rRNAs e tRNAs, como o tRNA-scan SE (LOWE & EDDY, 1997) e o RNAmmer (LAGESEN *et al*., 2007). Outros programas também podem ser acoplados ao processo, por exemplo, para a predição de regiões de transferência horizontal de genes (HGT – do inglês "*horizontal gene transfer"*, tais como ilhas de patogenicidade e regiões de fagos). Essas regiões podem ser identificadas pela diferença da composição de códons e do conteúdo G+C em relação ao restante do genoma e, frequentemente, são acompanhadas de integrases, transposases e

sequências de inserção (IS – do inglês “*insertion sequence”*) (RICHARDSON & WATSON, 2012).

2.2.2 Comparação genômica

O sequenciamento genômico de bactérias demonstrou que a variabilidade gênica entre as espécies é maior que o esperado e que até mesmo estirpes de uma mesma espécie podem diferir em 25% em seu conteúdo gênico (PALLEN & WREN, 2007).

Essa variabilidade genética está relacionada ao fato de os genomas bacterianos serem dinâmicos, pois eles apresentam regiões de duplicações, inversões, transposições, recombinações, inserções e deleções, as quais interferem nas capacidades metabólicas desses organismos e, consequentemente, nos seus estilos de vida. Eles podem ainda adquirir genes por transferência horizontal de genes (HGT), talvez a maior fonte de diversidade para os genomas bacterianos, que é geralmente mediada por plasmídeos, ilhas de patogenicidade e fagos (PALLEN & WREN, 2007; GUZMÁN *et al*., 2008).

Dessa forma, é esperado que uma espécie de bactéria não possa ser descrita por um conteúdo gênico fixo, pois novos genes devem ser encontrados em cada nova estirpe isolada. A esse montante de genes encontrados é dado o nome de pangenoma (FIGURA 2.7), que consiste de um núcleo de genes comum a todas as estirpes (chamado de *core* genoma) e de um conjunto de genes que diferencia as estirpes umas das outras (chamado de genoma acessório) (MEDINI *et al*., 2008; LAING *et al*., 2010; VAN TONDER *et al*., 2014).

Ao *core* genoma são atribuídos os genes relacionados com a informação e aqueles considerados essenciais, provavelmente herdados de um ancestral comum e pouco propensos a HGT. Ao genoma acessório são atribuídos os genes relacionados com a adaptação e a colonização de determinados ambientes, portanto responsáveis pelo estilo de vida dos organismos. Dentro desse grupo estão também os genes específicos de cada estirpe, provavelmente adquiridos por HGT (MEDINI *et al*., 2008; SCHLEIFER, 2009; KANG *et al*., 2014). Embora inicialmente utilizado para espécies, alguns autores tem usado o conceito de pan e *core* genoma para comparações em nível de gênero (LEFÉBURE & STANHOPE, 2007;

LUKJANCENKO *et al*., 2010). Eles também foram utilizados para descrever conjuntos gênicos de todo o domínio *Bacteria* (LAPIERRE & GOGARTEN, 2009).

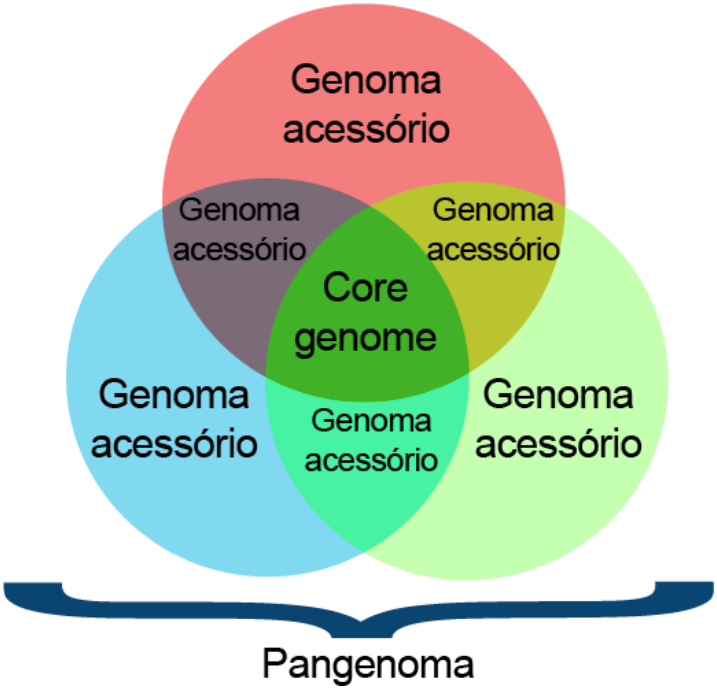


FIGURA 2.7: COMPARAÇÃO DE CONJUNTOS GÊNICOS

O *core* genoma se refere à intersecção de todos os conjuntos e o pangenoma à união deles.

FONTE: o autor (2015)

A diversidade gênica de um grupo taxonômico pode ser medida pela análise de homologia entre os conjuntos gênicos/proteômicos presentes nos genomas avaliados. Uma das estratégias utilizadas para isso é o BLAST 'todos contra todos' (*All-against-all* BLAST, do inglês), que consiste em confrontar esses conjuntos para avaliar quantas e quais proteínas são compartilhadas entre os organismos. Essa análise pode ser visualizada em forma de matriz de homologia, chamada de matriz BLAST (BLAST *matrix*) (FIGURA 2.8A). A partir do BLAST todos contra todos também é possível verificar os genes comuns a todos os genomas (*core* genoma), bem como todo o conjunto gênico/proteômico analisado (pangenoma) (BINNEWIES *et al*., 2006; LUKJANCENKO *et al*., 2012).

Evidências da dinâmica genômica podem ser observadas por comparações genômicas estruturais. A análise da ordem dos genes, ou estrutura dos genomas, pode indicar regiões de translocações ou inversões, bem como a análise de presença ou ausência de genes pode indicar aquisição/perda. Uma abordagem utilizada para visualizar a comparação estrutural de genomas é através do BLAST atlas, na qual genomas são representados em círculos para indicar neles a ausência

ou a presença de regiões genômicas homólogas a uma referência (FIGURA 2.8B) (BINNEWIES *et al*., 2006; EDWARDS & HOLT, 2013).

A disponibilização de sequências genômicas também tem permitido novas metodologias para a classificação taxonômica de bactérias. Abordagens laboratoriais, como a hibridização DNA-DNA (DDH, do inglês "DNA-DNA *hibridization*") e a identidade do gene 16S rRNA, têm sido substitídas pela utilização da sequência genômica completa *in silico*. Isso inclui o método GGDH *in silico* (do inglês "*Genome-to-Genome Distance Hybridization"*, ou distância de hibridização entre genomas) análogo à DDH, porém realizada por alinhamento de sequências. Esse método considera que duas espécies são distintas se seus genomas tiverem uma hibridização inferior a 70% (mesmo valor de corte utilizado na técnica de DDH) (THOMPSON *et al*., 2013). Outro método utilizado é a ANI (do inglês "*Average Nucleotide Identity"*, ou identidade média de nucleotídeos – GORIS *et al*., 2007), que mede a distância genética entre genomas através de regiões homólogas obtidas com o programa BLAST. Dessa forma, duas espécies são distintas se a ANI entre os genomas for inferior a 95% (THOMPSON *et al*., 2013; LAND *et al*., 2015).

2.3 Estudos genômicos de *Herbaspirillum*

Devido à importância agrícola do gênero *Herbaspirillum*, trabalhos genômicos envolvendo esse gênero já foram realizados. Com isso, a estirpe SmR1 de *Herbaspirillum seropedicae* teve seu genoma completamente finalizado, o qual consiste em um único cromossomo circular, de aproximadamente 5,5 Mb (WEISS, 2010, PEDROSA *et al.* 2011). Características gerais desse genoma podem ser vistas na TABELA 2.6.

Com relação a aspectos de seu metabolismo, essa bactéria apresenta o agrupamento de genes *nif* (*nifA*, *nifB*, *nifZ*, *nifZ1*, *nifH*, *nifD*, *nifK*, *nifE*, *nifN*, *nifX*, *nifQ*, *nifW*, *nifV*, *nifU* e *nifS*), relacionado à fixação de nitrogênio (PEDROSA *et al.,* 2011). Além disso, genes relacionados com os sistemas T2SS (sistema de secreção do tipo II), T5SS (sistema de secreção do tipo V), T4SS *pili* (sistema de secreção do tipo IV *pili*), T1SS (sistema de secreção do tipo I), T3SS (sistema de secreção do tipo III) e T6SS (sistema de secreção do tipo VI), envolvidos na interação planta/bactéria, também estão presentes no genoma dessa bactéria (PEDROSA *et al.,* 2011; MONTEIRO *et al*., 2012).

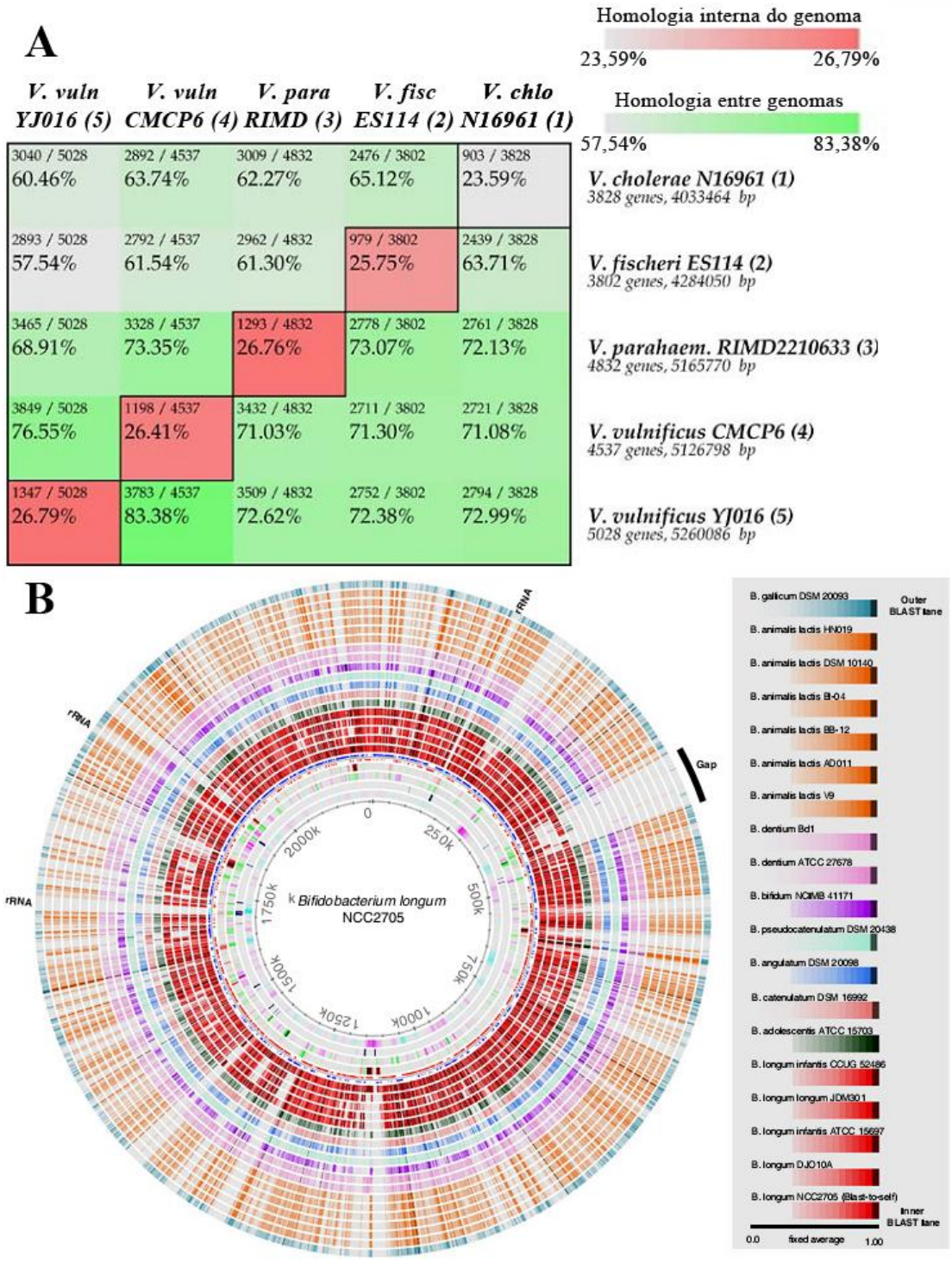


FIGURA 2.8: EXEMPLOS DE COMPARAÇÕES GENÔMICAS

Em A, é mostrada uma matriz BLAST comparando genomas de bactérias do gênero *Vibrio*, onde a homologia é representada como gráfico de calor. A diagonal do centro em vermelho (“Homologia interna do genoma”) representa o número de parálogos de cada genoma. Em B, é mostrado um BLAST atlas do gênero *Bifidobacterium*, que tem como referência *B. longum* NCC2705.

FONTE: BINNEWIES *et al*. (2006) e LUKJANCENKO *et al*. (2012)

TABELA 2.6: CARACTERÍSTICAS GERAIS DO GENOMA DE *H. seropedicae* SmR1

Tamanho (pb)	5.513.887
G+C%	63,4
Número total de genes	4804
Número total de CDS	4735
Tamanho médio das ORFs (pb)	1028,8
Regiões codificantes de proteínas (%)	88,3
rRNA operons	3
tRNAs	55
Genes com atribuição funcional	3108
Genes com predição funcional	497
Função desconhecida	1130
Regiões de possível transferência horizontal	18

FONTE: PEDROSA *et al.* (2011)

Com o avanço das tecnologias de sequenciamento genômico, novos genomas de bactérias do gênero *Herbaspirillum* foram sequenciados e publicados na forma de *draft*. Entre eles, os genomas das estirpes Os34 e Os45 de *H. seropedicae*, isoladas de raízes de arroz, cada um contendo 253 e 144 *contigs*, respectivamente. Nesses genomas foram encontrados genes relacionados com a fixação biológica de nitrogênio, síntese de ácido indol-acético (IAA, do inglês "*indol acetic acid*"), produção de sideróforos, ACC (1-aminociclopropano-1-carboxilato) deaminase e os sistemas T3SS, T6SS e T4SS *pili* (YE *et al*., 2012; ZHU *et al*., 2012).

A segunda espécie de *Herbaspirillum* a ter seu genoma publicado foi *H. lusitanum* estirpe P6-12, o qual consiste em um *draft* de 820 *contigs* e 37 *scaffolds*. Nele foi encontrado um gene que codifica para uma subunidade de RuBisCO, mas sua função não foi determinada. Genes *nif* estão ausentes (WEISS *et al.,* 2012). Atualmente esse genoma possui uma versão atualizada que compreende 20 *contigs* (WEISS, Informação verbal).

Em seguida, foi publicado o *draft* genômico de *Herbaspirillum huttiense* subsp. *putei* estirpe IAM 15032, que consiste de um *scaffold* e 73 *contigs*, nos quais foram encontrados genes que codificam para proteínas dos sistemas T1SS, T4SS e T5SS, enzimas responsáveis pelo metabolismo de celulose e a ACC deaminase (DE SOUZA *et al.*, 2013).

Outra espécie que teve seu *draft* genômico publicado foi *H. frisingense* estirpe GSF30, o qual é composto por 93 *contigs*. A sequência genômica desse organismo

apresenta os genes *nif,* os genes relacionados com o metabolismo de celulose e os genes relacionados com o T6SS (STRAUB *et al*., 2013)**.**

Cinco estirpes de *Herbaspirillum*, sem espécie descrita, também tiveram seus *drafts* genômicos publicados: a estirpe GW103, obtida da rizosfera de caniço-de-água (*Phragmites australis*), com *draft* composto por 1 *scaffold* e 6 *contigs* (LEE *et al*., 2012); as estirpes CF444 e YR522, obtidas da endosfera de *Populus deltoides* (algodão americano) com *drafts* compostos por 125 e 168 *contigs*, respectivamente (BROWN *et al*., 2012); a estirpe JC206, posteriormente classificada como *Herbaspirillum massiliense*, isolada de fezes de um paciente saudável, cujo *draft* compreende 27 *contigs* (LAGIER *et al*., 2012); e a estirpe RV1423, isolada de água contaminada com hidrocarbonetos, com *draft* formado por 131 *contigs* (JAUREGUI *et al*., 2014).

Além desses genomas, estão presentes no banco de dados de genomas do NCBI 3 *drafts* genômicos de baixa qualidade de estirpes ainda não classificadas: *Herbaspirillum* sp. B501 (1.546 *contigs*), *Herbaspirillum* sp. B65 (1.249) e *Herbaspirillum* sp. B39 (1.174). Essas estirpes foram obtidas de análises metagenômicas de arroz.

O Núcleo de Fixação Biológica de Nitrogênio (NFN) da Univesidade Fereral do Paraná (UFPR) tem se esforçado para sequenciar os genomas de todas as espécies de *Herbaspirillum*, bem como algumas estirpes de relevância para o conhecimento da diversidade gênica do grupo.

Dessa forma, a estirpe clínica AU14040 (obtida do trato respiratório humano) de *H. seropedicae* teve um *draft* genômico produzido (52 *contigs*), no qual não foram encontrados os genes *nif*, nem os genes relacionados ao T3SS, presentes em *H. seropedicae* SmR1. Por outro lado, foram encontrados um *operon* de síntese de parede celular, ausente na estirpe SmR1, e transposases, que podem estar envolvidas com a mudança de *habitat* dessa bactéria. Recentemente, sua sequência genômica foi finalizada (FAORO, *in prep*).

A estirpe M1 de *Herbaspirillum rubrisubalbicans* já tem seu genoma finalizado. De maneira geral, o conjunto e a ordem dos genes são similares ao que é observado no genoma de *H. seropedicae* SmR1. Com isso, o genoma de *H. rubrisubalbicans* M1 também apresenta genes relacionados aos T1SS, T2SS, T3SS, T5SS, T6SS e os genes *nif*. Além disso, foram encontrados genes relacionados com a biossíntese

de celulose e várias regiões de repetição, principalmente transposases e integrases (CARDOSO, 2011; MONTEIRO *et al*., 2012; SOUZA, *in prep*).

Um *draft* genômico da estirpe N3 de *Herbaspirillum hiltneri* também foi produzido (160 *contigs*), onde foram encontrados os T1SS, T3SS, T4SS, T5SS e genes que codificam para a RuBisCO e para a ACC deaminase. Recentemente esse genoma também foi finalizado (GUIZELINI, *in prep*).

A EMBRAPA (Empresa Brasileira de Pesquisa Agropecuária), também tem sequenciado o genoma de bactérias de interesse agrícola, entre eles, foram produzidos *drafts* genômicos das estirpes BR11417 e BR11335 de *H seropedicae*, além da estirpe BR11504 de *H. rubrisubalbicans* (BALDANI, Informação verbal).

2.3.1 Genômica comparativa de Herbaspirillum

Como *H. seropedicae* SmR1 é o *Herbaspirillum* melhor estudado e é o único até o presente momento que possui genoma completo publicado, ele é usado como referência para o grupo e, em geral, os genomas dos demais *Herbaspirillum* são comparados a ele. Com isso, as primeiras comparações genômicas de *Herbaspirillum* ocorreram entre o genoma de *H. seropediace* SmR1 e o *draft* genômico de *H. rubrisubalbicans* M1. Essas comparações mostraram que esses dois genomas são estruturalmente similares, mas apresentam uma inversão na região central do cromossomo (FIGURA 2.9A) (CRUZ *et al.*, 2010; CARDOSO, 2011).

Enquanto a montagem genômica de *H. rubrisubalbicans* M1 estava em andamento, Monteiro e colaboradores (2012) utilizaram uma abordagem diferente e compararam esses dois genomas através da hibridização subtrativa por supressão (SSH, do inglês “*Suppression Subtactive Hybridization*”). Os resultados obtidos mostraram a presença de um conjunto de genes relacionados à biossíntese de celulose, além de genes relacionados com lipopolissacarídeos (LPS) e adesinas, presentes no genoma de *H. rubrisubalbicans* M1, mas ausentes no genoma de *H. seropedicae* SmR1 (FIGURA 2.9B).

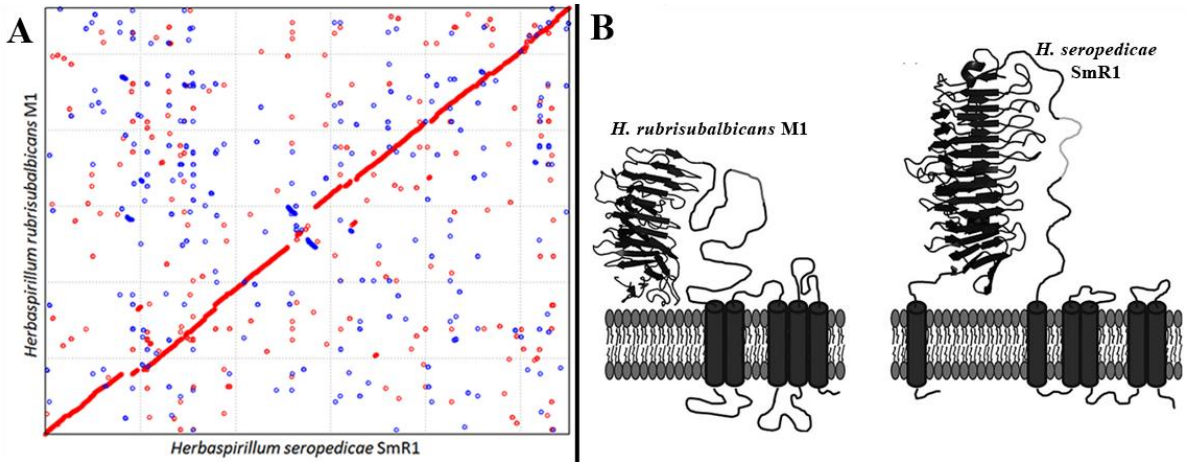


FIGURA 2.9: COMPARAÇÕES ENTRE OS GENOMAS DE *H. seropedicae* SmR1 E *H. rubrisubalbicans* M1

Em A, é mostrado um gráfico *dotplot*, produzido com os programas PROmer e MUMmerplot (KURTZ *et al*, 2004), entre os genomas de *H. seropedicae* SmR1 e *H. rubrisubalbicans* M1, onde pontos/linhas representam homologia entre os genomas. Em B, são mostrados os modelos de adesinas que são diferentes para *H. rubrisubalbicans* M1 e *H. seropedicae* SmR1.

FONTE: o autor (2015) e MONTEIRO *et al*. (2012)

Diante dos dados genômicos de *Herbaspirillum* disponibilizados nos últimos anos, Straub e colaboradores (2013) publicaram o que pode ser considerada de fato a primeira comparação genômica de *Herbaspirillum*, a qual envolveu 10 genomas de bactérias desse gênero (e mais 4 bactérias endofíticas) e os dados genômicos de *H. rubrisubalbicans* M1 obtidos da SSH. Embora focado no genoma de *H. frisingense* GSF30, o trabalho concluiu que os genomas das três espécies fixadoras de nitrogênio (*H. seropedicae*, *H. rubrisubalbicans* e *H. frisingense*) apresentam aspectos similares entre si e com os genomas de *H. huttiense* subsp. *putei* e *Herbaspirillum* sp. GW103. A comparação de genes por categorias pode ser visualizada na FIGURA 2.10A.

Análises filogenéticas baseadas no conjunto gênico desses genomas também foram realizadas. Elas reforçaram a relação já observada entre as *Herbaspirillum* spp. fixadoras de nitrogênio, mas também posicionaram as estirpes GW103 e CF444 dentro do grupo, de modo a sugerir que elas são estirpes de *H. huttiense* e *H. lusitanum*, repectivamente (FIGURA 2.10B) (WEISS, 2014).

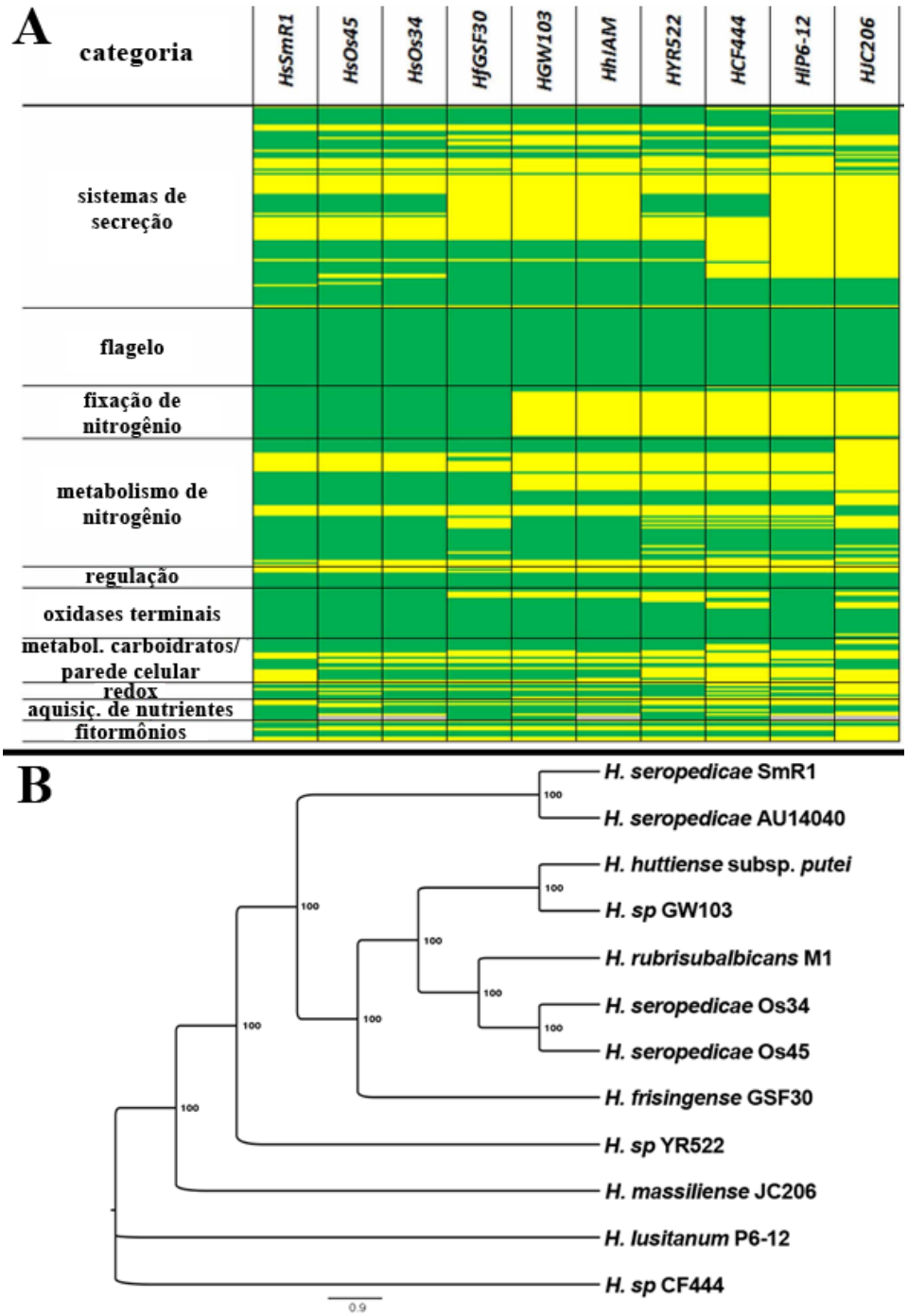


FIGURA 2.10: COMPARAÇÕES GENÔMICAS DE *Herbaspirillum* JÁ REALIZADAS

Em A, está representada uma matriz de ausência (amarelo) e presença (verde) de genes por categoria para vários genomas de *Herbaspirillum*. As siglas representam, respectivamente, *H. seropedicae* SmR1, *H. seropedicae* Os45, *H. seropedicae* Os34, *H. frisingense* GSF30, *Herbaspirillum* sp. GW103, *H. huttiense* subsp. *putei*, *Herbaspirillum* sp. YR522, *Herbaspirillum* sp. CF444, *H. lusitanum* P6-12 e *H. massiliense* JC206. Em B, é mostrada uma árvore filogenética baseada na concatenação do *core* genoma (768 genes) dos organismos analisados.

FONTE: adaptado de STRAUB *et al*. (2013) e PEDROSA *et al*. (2014)

3 JUSTIFICATIVA

Embora realizados estudos prévios, faltam estudos genômicos comparativos que esclareçam aspectos evolutivos, fisiológicos e taxonômicos do gênero *Herbaspirillum*. A diversidade de ambientes onde organismos desse gênero foram encontrados, aliada à importância econômica que algumas espécies representam, fornecem um amplo campo de pesquisa a respeito da relação entre as espécies do grupo.

Para isso, é também necessário que as sequências genômicas das espécies *H. autotrophicum*, *H. chlorophenolicum* e *H. rhizosphaerae* sejam obtidas, pois essas espécies apresentam características distintas das *Herbaspirillum* spp. comumente estudadas e devem ampliar o conhecimento que se tem a respeito do gênero através das análises genômicas comparativas.

Essas análises podem ser utilizadas para obter informações a respeito da conservação genética, da plasticidade genômica, da diversidade metabólica, da história evolutiva do gênero e grupos de genes. Além disso, essas análises podem resolver inconsistências taxonômicas e também identificar genes envolvidos em processos de interesse (como promoção do crescimento vegetal ou degradação de poluentes).

4 OBJETIVOS

4.1 Objetivo geral

- Comparar genomas de *Herbaspirillum* spp. e estabecer o conjunto gênico que está presente em todas as estirpes analisadas e a diversidade gênica do gênero

4.2 Objetivos específicos

- Montar e anotar o *draft* genômico da estirpe IAM 14942 de *H. autotrophicum*;
- Montar e anotar o *draft* genômico da estirpe CPW301 de *H. chlorophenolicum*;
- Montar e anotar o *draft* genômico da estirpe UMS-37 de *H. rhizosphaerae*;
- Anotar automaticamente os demais genomas de *Herbaspirillum* disponíveis;
- Comparar os genomas de *Herbaspirillum* spp. para obter o *core* e o pangenoma;
- Agrupar e classificar as estirpes de *Herbaspirillum* spp. com base no *core* e no pangenoma;
- Comparar genes e agrupamentos gênicos específicos, tais como o T3SS, T6SS, regiões de fago, genes *nif*, genes que codificam proteínas RuBisCO, RuBisCO-*like* e proteínas envolvidas nas vias de degradação de fenol;
- Utilizar a sequência genômica para a classificação taxonômica de estirpes de *Herbaspirillum* spp.

5 METODOLOGIA

5.1 Obtenção e avaliação dos conjuntos de dados de sequenciamento

Os genomas de *H. autotrophicum* IAM 14942, *H. rhizosphaerae* UMS-37 e *H. chlorophenolicum* CPW301 foram sequenciados nas plataformas SOLiD4 (Life Technologies - https://www.lifetechnologies.com/br/en/home.html) e Illumina MiSeq (Illumina - http://www.illumina.com) pelo Núcleo de Fixação Biológica de Nitrogênio (NFN) da UFPR. Os *reads* SOLiD (50 pb – pares de base) foram obtidos sem nenhum tipo de pareamento e, por isso, chamados de 'fragmentos' (*single-end*), enquanto os *reads* Illumina (~250 pb) foram obtidos na forma pareada (*pair-end*), porém com uma sobreposição esperada de 50 bases entre os pares (e não separados por centenas de bases como no conceito de *pair-end* tradicional).

Os *reads* obtidos por sequenciamento na plataforma SOLiD foram avaliados quanto à sua qualidade pelo programa *QualityAssessment* 0.5 (RAMOS *et al*., 2011), que é uma aplicação em JAVA e permite realizar o processo de *trimming* de qualidade e recalcular a cobertura genômica diante do subconjunto gerado por esse processo. Para o *trimming* de qualidade foram removidos os *reads* com qualidade média inferior a *phred* 20, valor considerado de baixa qualidade (ANDREWS, 2010). O *trimming* de tamanho foi realizado com auxílio de *scripts* desenvolvidos em linguagem Perl (https://www.perl.org), cedidos gentilmente pelo Dr. Vinicius Weiss do programa de Pós-gradução em Bioinformática da UFPR, com o qual foram descartadas as 15 últimas bases dos *reads*.

A qualidade dos *reads* Illumina foi avaliada pelo programa FastQC (ANDREWS, 2010), que rapidamente permite verificar se os *reads* apresentam algum tipo de problema e também fornece gráficos que ajudam a observar a qualidade do conjunto de dados. Para o *trimming* dos *reads* Illumina foi utilizada a plataforma CLC *Genomic Workbench* 6.0 (CLC Bio - http://www.clcbio.com), que é um ambiente para análise e visualização de dados provenientes de sequenciadores de nova geração, além de incorporar ferramentas típicas para a análise dessas sequências. Com essa plataforma de análise foram removidas as 50 últimas bases dos *reads* e foram removidos os *reads* com valor de qualidade média inferior a 0,5, segundo os parâmetros padrão do programa.

5.2 Montagem e anotação genômica

A montagem genômica com os *reads* provenientes da plataforma SOLiD foi realizada com o *pipeline de novo* (SOLiD *de novo acessory tools* 2.0 - Life Technologies), desenvolvido para otimizar as montagens genômicas realizadas com *reads* provenientes dessa plataforma. Entre os programas utilizados no *pipeline*, é possível destacar 4 deles: SAET (do inglês "*SOLiD Accuracy Enhancement Tool*"), que realiza a correção dos *reads* que apresentam problemas; *Preprocessor*, que cria e prepara os arquivos para o montador genômico; *Velvet* (ZERBINO, 2010), que é o programa responsável pela montagem de sequências curtas, baseado na construção de grafos de-Brujin; e ASiD (do inglês "*Assembly Assistent for* SOLiD"), que realiza o fechamento de *gaps* dentro de *scaffolds*. Como os *reads* SOLiD eram fragmentos e não seriam formados *scaffolds*, o programa ASiD foi desativado.

Durante a execução do *pipeline de novo*, o parâmetro *hsize* (referente ao programa *Velvet*) foi modificado diversas vezes para a obtenção de diferentes montagens. Esse parâmetro corresponde ao *k-mer* utilizado para construir as tabelas de dispersão (*hashes*) e gerar os grafos de-Brujin. É também considerado o parâmetro mais importante na execução do programa (ZERBINO, 2010).

Para a montagem genômica, tanto com os *reads* provenientes da plataforma SOLiD quanto com os *reads* provenientes da plataforma Illumina, foi utilizada a plataforma CLC *Genomic Workbench* 6.0. O programa para montagem genômica presente na plataforma CLC também utiliza grafos de-Brujin, onde o conceito de *words* equivale aos *hashes* do *pipeline de novo*/*Velvet*. Dessa forma, o parâmetro *wsize* (*word size*) foi alterado para produzir diferentes montagens. A plataforma CLC também foi utilizada para o mapeamento de *reads* SOLiD em genomas de outras estirpes de *Herbaspirillum* spp. A obtenção desses genomas será tratada adiante no tópico 5.3. Os parâmetros para o mapeamento foram 0,8 de similaridade do *read* em relação à referência e 0,5 de tamanho de *read* alinhado em relação à referência, conforme a opção padrão do programa de mapeamento.

Os *reads* provenientes da plataforma Illumina foram também submetidos à montagem genômica pelo programa Newbler 4.0 (também chamado de GS Assembler; 454 *Sequencing* – http://www.454.com), que foi desenvolvido inicialmente para a montagem genômica com dados provenientes da plataforma Roche/454, mas posteriormente foi adaptado para realizar montagens híbridas

incluindo arquivos FASTA e FASTQ. O programa conta com um ambiente gráfico amistoso, de fácil utilização, e gera diversos arquivos de saída contendo estatísticas de montagem, além da saída em formato ACE para visualização da montagem. O parâmetro "*overlap length*" do programa foi reduzido de 40 (padrão) para 20, visto que essa sobreposição é suficiente para unir *contigs* (CARDOSO, 2011).

O programa Newbler 4.0 também foi utilizado para fazer montagens híbridas com os *contigs* das montagens previamente obtidas, pelo uso dos demais programas. Para isso, os *contigs* obtidos previamente foram fragmentados em 500 pb (tamanho adaptado aos *reads* 454, para os quais o programa foi criado), com sobreposição de 50 pb (para auxiliar a remontagem) e, com isso, transformados em *reads* falsos. O processo de fragmentação foi realizado com um *script* desenvolvido em linguagem de programação Python 2.x (https://www.python.org).

Os arquivos ACE gerados pelo programa Newbler foram visualizados no programa Consed 20.0 (GORDON *et al*., 1998), que é utilizado para finalização de montagens genômicas através de edição manual das sequências. A interface gráfica e a forma como os arquivos ACE são gerados permite encontrar regiões repetitivas no genoma e resolvê-las, se possível (CARDOSO, 2011).

Os *contigs* obtidos nas montagens realizadas foram mapeados em relação a um genoma de referência com a utilização do pacote de programas MUMmer 3.0 (MUM do inglês "*Maximal Unique Match*"– KURTZ *et al*, 2004). Dentro desse pacote foram utilizados os programas PROmer (programa para alinhamento de sequências ao nível de proteínas, mais sensível que o alinhamento ao nível de nucleotídeo) e MUMerplot (programa para gerar gráficos do tipo *dotplot* baseados nos alinhamentos realizados pelo PROmer).

A anotação dos genomas foi realizada com a plataforma RAST 2.0 (AZIZ *et al*., 2008). Essa plataforma funciona *on line* e identifica os genes com as ferramentas tRNA-scan-SE (LOWE & EDDY, 1997), *search-for-rnas* (LARSEN, não publicado) e GLIMMER 2.0 (DELCHER *et al*., 1999). Os genes identificados são anotados com base na homologia em relação a genes presentes em modelos de subsistemas (inter-relações gênicas que podem ser entendidas como vias metabólicas) presentes no banco de dados SEED (OVERBEEK *et al*., 2014). Foram utilizados os parâmetros padrão da plataforma.

Os genes foram também anotados pela ferramenta KAAS (MORIYA *et al*., 2007). Esse programa, também disponível *on line*, identifica as funções dos genes

através de busca por similaridade no banco de dados KEGG (KANEHISA *et al*., 2012) e utiliza por padrão o método BBH (do inglês "*bi-directional best hit*"). Como as análises são realizadas com base no banco de dados KEGG, onde os genes já possuem sua função assinada, os genes recém-anotados são posicionados em mapas metabólicos, de forma semelhante ao que a plataforma RAST faz.

A lista de organismos passada ao programa KAAS para ser realizada a anotação foi: eco, sty, hin, pae, nme, hpy, rpr, mlo, bsu, sau, lla, spn, cac, mge, mtu, ctr, bbu, syn, aae, mja, afu, pho, ape, mmt, rso, cti, bma, bpe, axy, vei, vap, ctt, har, hse, bja, azl, cch, put, aav, xcb. Essa lista corresponde aos organismos utilizados pela opção do programa para a anotação de procariotos, com adição de organismos com os quais as proteínas de *Herbaspirillum* spp. tiveram o melhor alinhamento quando submetidas à análise com o programa BLAST contra o banco de dados NR (do inglês "*non-redundant*", ou não redundante) do NCBI (CRUZ *et al*., 2012).

As funções dos genes também foram identificadas localmente por pesquisas de similaridade através do programa BLAST (ALTSCHUL *et al*., 1997) contra o banco de dados COG (do inglês "*Cluster of Orthologous Groups*" - TATUSOV *et al*., 2003), que possui um sistema de classificação funcional baseado em 4 categorias principais (Armazenamento e processamento da informação, sinalização e processos celulares, metabolismo, e proteínas pobremente caracterizadas) e outras 25 subcategorias. O banco de dados COG foi obtido da página de *internet* 'COGs *Phylogenetic classification of proteins encoded in complete genomes*' (http://www.ncbi.nlm.nih.gov/COG).

Para a confecção de árvores filogenéticas de proteínas específicas, foi utilizada a versão *on line* do BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) e o banco de dados NR do NCBI, do qual foram extraídas sequências homólogas. As sequências em multi-FASTA foram submetidas à plataforma MEGA 6.0 (do inglês "*Molecular Evolutionary Genetics Analysis*" – TAMURA *et al*., 2013), dentro da qual foram alinhadas com o programa MUSCLE (do inglês "*MUltiple Sequence Comparison by Log-Expectation*" – EDGAR, 2004), com parâmetros padrão do programa, e as árvores foram produzidas pelo método *Maximum Likelihood*, com 1.000 replicatas de *bootstrap*.

5.3 Identificação de homólogos e comparação genômica

Para a comparação genômica de estirpes de *Herbaspirillum* spp. foram utilizados os genomas de: *H. autotrophicum* IAM 14942, *H. chlorophenolicum* CPW301 e *H. rhizosphaerae* UMS-37, montados e anotados neste trabalho; *H. rubrisubalbicans* M1, cedido gentilmente pelo Dr. Emanuel Maltempi de Souza (NFN – UFPR); *H. seropedicae* AU14040, cedido gentilmente pelo Dr. Helisson Faoro (NFN – UFPR; Intituto Carlos Chagas – Fiocruz Curitiba, PR); *H. hiltneri* N3, cedido gentilmente pelo MSc. Dieval Guizelini (NFN – UFPR; Setor de Educação Profissional e Tecnológica – UFPR); *H. lusitanum* P6-12, cuja montagem atualizada foi cedida gentilmente pelo Dr. Vinicius Weiss (Programa de Pós-graduação em Bioinformática – UFPR); *H. rubrisubalbicans* BR11504, *H. seropedicae* BR11335 e *H. seropedicae* BR11417, gentilmente cedidos pelo Dr. José Ivo Baldani (EMBRAPA Agrobiologia – Seropédica, RJ) e finalizados pelo Dr. Roberto Raittz (Setor de Educação Profissional e Tecnológica – UFPR; Programa de Pós-graduação em Bioinformática – UFPR). Os genomas já publicados foram obtidos do NCBI Genbank: *H seropedicae* SmR1 (NC_014323.1), *H. seropedicae* Os34 (AMSB00000000.1), *H. seropedicae* Os45 (AMSA00000000.1), *H. frisingense* GSF30 (NZ_AEEC00000000.2), *H. huttiense* subsp. *putei* IAM 15032 (referido nesse trabalho apenas como *H. huttiense* subsp. *putei*; ANJR00000000.1), *Herbaspirillum* sp. YR522 (NZ_AKJA00000000.1), *Herbaspirillum* sp. CF444 (NZ_AKJW00000000.1), *Herbaspirillum* sp. GW103 (NZ_AJVC00000000.1) e *Herbaspirillum* sp. JC206 (ou *H. massiliense* JC206; NZ_CAHF00000000.1), que podem ser encontrados pelos números de acesso no banco de dados genômico do NCBI.

As informações sobre esses genomas podem ser vistas na TABELA 5.1. Muitas vezes neste trabalho, principalmente nas imagens, esses organismos serão referenciados por uma abreviatura que também se encontra na TABELA 5.1.

Os genomas e *drafts* genômicos de *Herbaspirillum*, em formato FASTA ou multi-FASTA, foram todos anotados pela plataforma RAST para a padronização da anotação, com exceção do genoma de *H. seropedicae* SmR1 (completamente finalizado e anotado). Os genes anotados foram extraídos em sequências FASTA de aminoácidos, que correspondem aos proteomas das estirpes analisadas. Esses proteomas foram utilizados para criar um banco de dados para o programa BLAST

e, posteriormente, submetidos a uma busca por similaridade do tipo BLAST todos contra todos (LUKJANCENKO & WASSENAAR, 2010).

TABELA 5.1**:** GENOMAS DE *Herbaspirillum* UTILIZADOS COMO CONJUNTO DE DADOS

Organismo (abreviatura)	**Ambiente**	**Número de contigs**	**Referência**
H. seropedicae SmR1 (HseroSmR1)	Endofítico – gramíneas	1	PEDROSA *et al*., 2011
H. seropedicae AU14040 (HseroAU14040)	Trato respiratório – humano	52	FAORO, *in prep*.
H. seropedicae Os34 (HseroOs34)	Rizosfera – arroz	252	YE *et al*., 2012
H. seropedicae Os45 (HseroOs45)	Rizosfera – arroz	144	ZHU *et al*., 2012
H. seropedicae BR11335 (HseroBR335)	Endofítico – gramíneas	1	BALDANI, não publicado
H. seropedicae BR11417 (HseroBR417)	Endofítico – gramíneas	1	BALDANI, não publicado
H. rubrisubalbicans BR11504 (Hrubr504)	Endofítico – gramíneas	1	BALDANI, não publicado
H. rubrisubalbicans M1 (HrubrM1)	Endofítico – cana-de-açúcar	1	SOUZA, *in prep.*
H. huttiense subsp. *putei* IAM 15032 (Hhputei)	Água de poço/água subterrânea	37	DE SOUZA *et al*., 2013
H. frisingense GSF30 (HfrisGSF30)	Endofítico – *Miscanthus*	93	STRAUB *et al*., 2013
H. lusitanum P6-12 (HlusiP612)	Nódulos de raiz – feijão	20	WEISS *et al*., 2012 (atualização não publicada)
H. hiltneri N3 (HhiltN3)	Superfície de raiz – trigo	160	GUIZELINI, *in prep*.
H. massiliense JC206 (HmassJC206)	Fezes – humano	27	LAGIER *et al*., 2012
H. sp GW103 (HGW103)	Rizosfera – *Phragmites*	6	LEE *et al*., 2012
H. sp YR522 (HYR522)	Endosfera – *Populus*	168	BROWN *et al*., 2012
H sp. CF444 (HCF444)	Endosfera – *Populus*	125	BROWN *et al*., 2012
H. autotrophicum IAM 14942 (Hauto14942)	Lago eutrófico	99	Este trabalho
H. chlorophenolicum CPW301 (HchloCPW301)	Sedimento/região industrial	192	Este trabalho
H. rhizosphaerae UMS-37 (HrhizUMS37)	Rizosfera – *Allium*	55	Este trabalho

FONTE: o autor (2015), com base nas referências da tabela

O arquivo de saída do programa, em formato tabulado, foi tratado para estabelecer a homologia das proteínas e agrupar as famílias de proteínas, baseadas nos dois critérios de Lukjancenko & Wassenaar (2010): 1- o alinhamento obtido deve ter cobertura igual ou superior a 50% em relação ao tamanho da maior proteína; e 2- a identidade entre os aminoácidos deve ser igual ou superior a 50%.

O tratamento dos dados e as verificações de homologia foram realizados com auxílio de *scripts* desenvolvidos em linguagem Python 2.x. Depois de verificadas as homologias, um *pipeline* também desenvolvido em linguagem Python foi utilizado para: determinar as proteínas únicas codificadas para cada genoma; determinar o número de proteínas homólogas compartilhadas entre os pares e criar a matriz BLAST; verificar a acumulação de proteínas comuns a todos os conjuntos, de forma a obter o *core* genoma ao final do processo; verificar a acumulação de todo o conjunto de proteínas encontrado, de modo a obter o pangenoma ao final do processo; verificar a ausência e a presença de cada proteína que compõe o pangenoma e criar uma matriz binária com as marcações ‘1’ para presença e ‘0’ para a ausência de cada proteína.

A matriz binária obtida do pangenoma foi submetida ao programa PAST (do inglês “PAleontological STatistics” - HAMMER *et al*., 2001), onde foi agrupada pelo método *single linkage* com distância de Kimura (os quais apresentaram melhores agrupamentos) e 10.000 replicatas de *bootstrap*, para criar uma árvore filogenética, chamada de árvore do pangenoma (*pan-genome tree*) (SNIPEN & USSERY, 2010).

Os dados genômicos foram também utilizados para a realização de uma comparação geral baseada em BLAST atlas. Para isso foi utilizado o programa BRIG (do inglês “BLAST *Ring Image Generator*” - ALIKHAN *et al*., 2011), que é capaz de gerar uma comparação circular referente aos cromossomos procarióticos e permite comparar grupos genômicos a um organismo central que serve como referência. A referência utilizada foi o genoma de *H. seropedicae* SmR1, em formato *genbank* (GBK) e os demais genomas de *Herbaspirillum* comparados à referência foram utilizados em formato FASTA de aminoácidos. Para isso, foram utilizados os parâmetros padrão do programa, porém o BLAST interno do programa foi adaptado para o algoritmo tBLASTn.

Agrupamentos gênicos específicos de *Herbaspirillum* (regiões de fago, T3SS, T6SS e genes *nif*) foram analisados com auxilio do programa Artemis (CARVER *et al*., 2012), que fornece um ambiente gráfico para visualizar, analisar e anotar

genomas. As identidades das proteínas codificadas pelos genes que compõe esses agrupamentos foram obtidas a partir do BLAST todos contra todos, realizado previamente.

5.4 Análises taxonômicas baseadas na sequência genômica

A análise ANI (*Average Nucleotide Identity*) foi realizada com a ferramenta ANI Calculator (http://enve-omics.ce.gatech.edu/ani/), que funciona com base na metodologia de Goris e colaboradores (2007). Essa metodologia consiste basicamente na fragmentação dos genomas em sequências de 1.000 bases e alinhamento entre eles com o algoritmo BLASTn. O cálculo de ANI corresponde à média de identidade entre alinhamentos com mais de 30% de identidade e que possuem tamanho de pelo menos 70% da sequência total. Valores superiores a 95% de ANI foram utilizados para definir que dois organismos pertencem à mesma espécie (GORIS *et al*., 2007).

A análise GGDH (*Genome-to-Genome Distance Hybridization*) foi realizada com a ferramenta GGDC 2.0 (do inglês "*Genome-to-Genome Distance Calculator*" – http://ggdc.dsmz.de/distcalc2.php), que faz uso da abordagem GBDP (do inglês "*Genome Blast Distance Phylogeny*" - MEIER-KOLTHOF *et al*., 2013). Essa abordagem consiste em alinhar dois genomas utilizando o algoritmo BLASTn, porém os fragmentos são obtidos *a posteriori* (ao contrário da ANI) e correspondem diretamente aos alinhamentos produzidos. As informações dos alinhamentos são então utilizadas em três fórmulas de distância.

A fórmula 1 é a razão entre a soma de todos os pares de sequências homólogas e a soma dos tamanhos dos genomas. A fórmula 2 é a razão entre a soma de todos os nucleotídeos idênticos e a soma de todos os pares de sequências homólogas. A fórmula 3 é a razão entre a soma de todos os nucleotídeos idênticos e a soma dos tamanhos dos genomas (FIGURA 5.1).

Entre elas, a fórmula 2 é recomendada pela ferramenta GGDC e sua característica é não ser suscetível a problemas gerados pelo uso de genomas incompletos, pois não leva em consideração o tamanho dos genomas (MEIER-KOLTHOF *et al*., 2013). Essa fórmula foi também usada por Colston e colaboradores (2014) para a análise de genomas de *Aeromonas*.

$$d_0(X,Y) = 1 - \frac{H_{XY} + H_{YX}}{\lambda(X,Y)} \qquad \text{Fórmula 1}$$

$$d_4(X,Y) = 1 - \frac{2 \cdot I_{XY}}{H_{XY} + H_{YX}} \qquad \text{Fórmula 2}$$

$$d_6(X,Y) = 1 - \frac{2 \cdot I_{XY}}{\lambda(X,Y)} \qquad \text{Fórmula 3}$$

Onde:

$XY :=$ pesquisa BLAST utilizando os genomas X (*subject*) e Y (*query*)

$I_{XY} :=$ soma dos pares de bases idênticos de todas as sequências homólogas

$H_{XY} :=$ tamanho total de todas as sequências homólogas

$\lambda(X,Y) :=$ soma do tamanho dos dois genomas

FIGURA 5.1**:** FÓRMULAS UTILIZADAS PARA O CÁLCULO DE GBDP
FONTE: MEIER-KOLTHOF *et al*. (2013)

5.5 Análises e conjunto de dados complementares

Outros genomas também foram usados em algumas análises, todos eles obtidos do NCBI Genbank, pelo número de acesso entre parênteses: *Collimonas fungivorans* Ter331 (LEVEAU *et al*., 2004; CP002745.1), *Oxalobacter formigenes* OXCC13 (WARD *et al*., não publicado; ACDQ00000000.1), *Oxalobacter formigenes* HOxBLS (EARL *et al*., não publicado; NZ_ACDP00000000.2), *Oxalobacteraceae bacterium* IMCC9480 (OH *et al*., 2011; NZ_AEPR00000000.1), e *Herbaspirillum* sp. RV1423 (JAUREGUI *et al*., 2014; CBXX000000000.1). Os *drafts* genômicos das estirpes B501, B65 e B39 de *Herbaspirillum* sp. não foram utilizados por terem baixa qualidade (em torno de mil *contigs*).

Sequências para análises específicas foram obtidas do NCBI Genbank através de busca por similaridade com uso do programa BLAST *on line* e do banco de dados NR do NCBI. Árvores filogenéticas e anotação de vias metabólicas foram realizadas do mesmo modo como descritas previamente no tópico 5.2.

Um resumo da metodologia aplicada neste trabalho é mostrado na FIGURA 5.2.

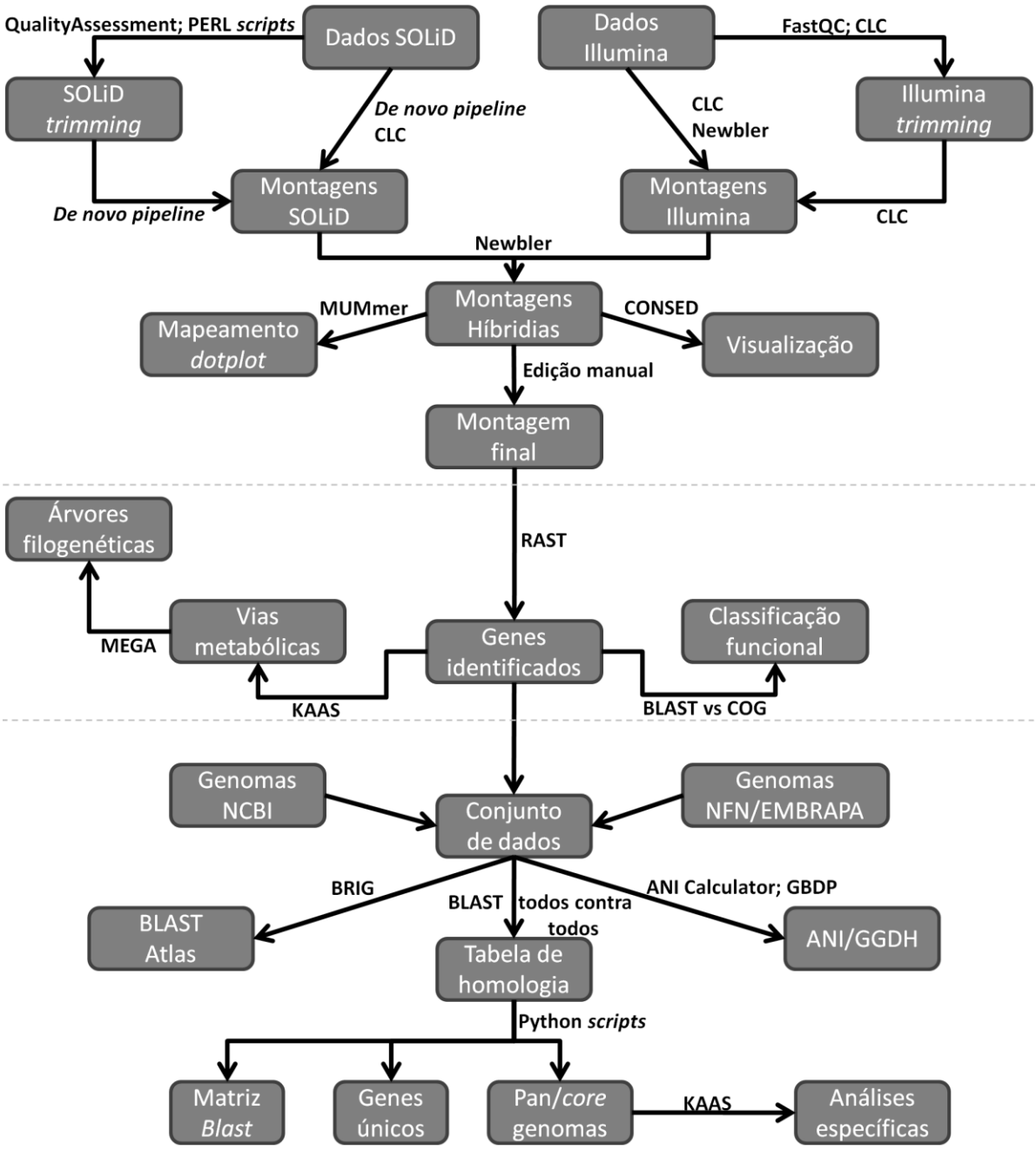


FIGURA 5.2**:** FLUXOGRAMA DA METODOLOGIA UTILIZADA

FONTE: o autor (2015)

6 RESULTADOS

6.1 Conjunto de dados de sequenciamento de DNA genômico e análise de qualidade

6.1.1 Dados de sequenciamento de DNA na plataforma SOLiD

Com a utilização da plataforma de sequenciamento SOLiD, foram obtidos 7.804.606 *reads* para o sequenciamento genômico de *H rhizosphaerae* UMS-37, 21.228.876 *reads* para o sequenciamento genômico de *H. chlorophenolicum* CPW301 e 20.607.513 *reads* para o sequenciamento genômico de *H. autotrophicum* IAM 14942. Foram obtidas leituras simples (fragmentos) não pareadas. O tamanho do genoma foi estimado em 5,0 Mb, para cada um dos três organismos, e a cobertura de bases foi estimada em 200 vezes para os genomas de *H. chlorophenolicum* CPW301 e *H. autotrophicum* IAM 14942 e em 80 vezes para o genoma de *H. rhizosphaerae* UMS-37. Essas estimativas foram baseadas nas sequências genômicas completas de *H. seropedicae* SmR1 (5,5 Mb – PEDROSA *et al*., 2011), *H. rubrisubalbicans* M1 (5,6 Mb – CARDOSO, 2011; SOUZA, *in prep*.) e montagens prévias (4,8 Mb).

Os *reads* obtidos na plataforma SOLiD, para os três genomas sequenciados neste trabalho, apresentaram comprimento de 50 pb e valores médios de qualidade por base inferiores a *phred* 30. A distribuição das médias de qualidade por base apresentou valores maiores nas bases iniciais e valores decrescentes até as bases finais, atingindo aproximadamente *phred* 16 (FIGURA 6.1).

6.1.2 Dados de sequenciamento de DNA na plataforma Illumina

Os *reads* produzidos pela plataforma Illumina MiSeq tiveram tamanho médio de 162 pb, 151 pb e 171 pb, e coberturas de 37x (1.140.064 *reads*), 55x (1.832.778 *reads*) e 43x (1.262.926 *reads*) para os genomas de *H. chlorophenolicum* CPW301, *H. rhizosphaerae* UMS-37 e *H. autotrophicum* IAM 14942, respectivamente (considerando um tamanho estimado para os genomas de 5,0 Mb). Esses sequenciamentos geraram *reads* pareados do tipo *pair-end*.

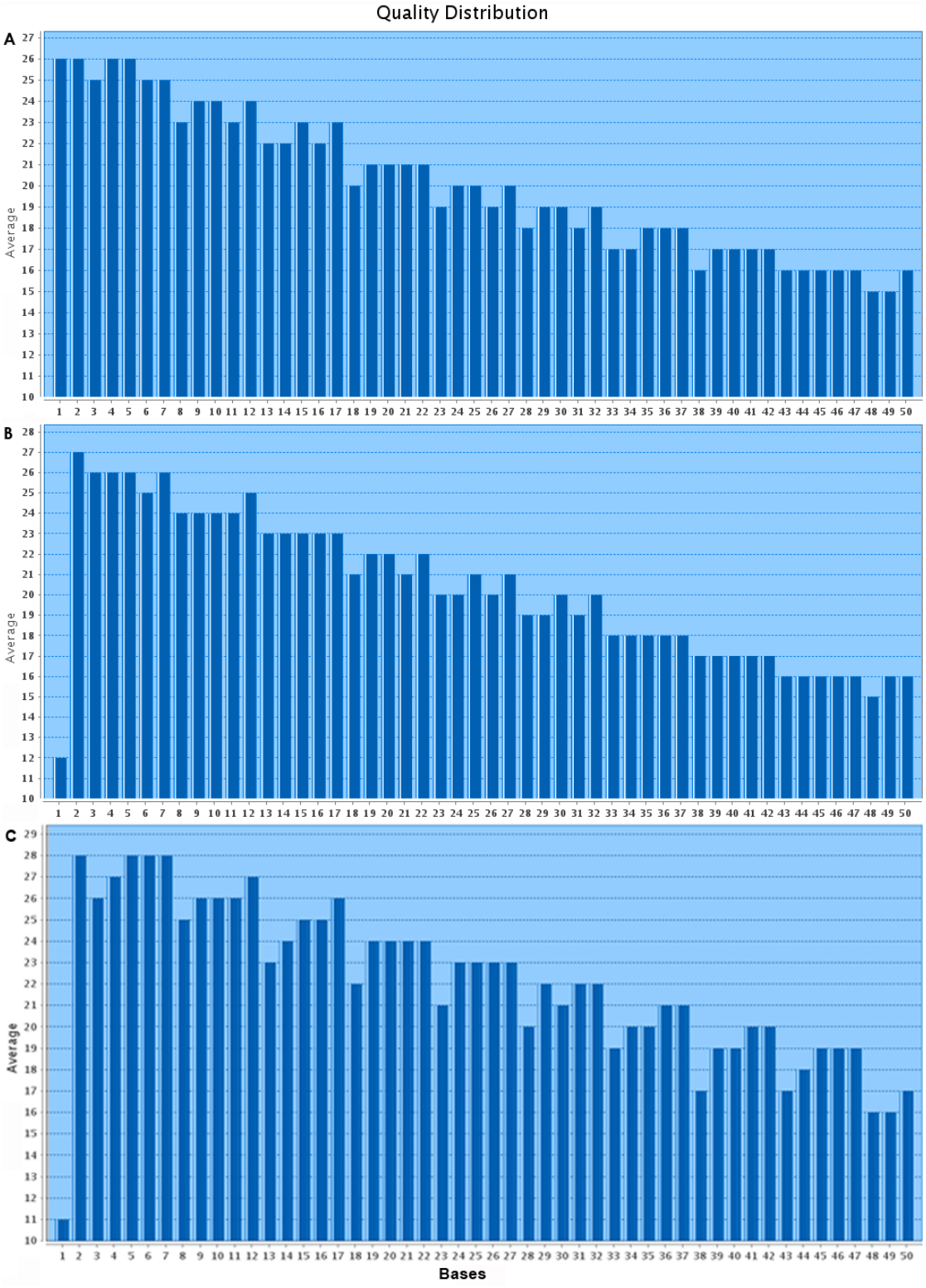


FIGURA 6.1: QUALIDADE MÉDIA POR BASE AO LONGO DOS *READS* PARA OS TRÊS GENOMAS SEQUENCIADOS NA PLATAFORMA SOLID

Em A, B e C são mostradas as médias de qualidade por base para os dados de sequenciamento dos genomas de *H. chlorophenolicum* CPW301, *H. rhizosphaerae* UMS-37 e *H. autotrophicum* IAM 14942, respectivamente. As médias das qualidades de bases são indicadas por valores *phred*. A distribuição de qualidades foi gerada com o programa QualityAssesment.

FONTE: o autor (2015)

Através da análise de qualidade, foi possível verificar que os *reads* obtidos com a plataforma de sequenciamento Illumina MiSeq apresentaram melhor qualidade em comparação com os *reads* obtidos com a plataforma SOLiD. A qualidade dos *reads* foi considerada boa, pois até a base 100 foi obtida média de qualidade *phred* superior a 30 e até a base 200 foi obtida média de qualidade *phred* superior a 20. Foi observado um decréscimo de qualidade dos *reads* em direção à extremidade 3', o qual já era esperado, conforme a característica da química de sequenciamento (FIGURA 6.2). As distribuições do conteúdo G+C e do tamanho dos *reads* podem ser vistas na FIGURA 6.3.

6.2 Montagem genômica

6.2.1 Montagem genômica com dados de sequenciamento da plataforma SOLiD

A baixa qualidade da primeira base dos *reads* SOLiD representa um grave problema, visto que esta plataforma de sequenciamento gera *reads* em formato *color space*, nos quais a decodificação de uma base (que pode ser errada) irá se propagar por todas as bases seguintes. Por esse motivo, os testes de montagem foram realizados com os dados de sequenciamento do genoma de *H. chlorophenolicum* CPW301, onde foram gerados *reads* de melhor qualidade.

Esse conjunto de dados foi submetido a testes de montagem de sequência genômica utilizando o *pipeline de novo* com diferentes valores de *hsize*, que derivam dos valores *hsize* 25 e 21 sugeridos para o *pipeline de novo* e para o programa *Velvet* (SOLID4 – Life Technologies; ZERBINO, 2010). O principal resultado usado para direcionar e avaliar a melhor montagem foi o número de *contigs* referente a um tamanho de genoma esperado de 5,0 Mb. Foi observado que as montagens com *hsizes* de tamanho maior produziram menor número de *contigs*, mas o tamanho dos genomas se tornou menor, provavelmente por utilizarem menos *reads*. Essas montagens também geraram *contigs* de maior tamanho. Por outro lado, as montagens com *hsizes* menores incluíram um número maior de *reads* e produziram genomas de tamanho maior, mas com um número de *contigs* maior e de tamanho médio menor (TABELA 6.1).

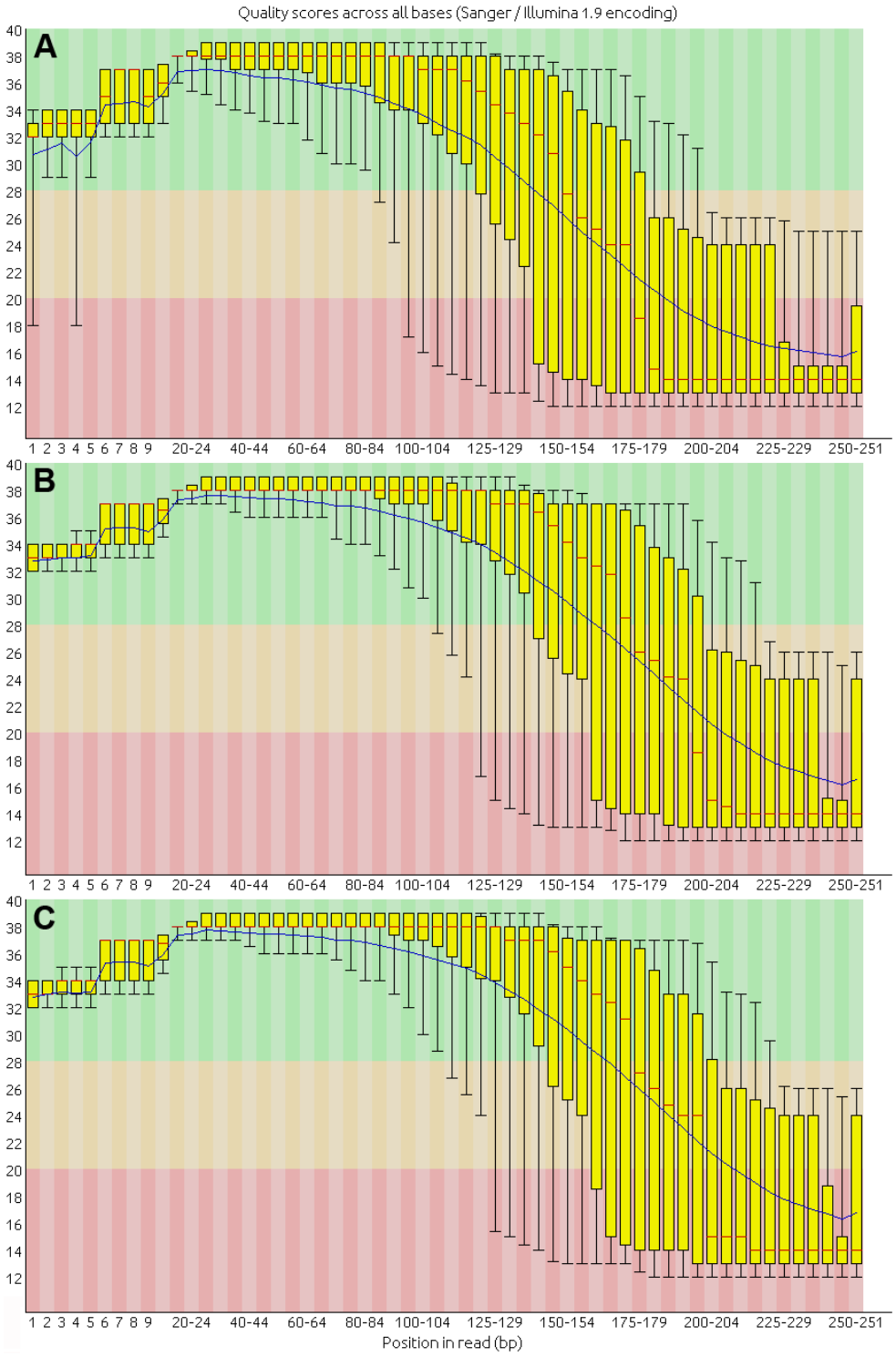


FIGURA 6.2: DISTRIBUIÇÃO DE QUALIDADE DO CONJUNTO DE DADOS DE SEQUENCIAMENTO GERADOS NA PLATAFORMA ILLUMINA MISEQ

A análise e o gráfico foram realizados com o programa FastQC. As barras amarelas mostram a distribuição das qualidades por base, com uma barra que compreende um intervalo de 10 a 90% dos *reads* (eixo 'y'), para cada base (eixo 'x'), e a linha azul mostra as médias de qualidade por base. A faixa vermelha inclui regiões de baixa qualidade, a faixa amarela inclui regiões de qualidade intermediária e a faixa verde inclui regiões de alta qualidade. Em A, os gráficos são referentes aos dados de sequenciamento genômico de *H. autotrophicum* IAM 14942. Em B, são referentes a *H. chlorophenolicum* CPW301. Em C, são referentes a *H. rhizosphaerae* UMS-37.

FONTE: o autor (2015)

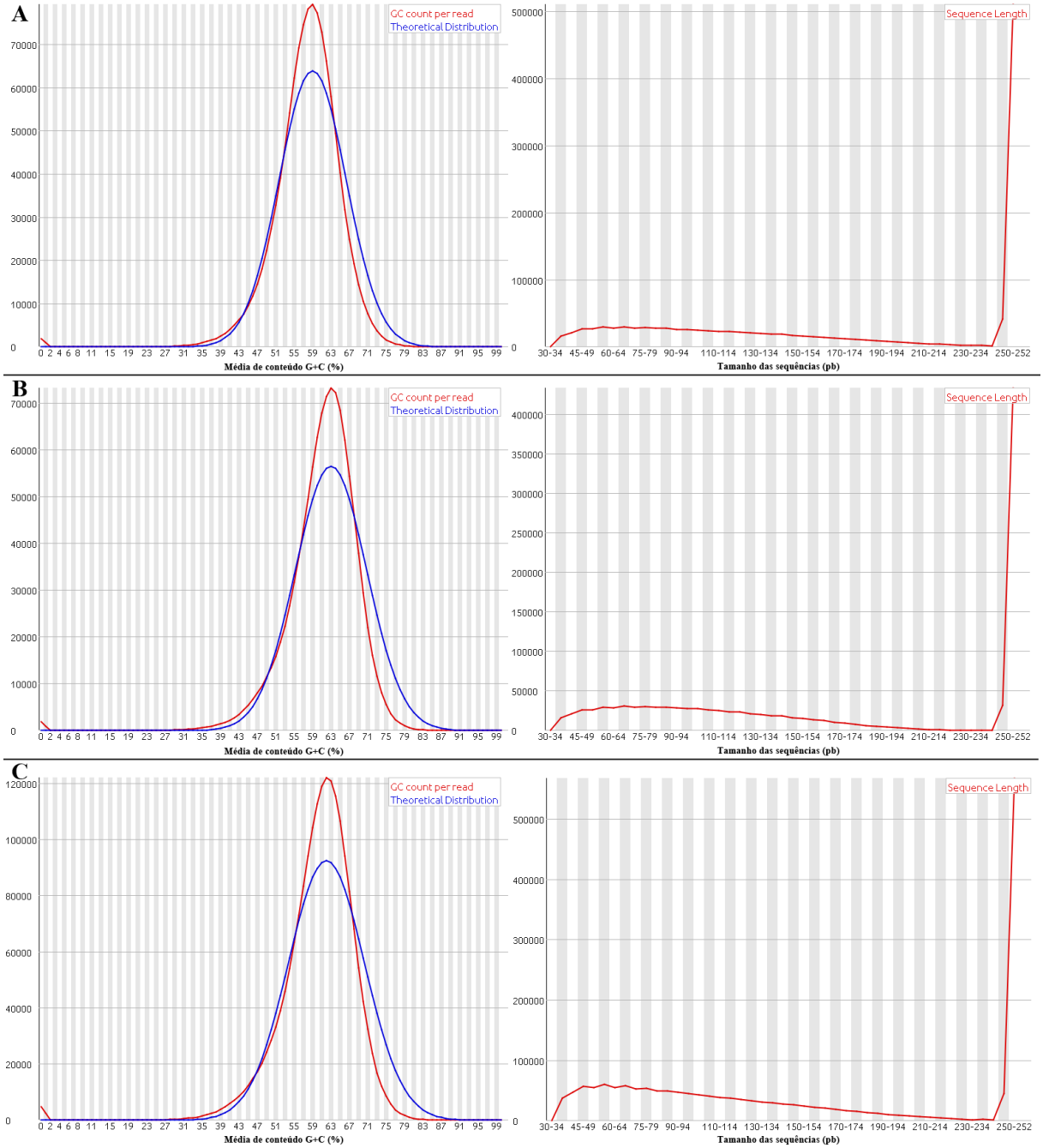


FIGURA 6.3: DISTRIBUIÇÃO DO TAMANHO E DO CONTEÚDO G+C NO CONJUNTO DE DADOS DE SEQUENCIAMENTO GERADO NA PLATAFORMA ILLUMINA

A análise e os gráficos foram realizados com o programa FastQC. Nos gráficos da esquerda, as linhas vermelhas representam a média de conteúdo G+C por *reads*, e a linha azul representa a distribuição teórica dessas bases. Nos gráficos da direira, as linhas vermelhas representam a distribuição de *reads* por tamanho. Em A, os gráficos são referentes aos dados de sequenciamento genômico de *H. autotrophicum* IAM 14942. Em B, são referentes a *H. chlorophenolicum* CPW301. Em C, são referentes a *H. rhizosphaerae* UMS-37.

FONTE: o autor (2015)

Para tentar minimizar o impacto da baixa qualidade dos *reads*, foi realizado tanto o *trimming* de qualidade quanto o descarte das 15 últimas bases da extremidade 3’ desses *reads*. Com isso, foram obtidos: 10.296.541 *reads* (102x de

cobertura) para *H chlorophenolicum* CPW103; 4.132.037 *reads* (41x de cobertura) para *H rhizosphaerae* UMS-37; 10.301.778 *reads* (103x de cobertura) para *H. autotrophicum* IAM 14942.

TABELA 6.1: MONTAGEM GENÔMICA DE *H. chlorophenolicum* CPW301 A PARTIR DE DADOS DE SEQUENCIAMENTO NA PLATAFORMA SOLID, COM USO DO *PIPELINE DE NOVO*

	***hsize* 17**	***hsize* 21**	***hsize* 23**	***hsize* 25**	***hsize* 27**
N50 (pb)	105	247	314	386	427
Maior *contig* (pb)	1267	1916	2532	3055	4433
Total de bases (Mb)	4,66	4,65	4,45	4,25	4,14
***Reads* usados (milhões)**	12	13	12,5	11,9	11,3
***Contigs* gerados**	14.888	14.702	13.432	12.448	12.127

FONTE: o autor (2015)

Os resultados obtidos para as montagens genômicas de *H. chlorophenolicum* CPW301 com os dados pós-*trimming* tiveram as mesmas características dos resultados obtidos sem a realização do *trimming*: *hsizes* maiores geraram menor número de *contigs*, porém genomas de tamanho menor. No entanto, o número de *contigs* aumentou e o tamanho dos genomas diminuiu em comparação às montagens sem *trimming*, sendo esses resultados considerados piores do que aqueles para as montagens obtidas sem *trimming* (TABELA 6.2).

O genoma de *Herbaspirillum chlorophenolicum* CPW301 foi também montado dentro da plataforma *CLC Genomics Workbench*. Da mesma forma que o parâmetro *hsize* foi modificado na utilização do *pipeline de novo*, o parâmetro *wsize* foi modificado na plataforma CLC, com valores derivados do *wsize* 25 padrão do programa, na tentativa de se obter a melhor montagem. Os melhores resutados obtidos utilizaram *wsize* 23 (cuja montagem teve maior valor de N50 e maior tamanho de genoma) e *wsize* 25 (cuja montagem utilizou maior número de *reads*). Como a plataforma CLC aceita *wsizes* pares, foi realizado um teste com *wsize* 24, cujo resultado combinou aspectos favoráveis dos testes com *wsizes* 23 e 25 (embora o N50 tenha diminuído), e por isso essa foi considerada a melhor

montagem obtida dentro dessa plataforma. Por outro lado, foi observado que, independentemente das montagens, o número de *contigs* das montagens realizadas com a plataforma CLC foi maior que o número de *contigs* gerados pelo *pipeline de novo*, embora os tamanhos dos genomas obtidos atingissem o valor esperado (TABELA 6.3).

TABELA 6.2: MONTAGEM GENÔMICA DE *H. chlorophenolicum* CPW301 A PARTIR DE DADOS DE SEQUENCIAMENTO NA PLATAFORMA SOLID, APÓS TRIMMING, COM USO DO PIPELINE DE NOVO

	***hsize* 17**	***hsize* 21**	***hsize* 23**	***hsize* 25**
N50 (pb)	213	363	336	308
Maior *contig* (pb)	1455	3770	3886	5550
Total de bases (Mb)	4,47	4,44	4,3	4,07
***Reads* usados (milhões)**	8,9	9	8,9	11,9
***Contigs* gerados**	20.752	16.986	18.032	18.545

FONTE: o autor (2015)

TABELA 6.3: MONTAGEM GENÔMICA DE *H. chlorophenolicum* CPW301 A PARTIR DE DADOS DE SEQUENCIAMENTO NA PLATAFORMA SOLID COM USO DE CLC GENOMICS WORKBENCH

Wsizes	***Reads* usados (mi) (%)**	***Contigs* gerados**	**Tamanho do genoma (Mb)**	**N50 (pb)**
17	8,1 (26%)	9.242	1,8	212
19	9,1 (29%)	10.861	2,6	215
21	18,1 (58%)	19.339	4,7	273
22	19,4 (63%)	19.783	4,9	285
23	19,9 (64%)	19.839	5	286
24	**20,2 (65%)**	**20.002**	**5**	**279**
25	20,2 (65%)	20.244	4,9	270
27	20,2 (65%)	20.610	4,8	253

FONTE: o autor (2015)

A plataforma CLC *Genomic Workbench* também foi utilizada para o mapeamento dos *reads* do sequenciamento genômico de *H. chlorophenolicum* CPW301 em relação aos genomas de outras estirpes de *Herbaspirillum* spp. disponíveis (presentes em bancos de dados públicos ou disponibilizados pelo NFN). Como resultado obtido, o número de *reads* mapeados nesses genomas foi muito baixo, inferior a 10%, e apenas cerca da metade dos genomas de referência foram cobertos. Para gerar *contigs*, as sequências consenso obtidas desses mapeamentos foram fragmentadas e, dessa forma, milhares de *contigs* foram obtidos (TABELA 6.4).

TABELA 6.4: TESTES DE MAPEAMENTO DO CONJUNTO DE DADOS DE SEQUENCIAMENTO GENÔMICO DE *H. chlorophenolicum* CPW301 NA PLATAFORMA SOLID EM SEQUÊNCIAS GENÔMICAS DE ESTIRPES DE *Herbaspirillum* spp.

Genoma de referência	***Reads* mapeados (mi) (%)**	***Contigs* gerados**	**Tamanho do genoma**
Herbaspirillum sp YR522	~1,6 (7,6%)	25.625	~2,8Mb
H. seropedicae SmR1	~1,8 (8,6%)	28.926	~3,2Mb
H. rubrisubalbicans M1	~1,66 (7,9%)	26.055	~2,8Mb
Herbaspirillum sp CF444	~1,2 (5,7%)	27.704	~2,7Mb
Herbaspirillum sp GW103	~1,8 (8,6%)	24.341	~2,8Mb
H. frisingense GSF30	~1,7 (8,1%)	22.111	~2,7Mb
H. lusitanum P6-12	~1,2 (5,5%)	25.780	~2,6Mb
H. massiliense JC206	~0,4 (2%)	27.743	~2,1Mb

FONTE: o autor (2015)

Devido ao fato das montagens produzidas não terem sido satisfatórias, mas possuírem atributos que em conjunto poderiam resultar em uma montagem melhor, foi proposta uma abordagem de montagens híbridas. Nessa abordagem, os *contigs* montados durante os testes realizados previamente foram fragmentados e utilizados como *reads* para novas montagens (chamadas de 'Pan'). Para isso foi utilizado o montador da plataforma CLC *Genomics Workbench* e realizados os seguintes testes: 1- somente usando *contigs* gerados pelo *pipeline de novo* (montagem 'Pan *de novo*'); 2- somente usando *contigs* gerados por montagem dentro da plataforma

CLC (montagem ‘Pan CLC’); 3- somente *contigs* gerados por mapeamento em genomas de referência (montagem ‘Pan map’); 4- por montagens híbridas dos três conjuntos anteriores (montagem ‘Pan *contigs’*).

A montagem ‘Pan *contigs’* com *wsize* 24 mostrou resultados absolutamente superiores às demais montagens (TABELA 6.5), pois produziu um número de *contigs* (13.940) similar ao obtido com o *pipeline de novo* (13.432, montagem ‘*Contigs de novo’*), mas com cerca de 1 Mb a mais de tamanho de genoma e também o maior valor de N50 entre todas as montagens obtidas.

TABELA 6.5: COMPARAÇÃO DOS RESULTADOS OBTIDOS PARA AS MONTAGENS HÍBRIDAS (PAN) EM RELAÇAO ÀS MONTAGENS OBTIDAS ANTERIORMENTE (*CONTIGS*)

Montagem	***Reads* usados (%)**	***Contigs* gerados**	**Tamanho do genoma (Mb)**	**N50**
Pan *contigs* (*Wsize* 24)	71%	13.940	5,5	540
Pan *contigs* (*wsize* auto)	69%	14.568	5,4	370
Contigs de novo* (*hsize* 23)	59%	13.432	4,45	314
Pan *de novo* (*Wsize* 24)	95%	14.135	4,7	423
Contigs* CLC (*Wsize* 24)	65%	20.002	5,0	279
Pan CLC (*Wsize* 24)	92%	17.239	5,4	380
Contigs* mapeados (*Herbaspirillum* sp. GW103)	8,6%	24.341	2,8	-
Pan map (*Wsize* 24)	44%	9.630	2,7	-

* Representam montagens originais, não híbridas, apenas para comparação

FONTE: o autor (2015)

Os 13.940 *contigs* obtidos na montagem ‘Pan *contigs’* foram mapeados contra o genoma completo de *H. seropedicae* SmR1 (FIGURA 6.4). Desses, 2.623 *contigs* apresentaram homologia com o genoma de referência, os quais cobriram apenas 1,8 Mb desse genoma. Dentro do conjunto de *reads* mapeados, a média de tamanho foi 687 pb, e N50 de 866 pb.

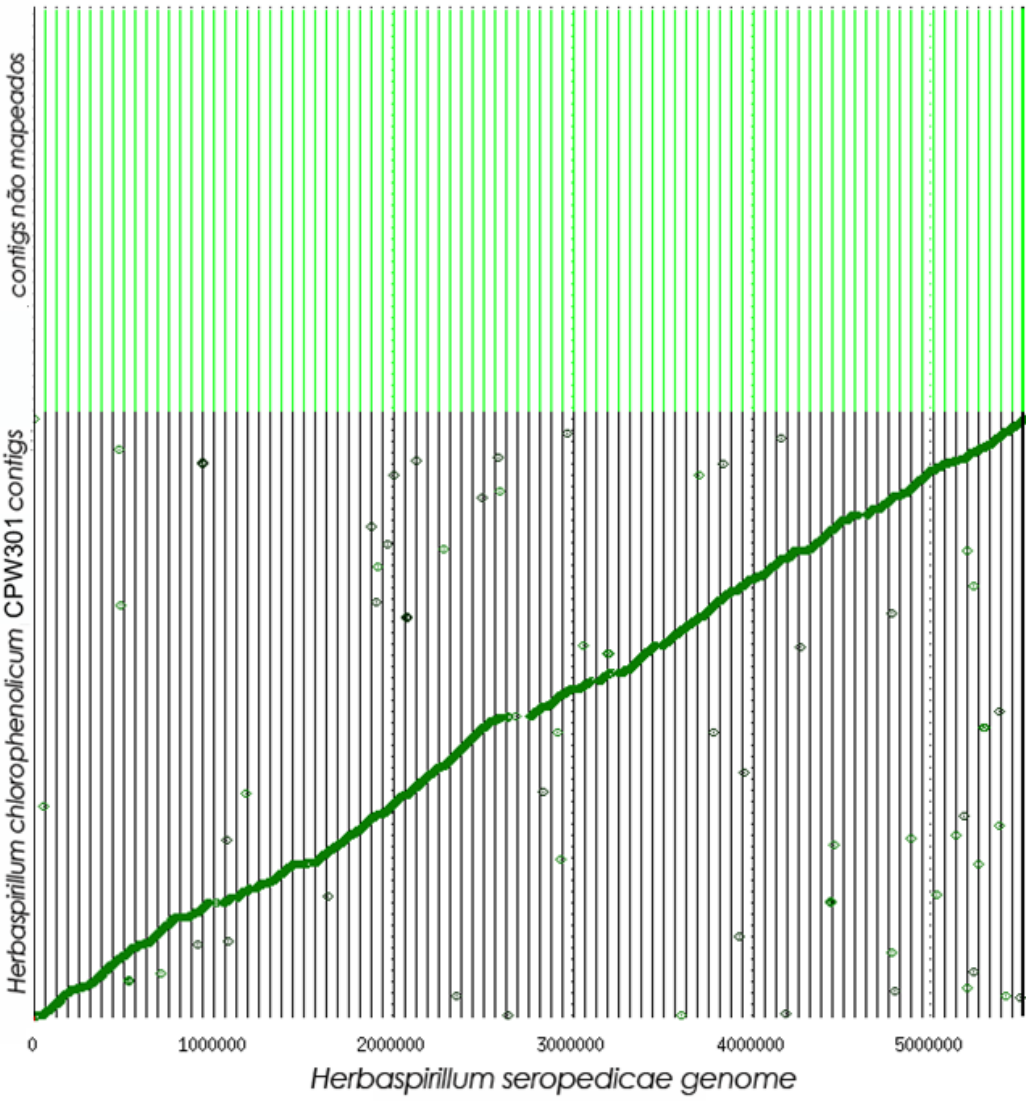


FIGURA 6.4: GRÁFICO DOTPLOT ENTRE OS CONTIGS DA MONTAGEM HÍBRIDA 'PAN CONTIGS' COM WSIZE 24 E O GENOMA COMPLETO DE *H. seropedicae* SmR1.

O alinhamento e o gráfico foram gerados com o pacote de programas MUMmer e está representado em verde. Acima do alinhamento está representada uma área de *contigs* que não foram mapeados.

FONTE: o autor (2015)

6.2.2 Montagem genômica com dados de sequenciamento da plataforma Illumina

O conjunto de dados proveniente da plataforma Illumina MiSeq foi automaticamente montado pelo programa Velvet 1.7 embutido na própria plataforma. Essas montagens apresentaram resultados melhores em relação ao que foi obtido com os dados de sequenciamento provenientes da plataforma SOLiD. Foram gerados 481, 457 e 1011 *contigs* para os genomas de *H. autotrophicum* IAM 14942, *H. rhizosphaerae* UMS-37 e *H. chlorophenolicum* CPW301 respectivamente. Essas

montagens apresentaram sequências longas, nas quais mais de 100 *contigs* possuíam tamanho superior a 10.000 pb (FIGURA 6.5).

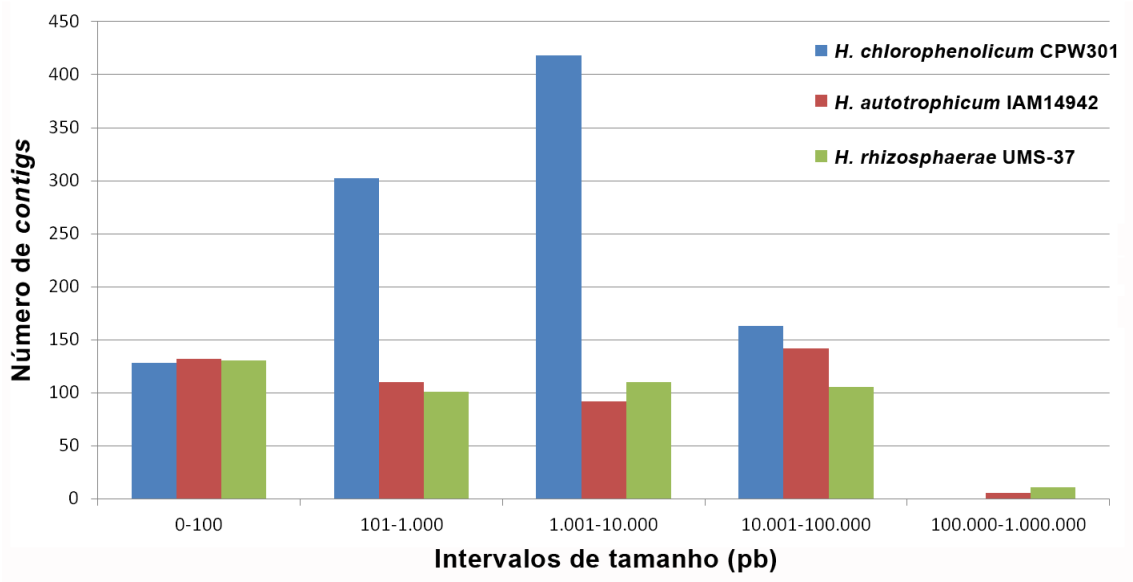


FIGURA 6.5: DISTRIBUIÇÃO DO TAMANHO DOS CONTIGS GERADOS PELA MONTAGEM AUTOMÁTICA ILLUMINA

FONTE: o autor (2015)

Embora a montagem tenha melhorado de forma significativa com os dados de sequenciamento da plataforma Illumina, o número de *contigs* obtidos, principalmente para *H. chlorophenolicum* CPW301, ainda não foi considerado suficiente para *drafts* genômicos de alta qualidade.

Dessa forma, novos testes de montagens foram realizados, utilizando novamente os dados de sequenciamento de *H. chlorophenolicum* CPW301 e a plataforma CLC. O primeiro teste foi realizado apenas com a ponta R1 dos pares (teste CLC R1 – C1), no qual foram obtidos 936 *contigs*, pouco menos do que a montagem automática Illumina (I), mas com tamanho de genoma de 5,37 Mb, contra 5,03 Mb da montagem automática Illumina. O segundo teste (CLC *pair-end* - C2) foi realizado com todo o conjunto dos reads *Illumina* (pares R1 e R2) e reduziu o número de *contigs* para 496, com 5,35 Mb de tamanho de genoma e um salto para um *contig* N50 de ~59 Kb de tamanho (contra ~20Kb da montagem C1) (TABELA 6.6).

Como a qualidade decresce ao longo do *read* Illumina, um processo de *trimming* foi realizado para a remoção das últimas 50 bases dos *reads* correspondentes ao sequenciamento de *H. chlorophenolicum* CPW301 (onde a

qualidade cai para *phred* menor que 25), assim como o descarte de *reads* com média de qualidade inferior a *phred* 20. Foram obtidos 1.071.393 *reads* com média de tamanho de 119,8 pb (27x de cobertura). Esses dados foram utilizados para uma nova montagem (CLC *trimmed* – C3) e geraram 570 *contigs*, 5,3 Mb de tamanho de genoma e N50 de ~45 Kb, valores inferiores ao obtido na montagem C2 (TABELA 6.6).

Após a obtenção dos 496 *contigs*, foram realizadas montagens híbridas, conforme realizado com as montagens obtidas com os dados provenientes da plataforma SOLiD. Para isso, os *contigs* das montagens I, C2 e PC foram fragmentados e remontados (montagem híbrida PC + I + C2). Essa montagem gerou 296 *contigs*, com tamanho de genoma de ~4,93 Mb e *contig* N50 com tamanho de ~44 Kb (TABELA 6.6).

TABELA 6.6: COMPARAÇÃO DAS MONTAGENS REALIZADAS PARA *H. chlorophenolicum* CPW301 (HC) COM OS DADOS DAS PLATAFORMAS ILLUMINA E SOLID

Montagens	***Contigs* gerados**	**Genoma (Mb)**	***Contig* N50**	**Maior *Contig* (kb)**
HC *contigs de novo* - *hsize* 23 (SOLiD)	13.432	4,45	314	2,5
HC *contigs* CLC - *wsize* 24 (SOLiD)	20.002	5	279	-
HC Pan *contigs* – (PC) (SOLiD)	13.940	5,5	540	5,8
HC Illumina – (I) (Illumina)	1.011	5,03	15.238	52
HC CLC R1 – (C1) (Illumina)	936	5,37	19.880	78,9
HC CLC *pair-end* – (C2) (Illumina)	**496**	**5,35**	**59.021**	**233,5**
HC CLC *trimmed* (C3) (Illumina)	570	5,3	45.113	183
HC hibridização (PC) + (I) + (C2) (híbrida)	296	4,93	44.191	264

FONTE: o autor (2015)

Alternativamente à plataforma CLC, os dados de sequenciamento Illumina de *H. chlorophenolicum* CPW301 foram submetidos ao montador Newbler (montagem 'Newbler *reads* Illumina'). O teste com esse montador gerou como resultado 3.271 *contigs*, com tamanho de genoma de 5,35 Mb. Um segundo teste realizado (montagem '*Reads* Illumina + PC') consistiu em uma montagem híbrida dos *reads* Illumina com a montagem PC, o que resultou na diminuição do número de *contigs* para 355 e tamanho de genoma de 5,31 Mb. Em um terceiro teste, foram

hibridizados os *reads* Illumina com *contigs* tanto da montagem PC quanto da montagem C2 (‘*Reads* Illumina + PC + C2’), o que resultou em 237 *contigs*, tamanho de genoma de 5,3 Mb e tamanho do *contig* N50 de ~63 Kb (TABELA 6.7).

TABELA 6.7: COMPARAÇÃO DAS MONTAGENS REALIZADAS PARA *H. chlorophenolicum* CPW301 (HC) COM A PLATAFORMA CLC GENOMICS WORKBENCH E COM O MONTADOR NEWBLER

	Contigs	**Genoma (Mb)**	**Tamanho médio dos *contigs***	**N50**	**Maior *contig* (Kb)**
HC **CLC** *pair-end* – (C2)	496	5,35	-	59.021	233,5
HC **Newbler** *reads* Illumina	3.271	5,35	1.649	2.100	13,8
Reads Illumina + PC (**Newbler**)	355	5,31	14.951	29.765	98,8
Reads Illumina + PC + C2 (**Newbler**)	237	5,30	28.834	63.293	150

FONTE: o autor (2015)

Essa terceira montagem híbrida realizada com o programa Newbler (*Reads* Illumina + PC + C2) foi visualizada no programa *Consed* para a avaliação de *gaps* que poderiam ser fechados manualmente. Nela foram encontradas várias repetições e foi observado que as repetições bifurcavam para *contigs* gerados a partir dos dados da plataforma SOLiD, ou da plataforma Illumina, mas nunca para *contigs* que incluíssem dados sobrepostos de ambas as plataformas (FIGURA 6.6). Essa divergência levou a acreditar que os dados provenientes da plataforma SOLiD e os dados provenientes da plataforma Illumina conflitavam entre si (não se complementavam), e isso poderia estar criando uma montagem quimérica. Como os *reads* provenientes da plataforma Illumina apresentavam maior confiabilidade em relação aos dados de sequenciamento da plataforma SOLiD, esses últimos foram retirados das montagens híbridas.

Além disso, já haviam sido observados indícios de ligação entre *contigs* montados pela plataforma CLC, mas que não foram unidos durante a montagem. Um novo teste consistiu em remontar os *contigs* obtidos na montagem C2, com a fragmentação dos *contigs* (C2 - remontagem), para uni-los. Com isso, foram obtidos 225 *contigs*, um genoma com tamanho de ~5,24 pb e tamanho de *contig* N50 de ~67 Kb. Após esse teste preliminar, os *reads* Illumina também foram adicionados à remontagem (*Reads* Illumina + C2).

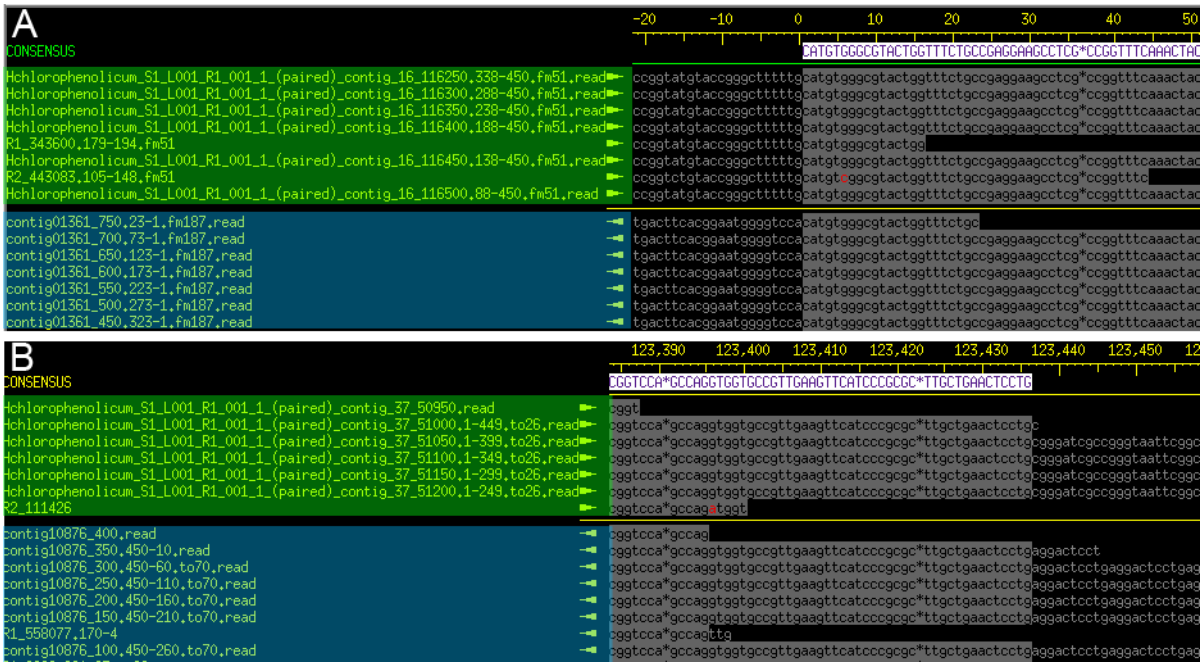


FIGURA 6.6: EXEMPLO DE UMA REPETIÇÃO CONFLITANTE ENTRE OS DADOS DE SEQUENCIAMENTO DA PLATAFORMA ILLUMINA E DA PLATAFORMA SOLID

Existe uma sequência consenso (o próprio *contig*, em branco) que bifurca em cada extremidade (representadas em A e B) para dois *contigs*. Esses *contigs* são formados ou por dados Illumina ou por dados SOLiD (representados em verde e azul, respectivamente). Por exemplo, em A (a extremidade 5') o *contig* bifurca para os *contigs* 51 (dados Illumina, em verde) e 187 (dados SOLiD, em azul). Essa separação entre os dados foi observada em diversos casos, o que leva a acreditar que não é simplesmente uma repetição genômica. A imagem foi obtida com o programa *Consed*.

FONTE: o autor (2015)

A melhora dos resultados através da montagem '*Reads* Illumina + C2' no programa Newbler foi considerada satisfatória (TABELA 6.8). O número de *contigs* caiu de 496 para 216, com perda de somente 0,05 Mb do genoma, o tamanho do *contig* N50 aumentou de 59 Kb para 109 Kb e o maior *contig* aumentou de 233,5 Kb para 336,8 Kb. Os *contigs* foram unidos aparentemente sem problemas de divergência de dados e essa montagem de 216 *contigs* foi considerada a melhor montagem obtida para *H. chlorophenolicum* CPW301, bem como o método de remontagem dos *contigs* Illumina, juntamente com os dados brutos da plataforma Illumina, foi considerado o melhor método para a montagem genômica com os dados obtidos.

TABELA 6.8: COMPARAÇÃO DAS MONTAGENS REALIZADAS PARA *H. chlorophenolicum* CPW301 (HC) COM O MONTADOR NEWBLER SOMENTE COM OS DADOS DE SEQUENCIAMENTO DA PLATAFORMA ILLUMINA

Montagem	***Contigs* gerados**	**Genoma**	**Média de tamanho dos *contigs***	***Contig* N50**	**Maior *contig* (kb)**
***Reads* Illumina + PC + C2**	237	5,30	28.834	63.293	150
C2 – remontagem Newbler	225	5,24	24.595	67.446	233,3
***Reads* Illumina + C2**	216	5,30	31.121	109.137	336,8

FONTE: o autor (2015)

6.2.3 Finalização da montagem genômica

Os *contigs* da montagem '*reads* Illumina + C2' foram mapeados no genoma de *H. seropedicae* SmR1. O gráfico *dotplot* dessa montagem em relação à referência mostrou de maneira mais clara a presença de inversões em relação ao gráfico produzido com as montagens automáticas Illumina (FIGURA 6.7A). Além disso, a cobertura em relação ao genoma de referência subiu para 2,1 Mb (38,1%) em relação a 1,8 Mb (32,6%) obtido com a montagem 'Pan *contigs*'.

A visualização e a investigação da montagem obtida mostraram que o número de *contigs* poderia ser reduzido a 85, se levado em conta a estrutura do genoma de referência para estabelecer a ordem dos *contigs* e também para resolver repetições. Foi também observado que as repetições constituíam de duplicações e a única repetição complexa consistia de quatro cópias de um *contig*. Usando a referência como padrão de estrutura, foi observado que duas dessas quatro cópias poderiam ser resolvidas na montagem (FIGURA 6.8).

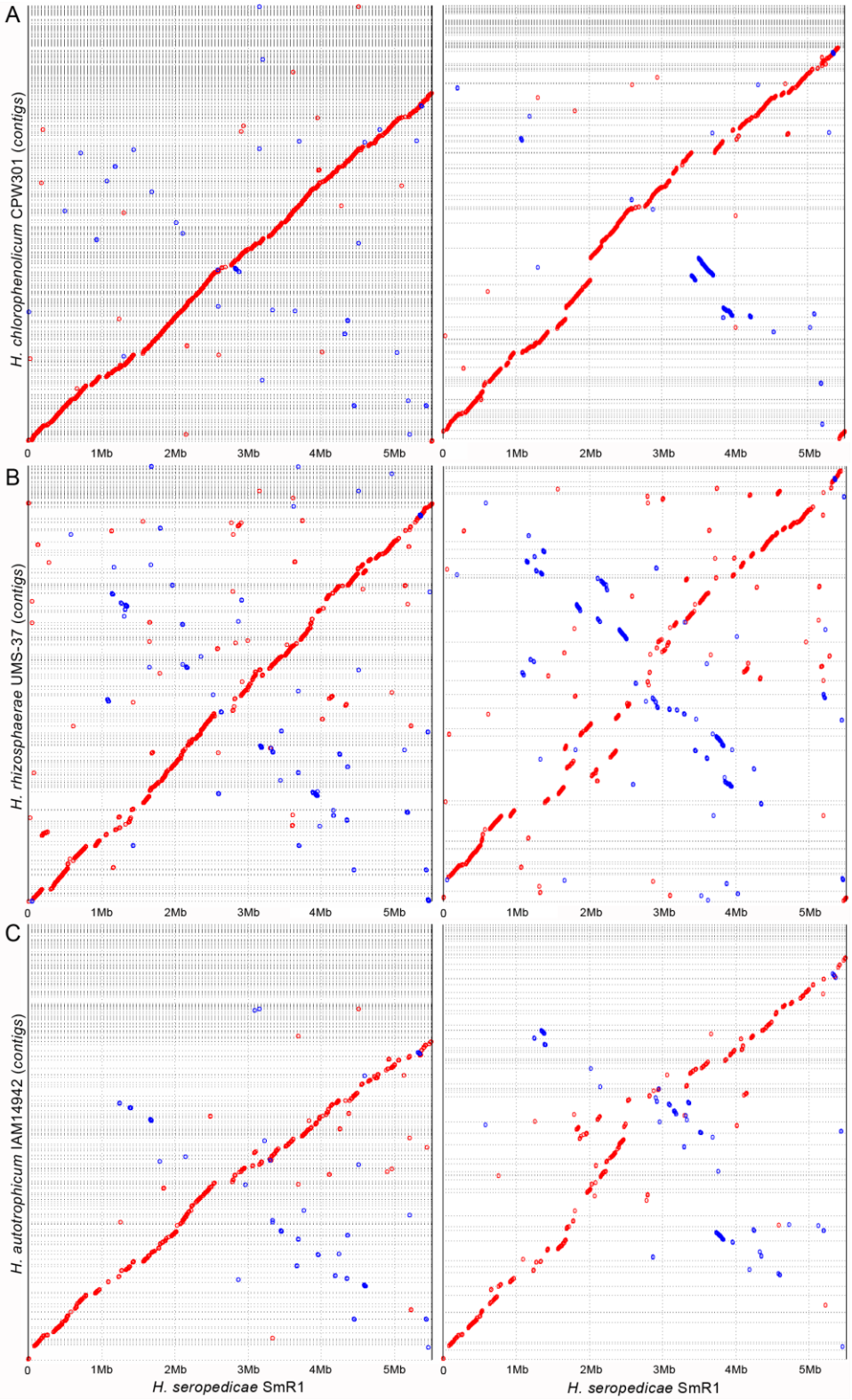


FIGURA 6.7: COMPARAÇÃO ENTRE OS GRÁFICOS DOTPLOT DAS MONTAGENS AUTOMÁTICAS ILLUMINA E DAS MONTAGENS FINAIS EM RELAÇÃO AO GENOMA DE *H. seropedicae* SmR1

À esquerda estão representadas as montagens automáticas Illumina e à direita as montagens finais de *H. chlorophenolicum* CPW301 (A), *H. rhizosphaerae* UMS-37 (B) e *H. autotrophicum* IAM 14942 (C). O genoma de *H. seropedicae* SmR1 foi usado como referência em todas as comparações.

FONTE: o autor (2015)

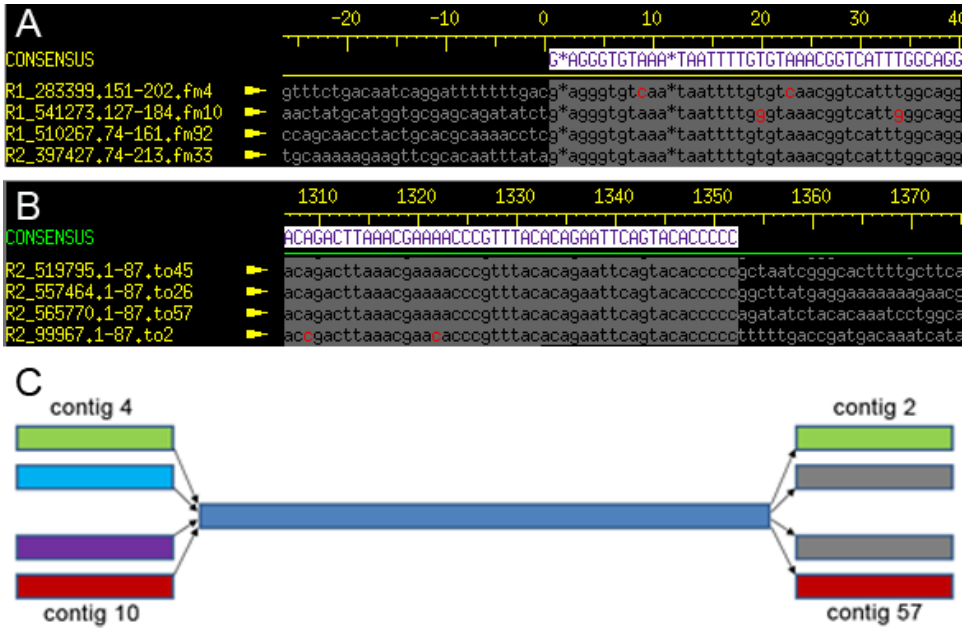


FIGURA 6.8: EXEMPLO DE REPETIÇÃO NA MONTAGEM GENÔMICA DE *H. chlorophenolicum* CPW301

Em 'A' e 'B' são mostradas as extremidades 5' e 3' do *contig*, respectivamente, e em 'C' é mostrado um modelo de como a repetição se comporta: 4 caminhos entram na repetição e 4 saem dela. Com base na referência foi possível supôr que o caminho que entra pelo *contig* 4 (.fm4) sai para o *contig* 2 (.to2), e o que entra pelo *contig* 10 (.fm10) sai pelo *contig* 57 (.to57), mas apenas *reads mate-pair* poderiam esclarecer esses caminhos.

FONTE: o autor (2015)

Entretanto, as tentativas de automatizar o fechamento de *gaps* falharam, assim como a tentativa de fechar os *gaps* manualmente, devido ao tempo despendido e às falhas do processo manual. Além disso, o fato dos *contigs* não formarem *scaffolds*, pela ausência de *reads mate-pair*, poderia induzir a erros de montagem se levada em conta somente a estrutura do genoma de referência. Esse processo seria ainda mais problemático para os outros dois genomas, pois as outras duas espécies são mais distantes evolutivamente da espécie usada como referência (FIGURA 6.7B e C).

Os *contigs* de *H. chlorophenolicum* CPW301 passaram por um filtro de tamanho e aqueles com tamanho menor que 200 pb foram descartados (os *reads* Illumina teriam tamanho suficiente para cobrir repetições menores que 200 bases, o que leva a acreditar que esses *contigs* representem alguma divergência de informação), o que resultou em 192 *contigs*, com conteúdo G+C de ~61,2%. Um resumo do processo de montagem e principais montagens envolvidas pode ser visualizado na FIGURA 6.9. O protocolo estabelecido para *H. chlorophenolicum* CPW301, com a montagem C2 e a montagem híbrida do tipo '*Reads* Illumina + C2'

pelo Newbler, foi também utilizado para as montagens genômicas de *H. autotrophicum* IAM 14942 e *H. rhizosphaerae* UMS-37 (FIGURA 6.7B e C).

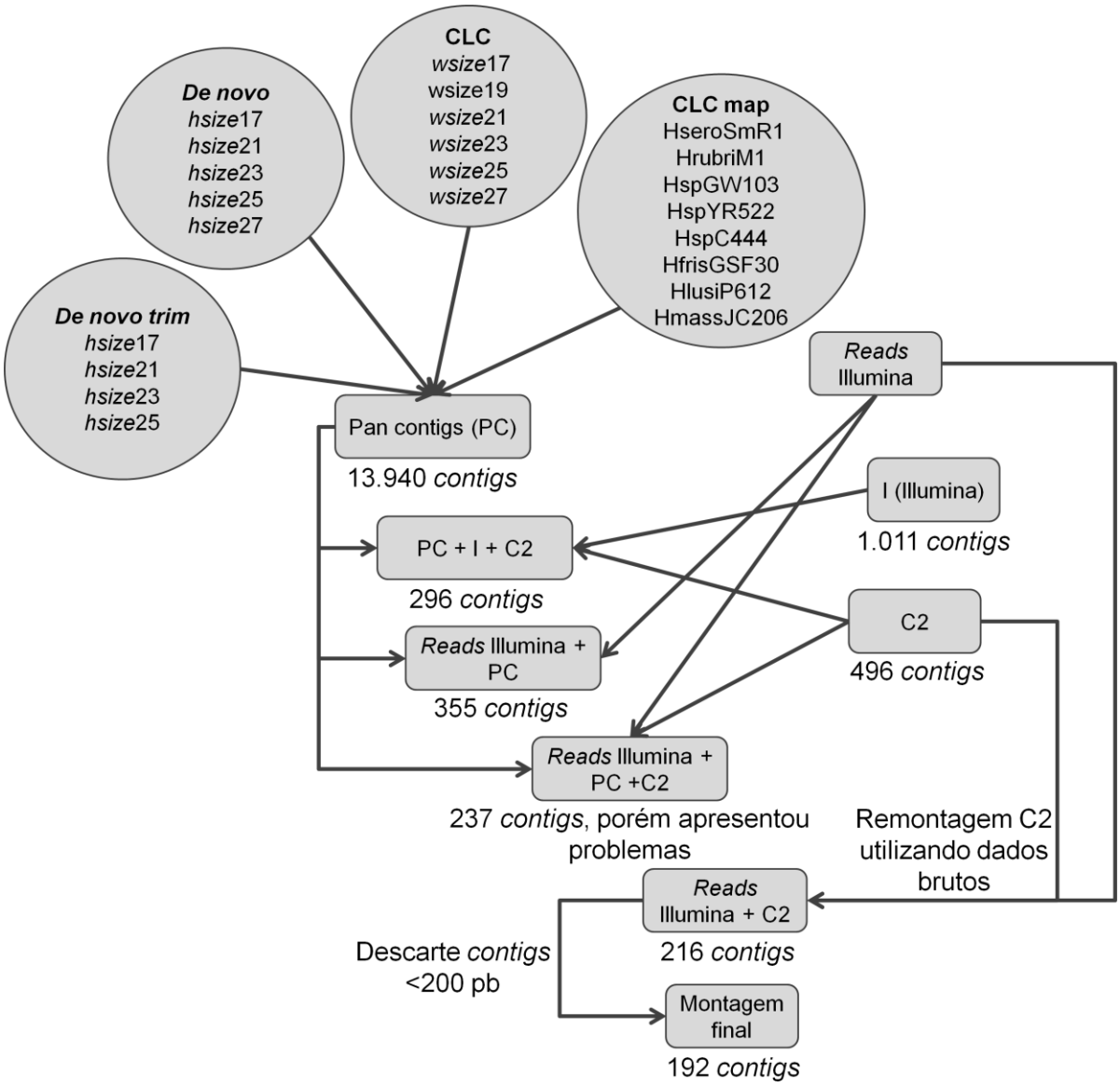


FIGURA 6.9: RESUMO DO PROCESSO DE MONTAGEM DO GENOMA DE *H. chlorophenolicum* CPW301

FONTE: o autor (2015)

As montagens C2 resultaram em 234 e 249 *contigs* para *H. autotrophicum* IAM 14942 e *H. rhizosphaerae* UMS-37, respectivamente. Após ser realizada a montagem híbrida, o número de *contigs* caiu para respectivamente 106 e 102, e após o filtro dos *reads* menores que 200 bases, para 99 *contigs* no genoma de *H. autotrophicum* IAM 14942 (conteúdo G+C de ~57,9%) e 55 *contigs* em *H. rhizosphaerae* UMS-37 (conteúdo G+C de 60,1%) (TABELA 6.9).

TABELA 6.9: DRAFTS GENÔMICOS FINAIS PARA AS MONTAGENS DE SEQUÊNCIA GENÔMICA DE *Herbaspirillum* spp.

Organismo	***Contigs***	**Tamanho do genoma**	**Conteúdo G+C**
H. rhizosphaerae UMS-37	55	5,266,109	~60.1%
H. chlorophenolicum CPW301	192	5,297,014	~61.2%
H. autotrophicum IAM 14942	99	6,009,835	~57,9%

FONTE: o autor (2015)

6.3 Anotação genômica de *H. autotrophicum* IAM 14942

6.3.1 Características gerais do genoma

Informações sobre o genoma de *H. autotrophicum* IAM 14942 podem ser visualizadas na TABELA 6.10. Foram anotados 2 genes que correspondem às subunidades 23S e 16S do RNA ribossomal (a subunidade 5S não foi anotada automaticamente), porém a análise do genoma leva a acreditar que existam duas cópias do operon rRNA (FIGURA 6.10). A categorização funcional segundo a plataforma RAST é mostrada na FIGURA 6.11.

6.3.2 Visão geral do metabolismo de H. autotrophicum IAM 14942

O metabolismo das *Herbaspirillum* spp*.* foi analisado com base nos testes bioquímicos realizados para a descrição das espécies e no que é descrito para *H. seropedicae* SmR1, entre outros *Herbaspirillum*. O genoma de *H. autotrophicum* IAM 14942 não apresenta o gene codificador da fosfofrutoquinase-1 (PFK-1, EC 2.7.1.11) na via de Embden-Meyerhoff-Parnas (via glicolítica – FIGURA 6.12) e nem o gene que codifica para a enzima 6-fosfogluconolactonase (EC 3.1.1.31 – FIGURA 6.12), utilizada na fase oxidativa da via das pentoses fosfato, o que impossibilitaria o metabolismo de açúcar por essas vias. Por outro lado, apresenta um gene que codifica para a 1-fosfofrutoquinase (*fruK*, EC 2.7.1.56) e para um sistema de

transporte PEP/PTS parcial (subunidade PTS-Fru-EIIB), responsáveis por gerar D-frutose-1,6P_2. Esses genes poderiam dar continuidade à via glicolítica. Já os genes relacionados com as vias metabólicas para a degradação de D-galactose e L-arabinose, não foram encontrados. Essa bactéria também não apresenta genes para importar monossacarídeos por sistemas de transporte ABC, com exceção de glicerol-3-fosfato.

TABELA 6.10:. INFORMAÇÕES GERAIS SOBRE O GENOMA DE *H. autotrophicum* IAM 14942

Característica	**Genoma (total)**	
	Valor	% do total[a]
Tamanho (bp)	6.009.835	100
Número de *contigs*	99	100
Conteúdo G+C (pb)	3.481.593	57,93
Regiões codificantes (pb)	5.293.194	88,07
Total de genes	5.639	100
RNAs	52	0,9
Proteínas codificadas	5.587	99,1
Genes com função atribuída ao COG	5.013	89,7

[a]O total é baseado no tamanho do genoma (em pares de bases) ou no número total de proteínas codificadas.

FONTE: o autor (2015)

O genoma apresenta o gene *otsB* para a síntese de trealose, que poderia funcionar como via para evitar o estresse osmótico. Genes envolvidos com a síntese e degradação de polímeros de glucose (genes *glgABX*) não foram encontrados no genoma parcial dessa bactéria.

Por outro lado, todos os genes da via da gluconeogênese estão presentes, inclusive os genes que codificam para a frutose-1,6-bifosfatase (EC 3.1.3.11) e para a fosfoenolpiruvato carboxiquinase (EC 4.1.1.32). O genoma parcial dessa bactéria também apresenta um gene que codifica para a enzima gliceraldeído-3-fosfato desidrogenase $NADP^+$-específica (EC 1.2.1.9 – FIGURA 6.12), a qual faz parte da

via glicolítica e que provavelmente é responsável por gerar NADPH na ausência da via das pentoses fosfato.

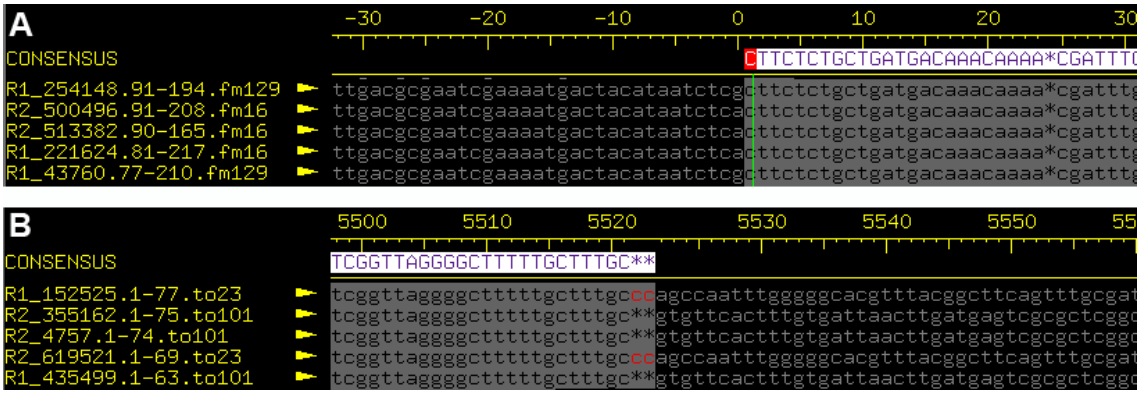


FIGURA 6.10: INDÍCIOS DE LIGAÇÃO DO OPERON 16S-23S-5S rRNA

Em ‘A’ e ‘B’ estão representadas as extremidades 5’ e 3’ do *contig* correspondente ao *operon* rRNA, respectivamente. É possível observar que a extremidade 5’ provém de dois caminhos, os *contigs* indicados por *fm16* e *fm129*, e a extremidade 3’ sai para outros dois caminhos, os *contigs* indicados por *to23* e *to101*. Isso leva a acreditar que existem duas cópias desse *operon*.

FONTE: o autor (2015)

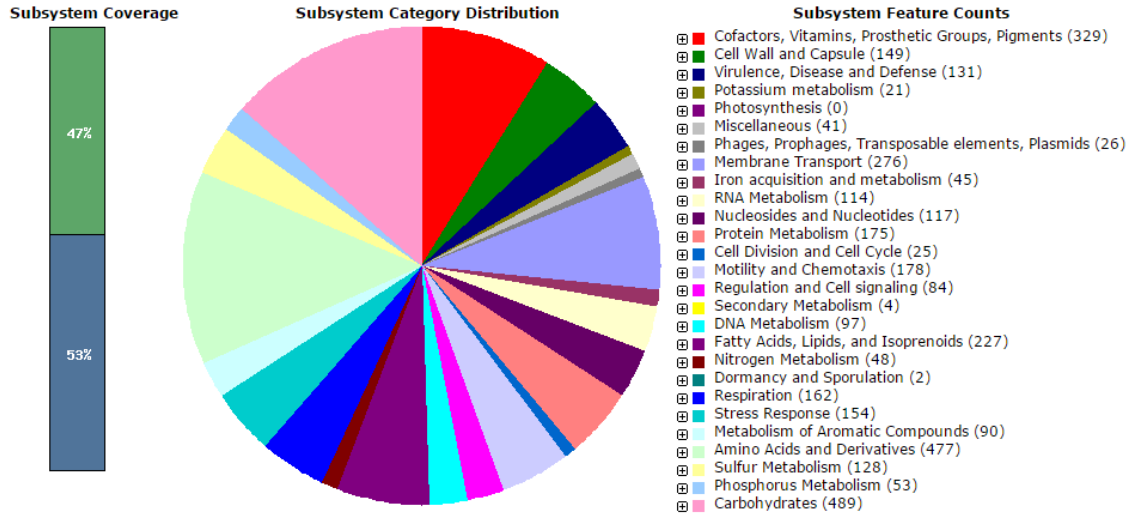


FIGURA 6.11: CATEGORIZAÇÃO FUNCIONAL DAS PROTEÍNAS DE *H. autotrophicum* IAM 14942 SEGUNDO A PLATAFORMA RAST

A barra da esquerda, em verde, representa a porcentagem de proteínas que possuem função atribuída pela plataforma RAST (47%).

FONTE: o autor (2015)

O genoma parcial de *H. autotrophicum* IAM 14942 apresenta os genes que codificam para as enzimas L-lactato desidrogenase (EC 1.1.1.27) e D-lactato desidrogenase – citocromo (EC 1.1.1.28). Também foram encontrados genes que codificam para a álcool desidrogenase (EC 1.1.1.1) e para a álcool desidrogenase –

citocromo (EC 1.1.2.8). Além disso, essa bactéria apresenta genes relacionados com o metabolismo de 1-butanol.

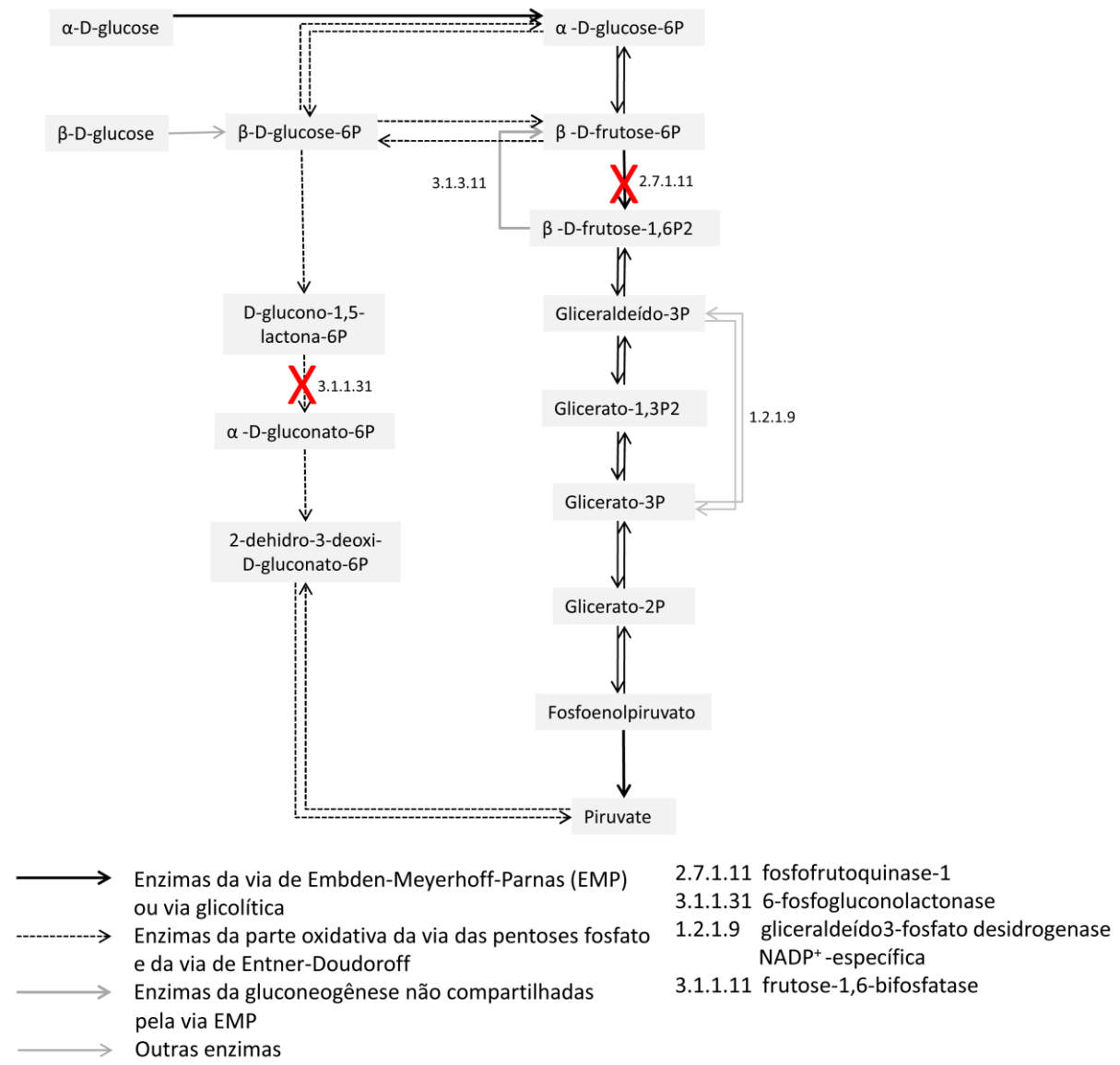


FIGURA 6.12: POSSÍVEIS VIAS PARA O METABOLISMO DE D-GLUCOSE EM *H. autotrophicum* IAM 14942

A via glicolítica e a parte oxidativa da via das pentoses fosfato estão incompletas, onde a ausência das enzimas fosfofrutoquinase-1 e 6-fosfogluconolactonase são representadas por "X".

FONTE: o autor (2015)

Estão também presentes os genes que codificam as enzimas necessárias para formar o complexo da piruvato desidrogenase e para converter piruvato a acetil-CoA, os genes que codificam as enzimas do ciclo do ácido cítrico, os genes que codificam para as proteínas que compõe as citocromo c oxidades aa3, bd e cbb3, assim como os genes que codificam para as proteínas que compõe a NADH desidrogenase, succinato desidrogenase, citocromo c redutase e ATP sintase.

6.3.3 Metabolismo de aminoácidos e nitrogênio em H. autotrophicum IAM 14942

Os 20 aminoácidos foram testados como fonte de carbono e energia para *H. autotrophicum* IAM 14942 (ARAGNO & SCHLEGEL, 1978) e, por isso, o metabolismo de aminoácidos foi investigado no genoma dessa bactéria. Foi verificado que ela apresenta o conjunto de genes que codificam para as enzimas das vias de degradação de L-asparagina, L-aspartato, L-glutamina, L-glutamato, L-alanina, L-serina, L-glicina, L-cisteína, L-arginina e L-prolina a piruvato ou intermediários do ciclo do citrato. Entre as vias de degradação de aminoácidos, *H. autotrophicum* IAM 14942 apresenta o conjunto de genes que codifica para as enzimas necessárias para a síntese e degradação de ureia, através do ciclo da ureia, com exceção do gene que codifica para a ornitina carbomiltransferase (EC 2.1.3.3). Além disso, foi observada a presença do *operon* que codifica para a urease, o qual contém os genes *ureABC*, e também a presença de genes que codificam para um transportador ABC completo para ureia (genes *urtABCDE*).

H. autotrophicum IAM 14942 apresenta em sua sequência genômica parcial os genes que codificam transportadores ABC para glutamato/aspartato, D-metionina e taurina. O genoma dessa bactéria também possui genes para o transporte de espermidina/putrescina e genes que codificam enzimas presentes na via de conversão de putrescina a espermina.

Como o gênero *Herbaspirillum* é conhecido pela fixação biológica de nitrogênio (BALDANI & BALDANI, 2005), o metabolismo de nitrogênio também foi analisado. Nas verificações *in silico* não foram encontrados os genes *nifHDK*, responsáveis por codificar para as proteínas que formam o complexo da nitrogenase, nem os genes relacionados às enzimas responsáveis pela redução dissimilatória do nitrato. Por outro lado, os genes relacionados às enzimas responsáveis pela redução assimilatória de nitrato (*nasA*), pela conversão de nitrito a amônia (genes *nirBD*) e genes relacionados com o transportador ABC para nitrato/nitrito/cianato estão presentes no genoma parcial dessa bactéria.

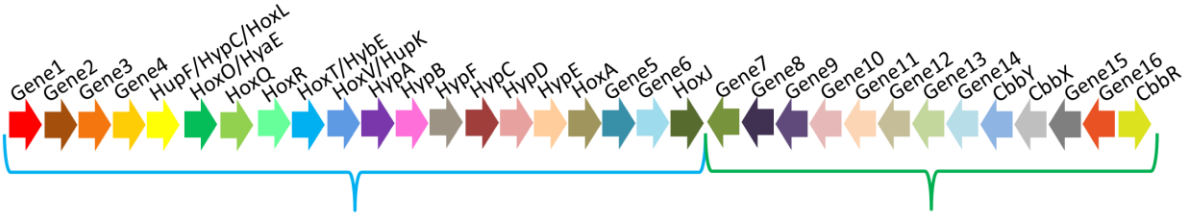
6.3.4 Fixação de carbono em H. autotrophicum IAM 14942

A espécie *H. autotrophicum* é descrita como capaz de fixar carbono autotroficamente (ARAGNO & SCHLEGEL, 1978). Dessa forma, foram encontrados os genes que codificam para as enzimas fosforibuloquinase (EC 2.7.1.19) e RuBisCO (EC 4.1.1.39), que pertencem ao ciclo de Calvin. *H. autotrophicum* IAM 14942 pode produzir ribulose-5P via transcetolase (EC 2.2.1.1) utilizando gliceraldeído-3P e sedoheptulose-7P. Esses genes estão em uma mesma região do genoma e formam um agrupamento gênico (FIGURA 6.13). Porém, os genes que codificam para as enzimas sedoheptulose-1,7-bifosfatase (EC 3.1.3.37) e sedoheptuloquinase (EC 2.7.1.14) não foram encontrados.

A espécie também foi descrita como capaz de utilizar hidrogênio como fonte de energia (ARAGNO & SCHLEGEL, 1978). Com isso, a análise do genoma parcial dessa bactéria revelou a presença dos genes *hoxAJLOQRTV* e dos genes *hypABCDEF*, relacionados com o metabolismo litoautotrófico. Esses genes formam um agrupamento vizinho ao agrupamento de genes responsáveis pela fixação de carbono (FIGURA 6.13). O genoma parcial dessa bactéria ainda apresenta os genes que codificam para a formato desidrogenase (EC 1.2.1.2) e formaldeído desidrogenase independente de glutationa (EC 1.2.1.46), que poderiam ser responsáveis pelo metabolismo organo-autotrófico.

6.3.5 Outras características metabólicas de H. autotrophicum IAM 14942

Aspectos metabólicos considerados relevantes, por terem sido observados em outros *Herbaspirillum*, também foram analisados. O genoma de *H. autotrophicum* IAM 14042 apresenta os genes *phbB*, *phbC/phaC* e *phaZ*, que codificam as enzimas responsáveis pelo metabolismo de poli(3-hidroxi-alcanoatos). Também foram encontrados genes que codificam para enzimas envolvidas na degradação de derivados de catecol e genes que codificam para enzimas relacionadas com a degradação de nitrobenzeno, vanilato e 4-nitrocatecol. Essa bactéria apresenta os genes *ppk* e *ppx*, que codificam enzimas envolvidas no metabolismo de polifosfato.



Gene1	Uptake hydrogenase small subunit precursor (EC 1.12.99.6)
Gene2	Uptake hydrogenase large subunit (EC 1.12.99.6)
Gene3	Ni,Fe-hydrogenase I cytochrome b subunit
Gene4	Hydrogenase maturation protease (EC 3.4.24.-)
Gene5	Uptake hydrogenase small subunit precursor (EC 1.12.99.6)
Gene6	Uptake hydrogenase large subunit (EC 1.12.99.6)
Gene7	Fructose-bisphosphate aldolase class II (EC 4.1.2.13)
Gene8	Phosphoglycerate kinase (EC 2.7.2.3)
Gene9	NAD-dependent glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12)
Gene10	Phosphoglycolate phosphatase (EC 3.1.3.18)
Gene11	Transketolase (EC 2.2.1.1)
Gene12	Phosphoribulokinase (EC 2.7.1.19)
Gene13	Fructose-1,6-bisphosphatase, type I (EC 3.1.3.11)
Gene14	Ribulose-phosphate 3-epimerase (EC 5.1.3.1)
Gene15	Ribulose bisphosphate carboxylase small chain (EC 4.1.1.39)
Gene16	Ribulose bisphosphate carboxylase large chain (EC 4.1.1.39)

FIGURA 6.13: ORGANIZAÇÃO DOS AGRUPAMENTOS DE GENES RELACIONADOS COM O COMPLEXO DA HIDROGENASE E COM A FIXAÇÃO DE CARBONO

Os agrupamentos gênicos são descritos como seus produtos protéicos. A proteínas Hox se referem às proteínas estruturais do complexo da hidrogenase, as proteínas Hyp são auxiliares à formação do complexo, e as proteínas Cbb são auxiliares à RuBisCO (‘*Ribulose-phosphate carboxylase*’). As demais proteínas codificadas pelos respectivos genes se encontram na legenda abaixo da figura.

FONTE: o autor (2015)

No genoma parcial dessa bactéria não foi encontrado o agrupamento de genes relacionados ao T3SS (sistema de secreção do tipo III), nem foram encontrados genes pertencentes ao T6SS (sistema de secreção do tipo VI), observados na sequência genômica de outros *Herbaspirillum*. Porém, foram encontrados os genes relacionados com os sistemas Tat (do inglês “*Twin arginine translocation*”) e Sec, para exportação de proteínas. Uma visão geral das características metabólicas de *H. autotrophicum* IAM 14942 pode ser visualizada na FIGURA 6.14.

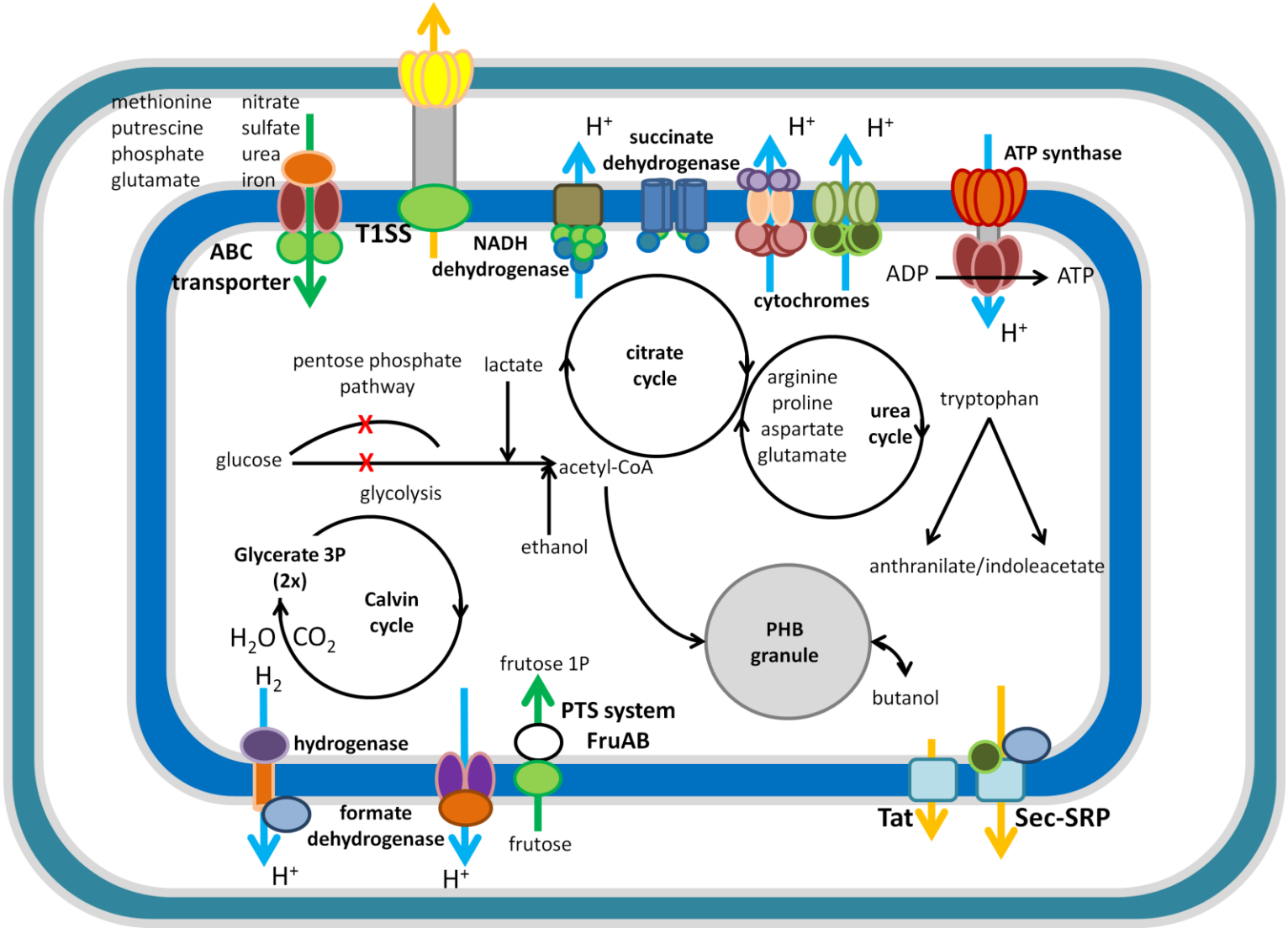


FIGURA 6.14: VISÃO GERAL DAS CARACTERÍSTICAS METABÓLICAS DE *H. autotrophicum* IAM 14942

FONTE: o autor (2015)

6.4 Anotação genômica de *H. chlorophenolicum* CPW301

6.4.1 Características gerais do genoma de H. chlorophenolicum CPW301

Informações sobre o genoma de *H. chlorophenolicum* CPW301 podem ser visualizadas na TABELA 6.11. Foram anotados 3 genes que correspondem às subunidades do 23S, 16S e 5S do RNA ribossomal (TABELA 6.11), mas a análise do *contig* referente a esses genes leva a acreditar que, assim como em *H. autotrophicum* IAM 14942, existam duas cópias do operon rRNA (FIGURA 6.15). A categorização funcional segundo a plataforma RAST é mostrada na FIGURA 6.16.

TABELA 6.11: INFORMAÇÕES GERAIS SOBRE O GENOMA DE *H. chlorophenolicum* CPW301

Característica	**Genoma (total)**	
	Valor	% do total[a]
Tamanho (bp)	5.297.014	100
Número de *contigs*	192	100
Conteúdo G+C (pb)	3.242.677	61,22
Regiões codificantes (pb)	4.531.296	85,54
Total de genes	4.817	100
RNAs	51	1,06
Proteínas codificadas	4.766	98,9
Genes com função atribuída ao COG	4.294	89,1

[a]O total é baseado no tamanho do genoma (em pares de bases) ou no número total de proteínas codificadas.

FONTE: o autor (2015)

6.4.2 Visão geral do metabolismo de H. chlorophenolicum CPW301

Assim como já foi reportado para outras estirpes de *Herbaspirillum* spp., a sequência genômica parcial de *H. chlorophenolicum* CPW301 também não apresenta o gene codificador da fosfofrutoquinase-1 (PFK-1, EC 2.7.1.11), mas

possui os genes relacionados com a via das pentoses fosfato e com a via de Entner-Doudoroff, o que permitiria metabolizar D-glucose e D-frutose a piruvato. Genes relacionados com hexoquinase (EC 2.7.1.1), sistema PTS manose-específico (EC 2.7.1.69) ou com manoquinase (EC 2.7.1.7) responsáveis pela conversão de D-manose a D-manose-6P, não foram encontrados. Genes relacionados com o metabolismo de ramnose, L-arabinose e D-galactose também não foram encontrados. No genoma parcial dessa bactéria estão presentes genes relacionados com sistemas de transporte ABC completos para ribose/D-xilose e glicerol-3-fosfato, além de subunidades dos sistemas ABC para o transporte de glucose/manose e D-xilose. Porém, genes relacionados com o metabolismo de xilose estão ausentes.

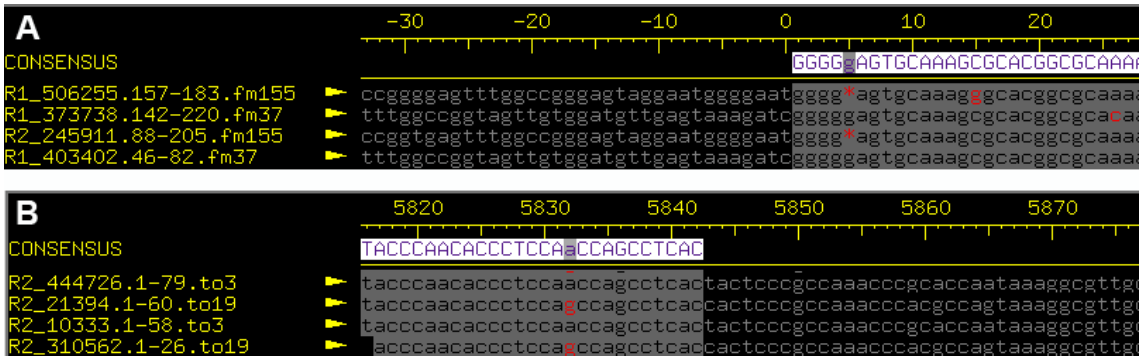


FIGURA 6.15: INDÍCIOS DE LIGAÇÃO DO OPERON 16S-23S-5S rRNA

Em ‘A’ e ‘B’ estão representadas as extremidades 5’ e 3’ do *contig* correspondente ao *operon* rRNA, respectivamente. É possível observar que a extremidade 5’ provém de dois caminhos, os *contigs* indicados por *fm37* e *fm155*, e a extremidade 3’ sai para outros dois caminhos, os *contigs* indicados por *to3* e *to19*. Isso leva a acreditar que existem duas cópias desse *operon*.

FONTE: o autor (2015)

A sequência genômica parcial de *H. chlorophenolicum* CPW301 apresenta o gene *otsB* para a produção de trealose, mas não apresenta genes envolvidos na produção de polímeros de glucose. Todos os genes da via da gluconeogênese estão presentes no genoma parcial dessa bactéria. Esse genoma apresenta também os genes que codificam as enzimas necessárias para formar o complexo da piruvato desidrogenase e para converter piruvato a acetil-CoA.

H. chlorophenolicum CPW301 não apresenta em sua sequência genômica o gene que codifica para a enzima L-lactato desidrogenase (EC 1.1.1.27), mas foi encontrado o gene para a álcool desidrogenase (EC 1.1.1.1). Todas as enzimas do ciclo do ácido cítrico estão presentes, além dos genes relacionados com a NADH desidrogenase, succinato desidrogenase, citrocromo C redutase e oxidase, complexos cbb3 e bd e ATP-sintase.

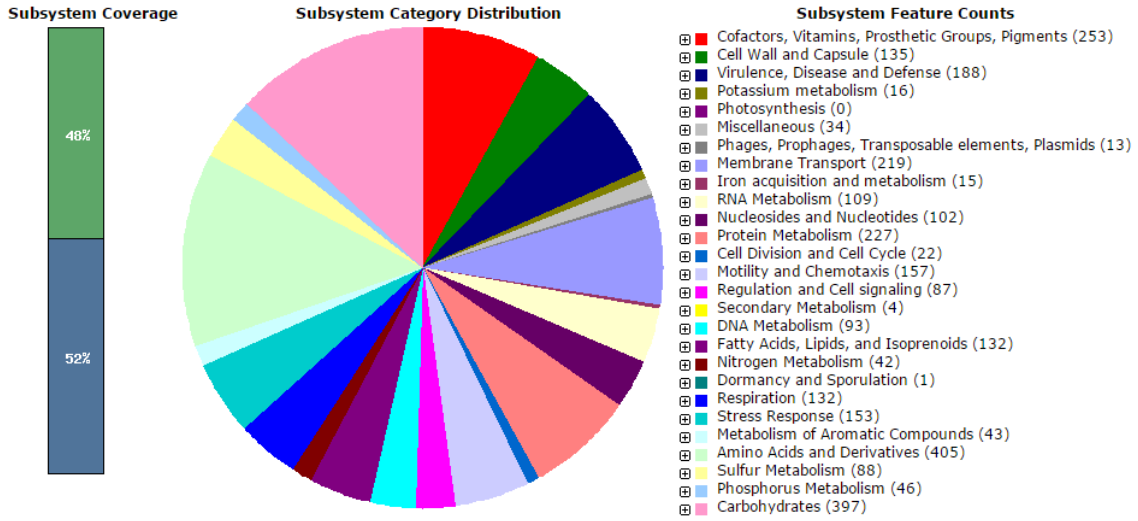


FIGURA 6.16: CATEGORIZAÇÃO FUNCIONAL DAS PROTEÍNAS DE *H. chlorophenolicum* CPW301 SEGUNDO A PLATAFORMA RAST

A barra da esquerda, em verde, representa a porcentagem de proteínas que possuem função atribuída pela plataforma RAST (48%).

FONTE: o autor (2015)

6.4.3 Metabolismo de aminoácidos e nitrogênio em H. chlorophenolicum CPW301

Assim como para *H. autotrophicum* IAM 14942, foi analisado o metabolismo de aminoácidos de *H. chlorophenolicum* CPW301, embora testes bioquímicos com aminoácidos não tenham sido realizados para essa espécie. As análises *in silico* mostraram que ela poderia degradar a piruvato ou intermediários do ciclo do ácido cítrico: L-alanina, L-aspartato, L-asparagina, L-glutamato, L-glutamina, L-glicina, L-serina, e L-cisteína. Embora apresente a urease, a arginase/agmatinase/ formimionoglutamate hidrolase (EC 3.5.3.1), responsável pela degradação de arginina a ureia, está ausente. Porém, a arginina poderia ser degradada a putrescina via agmatinase (EC 3.5.3.11). Além disso, o genoma de *H. chlorophenolicum* CPW301 apresenta os genes que codificam transportadores ABC para glutamato/aspartato, D-metionina e cistina, além de possuir genes para o transporte de ureia e aminoácidos de cadeia ramificada.

O metabolismo de nitrogênio também foi analisado, de modo que *H. chlorophenolicum* CPW301 não apresenta em seu genoma os genes *nifHDK*, responsáveis pela fixação biológica de nitrogênio e também seria incapaz de reduzir nitrato a nitrito. Porém, seria capaz de realizar a redução assimilatória de nitrato

(gene *nasA*), a conversão de nitrito a amônia (genes *nirBD*) e transportar nitrato/nitrito/cianato por transportador do tipo ABC.

6.4.4 Degradação de fenol e 4-clorofenol em H. chlorophenolicum CPW301

A espécie *H. chlorophenolicum* é descrita como capaz de degradar fenol e 4-clorofenol (IM *et al*., 2004). No genoma da estirpe CPW301 foram encontrados genes que codificam enzimas envolvidas na degradação de benzeno e fenol, de modo que esses compostos poderiam ser degradados a piruvato e acetil-CoA. O genoma dessa bactéria também apresenta o conjunto de genes para a degradação de tolueno, mas não para a degradação de xileno. Ainda foram encontrados genes envolvidos com a degradação de salicilato.

Conforme as análises *in silico*, aparentemente *H. chlorophenolicum* CPW301 degrada 4-clorofenol via 4-clorocatecol e, posteriormente, via clivagem *meta* (ARORA & BAE, 2014), mas a primeira enzima dessa via, 4-clorofenol-2-monoxigenase (EC 1.14.13.20), não foi encontrada. A enzima posterior dessa via, catecol-2,3-dioxigenase (EC 1.13.11.2), responsável por gerar semi-aldeído 5-cloro-2-hidoximucônico (5C2HMS), foi encontrada, mas não a enzima subsequente (sem nome, EC 1.97.1.-). Três enzimas completam a via de degradação desse composto a piruvato e acetaldeído (EC 3.7.1.9; EC 4.2.1.80; e EC 4.1.3.39), das quais apenas a última não foi encontrada.

A árvore filogenética da catecol-2,3-dioxigenase de *H. chlorophenolicum* CPW301 e suas proteínas homólogas obtidas via BLAST contra o banco de dados NR do NCBI, e com adição da catecol-2,3-dioxigenase de *Herbaspirillum sp* GW103, revelou que essa proteína é evolutivamente relacionada com a homóloga encontrada na estirpe RV1423 de *Herbaspirillum* e com a homóloga encontrada em *Xenophilus azovorans* (FIGURA 6.17).

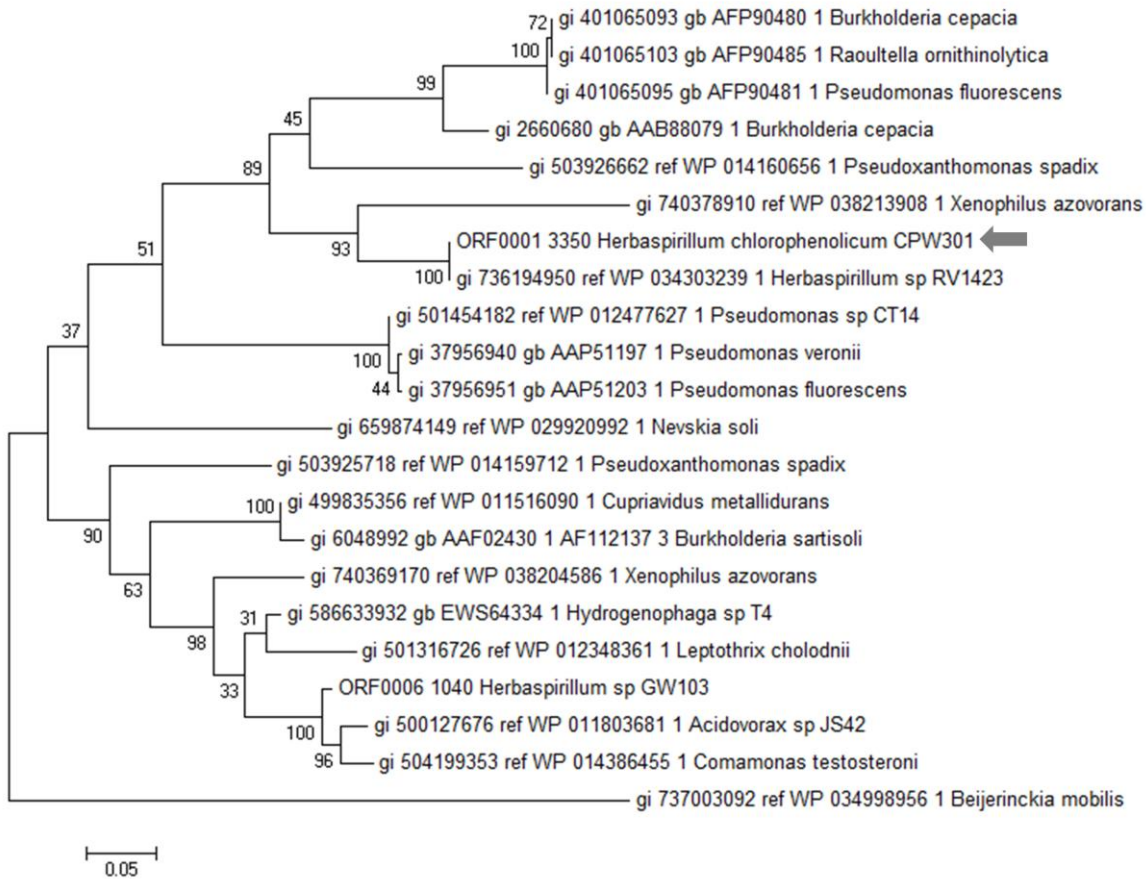


FIGURA 6.17: ÁRVORE FILOGENÉTICA DA CATECOL-2,3-DIOXIGENASE

A árvore foi construída dentro do programa MEGA 6.0, com alinhamento produzido pelo MUSCLE, com método de construção da árvore *Maximum Likelihood* e com 1000 replicatas de *bootstrap*.

FONTE: o autor (2015)

6.4.5 Outras características metabólicas de H. chlorophenolicum CPW301

Aspectos metabólicos de interesse em *Herbaspirillum* também foram analisados. O genoma dessa bactéria apresenta os genes para o metabolismo de poli(3-hidroxi-alcanoatos – genes *phbBC*, *phaC*, *phaZ*). Também apresenta dois genes para duas RubisCOs-*like*, além de possuir em sua sequência genômica genes relacionados com o T3SS, que por sua vez é encontrado em *Herbaspirillum* associados com plantas (STRAUB *et al*., 2013). Uma visão geral das características metabólicas dessa bactéria pode ser vista na FIGURA 6.18.

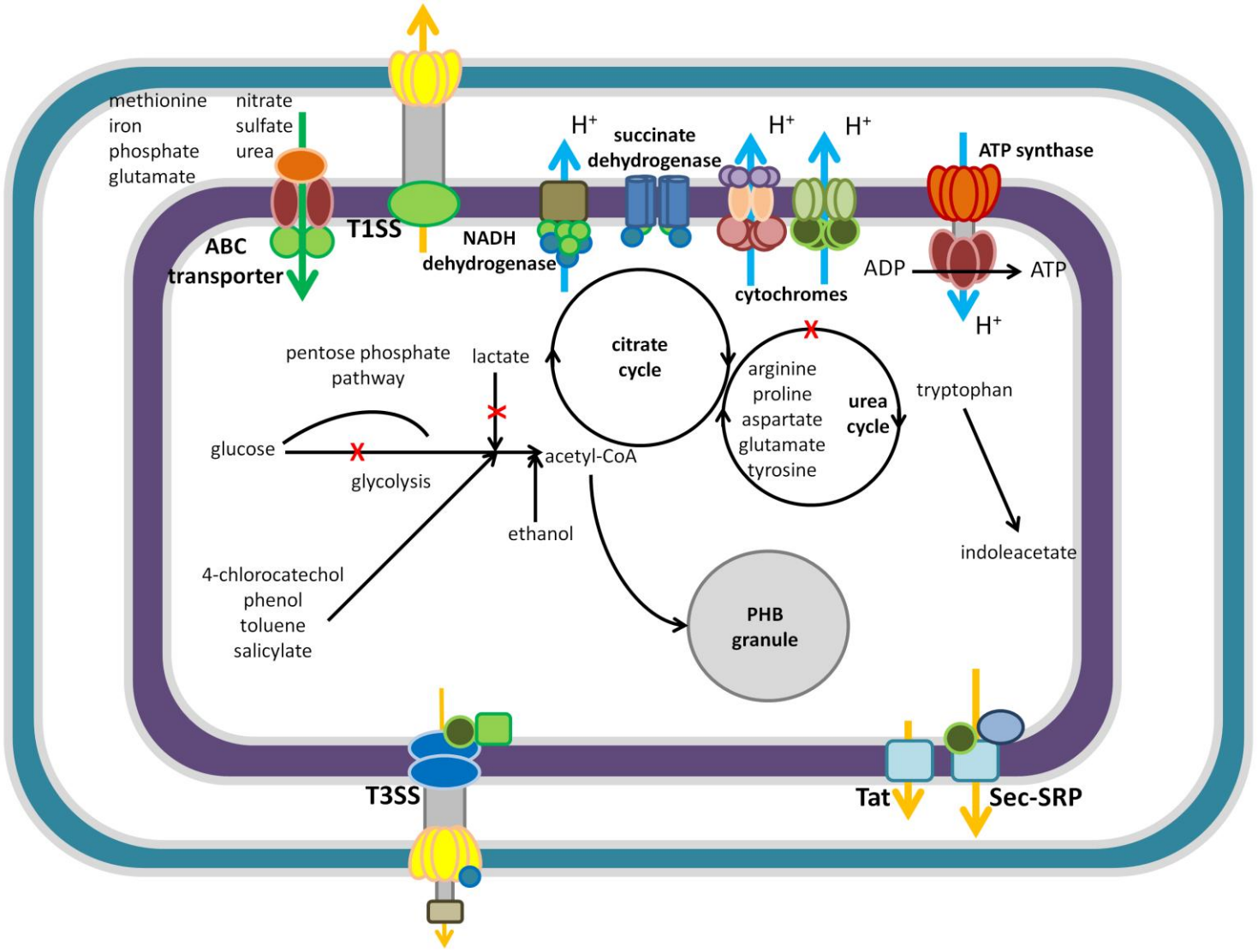


FIGURA 6.18: VISÃO GERAL DAS CARACTERÍSTICAS METABÓLICAS DE *H. chlorophenolicum* CPW301

FONTE: o autor (2015)

6.5 Anotação genômica de *H. rhizosphaerae* UMS-37

6.5.1 Características gerais do genoma

Informações sobre o genoma de *H. rhizosphaerae* UMS-37 podem ser visualizadas na TABELA 6.12. Foram anotados três genes que correspondem às subunidades do 23S, 16S e 5S do RNA ribossomal, porém a análise do genoma leva a acreditar que, diferentemente dos genomas dos dois *Herbaspirillum* descritos anteriormente, existam três cópias do operon rRNA (FIGURA 6.19). A categorização funcional segundo a plataforma RAST é mostrada na FIGURA 6.20.

TABELA 6.12: INFORMAÇÕES GERAIS SOBRE O GENOMA DE *H. rhizosphaerae* UMS-37

Característica	**Genoma (total)**	
	Valor	% do total[a]
Tamanho (bp)	5.266.109	100
Número de *contigs*	55	100
Conteúdo G+C (pb)	3.165.774	60,12
Regiões codificantes (pb)	4.610.050	87,54
Total de genes	4.864	100
RNAs	51	1,05
Proteínas codificadas	4.813	98,95
Genes com função atribuída ao COG	4.364	89,7

[a]O total é baseado no tamanho do genoma (em pares de bases) ou no número total de proteínas codificadas.

FONTE: o autor (2015)

6.5.2 Visão geral do metabolismo de H. rhizosphaerae UMS-37

O genoma de *H. rhizosphaerae* UMS-37 também não apresenta a fosfofrutoquinase-1 (PFK-1, EC 2.7.1.11), como os demais *Herbaspirillum*, porém a rota para o metabolismo de D-glucose, D-frutose e D-manose pode desviar para a

via das pentoses fosfato. O gene que codifica para a D-galactose 1-desidrogenase (EC 1.1.1.48) não foi encontrado. Esse genoma apresenta genes relacionados com transportadores ABC completos para: glucose/manose; ribose/D-xilose; e D-xilose. Porém, os genes relacionados com hexoquinase (EC 2.7.1.1), sistema PTS manose-específico (EC 2.7.1.69) ou com manoquinase (EC 2.7.1.7), responsáveis pela conversão de D-manose a D-manose-6P, não foram encontrados. O gene que codifica para a L-iditol 2-desidrogenase, responsável por converter sorbitol a D-frutose, está presente no genoma parcial dessa bactéria. Para a produção de trealose foram encontrados tanto o gene *ostB* quanto os genes *treXYZ*, mas o genoma dessa bactéria não apresenta os genes relacionados com a produção de polímeros de glucose. O gene relacionado com a beta-galactosidase (EC 3.2.1.23) também não foi encontrado.

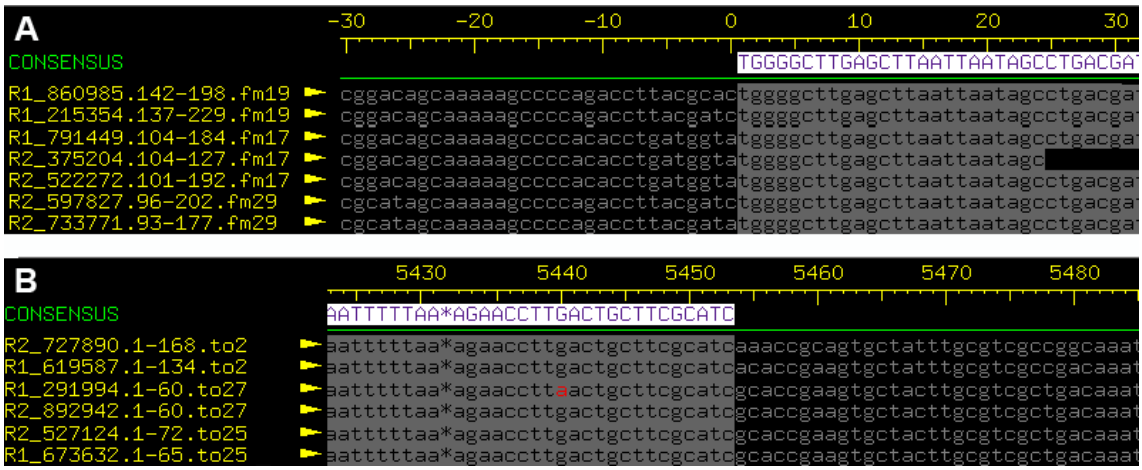


FIGURA 6.19: INDÍCIOS DE LIGAÇÃO DO OPERON 16S-23S-5S rRNA

Em ‘A’ e ‘B’ estão representadas as extremidades 5’ e 3’ do *contig* correspondente ao *operon* rRNA, respectivamente. É possível observar que a extremidade 5’ provém de três caminhos, *contigs* indicados por *fm17*, *fm19* e *fm29*, e a extremidade 3’ sai para outros três caminhos, *contigs* indicados por *to2*, *to25* e *to27*. Isso leva a acreditar que existem três cópias desse *operon*.

FONTE: o autor (2015)

Os genes que codificam para todas as enzimas do ciclo do ácido cítrico estão presentes, juntamente com os genes que codificam para a NADH desidrogenase, succinato desidrogenase, citrocromo C redutase e oxidase, complexo bd e também ATPase. No entanto, não foram encontrados genes que codificam para o complexo cbb3.

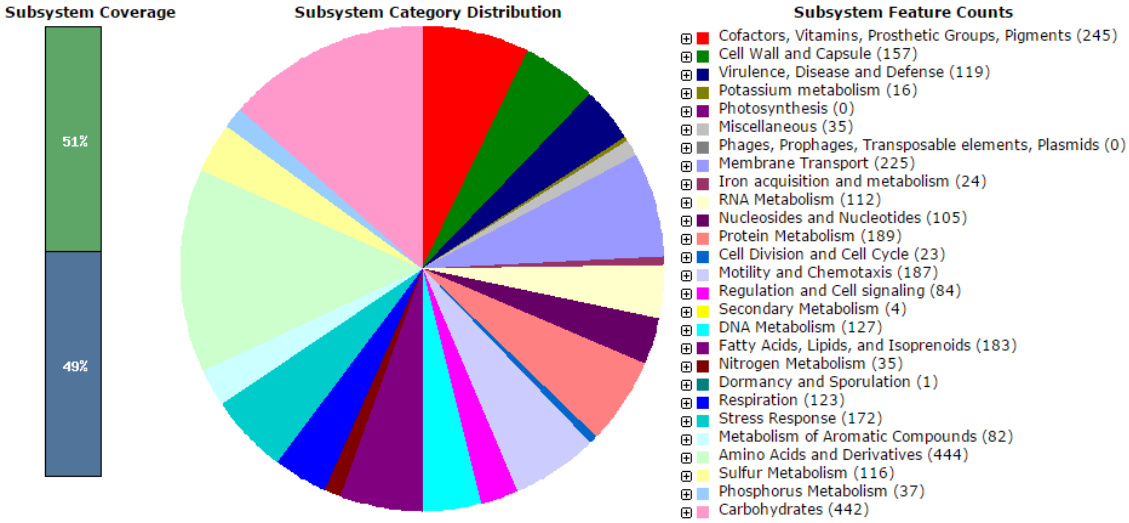


FIGURA 6.20: CATEGORIZAÇÃO FUNCIONAL DAS PROTEÍNAS DE *H. rhizosphaerae* UMS-37 SEGUNDO A PLATAFORMA RAST

A barra da esquerda, em verde, representa a porcentagem de proteínas que possuem função atribuída pela plataforma RAST (51%).

FONTE: o autor (2015)

6.5.3 Metabolismo de amonoácidos e de nitrogênio em H. rhizosphaerae UMS-37

Através da análise genômica é sugerido que essa bactéria degrade L-alanina, L-aspartato, L-asparagina, L-glutamato, L-glutamina, L-glicina, L-serina, L-cisteína e L-arginina a piruvato ou intermediários do ciclo do ácido cítrico. A arginina poderia ser degradada via ciclo da ureia, de maneira que genes relacionados a esse ciclo, bem como à urease, foram encontrados. *H. rhizosphaerae* UMS-37 poderia degradar arginina também via agmatinase (EC 3.5.3.11). No genoma parcial desse organismo foram encontrados genes que codificam transportadores ABC para: glutamato/aspartato; D-metionina; aminoácidos de cadeia ramificada; e o transportador de ureia.

Os genes que codificam para a nitrogenase não foram encontrados no genoma parcial de *H.rhizosphaerae* UMS-37. Da mesma forma que os *Herbaspirillum* não fixadores de nitrogênio, o genoma dessa bactéria apresenta os genes *nasAB* e *nirBD* e os genes relacionados com o transportador ABC para nitrato/nitrito/cianato.

6.5.4 Metabolismo de compostos fenólicos em H. rhizosphaerae UMS-37

Assim como para *H. chlorophenolicum* CPW301, a via de degradação de 4-clorofenol foi analisada no genoma de *H. rhizosphaerae* UMS-37, embora ela não tenha sido descrita como uma bactéria que degrada 4-clorofenol. No genoma de *H. rhizosphaerae* UMS-37 foram encontrados genes relacionados com a via de degradação desse composto, que seguiria pela clivagem *orto*, com presença do gene que codifica a enzima catecol-1,2-dioxigenase (EC 1.13.11.1) para gerar 3-cloromuconato (ARORA & BAE, 2014). Na etapa seguinte da via, o 3-cloromuconato seria convertido a uma toxina chamada protoanemonina pela enzima muconato cicloisomerase (EC 5.5.1.1). Foi verificado que essa enzima está presente também no genoma de outros *Herbaspirillum*. A árvore filogenética da muconato cicloisomerase e suas proteínas homólogas, obtidas via BLAST contra o banco de dados NR do NCBI, e com adição das enzimas homólogas em *Herbaspirillum*, revelou que essa proteína é evolutivamente relacionada com as homólogas encontradas em *Herbaspirillum* em geral, principalmente com a homóloga presente na estirpe CF444 e com as homólogas encontradas em *Collimonas arenae* e *Azohydromonas australica* (FIGURA 6.21).

H. rhizospharae UMS-37 também apresenta o gene que codifica para a enzima salicilato hidroxilase (EC 1.14.13.1), responsável pela conversão de salicilato a catecol, e os genes que codificam as enzimas responsáveis pela degradação desse composto.

6.5.5 Outras características metabólicas de H. rhizosphaerae UMS-37

Como as rizobactérias são conhecidas por produzirem antibióticos (BENEDUZI *et al*., 2012), algumas vias para a produção desses compostos foram analisadas. Além da protoanemonina, foi encontrado no genoma de *H. rhizosphaerae* UMS-37 o gene que codifica para a penicilina amidase (EC 3.5.1.11), que converte penicilina em ácido 6-aminopenicilânico, e o gene que codifica para a beta-lactamase classe A (EC 3.5.2.6).

Outras características observadas em *H. rhizosphaerae* UMS-37 são a presença dos sistemas de exportação de proteínas Sec e Tat, bem como a presença

de genes responsáveis pela produção de poli(3-hidroxi-alcanoatos). Uma visão geral das características metabólicas desse organismo pode ser vista na FIGURA 6.22.

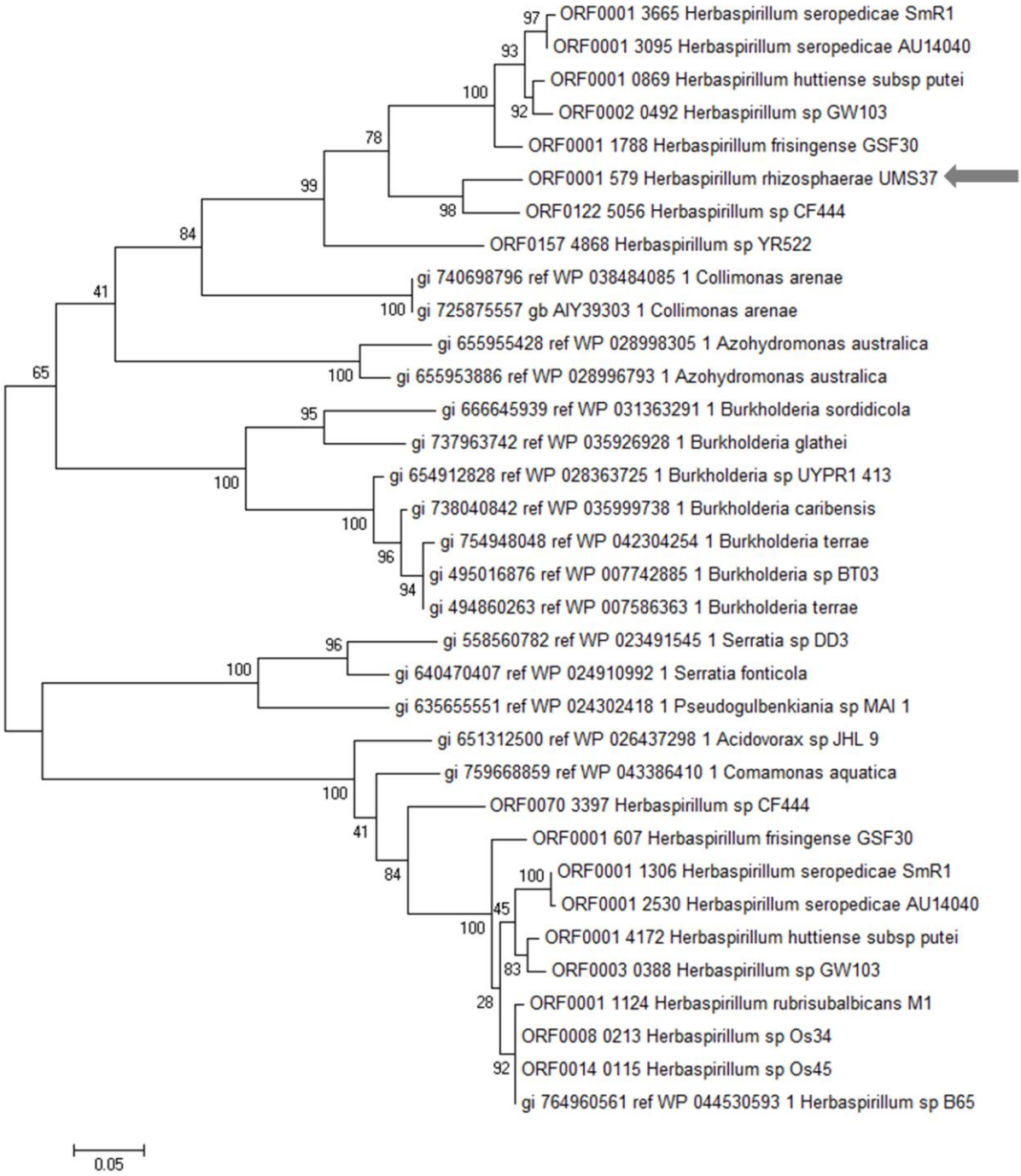


FIGURA 6.21: ÁRVORE FILOGENÉTICA DE MUCONATO CICLOISOMERASES

A árvore foi construída dentro do programa MEGA 6.0, com alinhamento produzido pelo MUSCLE, com método de construção da árvore *Maximum Likelihood* e com 1.000 replicatas *bootstrap*.

FONTE: o autor (2015)

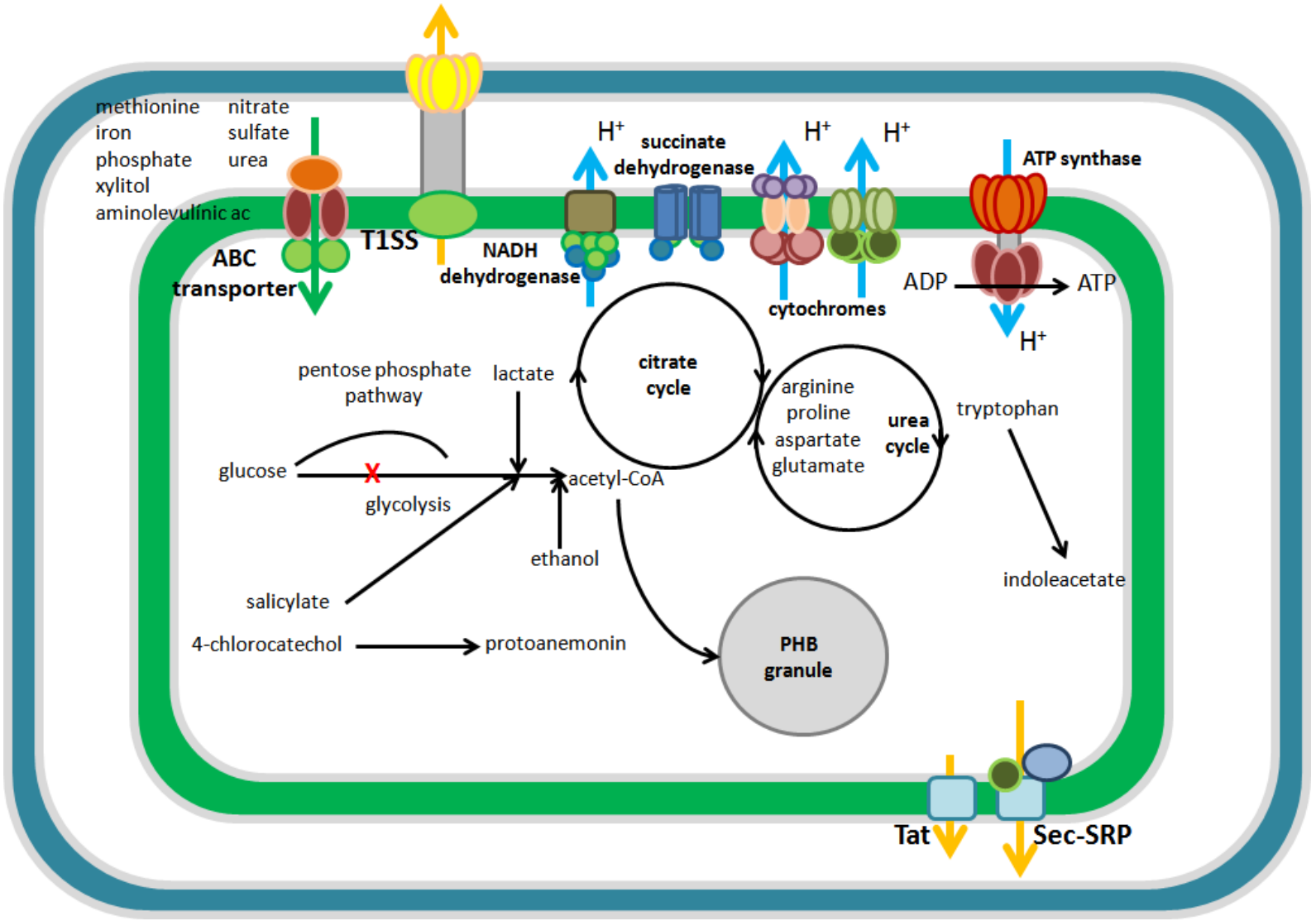


FIGURA 6.22: VISÃO GERAL DAS CARACTERÍSTICAS METABÓLICAS DE *H. rhizosphaerae* UMS-37

FONTE: o autor (2015)

6.6 COMPARAÇÃO GENÔMICA

6.6.1 Core e pangenoma de Herbaspirillum

Foi verificado que o número de genes nos 19 genomas de estirpes de *Herbaspirillum* spp. analisados varia entre 4.031 (*Herbaspirillum massiliense* JC206) e 5.587 genes (*H. autotrophicum* IAM 14942) (FIGURA 6.23), com média de 4.897 genes por genoma. Esses 19 genomas foram posteriormente comparados através do BLAST todos contra todos para a identificação de homólogos. Dos resultados obtidos, ao desconsiderar os genomas das estirpes de *H. seropedicae*, foi verificado que em média 762 genes (15%) são únicos de cada genoma analisado. Os genomas de *Herbaspirillum massiliense* JC206 e *H. autotrophicum* IAM 14942 são os que apresentam maior número de genes únicos (1.738 e 1.981 genes, respectivamente). Para a espécie *H. seropedicae*, 203 genes (4%) diferenciam as estirpes entre si.

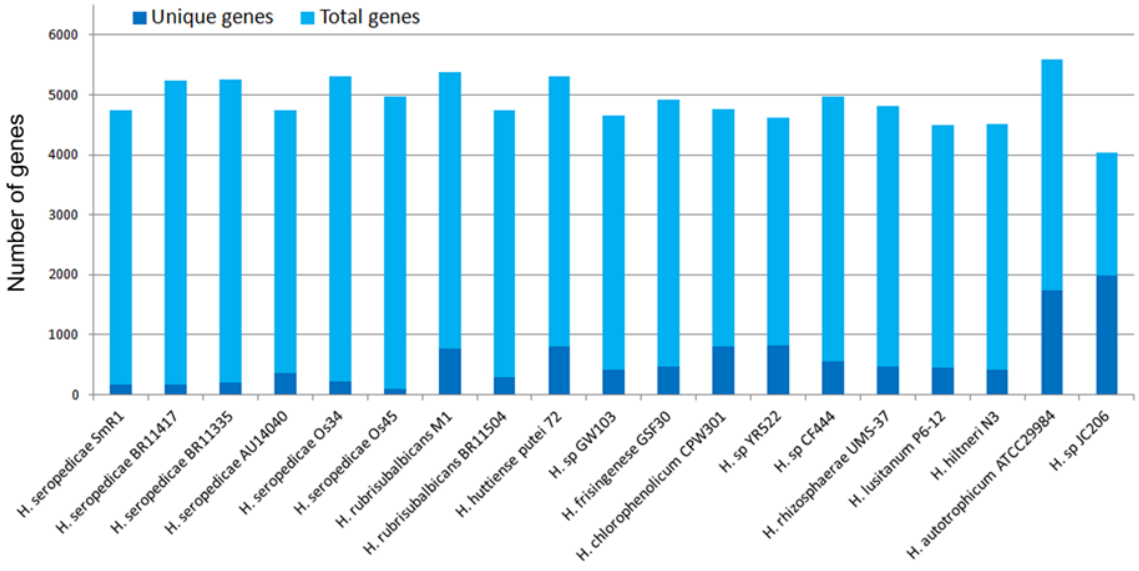


FIGURA 6.23: TOTAL DE GENES E GENES ÚNICOS PARA CADA GENOMA DE *Herbaspirillum* spp.

Os resultados foram obtidos a partir do BLAST todos contra todos.

FONTE: o autor (2015)

Na análise de homologia entre os pares de genomas, representada pela matriz BLAST (FIGURA 6.24), foi verificado que *H. massiliense* JC206, *H. autotrophicum* IAM 14942, *Herbaspirillum sp* YR522 e *H. chlorophenolicum* CPW301 se mostraram divergentes dos demais *Herbaspirillum*, pois não compartilham mais

de 70% do conjunto proteômico com nenhum outro genoma. Por outro lado, é possível observar a forte relação entre as estirpes de *H. seropedicae* e delas com as estirpes de *H. rubrisubalbicans* e *H. huttiense* subsp. *putei*. As estirpes Os34 e Os45 apresentaram maior número de proteínas homólogas em relação às estirpes M1 e BR11504 (abreviada como BR504) de *H. rubrisubalbicans* (entre 77 e 80%) do que em relação às demais estirpes de *H. seropedicae* (entre 71 e 74%) (FIGURA 6.24).

Também foi observada a relação entre *H. rhizosphaerae* UMS-37, *H. lusitanum* P6-12 e *H. hiltneri* N3. Nesse contexto, ainda foi possível observar que a estirpe GW103 apresenta maior número de proteínas homólogas com *H. seropedicae* Os34 (82,2%) e *H. huttiense* subsp. *putei* (81%), enquanto a estirpe CF444 apresenta maior número de proteínas homólogas com *H. rhizosphaerae* UMS-37 (80%). Esses resultados mostram a divisão das estirpes analisadas em dois grupos, com relação às proteínas homólogas.

A partir da comparação entre proteomas, utilizando a abordagem BLAST todos contra todos, foram estabelecidos os grupos de famílias de proteínas para o gênero *Herbaspirillum* e, a partir das famílias determinadas, o *core* genoma do gênero (genes comuns a todos os genomas analisados), compreendendo 1.412 genes (29% da média de genes), e o pangenoma do gênero (todo o conjunto de famílias encontradas), correspondendo a 19.945 genes (4 vezes o número médio de genes) (FIGURA 6.25).

Os conjuntos proteômicos das estirpes de *Herbaspirillum* spp. foram ordenados por grau de homologia em relação ao proteoma de *H. seropedicae* SmR1 determinado pela matriz BLAST. Foi observada uma tendência de estabilização tanto do gráfico do *core* quanto do pangenoma, até a adição do genoma de *H. hiltneri* N3 (identificado na FIGURA 6.25 pelo número ‘17’). Entretanto, nova variação foi observada no gráfico a partir da adição dos proteomas de *H. autotrophicum* IAM 14942 e *H. massiliense* JC206 (genomas ‘18’ e ‘19’ da FIGURA 6.25). Nessa mesma análise, foi observado que o *core* genoma para a espécie *H. seropedicae* corresponde a 3.381 genes e o pangenoma a 7.777 genes.

Para verificar quais classes de proteínas pertenciam ao *core* e ao pangenoma, foi realizada a classificação funcional segundo o COG desses dois conjuntos. Essa classificação mostrou que o pangenoma possui maior número de proteínas não classificados ou com função desconhecida (categorias COG R, S e X - FIGURA 6.26) em relação ao *core* genoma. Também houve aumento das categorias G, P e K

(metabolismo de carboidratos, metabolismo de íons inorgânicos e transcrição, respectivamente – FIGURA 6.26) no pangenoma, o que demonstra que a diversidade gênica de *Herbaspirillum* corresponde a essas classes.

	H. sero SmR1	*H. sero* BR417	*H. sero* BR335	*H. sero* AU14040	*H. sero* Os34	*H. sero* Os45	*H. rubr* M1	*H. rubr* BR504	*H. hutt putei*	*H. sp* GW103	*H. fris* GSF30	*H. chlo* CPW301	*H. sp* YR522	*H. sp* CF444	*H. rhiz* UMS-37	*H. lusi* P6-12	*H. hilt* N3	*H. auto* IAM 14942	*H. mass* JC206
H. sero SmR1	100	93,8	92,3	85,4	80,4	77,4	78,8	77,2	77,9	76,2	77,1	63,8	65,4	63,9	62,1	61,5	61,5	58,0	38,0
H. sero BR417	84,9	100	94,9	78,2	73,4	70,5	70,8	70,7	71,8	69,4	70,2	58,1	59,4	57,7	56,2	55,7	55,7	52,5	34,1
H. sero BR335	83,3	94,7	100	76,7	72,0	68,4	69,6	69,6	70,6	68,0	68,9	56,8	58,2	56,4	54,9	54,2	54,4	51,4	33,2
H. sero AU14040	84,8	85,7	84,2	100	80,3	77,3	76,8	76,1	78,9	77,6	77,9	65,3	65,7	64,1	63,2	62,5	61,7	59,0	37,7
H. sero Os34	72,0	72,4	71,3	72,4	100	91,4	78,0	77,0	73,3	73,1	72,0	59,9	60,0	59,9	57,8	56,5	56,9	53,7	34,0
H. sero Os45	73,4	73,9	72,6	73,9	97,3	100	80,1	79,0	75,5	74,9	73,9	61,4	61,6	61,8	59,4	58,0	58,7	54,9	35,2
H. rubr M1	70,4	69,9	69,0	69,3	78,5	76,2	100	81,8	71,0	70,4	69,9	57,6	59,0	58,9	56,7	54,9	55,6	52,5	33,5
H. rubr BR504	76,5	77,3	76,3	76,2	85,6	82,7	90,0	100	78,2	76,5	75,8	63,3	64,7	64,4	62,2	60,1	60,9	57,3	36,7
H. hutt putei	69,2	70,5	69,5	71,0	73,3	71,1	70,2	70,4	100	71,7	70,0	57,8	59,8	58,4	57,9	56,2	52,6	52,6	33,1
H. sp GW103	76,4	76,7	75,5	78,6	82,2	79,5	78,7	77,7	81,0	100	80,4	66,0	67,7	65,8	65,4	63,1	63,3	59,4	38,1
H. fris GSF30	74,4	74,6	73,8	75,8	77,6	75,2	74,9	73,8	75,5	77,2	100	64,3	66,7	64,9	62,0	63,3	58,6	59,4	36,3
H. chlo CPW301	63,1	63,4	62,5	64,7	66,2	63,9	63,2	62,8	64,1	64,9	65,8	100	64,9	65,6	63,9	63,4	62,7	62,8	37,2
H. sp YR522	65,7	65,9	64,7	66,7	67,6	65,5	65,8	65,6	67,4	67,7	69,7	66,4	100	63,9	63,5	63,0	63,4	59,4	37,3
H. sp CF444	61,5	61,3	60,3	62,3	64,8	62,8	63,0	62,4	62,8	62,8	64,8	63,9	61,0	100	80,0	70,4	72,2	63,0	37,4
H. rhiz UMS-37	61,5	61,5	60,5	62,9	63,8	61,8	62,1	61,8	64,0	64,1	66,1	63,7	62,1	82,3	100	73,9	76,0	65,8	38,5
H. lusi P6-12	65,6	65,4	64,2	66,9	67,3	65,2	65,0	64,3	66,3	66,5	68,0	68,3	65,9	77,6	79,2	100	82,2	69,4	41,6
H. hilt N3	64,5	64,6	63,7	65,3	66,9	64,9	65,0	64,3	65,9	66,4	68,5	66,6	65,7	78,7	80,4	81,2	100	67,7	40,4
H. auto IAM 14942	50,1	50,2	49,5	51,0	51,6	49,7	50,1	49,7	50,7	50,8	52,4	54,6	50,4	56,2	57,4	56,3	55,7	100	33,1
H. mass JC206	42,3	42,2	41,6	42,5	42,6	41,6	42,0	41,6	41,5	42,4	42,1	42,1	41,6	43,5	43,4	43,7	43,0	43,2	100

FIGURA 6.24: MATRIZ BLAST DE SIMILARIDADE ENTRE OS CONJUNTOS DE PROTEOMAS DE ESTIRPES DE *Herbaspirillum* spp.

Os números dentro dos quadros representam a porcentagem de proteínas homólogas entre dois genomas, que também é representada com uma coloração (quanto mais clara, maior o número de proteínas homólogas compartilhadas entre duas estirpes). Os conjuntos proteômicos verticais são as referências de cada análise e os horizontais aqueles aos quais as referências estão sendo comparadas. Por exemplo, a primeira linha representa a comparação do conjunto proteômico de *H. seropedicae* SmR1 (abreviado como *H. sero* SmR1) contra cada um dos demais conjuntos.

FONTE: o autor (2015)

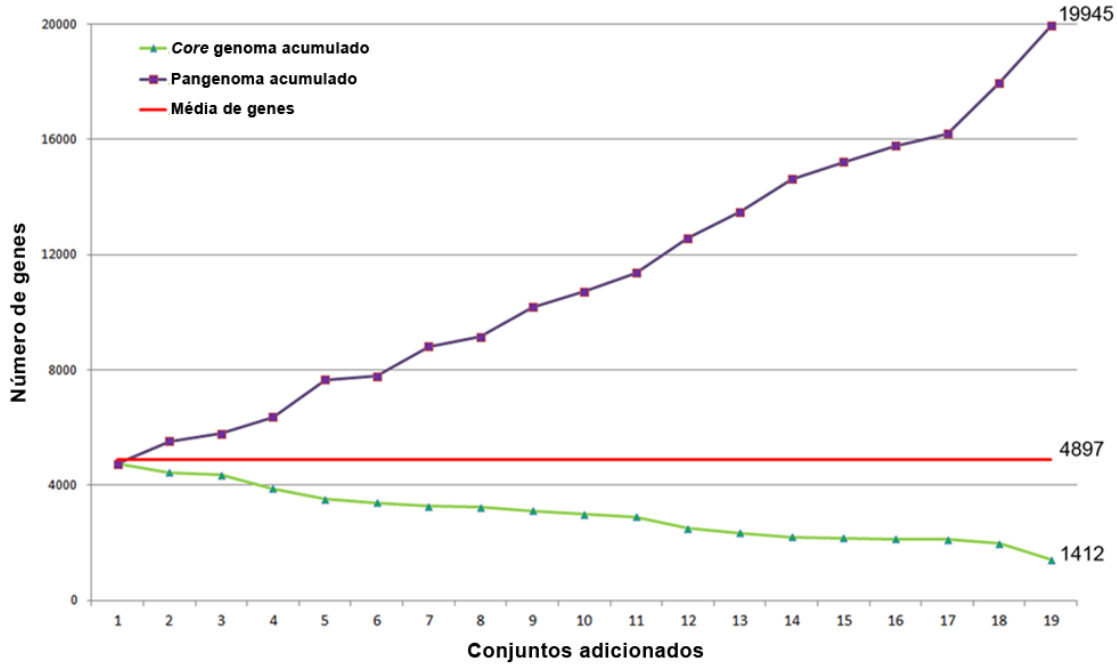


FIGURA 6.25: GRÁFICO DO CORE E DO PANGENOMA CUMULATIVOS PARA OS GENOMAS DE *Herbaspirillum* spp.

A numeração corresponde à ordem de entrada dos organismos na comparação (a mesma mostrada na FIGURA 6.23).

FONTE: o autor (2015)

A presença/ausência das proteínas que compõe o pangenoma de *Herbaspirillum* foi utilizada para agrupar os conjuntos proteômicos desses organismos em forma de árvore filogenética (também chamada de árvore do pangenoma). Foi observada a divisão dos *Herbaspirillum* em dois filogrupos distintos: O primeiro (filogrupo 1), corresponde às estirpes das espécies *H. seropedicae*, *H. rubrisubalbicans*, *H. frisingense*, *H. huttiense*, *H. chlorophenolicum* e às estirpes YR522 e GW103; o segundo (filogrupo 2), é formado pelas estirpes das espécies *H. lusitanum*, *H. rhizosphaerae*, *H. hiltneri*, *H. autotrophicum*, *H. massiliense* e pela estirpe CF444 (FIGURA 6.27). A divisão dos genomas de *Herbaspirillum* spp. em dois filogrupos, concorda com os resultados obtidos a partir da matriz BLAST.

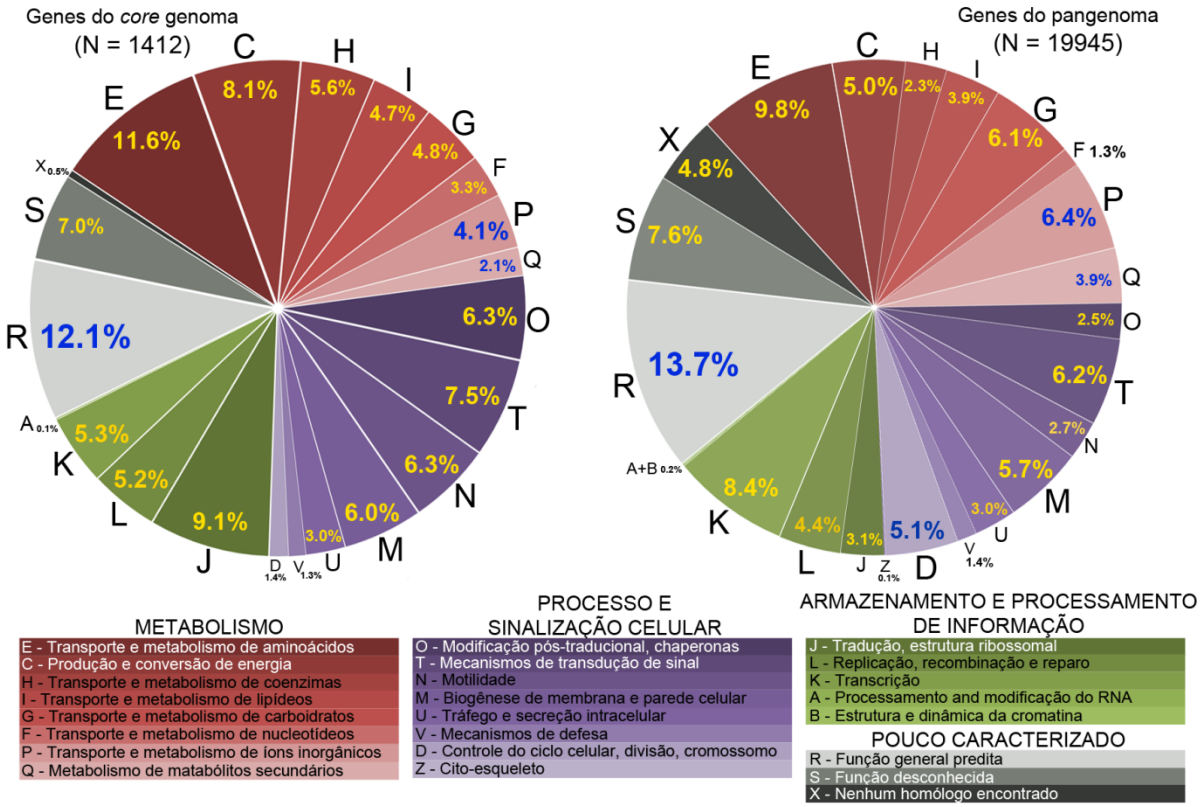


FIGURA 6.26: CATEGORIZAÇÃO FUNCIONAL DO CORE E DO PANGENOMA SEGUNDO O COG

As porcentagens de cada categoria COG, representadas por uma letra e uma cor, são mostradas no gráfico. O *core* genoma está representado à esquerda, o pangenoma à direita, e abaixo deles está a descrição de cada categoria COG.

FONTE: o autor (2015)

Nessa árvore, foi também observado que as estirpes Os34 e Os45 de *H. seropedicae* estão agrupadas juntamente com as estirpes de *H. rubrisubalbicans* e, em outro nível, com *H. huttiense* subsp. *putei* e *Herbaspirillum sp* GW103. Por outro lado, as demais estirpes de *H. seropedicae* formaram um grupo isolado das estirpes Os34 e Os45 (FIGURA 6.27).

A árvore indica que *H. seropedicae*, *H. rubrisubalbicans*, *H. huttiense*, *Herbaspirillum sp* GW103 e *H. frisingense* possuem um ancestral comum. Assim, esse grupo corresponde aos organismos fixadores de nitrogênio, incluindo as três espécies capazes de fixar o nitrogênio atmosférico (ramo dos fixadores de nitrogênio). Enquanto isso, *H. chlorophenolicum* CPW301 e *Herbaspirillum sp* YR522, embora façam parte do filogrupo 1, formam um subgrupo separado (FIGURA 6.27).

No filogrupo 2, foi observada a formação de três pares: 1- *Herbaspirillum sp* CF444 e *H. rhizosphaerae* UMS-37; 2- *H. lusitanum* P6-12 e *H. hiltneri* N3; e 3- *H. autotrophicum* IAM 14942 e *H. massiliense* JC206. Os dois primeiros pares se

relacionam entre si em um ramo, enquanto o terceiro está isolado dos demais (FIGURA 6.27).

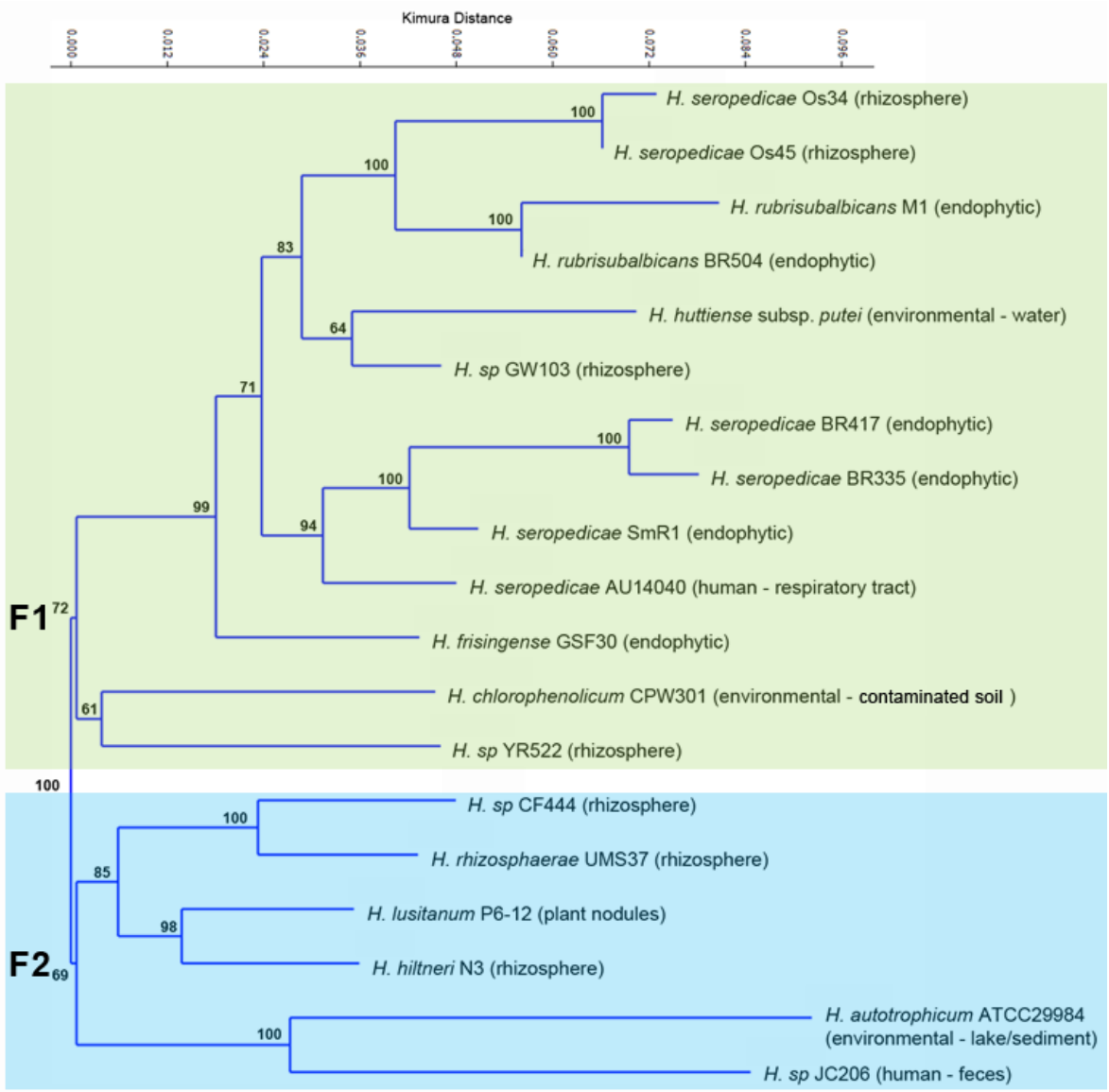


FIGURA 6.27: DENDOGRAMA BASEADO NA PRESENÇA/AUSÊNCIA DE GENES QUE CODIFICAM PROTEÍNAS DO PANGENOMA.

Esse dendograma foi construído com o programa PAST, pelo método de *single linkage*, com a utilização da distância de Kimura e 10.000 replicatas de *bootstrap*. 'F1' indica filogrupo 1 e 'F2' indica filogrupo 2.

FONTE: o autor (2015)

Os genomas das estirpes, cujas espécies já foram descritas, foram separadas em seus respectivos filogrupos para a análise do pan e *core* genoma de cada filogrupo. Foi considerado como filogrupo 1 o conjunto de genomas formado por *H. seropedicae* SmR1, *H. rubrisubalbicans* M1, *H. huttiense* subsp. *putei*, *H. frisingense* GSF30 e *H. chlorophenolicum* CPW301. Foi considerado como filogrupo 2 o

conjunto de genomas formado por *H. lusitanum* P6-12, *H. rhizosphaerae* UMS-37, *H. hiltneri* N3 e *H. autotrophicum* IAM 14942. Os demais genomas foram removidos da análise para equilibrar os conjuntos e para classificação *a posteriori* com base nos *core* genomas.

Foi verificado que o *core* genoma do filogrupo 1 corresponde a 2.683 genes e o *core* genoma do filogrupo 2 corresponde a 2.840 genes (FIGURA 6.28). Os dois *core* genomas se sobrepõe em 2.194 genes, de forma que 111 genes são exclusivos do filogrupo 1 e 133 genes são exclusivos do filogrupo 2.

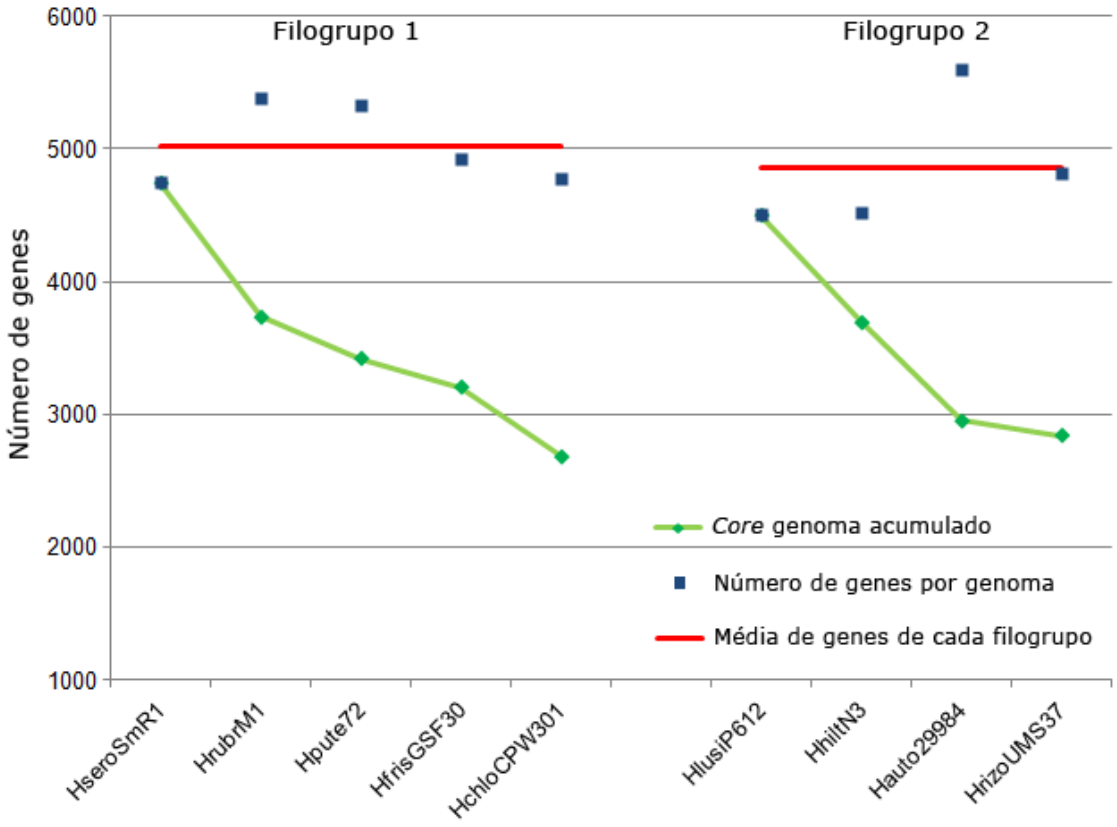


FIGURA 6.28: GRÁFICO CUMULATIVO PARA O CORE GENOMA DOS DOIS FILOGRUPOS DO GÊNERO *Herbaspirillum*

O *core* genoma do filogrupo 1 corresponde a 2.683 genes e do filogrupo 2 a 2.840 genes.

FONTE: o autor (2015)

Na TABELA 6.13 é mostrada a classificação geral dos genes exclusivos de cada filogrupo. Foi observado que cada filogrupo possui 15 reguladores exclusivos. No filogrupo 1, esses reguladores são em sua maioria proteínas LysR, enquanto no filogrupo 2 existe maior variedade de proteínas reguladoras (DeoR, GntR, IclR, MarR, entre outras). Os filogrupos também se diferenciam pela presença de proteínas específicas do sistema *pili* tipo IV para cada um deles. Além disso, os

organismos do filogrupo 1 compartilham acetiltransferases que não são encontradas no filogrupo 2, bem como uma proteína anotada como ‘proteína secretada’, uma proteína EpsM (relacionada com a secreção de uma proteína colérica) e genes para a biossíntese de EPS (exopolissacarídeos). Os organismos do filogrupo 2 compartilham um gene relacionado com o transporte de hemina, além de uma proteína anotada como ‘proteína exportada’, uma proteína de ciclo circadiano (KaiC) e maior variedade de classes de enzimas.

TABELA 6.13: CLASSIFICAÇÃO DOS GENES EXCLUSIVOS DE CADA FILOGRUPO

Categoria	**Exclusivas do filogrupo 1**	**Exclusivas do filogrupo 2**
Proteínas hipotéticas	25	45
Proteínas de membrana, transportadores e lipoproteínas	24	19
Reguladores de transcrição, fatores sigma, sistemas de dois componentes	15	15
Transferases	10	4
Proteínas do sistema *pili* tipo IV	5	3
Biossíntese de EPS e glucanas	3	0
Proteínas secretadas ou exportadas, toxinas	2	1
Outras proteínas	27	46
Total	**111**	**133**

FONTE: o autor (2015)

Os conjuntos dos *core* genomas foram utilizados para uma classificação quantitativa das estirpes *H. seropedicae* AU14040, *H. seropedicae* Os34, *H. seropedicae* Os45, *Herbaspirillum* sp. GW103, *Herbaspirillum* sp. YR522, *Herbaspirillum* sp. CF444 e *H. massiliense* JC206. Nessa análise, foi verificado o número de proteínas homólogas de cada genoma em relação ao *core* genoma de cada filogrupo. Na representação gráfica (FIGURA 6.29), a área em verde, dividida pela linha que liga o ponto de intersecção com o ponto de união (número máximo de proteínas) entre os *core* genomas dos filogrupos, foi atribuída ao gênero *Herbaspirillum*.

Nessa representação, a diagonal que corta a área que representa o gênero *Herbaspirillum*, é usada como limite divisório para classificação de estirpes de

Herbaspirillum spp. no filogrupo 1 (abaixo da diagonal) e no filogrupo 2 (acima da diagonal) (FIGURA 6.29).

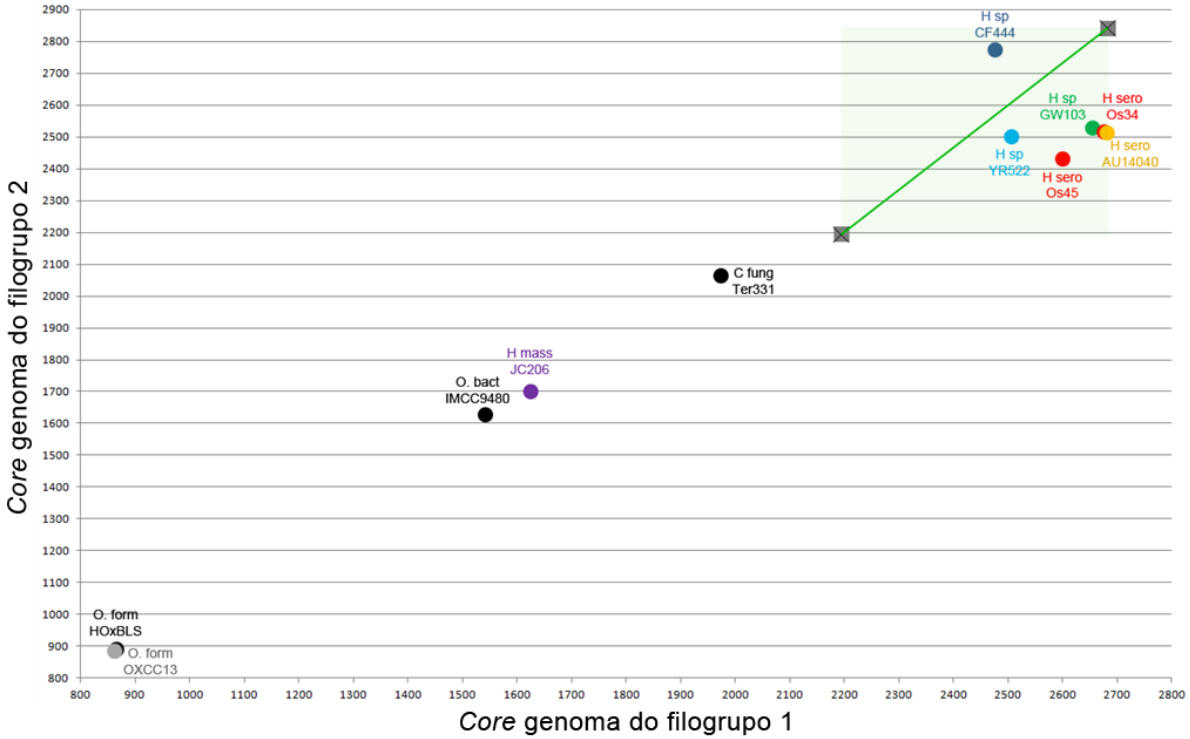


FIGURA 6.29: RELAÇÃO ENTRE O NÚMERO DE PROTEÍNAS COMPARTILHADAS NO CORE GENOMA DOS FILOGRUPOS 1 E 2 E A CLASSIFICAÇÃO DE ESTIRPES NO GÊNERO *Herbaspirillum*

A área verde marcada no gráfico, delimitada pela intersecção dos conjuntos (2.194 genes) e pelo *core* genoma total dos filogrupos (2.683 e 2.840 para os filogrupos 1 e 2, respectivamente), representa a área correspondente ao gênero *Herbaspirillum*. A linha que corta essa área divide o conjunto entre os filogrupos 1 (abaixo dela) e 2 (acima dela). Os pontos representam o número de genes que cada genoma analisado compartilha com cada um dos filogrupos.

FONTE: o autor (2015)

Com isso, as estirpes Os34, Os45 e AU14040 de *H. seropedicae*, e as estirpes GW103 e YR522 foram posicionadas dentro do gênero *Herbaspirillum* e dentro do filogrupo 1. Por outro lado, a estirpe CF444 foi posicionada dentro do gênero *Herbaspirillum*, mas no filogrupo 2 (FIGURA 6.29).

A análise conseguiu separar as estirpes de *Herbaspirillum* spp. de *Collimonas fungivorans* Ter331, mas demonstrou que *H. massiliense* JC206 não faria parte do gênero *Herbaspirillum*. Genomas de outras bactérias também foram testados: *Oxalobacter formigenes* estirpes HOxBLS e OXCC13 e o genoma de uma bactéria classificada apenas como *Oxalobateracae bacterium* IMCC9480. Porém, nenhuma delas se mostrou próxima ao gênero *Herbaspirillum* (FIGURA 6.29).

6.6.2 Análise de genes e agrupamentos gênicos específicos

O genoma de *H. seropedicae* SmR1 serviu de referência para comparações entre os genomas de estirpes de *Herbaspirillum* spp. através do BLAST atlas (FIGURA 6.30), no qual é possível identificar a presença/ausência de genes homólogos em relação à ordem dos genes no genoma de *H. seropedicae* SmR1. É possível observar também a presença de grupos de genes relacionados a fagos, sistemas de secreção e fixação biológica de nitrogênio em outras estirpes de *Herbaspirillum* spp. (FIGURA 6.30). As falhas representam a ausência de proteínas homólogas nos demais genomas em relação à referência, na qual muitas dessas falhas vêm acompanhadas pela queda de conteúdo G+C (FIGURA 6.30). A comparação referente a determinados agrupamentos gênicos pode ser visualizada em detalhes na FIGURA 6.31.

Os genes *nif* são responsáveis pela fixação biológica de nitrogênio e estão presentes em *H. seropedicae* SmR1 (PEDROSA *et al*., 2011). Esses genes também foram encontrados nas demais estirpes de *H. seropedicae* (com exceção da estirpe clínica AU14040), nas estirpes de *H. rubrisubalbicans* e em *H. frisingense* GSF30 (FIGURA 6.31B).

A comparação da estrutura desse agrupamento de genes mostrou que a disposição dos genes é idêntica entre os *Herbaspirillum* que os possuem. Foi também observado que a identidade em nível de proteína é, em geral, superior a 90%, o que demonstra sua conservação (FIGURA 6.32).

O agrupamento de genes relacionado ao T3SS está potencialmente relacionado com a interação planta/bactéria em *H. seropedicae* SmR1 (MONTEIRO *et al*., 2012) e foi considerado necessário para a interação de *H.rubrisubalbicans* M1 com gramíneas (SCHMIDT *et al*., 2012). Esse agrupamento foi encontrado em todas as estirpes de *H. seropedicae* (com exceção da estirpe AU14040), nas estirpes de *H. rubrisubalbicans*, em *H. hiltneri* N3, em *H. chlorophenolicum* CPW301 e nas estirpes YR522 e CF444 (FIGURA 6.31C). A estirpe YR522 apresenta dois agrupamentos gênicos relacionados ao sistema de secreção do tipo III, mas somente um apresenta homologia com os encontrados nas demais estirpes de *Herbaspirillum* spp. (FIGURA 6.33).

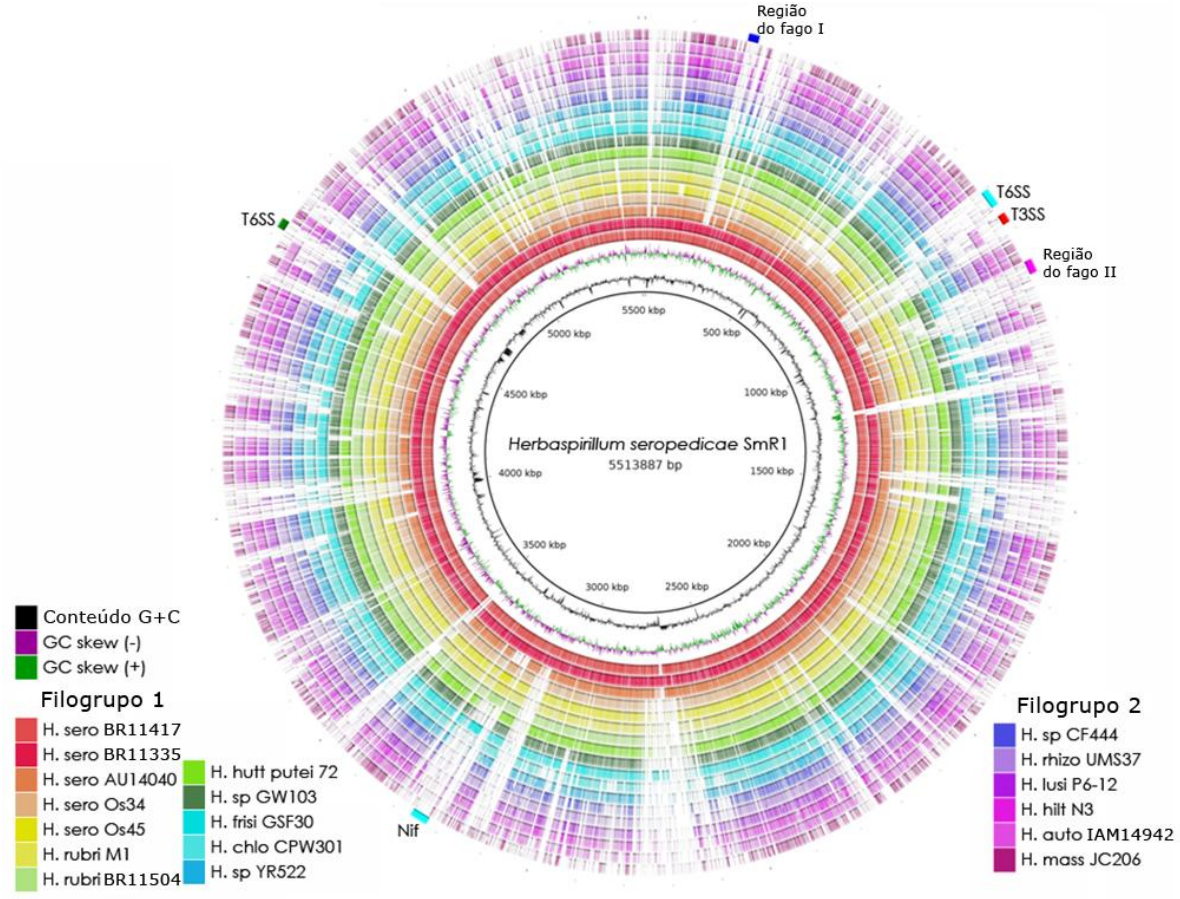


FIGURA 6.30: BLAST ATLAS DOS GENOMAS DE *Herbaspirillum*

De dentro para fora temos: o genoma de referência (círculo interno); variação do conteúdo G+C; GC *skew*; conjuntos proteômicos dos demais genomas em homologia com a referência (coloridos conforme mostra a figura); e agrupamentos gênicos de interesse (duas regiões de fago, dois sistemas de secreção do tipo VI, sistema de secreção do tipo III, e genes *nif*) (círculo mais externo). As falhas nos círculos referentes aos conjuntos proteômicos representam proteínas presentes na referência, mas ausentes nesses conjuntos. A comparação foi feita com uso do programa BRIG.

FONTE: o autor (2015)

De maneira geral, a estrutura do agrupamento é conservada, com exceção da presença de proteínas hipotéticas que, em alguns casos, nem sequer apresentam homologia entre os genomas. A porcentagem de identidade das proteínas varia bastante, mas foi observada maior conservação entre as proteínas de *H. rubrisubalbicans* M1 e as proteínas das estirpes Os34 e Os45 de *H. seropedicae* (FIGURA 6.33).

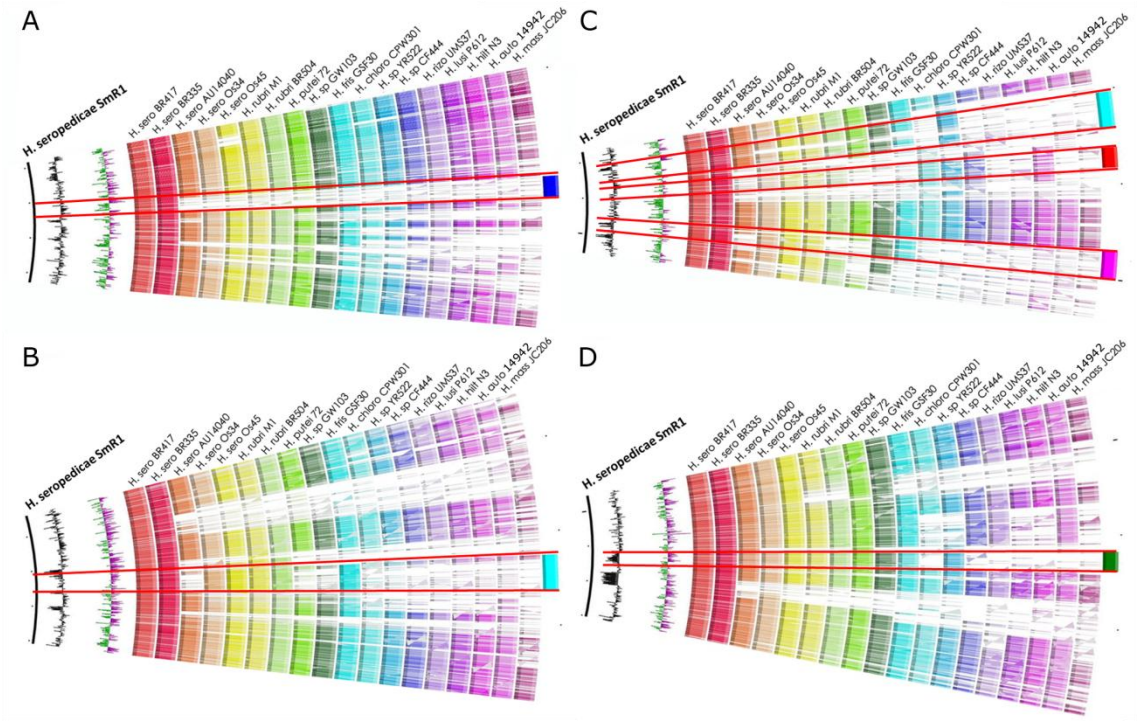


FIGURA 6.31: DETALHE DO BLAST ATLAS DE ALGUNS AGRUPAMENTOS GÊNICOS EM ESTIRPES DE *Herbaspirillum* spp.

A ordem dos círculos é a mesma da FIGURA 6.30. Em A é mostrada a região do fago I; em B, o agrupamento de genes *nif*; em C, um dos agrupamentos gênicos relacionados ao sistema de secreção do tipo VI (azul), os genes do sistema de secreção do tipo III (vermelho), a região do fago II (rosa); em D, outro agrupamento gênico relacionado ao sistema de secreção do tipo VI. As linhas vermelhas servem para facilitar a visualização dos agrupamentos ao longo dos círculos. A comparação foi feita com o programa BRIG.

FONTE: o autor (2015)

O agrupamento de genes do T6SS está relacionado com a liberação de proteínas efetoras dentro da célula eucariótica hospedeira e é sugerido que ele esteja envolvido no reconhecimento planta/bactéria em *H. seropedicae* SmR1, porém também é descrito que esse sistema apresenta propriedades antimicrobianas e está associado com a competição entre bactérias (MACINTYRE *et al*., 2010; PEDROSA *et al*., 2011).

Esse agrupamento foi encontrado em todos os genomas de organismos que fazem parte do filogrupo 1, com exceção de *H. chlorophenolicum* CPW301. Porém, não foi encontrado em nenhum genoma de organismos do filogrupo 2 (FIGURA 6.31C e D).

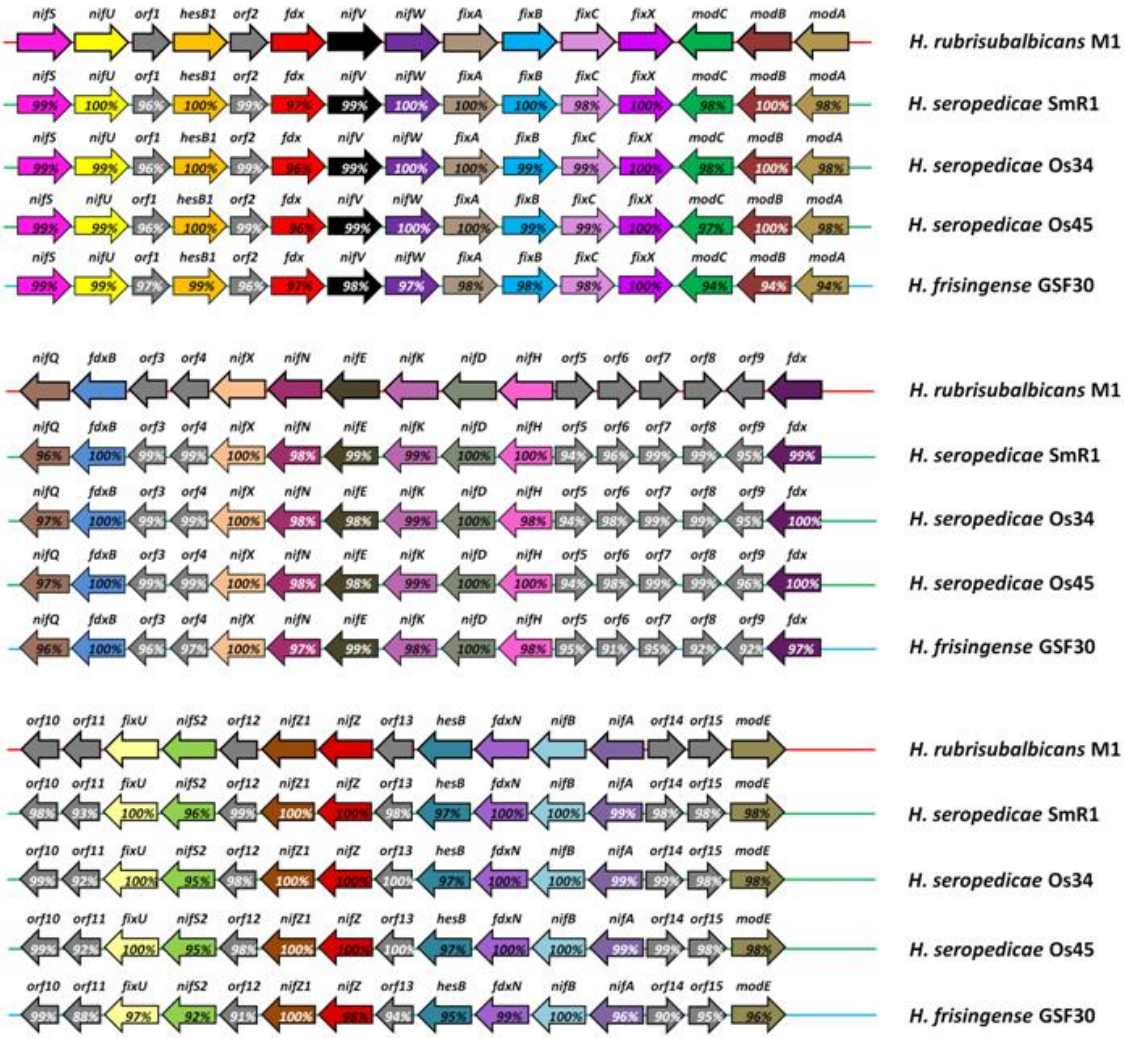


FIGURA 6.32: ESTRUTURA DO AGRUPAMENTO DE GENES *nif* EM *Herbaspirillum*

O agrupamento está dividido em três partes devido ao seu tamanho. As setas representam os genes em ordem e seus sentidos. A referência é *H. rubrisubalbicans* M1 e as porcentagens de identidade das proteínas dos demais genomas são dadas em relação à referência. Setas em cinza se referem a genes que codificam proteínas hipotéticas.

FONTE: o autor (2015)

H. seropedicae SmR1 apresenta dois desses agrupamentos gênicos, os quais são homólogos entre si e homólogos aos agrupamentos encontrados nos demais *Herbaspirillum*. No entanto, os outros *Herbaspirillum* apresentam somente um desses agrupamentos. A organização desses genes difere entre os genomas pela presença de genes que codificam proteínas hipotéticas posicionadas na região central do agrupamento e que não apresentam homologia entre os genomas. Por outro lado, as proteínas de função conhecida são conservadas e apresentam identidade maior que 80% em sua maioria (FIGURA 6.34).

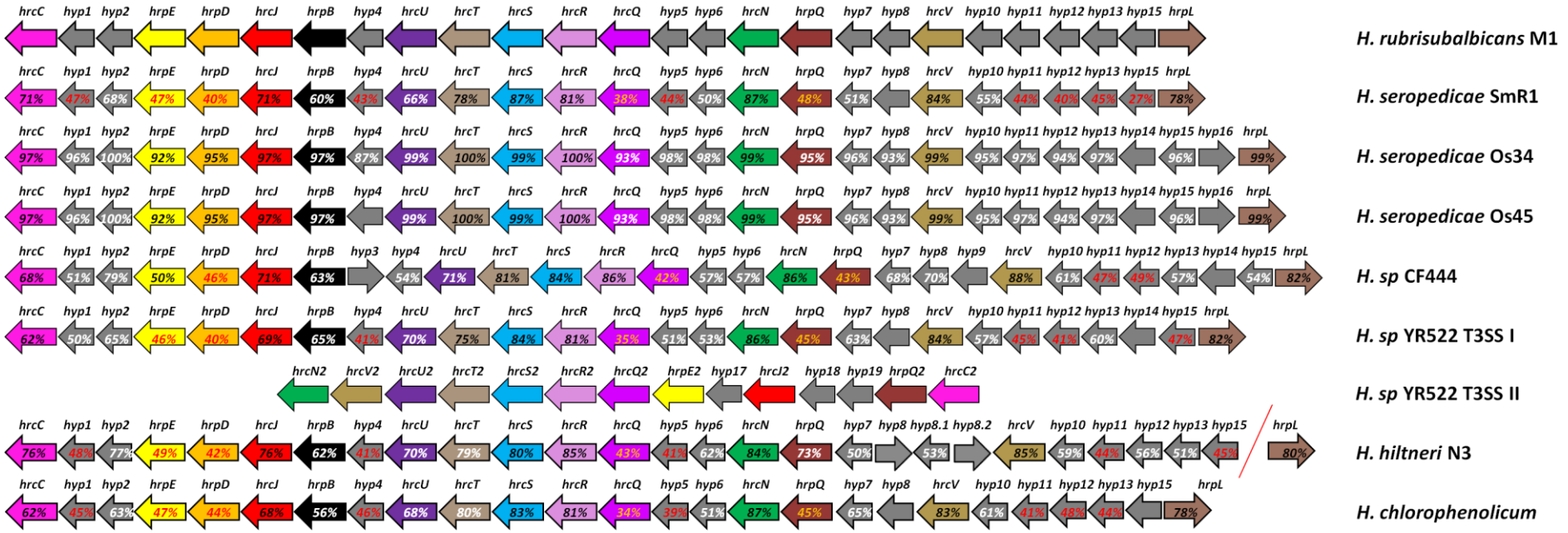


FIGURA 6.33: ESTRUTURA DO AGRUPAMENTO GÊNICO RELACIONADO AO T3SS EM ESTIRPES DE *Herbaspirillum* spp.

As setas representam os genes em ordem e o sentido da transcrição. A referência é *H. rubrisubalbicans* M1 e as porcentagens de identidade das proteínas dos demais genomas são dadas em relação à referência. Porcentagens menores que 50% são marcadas em vermelho. A barra em *H. hiltneri* N3 representa quebra de *contig*. Setas em cinza se referem à codificação de proteínas hipotéticas.

FONTE: o autor (2015)

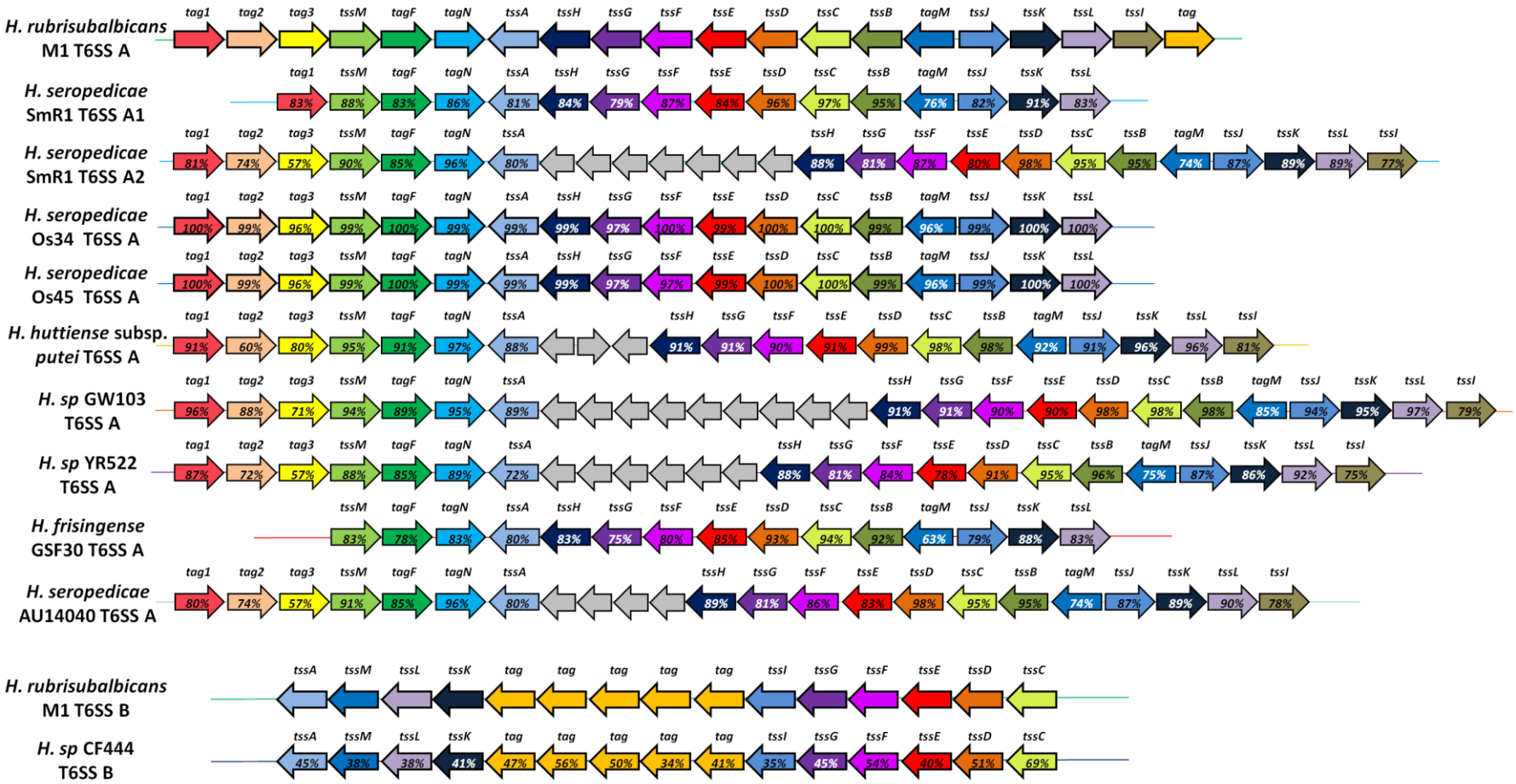


FIGURA 6.34: ESTRUTURA DO AGRUPAMENTO GÊNICO RELACIONADO AO T6SS EM ESTIRPES DE *Herbaspirillum* spp.

As setas representam os genes em ordem e o sentido da transcrição. A referência é *H. rubrisubalbicans* M1, que apresenta dois agrupamentos (T6SS A e T6SS B). As porcentagens de identidade das proteínas dos demais genomas são dadas em relação à referência. Setas em cinza se referem à codificação de proteínas hipotéticas.

FONTE: o autor (2015)

H. rubrisubalbicans M1 apresenta um segundo agrupamento, não homólogo aos dos *Herbaspirillum* do filogrupo 1. Um agrupamento de estrutura idêntica foi encontrado em *Herbaspirillum sp* CF444, mas a identidade entre a maioria das proteínas é menor que 50% (FIGURA 6.34).

H. seropedicae SmR1 apresenta duas regiões de genes correspondentes a fagos (PEDROSA *et al*., 2011), que são dois importantes vestígios evolutivos no genoma dessa bactéria. A região do fago I de *H. seropedicae* SmR1 foi encontrada somente nas estirpes BR11335 e BR11417 dessa mesma espécie e por isso não acrescenta maiores informações à comparação genômica (FIGURA 6.31A). No entanto, a região do fago II foi encontrada nas demais estirpes de *H. seropedicae* (com exceção da estirpe AU14040), na estirpe M1 de *H. rubrisubalbicans*, em *H. frisingense* GSF30 e em *Herbaspirillum sp* GW103 (FIGURA 6.31C).

A disposição dos genes na região do fago II é conservada na maioria dos genomas de *Herbaspirillum*, embora muitos deles apresentem variação em relação a genes que codificam para proteínas hipotéticas. No entanto, o agrupamento encontrado em *H. frisingense* GSF30 apresenta estrutura diferente dos demais, embora a maioria das proteínas codificadas por esse agrupamento apresente identidade superior a 80% em relação à referência (FIGURA 6.35).

Outro agrupamento analisado, os genes *wss*, está relacionado com o processo de colonização da planta hospedeira por *H. rubrisubalbicans* M1 (MONTEIRO *et al*., 2012). Esse agrupamento está ausente em *H. seropedicae* SmR1 e, portanto, *H. rubrisubalbicans* M1 foi utilizado como referência para o BLAST atlas. Foi observada a presença de genes *wss* também na estirpe BR11504 de *H. rubrisubalbicans*, nas estirpes Os34 e Os45 de *H. seropedicae*, em *H. huttiense* subsp. *putei*, em *H. frisingense* GSF30 e em *Herbaspirillum* sp GW103 (FIGURA 6.36A).

A estrutura do agrupamento é similar entre os genomas, mas em *H. huttiense* subsp. *putei* o primeiro gene (*yhj*Q) apresenta uma mudança de fase de leitura e está dividido em duas partes. Em *H. frisingense* GSF30, o agrupamento está quebrado por uma falha de sequenciamento ou montagem, de forma que o penúltimo gene (*wssH*) está em outro *contig*, e o último gene (*wssI*) não foi encontrado. A identidade entre as proteínas codificadas pelos demais genes do agrupamento é superior a 80% para a maioria delas, o que mostra que elas são conservadas entre os organismos analisados (FIGURA 6.36B).

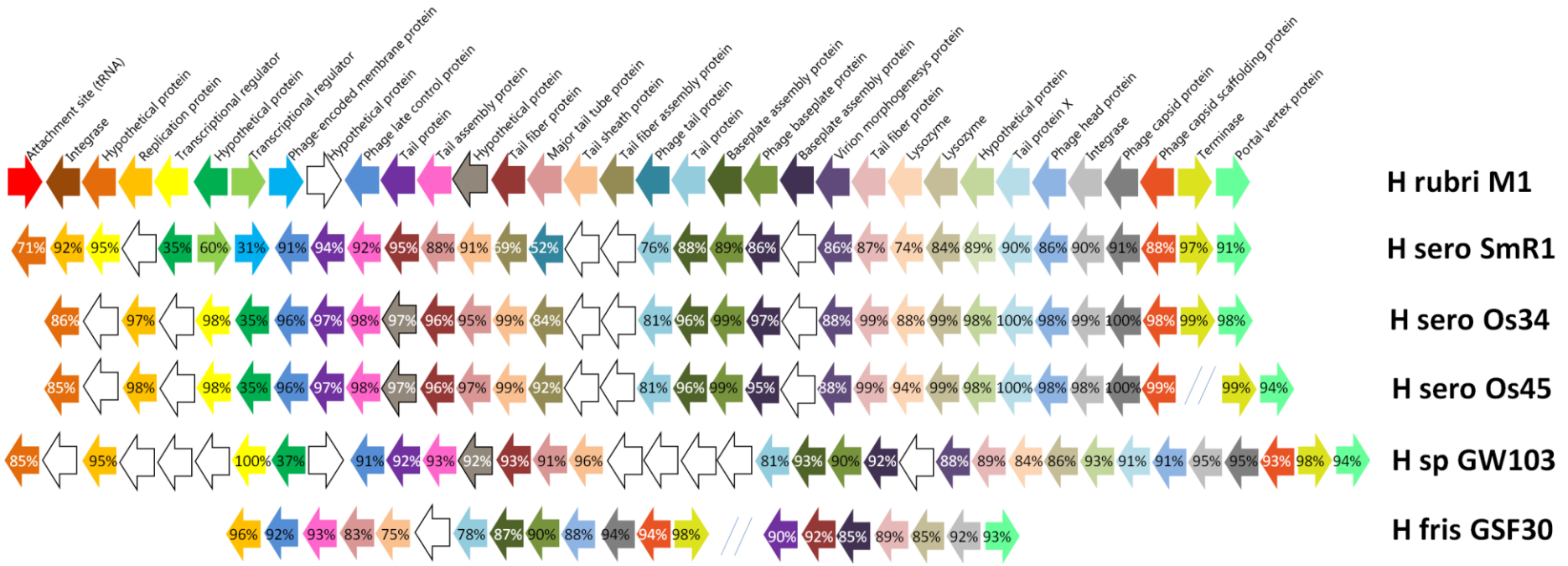


FIGURA 6.35: AGRUPAMENTO GÊNICO RELACIONADO À REGIÃO DO FAGO II EM ESTIRPES DE *Herbaspirillum* spp.

As setas representam os genes em ordem e o sentido de transcrição. A referência é *H. rubrisubalbicans* M1 e as porcentagens de identidade das proteínas dos demais genomas são dadas em relação à referência. Setas em branco se referem à codificação de proteínas hipotética. As barras em *H. seropedicae* Os45 se referem à quebra de *contig*. As barras em *H. frisingense* GSF30 se referem à quebra do agrupamento em duas regiões distintas no genoma.

FONTE: o autor (2015)

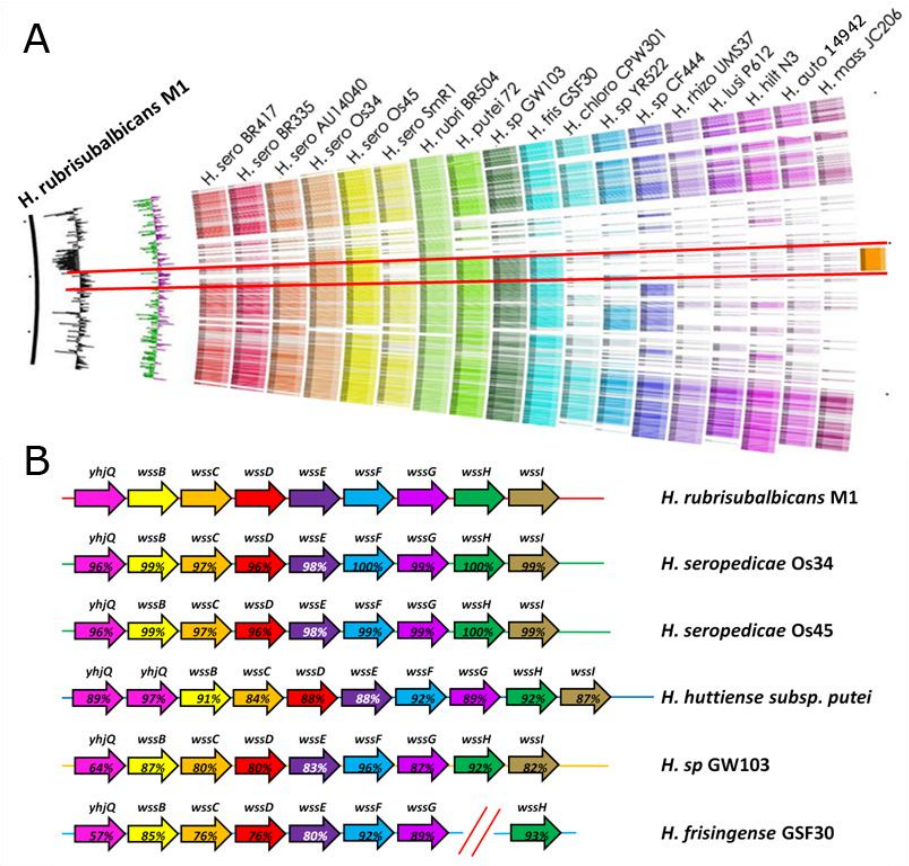


FIGURA 6.36: COMPARAÇÃO GLOBAL E ESTRUTURA DO AGRUPAMENTO DE GENES *wss* EM ESTIRPES DE *Herbaspirillum* spp.

Em A, é representada a comparação global do agrupamento, na qual a ordem dos círculos é a mesma da FIGURA 6.30, com a diferença que a referência é *H. rubrisubalbicans* M1. Em B, é mostrada a estrutura do agrupamento, onde as setas representam os genes em ordem e seus sentidos. As porcentagens de identidade das proteínas são dadas em relação à referência. As barras em *H. frisingense* GSF30 indicam quebra de *contig*. As comparações fram realizadas com o programa BRIG.

FONTE: o autor (2015)

A partir da configuração da árvore do pangenoma e da presença/ausência de genes foram propostos quais seriam os pontos de aquisição dos agrupamentos gênicos apresentados anteriormente, ao longo da evolução dos *Herbaspirillum*.

Dessa forma, foi atribuído que o agrupamento de genes *nif* foi adquirido pelo ancestral que deu origem às espécies *H. rubrisubalbicans*, *H. seropedicae*, *H. frisingense* e *H. huttiense* (ramo dos fixadores de nitrogênio). Posteriormente, esse agrupamento teria sido perdido pelo ancestral de *H. huttiense* e *Herbaspirillum sp* GW103, bem como pela estirpe AU14040 de *H. seropedicae* (FIGURA 6.37).

A região do fago II teria sido adquirida por esse mesmo ancestral do ramo dos fixadores de nitrogênio, posteriormente perdida por *H. huttiense* subsp. *putei*, pela estirpe BR11504 de *H. rubrisubalbicans* e pela estirpe AU14040 de *H. seropedicae*

(FIGURA 6.37). Nesse contexto filogenético, *H. frisingense* teria herdado essa região do ancestral do ramo, embora pela sintenia da região a hipótese seja de origem distinta aos demais *Herbaspirillum* (FIGURA 6.35). A região do fago 1 teria sido adquirida somente pelo ancestral das estirpes SmR1, BR11335 e BR11417 de *H. seropedicae*.

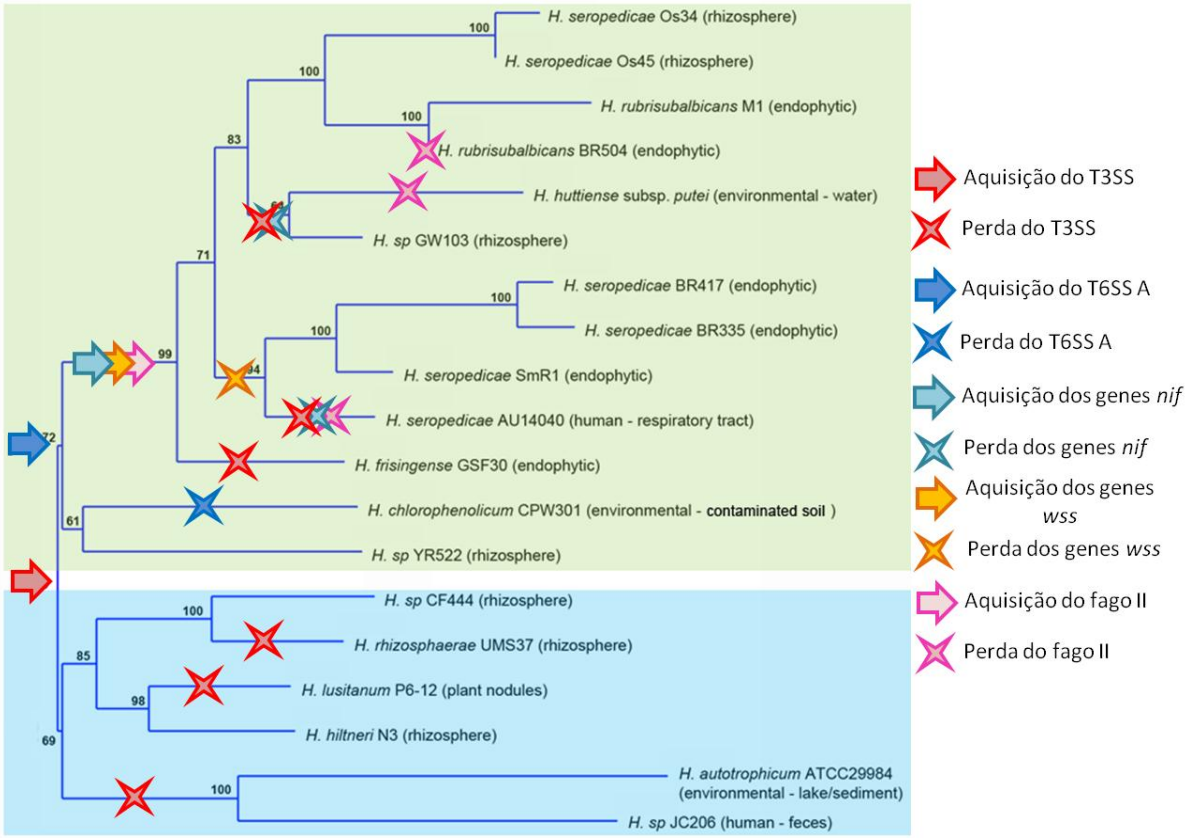


FIGURA 6.37: DINÂMICA DA AQUISIÇÃO E PERDA DE ALGUNS AGRUPAMENTOS GÊNICOS POR ESTIRPES DE *Herbaspirillum* spp.

A árvore é a mesma da FIGURA 6.27, com adição dos pontos de aquisição e perda de genes para os agrupamentos referentes ao T3SS, T6SS A, genes *nif*, genes *wss* e região do fago II.

FONTE: o autor (2015)

O ancestral do ramo dos fixadores de nitrogênio teria ainda adquirido o agrupamento gênico para a biossíntese de celulose, que teria sido perdido pelo ancestral de *H. seropedicae* (FIGURA 6.37). O agrupamento de genes relacionado com o T3SS, por estar presente em ambos os filogrupos 1 e 2, foi atribuído ao ancestral do gênero e teria sido perdido pelo ancestral de *H. massiliense* e *H. autotrophicum*, por *H. rhizosphaerae*, por *H. lusitanum*, por *H. frisingense*, pelo ancestral de *H. huttiense* e *Herbaspirillum* sp. GW103 e pela estirpe AU14040 de *H.*

seropedicae (FIGURA 6.37). *Herbaspirillum* sp. YR522 teria adquirido um novo T3SS.

O agrupamento de genes relacionado com o T6SS poderia, de forma equivalente: A - estar presente no ancestral do gênero e ter sido perdido pelo ancestral do filogrupo 2; ou B - ter sido adquirido somente pelo ancestral do filogrupo 1. Independentemente disso, *H. chlorophenolicum* teria perdido esse agrupamento e *H. seropedicae* SmR1 o teria duplicado. Apenas para fim de representação, foi considerada a hipótese B (FIGURA 6.37). A segunda cópia desse agrupamento presente nas estirpes de *H. rubrisubalbicans*, com identidade das proteínas menor que 50% em relação a *Herbaspirillum* sp. CF444, teriam origens distintas nesses dois ramos.

O genoma de *H. seropedicae* SmR1 apresenta genes para a degradação de compostos fenólicos (PEDROSA *et al*., 2011), enquanto *H. chlorophenolicum* CPW301 foi descrito como capaz de degradar fenol e 4-clorofenol (IM *et al*., 2004), o que faz dos *Herbaspirillum* potenciais organismos biorremediadores. Foi observado por análise *in silico* que estirpes de *Herbaspirillum* spp. podem degradar catecol (provavelmente fenol também) pelas vias de clivagem *orto* (com a formação de piruvato) e *meta* (com a formação de succinil-CoA) (FIGURA 6.38). Dentre eles, somente *H. chlorophenolicum* CPW301 e *Herbaspirillum* sp. GW103 possuem os genes que codificam as enzimas envolvidas na degradação pela clivagem *orto*. *Herbaspirillum sp* GW103 possui também os genes para degradar catecol pela clivagem *meta*, que é a via amplamente distribuída em *Herbaspirillum*, também utilizada por: *H. seropedicae* estirpes SmR1, AU14040, Os34 e Os45; *H. rubrisubalbicans* M1; *H. huttiense* subsp. *putei*; *H. frisingense* GSF30; *Herbaspirillum sp* estirpes YR522 e CF444; e *H. rhizosphaerae* UMS-37. Os demais (*H. hiltneri* N3, *H. lusitanum* P6-12, *H. autotrophicum* IAM 14942, *H. massiliense* JC206) não apresentam nenhuma das vias.

Foi verificado ainda que em estipes de *Herbaspirillum* spp. são encontradas duas vias para a degradação de 4-clorocatecol (possivelmente para a degradação de 4-clorofenol também) (FIGURA 6.39). O gene que codifica para a catecol-2,3-dioxigenase (EC 1.13.11.2), na degradação via clivagem *meta*, foi encontrado em *H. chlorophenolicum* CPW301, e nos genomas de *H. seropedicae* AU14040 e *H. sp* GW103. Em *H. chorophenolicum* CPW301 e *H. sp* GW103 os genes são anotados como *dmpB*, e a identidade ao nível de proteína entre eles é de 72%. O gene de *H.*

seropedicae AU14040 é anotado como *catE*, e a proteína codificada por ele não apresenta homologia com as encontradas em *H. chlorophenolicum* CPW301 e *H. sp* GW103.

H. seropedicae AU14040 e *Herbaspirillum sp* GW103 podem degradar 4-clorocatecol também pela clivagem *orto* e produzir protoanemonina. Essa via é amplamente difundida em *Herbaspirillum* e foi encontrada em *H. seropedicae* estirpes SmR1, Os34 e Os45, *H. rubrisubalbicans* M1, *H. huttiense* subsp. *putei*, *H. frisingense* GSF30, *H. sp* YR522, *H. sp* CF444 e *H. rhizosphaerae* UMS-37. Alguns deles apresentam dois genes que codificam a muconato cicloisomerase (EC 5.5.1.1), mas as duas proteínas não são homólogas entre si (ver subtópico 6.5.4).

Com isso, somente *H. hiltneri* N3, *H. lusitanum* P6-12, *H. autotrophicum* IAM 14942 e *H. massiliense* JC206 não degradariam 4-clorocatecol. Foi também verificado que todos os *Herbaspirillum* apresentam o gene carboxi-metil enebutenolidase (EC 3.1.1.45), o que deixa em aberto a possível degradação de 4-clorocatecol por uma terceira via (FIGURA 6.39).

No genoma de *H. lusitanum* P6-12 havia sido descrita uma proteína RuBisCO-*like* (WEISS *et al*., 2012), que mais tarde passou a ser encontrada em outros genomas de *Herbaspirillum* e provavelmente está relacionada com o metabolismo de enxofre (STRAUB *et al*., 2013). A comparação genômica mostrou a presença de uma proteína paráloga a ela com 48% de identidade. Homólogas dessas proteínas também foram encontradas em *H. seropedicae* estirpes Os34 e Os45, *Herbaspirillum* sp. GW103, *H. frisingense* GSF30, *H. chlorophenolicum* CPW301, *H. hiltneri* N3, *Herbaspirillum* sp. YR522 e *H. autotrophicum* IAM 14942. No entanto, somente *H. hiltneri* N3 e *H. chlorophenolicum* CPW301 também apresentam as duas proteínas.

Através de pesquisa de similaridade BLAST contra o banco de dados NR do NCBI, foi verificado que essas proteínas são homólogas a proteínas encontradas no gênero *Burkholderia*. Também foi verificado que existem homólogas em *H. rubrisubalbicans* (não detectada pelo BLAST todos contra todos local nas estirpes analisadas neste trabalho) e *Herbaspirillum* estirpes B39 e RV1423.

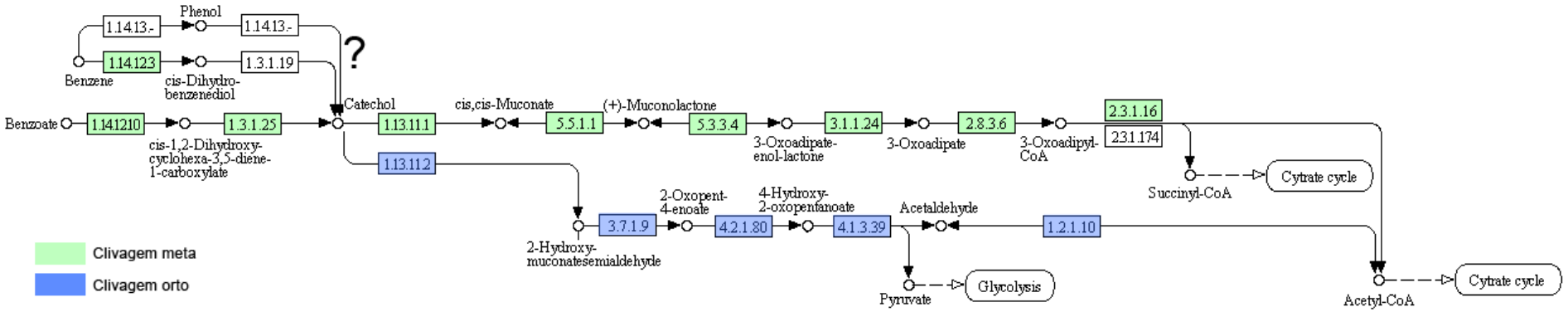


FIGURA 6.38: VIAS PARA A DEGRADAÇÃO DE FENOL/CATECOL EM ESTIRPES DE *Herbaspirillum* spp.

A via marcada em verde é amplamente distribuída no gênero *Herbaspirillum* e degrada catecol/fenol/benzoato a succinil-CoA e acetil-CoA pela clivagem *meta*. A via em azul é encontrada somente em *H. chlorophenolicum* CPW301 e *Herbaspirillum* sp. GW130 e degrada esses compostos a piruvato e acetil-CoA pela clivagem *orto*. A análise foi realizada com o programa KAAS. A interrogação indica que os genes para a degradação de fenol não aparecem na análise, mas estão descritos na literatura para *H. seropedicae* SmR1 (PEDROSA *et al*., 2011).

FONTE: o autor (2015)

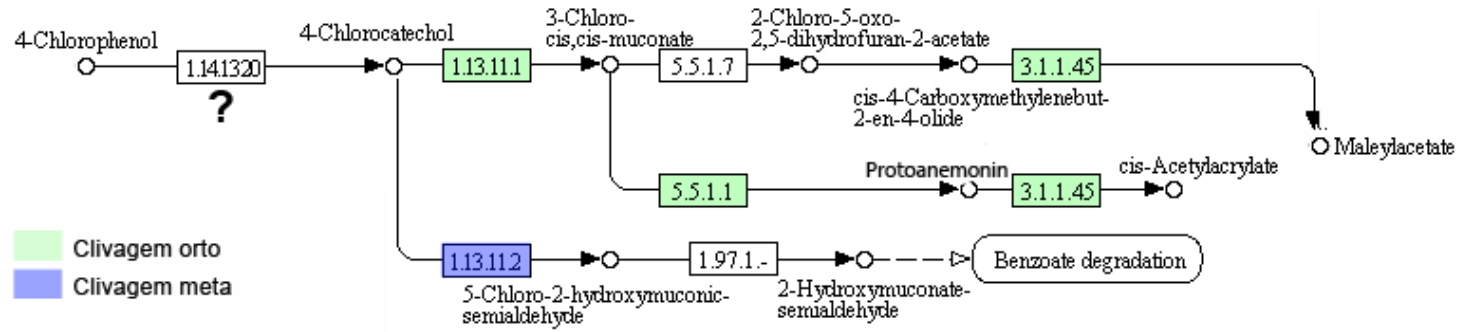


FIGURA 6.39: VIAS PARA A DEGRADAÇÃO DE 4-CLOROFENOL/4-CLOROCATECOL PARA ESTIRPES DE *Herbaspirillum* spp.

A via marcada em verde é amplamente distribuída em estirpes de *Herbaspirillum* spp. e degrada 4-clorocatecol/4-clorofenol para formar protoanemonina (clivagem *orto*), enquanto a via em azul degrada esse composto a semi-aldeído 5-cloro-2-hidoximucônico (clivagem *meta*). As análises foram feitas com o progama KAAS. A interrogação sugere que 4-clorofenol pode ser degradado, embora o gene para isso não apareça na análise.

FONTE: o autor (2015)

A proteína RuBisCO-*like* de *H. autotrophicum* IAM14942 não corresponde à RuBisCO responsável pela fixação de carbono, de forma que esse organismo apresenta as duas enzimas. Através de pesquisa de similaridade BLAST contra o banco de dados NR do NCBI, foi verificado que a RuBisCO fixadora de carbono presente em *H. autotrophicum* IAM14942 é homóloga à encontrada em *Methyloversalitis universalis.* Também foi verificada a presença dessa proteína em uma estirpe de *Herbaspirillum* nomeada 'TSA66' (FIGURA 6.40).

A análise filogenética dessas proteínas mostra que, de fato, a RuBisCO fixadora de carbono forma um grupo isolado das demais proteínas RuBisCO-*like*. Da mesma forma, as duas RuBisCOs-*like* homólogas às encontradas em *H. lusitanum* P6-16 formam dois grupos distintos (FIGURA 6.40).

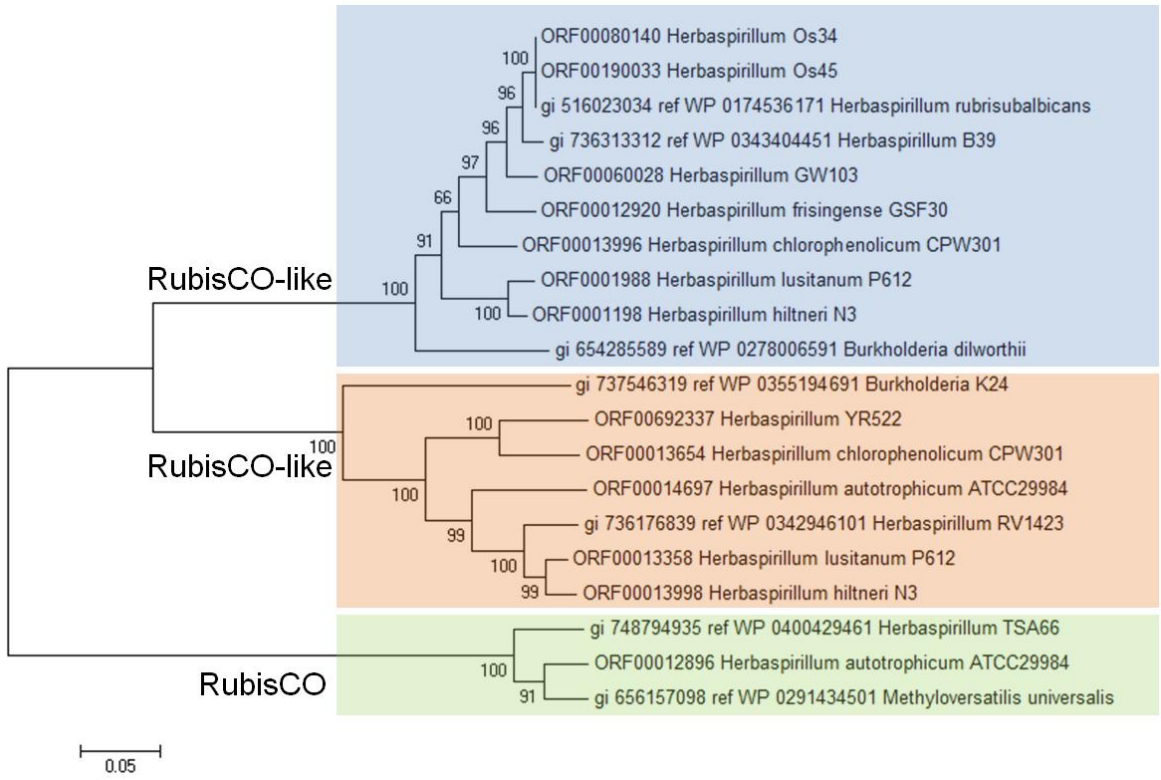


FIGURA 6.40: ÁRVORE FILOGENÉTICA DAS RUBISCOS ENCONTRADAS EM ESTIRPES DE *Herbaspirillum* spp.

A árvore foi construída com o programa MEGA 6.0, com alinhamento produzido pelo MUSCLE, pelo método *Maximum Likelihood* com 1.000 replicatas de *bootstrap*. Em verde está marcado o grupo da RuBisCO fixadora de carbono encontrada em *H. autotrophicum* IAM 14942. Em azul e vermelho estão marcados os grupos formados pelas RuBisCOs-*like* das demais estirpes de *Herbaspirillum* spp.

FONTE: o autor (2015)

6.6.3 Genes que segregam Herbaspirillum associativos de Herbaspirillum ambientais

A filogenia das estirpes de *Herbaspirillum* spp. não apresenta relação com o ambiente em que esses organismos vivem. Tanto no filogrupo 1 quanto no filogrupo 2 existem *Herbaspirillum* que foram isolados de plantas ou rizosfera de plantas. Nesta análise, foram considerados associativos os organismos: *H. seropedicae* estirpes SmR1, Os34 e Os45; *H. rubrisubalbicans* M1; *H. frisingense* GSF30; *Herbaspirillum* estirpes GW103, YR522, CF444; *H. lusitanum* P6-12; *H. rhizosphaerae* UMS-37; e *H. hiltneri* N3. Os demais *Herbaspirillum* foram incluídos no grupo dos ambientais: *H. huttiense* subsp. *putei*, *H. chlorophenolicum* CPW301, *H. autotrophicum* IAM14942 e *H. massiliense* JC206.

Com isso, não foi encontrado nenhum gene que estivesse presente em todos os *Herbaspirillum* associativos analisados e, ao mesmo tempo, ausente em todos os ambientais. Porém, alguns genes foram separados para a análise: aqueles presentes em todos os associativos e em apenas um dos ambientais (11 associativos e 1 ambiental, 17 genes), e aqueles encontrados em pelo menos 10 associativos e ausentes em todos os ambientais (10 associativos e 0 ambientais, 7 genes).

Com exceção das proteínas anotadas como hipotéticas, as proteínas codificadas por esses genes foram: fosfodiesterase transmembrana; exoribonuclease BN transmembrana; permease transmembrana; componente permease de transportador tipo ABC para açúcar; componente ATPase de transportador tipo ABC para aminoácido de cadeia ramificada; trans-hidrogenase transmembrana NAD(P) (subunidade alfa); regulador transcricional da família GntR; sistema de dois componentes para regulação de resposta; sistema de dois componentes sensor de histidina quinase; proteína LrgA efetora de mureína hidrolase; ureído-glicolato hidrolase; amidase; L-aminopeptidase/D-esterase; D-arabinitol 4-desidrogenase; etanolamina amônia-liase subunidade menor; etanolamina amônia-liase subunidade maior; diguanilato ciclase/fosfodiesterase com sensor PAS/PAC; nucleosídeo-difosfato-açúcar epimerase; e di-hidroxiácido desidratase/fosfogluconato desidratase.

6.6.4 Análises taxonômicas baseadas em sequência genômica

A disponibilidade de sequências genômicas completas também foi utilizada para verificar a classificação taxonômica das estirpes de *Herbaspirillum* spp. pelos métodos de ANI (identidade média de nucleotídeos) com linha de corte de 95% (GORIS *et al*., 2007) e GGDH (distância de hibridização entre dois genomas) com linha de corte de 70% (THOMPSON *et al*., 2013). A análise ANI apresentou identidade superior a 95% entre as estirpes SmR1, BR11417, BR11335 e AU14040 de *H. seropedicae*, o que demonstra que essas estirpes são de uma mesma espécie. No entanto, as estirpes Os34 e Os45 apresentaram identidade de 88,6% e 88,5% respectivamente e, dessa forma, não pertencem à espécie *H. seropedicae*. A análise de GGDH também demonstra isso, pois as estirpes Os34 e Os45 hibridizam apenas 30% com a estirpe SmR1, muito abaixo de 70% (FIGURA 6.41).

Pela análise ANI é mais coerente classificar as estirpes Os34 e Os45 como pertencentes à espécie *H. rubrisubalbicans*, visto que apresentaram identidade de 97,1% e 97,2% em relação à estirpe M1 dessa espécie, respectivamente. Isso também é confirmado pela análise GGDH, onde os genomas hibridizaram 72% e 72,6% com a estirpe M1, respectivamente (FIGURA 6.41).

A estirpe GW103 apresentou identidade de 90,8% em relação à *H. huttiense* subsp. *putei*, segundo a análise ANI, e hibridização de 38,1% segundo a análise GGDH, o que demonstra que essa estirpe é uma nova espécie. Da mesma maneira, as estirpes YR522 e CF444 não apresentaram nenhum indício de pertencer a espécies já descritas e também são, portanto, espécies novas. A relação da estirpe YR522 é mais próxima com *H. chlorophenolicum* CPW301 (85,3% de ANI e 24,8% de GGDH), como já observado na árvore do pangenoma, e a relação da estirpe CF444 é maior com *H. hiltneri* N3 (90,8% de ANI e 37,7% de GGDH) (FIGURA 6.41), enquanto nas análises anteriores essa estirpe se mostrou mais próxima de *H. rhizosphaerae* UMS-37 (ver FIGURA 6.24).

	H. sero SmR1	*H. sero* BR417	*H. sero* BR335	*H. sero* AU14040	*H. sero* Os34	*H. sero* Os45	*H. rubr* M1	*H. rubr* BR504	*H. hutt putei*	*H. sp* GW103	*H. fris* GSF30	*H. chlo* CPW301	*H. sp* YR522	*H. sp* CF444	*H. rhiz* UMS-37	*H. lusi* P6-12	*H. hilt* N3	*H. auto* IAM 14942	*H. mass* JC206
H. sero SmR1	100	100	100	89,7	30,0	30,0	29,9	29,9	30,4	30,6	29,8	25,5	23,9	22,3	22,2	22,3	22,5	21,2	21,1
H. sero BR417	100	100	100	89,6	29,9	30,0	29,9	29,9	30,4	30,5	29,8	25,5	23,9	22,3	22,2	22,3	22,4	21,1	20,9
H. sero BR335	100	100	100	83,3	30,3	30,3	30,2	30,2	30,7	30,9	30,1	25,7	24,2	22,5	22,4	22,5	22,6	21,3	21,0
H. sero AU14040	98,9	98,9	98,9	100	30,1	30,0	30,1	30,0	30,8	30,7	29,8	25,5	23,9	22,2	22,3	22,2	22,3	21,1	21,0
H. sero Os34	88,6	88,5	88,6	88,4	100	99,2	72,0	72,0	30,8	31,4	29,9	25,0	23,6	22,1	22,1	22,2	22,1	20,6	21,1
H. sero Os45	88,4	88,5	88,5	88,4	99,9	100	72,6	72,4	30,7	31,3	29,8	24,9	23,6	22,1	22,1	22,1	22,1	20,5	20,9
H. rubr M1	88,5	88,4	88,5	88,4	97,1	97,2	100	94,6	31,0	31,4	29,9	25,1	23,7	22,3	22,1	22,2	22,1	20,7	21,1
H. rubr BR504	88,5	88,3	88,5	88,4	97,2	97,1	99,5	100	30,9	31,4	29,9	25,1	23,7	22,2	22,1	22,2	22,0	20,7	20,7
H. hutt putei	88,7	88,8	88,8	88,9	88,5	88,5	88,6	88,6	100	38,1	29,8	25,6	24,1	22,4	22,5	22,6	22,2	20,9	21,0
H. sp GW103	88,7	88,6	88,8	88,7	88,7	88,7	88,7	88,7	90,8	100	29,6	25,5	23,6	22,2	22,2	21,9	22,0	21,0	21,0
H. fris GSF30	88,4	88,4	88,4	88,4	88,2	88,1	88,1	88,1	88,2	88,1	100	25,7	24,2	22,5	22,5	22,6	22,5	21,0	21,0
H. chlo CPW301	85,9	85,9	85,9	85,8	85,5	85,4	85,5	85,5	86,0	85,8	86,0	100	24,8	23,0	23,0	23,1	22,9	21,6	21,8
H. sp YR522	84,8	84,8	84,7	84,8	84,5	84,5	84,4	84,6	84,5	84,5	84,9	85,3	100	22,2	22,0	22,1	22,1	20,7	20,4
H. sp CF444	83,4	83,4	83,5	83,4	83,2	83,2	83,1	83,1	83,5	83,2	83,5	84,0	83,2	100	29,5	28,9	37,7	22,5	21,3
H. rhiz UMS-37	83,5	83,5	83,4	83,5	83,2	83,2	83,3	83,3	83,6	83,3	83,5	84,1	83,3	88,3	100	40,1	29,0	22,2	20,8
H. lusi P6-12	83,5	83,5	83,5	83,5	83,3	83,2	83,4	83,3	83,5	83,3	83,5	84,1	83,2	87,8	91,3	100	28,7	22,4	20,7
H. hilt N3	83,3	83,2	83,3	83,4	83,2	83,1	83,2	83,1	83,4	83,2	83,4	84,1	83,1	90,8	87,9	87,7	100	22,7	20,8
H. auto IAM 14942	82,1	82,1	82,0	82,1	82,0	82,1	82,0	82,1	82,1	82,1	82,1	82,5	82,0	83,8	83,5	83,6	84,1	100	19,8
H. mass JC206	82,1	82,1	82,1	82,1	81,7	81,7	81,8	81,7	82,0	82,1	82,0	82,4	81,6	82,2	81,7	81,9	82,0	81,2	100

FIGURA 6.41: ANI E GGDH ENTRE GENOMAS DE ESTIRPES DE *Herbaspirillum* spp.

Nessa imagem, são mostradas duas matrizes triangulares, separadas por um eixo diagonal que compara os genomas com eles mesmos (valores 100%, em branco). A matriz mostrada à direita desse eixo (mais escura) representa os valores de GGDH. A matriz à esquerda (mais clara) representa os valores de ANI. Os números no interior dos quadrados representam a porcentagem de identidade entre dois genomas conforme a análise realizada. Os valores vêm acompanhados de uma coloração, que quanto mais clara, maior a identidade entre dois genomas.

FONTE: o autor (2015)

7 DISCUSSÃO

7.1 Avaliação dos conjuntos de dados de sequenciamento

A plataforma SOLiD foi utilizada com sucesso pelo grupo NFN para a produção dos *drafts* genômicos de *H. lusitanum* P6-12 e *H. huttiense subsp putei* (WEISS *et al.,* 2012; DE SOUZA *et al.,* 2013), para os quais foram obtidas coberturas de aproximadamente 1.000 vezes e *reads* pareados. Em comparação a isso, é possível afirmar que a quantidade de *reads* obtidos (entre 80 e 200 vezes de cobertura) para os genomas de *H. chlorophenolicum* CPW301, *H. autotrophicum* IAM 14942 e *H. rhizosphaerae* UMS-37 foi muito baixa utilizando a mesma plataforma.

Ao longo desses *reads* foi observado decréscimo de qualidade, o que já era esperado em *reads* obtidos da plataforma SOLiD, assim como ocorre em outras plataformas. Porém, ao levar em conta o tamanho reduzido dos *reads* (50 pb), a perda da confiabilidade das últimas bases torna a informação ainda mais limitada. Além disso, a baixa qualidade da primeira base representou um grave problema, devido ao processo de decodificação SOLiD. Por isso, os únicos dados de sequenciamento que puderam ser considerados como modelo para o início do processo de montagem genômica foram os de *H. chlorophenolicum* CPW301 (FIGURA 6.1).

Em geral, foi observado que a média de tamanho dos *reads* Illumina, entre 150 e 180 pb (FIGURA 6.3), eram equivalentes ao tamanho de muitos *contigs* obtidos nas montagens realizadas com os *reads* SOLiD. Por exemplo, na montagem SOLiD com dados brutos e *hsize* 23, a média de tamanho dos *contigs* obtidos foi 335 pb, enquanto que a união dos pares Illumina equivalem a cerca de 300 bases. O fato dos *reads* Illumina serem dados brutos os tornaram também mais confiáveis do que os *contigs* SOLiD (usados como *reads*) para as montagens ‘Pan’, pois não apresentam propagação de erros. Aliado a isso, a qualidade dos *reads* provenientes da plataforma Illumina foi bastante satisfatória (FIGURA 6.2), além de dispensar a decodificação SOLiD que, pelo erro de codificação de uma base, pode gerar um *read* total ou parcialmente falso.

7.2 Montagem e anotação genômica

A baixa cobertura para os dados SOLiD, ausência de *reads* pareados e a baixa qualidade dos *reads* obtidos são refletidas nas montagens genômicas obtidas com o *pipeline de novo*, as quais compreenderam milhares de *contigs* sem que o tamanho esperado do genoma (5 Mb) fosse atingindo (TABELA 6.1). Ao aumentar o tamanho do *hsize*, foi observada a alteração do tamanho do *contig* N50 que, embora aumente também, não representa uma melhora real, pois o tamanho do conjunto de *contigs* diminui. Dessa maneira, escolher uma melhor montagem seria algo totalmente arbitrário. Por isso, apenas para fins de comparações com outras montagens, a montagem com *hsize* 23 foi utilizada como referência entre as montagens SOLiD produzidas com o *pipeline de novo*.

A redução do conjunto de dados (menor cobertura) e do tamanho dos *reads*, através do processo de *trimming*, mostrou ter um impacto negativo nas montagens realizadas (TABELA 6.2). A utilização de um maior número de *reads*, de tamanho maior, porém de qualidade inferior, mostrou ser mais vantajoso nos quesitos avaliados, embora não seja descartado que essas montagens possam ter incoerências devido a problemas de qualidade.

Mesmo com a utilização da plataforma CLC *Genomic Workbench*, os resultados obtidos não foram proveitosos (TABELA 6.3), visto que o maior problema estava no conjunto de dados utilizados e não nas ferramentas de montagem genômica. Para o mapeamento de *reads* também eram esperados resultados melhores, como por exemplo, *reads* cobrindo praticamente toda a referência, principalmente para os *reads* SOLiD, visto que quanto mais curto o *read*, mais fácil é seu encaixe em uma referência. No entanto, foi observada pouca similaridade entre os *reads* de *H. chorophenolicum* CPW301 e os genomas de *Herbaspirillum* spp., o que aparentemente é explicado pela baixa qualidade dos *reads* (TABELA 6.4).

Para tentar diminuir os problemas relacionados ao tamanho e à baixa qualidade dos *reads*, os *contigs* das montagens previamente realizadas passaram a ser utilizados como *reads* nas montagens ‘Pan’. Esses *contigs* já montados, embora pudessem apresentar o problema de propagação de erros (ao serem reutilizados como *reads*), foram obtidos de um consenso de vários *reads* empilhados, ou seja, eles poderiam também ter maior confiabilidade de sequência em relação aos *reads* brutos. O tamanho maior desses *reads* provavelmente foi responsável por facilitar a

montagem (TABELA 6.5), ainda que não pudesse resolver repetições e, consequentemente, produzir *drafts* de qualidade.

O resultado do mapeamento dos *contigs* obtidos na montagem 'Pan *contigs'*, em relação ao genoma de *H. seropedicae* SmR1 (FIGURA 6.4), foi similar aos resultados dos mapeamentos obtidos com os *reads* brutos, pois tiveram uma baixa cobertura (1,8 Mb) em relação ao genoma de referência. Porém, isso também poderia demonstrar que os *contigs* montados provavelmente não compreenderam o genoma como um todo, mas que faltaria cobertura para a montagem de várias regiões.

Com relação à qualidade das sequências e aos resultados prévios para a montagem genômica com dados provenientes da plataforma Illumina, mesmo com cobertura inferior, foi observada a vantagem do uso de *reads* longos dessa plataforma em relação ao uso de *reads* curtos da plataforma SOLiD (FIGURA 6.5).

A montagem C2 de *H. chlorophenolicum* CPW301, realizada com o conjunto de dados da plataforma Illumina (TABELA 6.6), foi igualada (em número de *contigs*) aos resultados prévios obtidos para *H. autotrophicum* IAM 14942 e *H. rhizosphaerae* UMS-37 obtidas com a montagem automática Illumina, embora ainda não fosse considerada satisfatória. A diminuição do número de *contigs* de 1.011 para 496 e o aumento do tamanho do genoma em ~0,3 Mb, para *H. chlorophenolicum* CPW301, demonstraram que as montagens automáticas Illumina ainda poderia ser melhoradas para os outros genomas.

A interferência da cobertura e do pareamento nas montagens foi observada pela diferença de resultados das montagens C1 (que utilizou somente a ponta R1) e C2 (que utilizou as duas pontas), o que demonstra a importância dessas variáveis para a montagem de um genoma (TABELA 6.6).

Assim como ocorreu para os dados de sequenciamento da plataforma SOLiD, o *trimming* das sequências mostrou ser prejudicial à montagem genômica, provavelmente devido à diminuição do tamanho dos *reads* e da diminuição da cobertura (TABELA 6.6). Mesmo sem a garantia de qualidade das sequências pós-*trimming*, é possível imaginar que a própria cobertura se encarregue de "corrigir" a extremidade dos *reads* dos dados brutos, de forma que essa informação seja consertada sem prejuízo de tamanho de *reads* e cobertura.

A montagem híbrida PC + I + C2, embora tenha apresentado menor número de *contigs*, apresentou também diminuição do tamanho do conjunto de *contigs*, o que

torna difícil afirmar se houve ou não ganho nessa montagem. Mesmo com a diminuição do tamanho do conjunto, o *contig* N50 diminuiu de tamanho, e o único aspecto positivo dessa montagem foi o tamanho do maior *contig* (~264 Kb, contra ~233,5 Kb da montagem C2 - TABELA 6.6).

Já as montagens realizadas com o montador Newbler mostraram que, embora incapaz de fazer bom uso dos *reads* provenientes da plataforma Illumina, isolados, esse montador apresenta bom desempenho com a união de *reads* Illumina aos *contigs* das montagens realizadas anteriormente. A montagem híbrida *Reads* Illumina + PC + C2 foi considerada superior à hibridização feita pela plataforma CLC *Genomics Workbench*, onde ocorreu perda do tamanho do genoma (TABELA 6.7). No entanto, o número de divergências encontradas entre os dados provenientes das duas plataformas de sequenciamento (FIGURA 6.6) levou a acreditar que as montagens realizadas com os dados de sequenciamento da plataforma SOLiD apresentassem algum tipo de erro, ou interferissem nos dados Illumina, e por isso foram abandonados.

Com essa exclusão, foi possível observar melhora real nas montagens obtidas, culminando com a montagem *Reads* Illumina + C2 e seus 216 *contigs* gerados (TABELA 6.8). A visualização prévia das montagens produzidas na plataforma CLC *Genomics Workbench* já havia demonstrado indícios de ligação de *contigs*, que por algum motivo não eram unidos pelo montador CLC. Uma das hipóteses seria o tamanho de *overlap* utilizado: se o *wsize* no CLC é 24, o *overlap* seria 23, enquanto no Newbler o *overlap* usado foi 20. No entanto, melhora nas montagens produzidas dentro do CLC com *wsizes* menores que 23 não foram observadas nos testes preliminares. Nesse contexto, as hibridizações das montagens pelo programa Newbler foram consideradas um processo de finalização automática para a união de *contigs*.

Ainda assim, o número de *contigs* poderia ser reduzido por processos manuais, mas as tentativas de uni-los foram consideradas inadequadas para genomas produzidos com *reads pair-end*, visto que somente os *reads mate-pair* poderiam solucionar repetições e garantir determinadas junções (FIGURA 6.8). No entanto, o método de hibridização *Reads* Illumina + C2 foi considerado proveitoso e aplicado aos outros dois genomas. Essa estratégia, desenvolvida neste trabalho, foi aplicada com êxito na montagem completa do genoma do cloroplasto de *Podocarpus lambetii* (VIEIRA *et al*., 2014; FAORO, Informação verbal), e também foi responsável por

melhoras significativas das montagens genômicas de *Azoarcus olearius* DQS-4 (MENEGAZZO, Informação verbal) e *H. hiltneri* N3 (GUIZELINI, Informação verbal).

O mapeamento da montagem genômica de *H. chlorophenolicum* CPW301 produzida com dados de sequenciamento da plataforma Illumina, em relação ao genoma de *H. seropedicae* SmR1, mudou muito se comparado à montagem 'Pan *contigs'* realizada com dados SOLiD, o que tornou mais visível as diferenças estruturais dos dois genomas. Na montagem 'Pan *contigs'*, o genoma estava mais fragmentado e por isso o alinhamento foi facilitado. Da mesma forma, foi observado o aumento de divergências nos gráficos *dotplot* realizados com as montagens automáticas Illumina frente à montagem final, devido justamente ao tamanho maior dos *contigs* utilizados (FIGURA 6.7).

Por outro lado, a cobertura de 38% (2.083.540 bases) da montagem final em relação à referência, foi pouco maior do que a obtida com os dados SOLiD. Isso elimina a hipótese de que regiões do genoma não haviam sido montadas e leva a acreditar que as sequêmcias genômicas são, em sua maior parte, distintas entre si. Os genomas dos outros dois organismos são taxonomicamente distantes da referência e por isso o perfil dos gráficos *dotplot* ficou mais fragmentado (FIGURA 6.7).

O conjunto de *contigs* obtido para cada um dos três genomas foi considerado satisfatório para o processo de anotação, com uma única objeção de que os *contigs* com menos de 200 bases deveriam ser descartados, visto que esses *contigs* não fazem sentido devido ao tamanho dos *reads* obtidos. Aparentemente as montagens atingiram seu limite de redução de número de *contigs* até onde os dados permitiram, resultando em genomas maiores que o esperado e com conteúdo G+C condizente com o observado dentro do gênero *Herbaspirillum* (~60% - TABELA 6.9).

Na anotação de *H. autotrophicum* IAM 14942 (FIGURA 6.14), a ausência do gene codificador da fosfofrutoquinase-1 (PFK-1, EC 2.7.1.11) na via de Embden-Meyerhoff-Parnas também já havia sido descrita para *H. seropedicae* SmR1 (PEDROSA *et al.*, 2011). Pedrosa e colaboradores (2011) sugeriram que *H. seropedicae* SmR1 supere esse bloqueio utilizando a via das pentoses fosfato e a via de Entner-Doudoroff para metabolizar D-glucose, D-frutose e D-manose a piruvato. Entretanto, o gene que codifica para a enzima 6-fosfogluconolactonase, utilizada na fase oxidativa da via das pentoses fosfato, não foi encontrado no genoma de *H. autotrophicum* IAM 14942 (FIGURA 6.12). Dentre as enzimas que

poderiam substituí-la, foi encontrado um gene de uma proteína anotada como ‘possível gluconolactonase’ que não apresenta homologia com nenhuma proteína presente nos demais *Herbaspirillum*.

Em *H. seropedicae* SmR1 foi descrita a presença de genes envolvidos com a síntese e degradação de polímeros de glucose (genes *glgABX*) e que serviriam de proteção contra estresse osmótico (PEDROSA *et al.*, 2011). Entretanto, esses genes não foram encontrados no genoma parcial de *H. autotrophicum* IAM 14942, o que reforça o fato de que essa bactéria não metaboliza glucose. Ainda assim, a bactéria poderia proteger-se do estresse osmótico produzindo trealose via gene *ostB*. Em *Herbaspirillum* em geral, essa síntese pode ocorrer tanto por genes *ots* quanto por genes *tre*, com exceção de *Herbaspirillum sp* CF444 e *H. massiliense* JC206 (STRAUB *et al*., 2013).

A fase oxidativa da via das pentoses fosfato é responsável pela produção de NADPH, o que seria afetado pela possível ausência da 6-fosfogluconolactonase. No entanto, no genoma parcial de *H. autotrophicum* IAM 14942 foi encontrado um gene que codifica para a enzima gliceraldeído-3-fosfato desidrogenase $NADP^+$-específica (EC 1.2.1.9 – FIGURA 6.12). Essa enzima converte gliceraldeído-3P a glicerato-3P com produção de NADPH, e é descrita como responsável pela produção de NADPH para biossíntese em *Streptococcus mutans* e em algumas espécies de bactérias autotróficas (BOYD *et al.*, 1995, FILLINGER *et al*., 2000). Nos demais *Herbaspirillum*, essa enzima está ausente, o que indica que *H. autotrophicum* IAM 14942 apresenta uma rota alternativa para a produção de NADPH que é independente da via das pentoses fosfato.

Outra forma de dar continuidade à via glicolítica seria através dos genes que codificam para a enzima 1-fosfofrutoquinase (*fruK*, EC 2.7.1.56) e para um sistema de transporte PEP/PTS parcial (subunidade PTS-Fru-EIIB). Esses genes estão ausentes nos genomas dos demais *Herbaspirillum*. Embora a análise *in silico* não possa certificar a funcionalidade das vias metabólicas analisadas, caso sejam viáveis, o metabolismo de açúcares em *H. autotrophicum* IAM 14942 provavelmente é diferente do que é conhecido para os demais *Herbaspirillum*. Da mesma forma, os genes relacionados com transporte e degradação de outros açúcares, como D-galactose e L-arabinose, também não foram encontrados. Por outro lado, a bactéria apresenta todos os genes da via da gluconeogênese, como também foi relatado para *H. seropedicae* SmR1 (PEDROSA *et al.*, 2011).

A presença de genes codificadores das enzimas L-lactato desidrogenase (EC 1.1.1.27) e D-lactato desidrogenase – citocromo (EC 1.1.28) são condizentes com a fisiologia da bactéria, pois *H. autotrophicum* IAM 14942 foi descrita como capaz de crescer na presença de lactato como única fonte de carbono e energia (ARAGNO & SCHLEGEL, 1978). Em desacordo com isso, está a presença da álcool desidrogenase (EC 1.1.1.1), da álcool desidrogenase – citocromo (EC 1.1.2.8) e dos genes relacionados ao metabolismo de 1-butanol, visto que na descrição da espécie é mencionada a incapacidade dessa bactéria crescer na presença de alcoóis como únicas fontes de carbono e energia (ARAGNO & SCHLEGEL, 1978). A presença de genes relacionados com o metabolismo de açúcares e alcoóis, em desacordo com a literatura, reflete a falta de estudos realizados para a compreensão da fisiologia dessa bactéria.

O fato de os genes que codificam as enzimas do complexo da piruvato desidrogenase e da conversão de piruvato a acetil-CoA, assim como as enzimas do ciclo do ácido cítrico, estarem presentes, pode demonstrar a preferência da espécie para aminoácidos e ácidos orgânicos como fontes de carbono e energia (ARAGNO & SCHLEGEL, 1978. BALDANI *et al*., 2014).

Aragno & Schlegel (1978) demonstraram que *H. autotrophicum* IAM 14942 pode crescer na presença de L-alanina, L-asparagina, L-aspartato, L-glutamato, L-glutamina, L-glicina, L-isoleucina, L-fenilalanina, L-prolina, L-triptofano e L-tirosina como únicas fontes de carbono e energia. Porém, nem todos os genes responsáveis pela degradação desses aminoácidos foram encontrados (ver subtópico 6.3.3). Entre as vias metabólicas observadas pelo conjunto proteômico do genoma, foram encontrados os genes relacionados com o ciclo da ureia. Essa é provavelmente também a via utilizada para a biossíntese de arginina, como sugerido para *H. seropedicae* SmR1 (PEDROSA *et al.*, 2011). Pedrosa e colaboradores (2011) também sugeriram uma via alternativa para o metabolismo de L-arginina em *H. seropedicae* SmR1 que envolve a conversão de L-arginina a agmatina e, posteriormente, a putrescina. Porém, o gene que codifica para a enzima agmatinase (EC 3.5.3.11), presente nessa via, não foi encontrado na sequência genômica parcial de *H. autotrophicum* IAM 14942.

Na sequência genômica dessa bactéria estão presentes genes que codificam transportadores ABC para espermidina/putrescina, ausentes em *H. seropedicae* SmR1. Em complemento a eles, foram também encontrados genes que codificam

enzimas presentes na via de conversão de putrescina a espermina, a qual conflui para a via de degradação de β-alanina/glutationa. A presença desses genes pode estar relacionada com o fato de a putrescina ser derivada da degradação de aminoácidos e encontrada em matéria orgânica em decomposição, visto que *H. autotrophicum* IAM 14942 foi isolada de um ambiente eutrófico. Essa abundância de matéria orgânica poderia justificar também o fato de o organismo utilizar ácidos orgânicos e aminoácidos como principais fontes de carbono e energia.

H. autotrophicum IAM 14942 foi descrita como incapaz de realizar a fixação biológica de nitrogênio e incapaz de reduzir nitrato a nitrito (ARAGNO & SCHLEGEL, 1978) e, como o esperado, os genes *nifHDK*, responsáveis por codificar para as proteínas que formam o complexo da nitrogenase, não estão presentes no genoma. Os genes relacionados com a redução dissimilatória do nitrato também não foram encontrados. Baseado em uma compilação de características metabólicas do gênero (JUNG *et al*., 2007) e no fato de que *H. lusitanum* P6-12 não apresenta os genes *nifHDK* (WEISS *et al*., 2012) é possível afirmar que a capacidade de reduzir nitrato a nitrito é atribuída somente aos *Herbaspirillum* fixadores de nitrogênio. A presença dos genes relacionados com a redução assimilatória de nitrato (*nasA*), pela conversão de nitrato a amônia (genes *nirBD*), e pelo transportador ABC para o transporte de nitrato/nitrito/cianato são também encontrados nos demais genomas de *Herbaspirillum*.

Aragno & Schlegel (1978) verificaram que o CO_2 é fixado autotroficamente através do ciclo de Calvin em *H. autotrophicum* IAM 14942. Nessa via, foram encontrados os genes que codificam para as enzimas fosforibuloquinase (EC 2.7.1.19), que converte ribose-5P em ribulose-5P, e RuBisCO (EC 4.1.1.39), que fixa carbono em ribulose-1,5P_2 para produzir duas moléculas de glicerato-3P (MARTIN & SCHNARRENBERGER, 1997), o que torna possível a fixação de carbono em *H. autotrophicum* IAM 14942.

A presença de genes relacionados com proteínas RuBisCO-*like* (RLP; também denominada forma IV da RuBisCO) foi identificada em outros *Herbaspirillum*, assim como já haviam sido descritas (WEISS *et al*., 2012). Entretanto, o papel das RLP ainda não é bem conhecido, mas para alguns organismos elas parecem estar envolvidas em uma via de recuperação de metionina, enquanto em bactérias sulfurosas verdes, estão envolvidas no metabolismo oxidativo de tiossulfato (TABITA *et al*., 2007). De qualquer forma, apesar da similaridade e possível homologia com

as RuBisCOs clássicas, as RLP parecem não apresentar atividade de carboxilase, embora essas enzimas catalizem reações de enolização similares como parte de seus mecanismos de reação (TABITA *et al*., 2007).

Os genes que codificam para as enzimas sedoheptulose-1,7-bifosfatase (EC 3.1.3.37) e sedoheptuloquinase (EC 2.7.1.14), as quais formam duas vias para produzir sedoheptulose-7P, não foram encontrados. Para cianobactérias, foi descrito uma única enzima que apresenta as atividades tanto da sedoheptulose-1,7-bifosfatase quanto da frutose-1,6-bifosfatase (MIYAGAWA *et al*., 2001). *H. autotrophicum* IAM 14942 apresenta dois genes que codificam para a frutose-1,6-bifosfatase. O primeiro deles é homólogo aos genes da frutose-1,6-bifosfatase dos demais *Herbaspirillum* e faz parte da via da gluconeogênese. O segundo está dentro do agrupamento gênico relacionado com a via de fixação de carbono e não apresenta homologia com nenhum gene dos demais *Herbaspirillum* (FIGURA 6.13). A presença desse segundo gene, dentro do agrupamento de genes relacionados com a fixação de carbono, sugere que a segunda frutose-1,6-bifosfatase substitua a sedoheptulose-1,7-bifosfatase.

Para o metabolismo litoautotrófico, foi descrito que *H. autotrophicum* IAM 14942 utiliza H_2 como fonte de energia (ARAGNO & SCHLEGEL, 1978). No genoma parcial dessa bactéria foram encontrados os genes *hoxAJLOQRTV* que codificam para as enzimas responsáveis pela formação do complexo da hidrogenase, bem como os genes *hypABCDEF* que são auxiliares à formação desse complexo (EBERZ *et al*., 1986, SCHIFFELS *et al*., 2013). A proximidade entre os genes que compõe o complexo da hidrogenase e os genes responsáveis pela fixação de carbono sugere que todos esses genes formem um grande agrupamento (FIGURA 6.13) e que eles tenham sido adquiridos em conjunto por transferência horizontal de genes, embora indícios dessa transferência não tenham sido encontrados.

O genoma parcial de *H. autotrophicum* IAM 14942 também apresenta o gene que codifica para a formato desidrogenase (EC 1.2.1.2), que em *Ralstonia eutropha* H16 é responsável pela obtenção de energia proveniente do ácido fórmico para fixar carbono organo-autotroficamente (POHLMANN *et al*., 2006).

A presença de grânulos citoplasmáticos de poli(3-hidroxibutirato) já havia sido descrita para a espécie (ARAGNO & SCHLEGEL, 1978), bem como para *H. seropedicae* SmR1 (PEDROSA *et al.*, 2011), o que justifica a presença dos genes *phb*B, *phbC/pha*C e *pha*Z para o metabolismo de poli(3-hidroxi-alcanoatos). Foi

verificado que todos os genomas de *Herbaspirillum* apresentam esses genes e que eles são homólogos entre as estirpes de *Herbaspirillum* spp.

Diferentemente de *H. seropedicae* SmR1, *H. autotrophicum* IAM 14942 não possui os genes envolvidos na degradação de benzoato e de fluorobenzeno. Para *H. seropedicae* SmR1, a presença desses genes foi associada à flexibilidade metabólica e defesa contra compostos tóxicos derivados da planta (PEDROSA *et al.*, 2011). Portanto, a ausência desses genes no genoma de *H. autotrophicum* IAM 14942 é justificada pelo ambiente não endofítico onde a espécie foi encontrada. Por outro lado, a presença de genes que codificam para enzimas envolvidas na degradação de derivados de catecol e outros compostos fenólicos pode indicar a diversidade de metabólitos aos quais a bactéria tem acesso em seu ambiente.

H. autotrophicum IAM 14942 apresenta os genes *ppk* e *ppx*, que codificam para enzimas envolvidas na síntese e degradação de polifosfato. Para *H. seropedicae* SmR1, foi sugerido que esses genes estejam relacionados com mecanismos de defesa adaptativa ao ambiente endofítico (PEDROSA *et al.*, 2011). No entanto, esses genes estão também associados a uma série de fatores de resistência a estresse (estresses oxidativo e osmótico, calor, falta de nutrientes – SEUFFERHELD *et al*., 2008). A presença desses genes em *Herbaspirillum* pode estar relacionada com a diversidade do grupo, pois esses genes poderiam ajudá-los a enfrentar diversos tipos de estresse e, com isso, permitir a colonização de novos ambientes. Isso poderia justificar o fato de os *Herbaspirillum* terem sido encontrados em diversos nichos (LAGIER *et al*., 2012, JAUREGUI *et al*., 2014), inclusive alguns bem particulares, como depósitos vulcânicos (LU *et al*., 2008).

A ausência dos genes relacionados ao T3SS, responsável pela liberação de proteínas efetoras dentro da célula eucariótica hospedeira (SCHMIDT *et al*., 2012), e dos genes pertencentes ao T6SS, que está envolvido com a interação de virulência à planta (STRAUB *et al*., 2013), provavelmente é devida ao fato de *H. autotrophicum* IAM 14942 não ter sido isolado de ambiente endofítico.

Entretanto, no genoma parcial de *H. autotrophicum* IAM 14942 foram encontrados os genes relacionados com o sistema de exportação de proteínas Tat (*Twin arginine target*). Em *Ralstonia eutropha* H16, esse sistema está envolvido com a exportação da formato desidrogenase através da membrana citoplasmática, para que seja realizado o metabolismo organo-autotrófico (POHLMANN *et al*., 2006).

Para *H. chlorophenolicum* CPW301 (FIGURA 6.18), embora tenha sido verificado que ela poderia metabolizar frutose e glucose, não constam na literatura testes bioquímicos realizados com esses açúcares. No entanto, foram testados manose, ramnose, ribose e xilose como únicas fontes de carbono e energia para essa bactéria, que se mostrou incapaz de crescer na presença desses substratos (IM *et al*., 2004). De fato, a sequência genômica dessa bactéria não apresenta genes relacionados com o metabolismo de manose, ramnose e xilose, mas aparentemente a ribose poderia ser convertida a gliceraldeído-3P ou frutose-6P pelas interconversões da via das pentoses fosfato. Com relação aos sistemas de transporte ABC, o sistema para o transporte de ribose/xilose talvez funcione apenas para ribose, o sistema parcial para D-xilose talvez não seja funcional e o sistema de transporte de glucose/manose ou funcione apenas para glucose ou também não seja funcional.

Assim como descrito para *H. autotrophicum* IAM 14942 e ao contrário de *H. seropedicae* SmR1, *H. chlorophenolicum* CPW301 não apresenta os genes para a síntese de polímeros de glucose. Essa bactéria apresenta o gene *otsB* para a síntese de trealose, mas não os genes *tre*, o que a torna mais uma exceção dentro do gênero *Herbaspirillum*.

Não são relatados na literatura testes bioquímicos com *H. chlorophenolicum* CPW301 que utilizem lactato ou etanol, mas pela sequência genômica é possível supôr que essa bactéria degrade somente etanol. Assim como no genoma de *H. seropedicae* SmR1 e *H. autotrophicum* IAM 14942, todas as enzimas do ciclo do ácido cítrico estão presentes, juntamente com os genes relacionados com a NADH desidrogenase, succinato desidrogenase, citrocromo C redutase e oxidase, compexos cbb3 e bd e com a ATP-sintase.

Também não são conhecidos os aminoácidos que *H. chlorophenolicum* CPW301 é capaz de degradar, mas a análise *in silico* presume que não sejam muitos (ver subtópico 6.4.3). Embora não apresente a arginase/agmatinase/ formimionoglutamate hidrolase (EC 3.5.3.1), o que bloqueia o ciclo da ureia, a arginina poderia ser degradada a putrescina via agmatinase (EC 3.5.3.11), assim como sugerido para *H. seropedicae* SmR1 (PEDROSA *et al*., 2011). Apesar disso, a sequência genômica dessa bactéria apresenta a urease e o transportador ABC para esse substrato, o que indica que esse composto pode ser degradado.

Conforme a descrição da espécie, *H. chlorophenolicum* CPW301 não possui genes *nif* e, como os demais *Herbaspirillum* não diazotróficos, seria também incapaz de reduzir nitrato a nitrito (IM *et al*., 2004), o que está em concordância com o genoma. Como ocorre em *Herbaspirillum*, essa bactéria apresenta os genes *nasAB* e *nirBD*, mas não os genes *nar*.

Embora *H. seropedicae* SmR1 também seja capaz de degradar fenol, foi observado que as duas bactérias seguem vias distintas para a degradação desse composto a partir do catecol, de modo que *H. seropedicae* SmR1 degrada fenol a succinil-CoA pela via da clivagem *meta* (PEDROSA *et al*., 2011) e *H. chlorophenolicum* CPW301 degrada fenol a piruvato ou acetil-CoA pela via da clivagem *orto*.

No genoma parcial de *H. chlorophenolicum* CPW301 foram encontrados genes envolvidos com a degradação de salicilato, passando por gentisato, que é considerada uma das vias mais importantes para a clivagem do anel de compostos aromáticos (CHAO & ZHOU, 2014). O salicilato apresenta diversos efeitos em procariotos: pode afetar a resistência ou a susceptibilidade de bactérias a antibióticos; pode aumentar a captação de ferro, funcionando como sideróforo; pode afetar a produção de fatores de virulência; e pode desfavorecer o patógeno bacteriano na colonização de plantas hospedeiras, pois ele está envolvido com a resistência sistêmica adquirida em plantas (PRICE *et al*., 2000). Até o presente momento, nenhum estudo sobre o efeito de salicilato em *H. chlorophenolicum* CPW301 foi realizado, mas a presença desses genes leva a duas hipóteses: 1- o salicilato é um intermediário da degradação de naftaleno, embora genes para a degradação de naftaleno não tenham sido encontrados; ou 2- ele pode aumentar a suscetibilidade a antibióticos produzidos por outros organismos, ou pode desfavorecer uma possível colonização de plantas e por isso deve ser degradado. A hipótese de a bactéria estar associada a plantas não é descartada, principalmente por apresentar o T3SS.

Na degradação do 4-clorofenol, a enzima 4-clorofenol-2-monoxigenase (EC 1.14.13.20) não foi encontrada. Porém ARORA & BAE (2014) se referem a essa enzima do modo genérico apenas como monoxigenase (EC 1.14.13.-), o que leva a acreditar que a reação possa ser realizada por outra enzima da mesma classe. Do mesmo modo, no genoma parcial de *H. chlorophenolicum* CPW301 foram encontradas as enzimas: 2-octaprenil-3-metil-6-metoxi-1,4-benzoquinol hidroxilase

(EC 1.14.13.-), nitrilotriacetato monoxigenase componentes A e B (EC 1.14.13.-) e vanilato O-demetilase oxidoredutase (EC 1.14.13.-). É provável que a bactéria utilize uma dessas enzimas para realizar essa primeira reação.

A enzima catecol-2,3-dioxigenase (EC 1.13.11.2), responsável por gerar 5C2HMS, foi encontrada nas estirpes GW103 e RV1423 de *Herbaspirillum*, o que sugere que a degradação desse composto não seja exclusiva de *H. chlorophenolicum* CPW301 dentro do gênero. A filogenia da enzima catecol-2,3-dioxigenase de *H. chlorophenolicum* CPW301 mostra maior identidade com sua homóloga em *Herbaspirillum* sp. RV1432 (isolada de água contaminada com hidrocarbonetos – JAREGUI *et al*., 2014) e com a de *Xenophilus azovorans* (isolada de meio enriquecido com 1-(4'-carboxifenilazo)-2-naftol, que também é um agente contaminante – BLÜMEL *et al*., 2001), enquanto a catecol-2,3-dioxigenase de *Herbaspirillum sp* GW103 (isolada da rizosfera de *Phragmites australis* – LEE *et al*., 2012) parece ter uma história evolutiva diferente (FIGURA 6.17). Isso seria justificado pelos diferentes ambientes em que *H. chlorophenolicum* CPW301 e *Herbaspirillum sp* GW103 foram encontradas e, por consequência, poderiam ter adquirido essa enzima de diferentes organismos por HGT. Arora & Bae (2014) também mencionam que a via de degradação do 4-clorofenol pode encerrar no 5C2HMS. Isso justificaria a ausência da enzima subsequente (EC 1.97.1.-).

Assim como em *H. autotrophicum* IAM 14942 e nos demais *Herbaspirillum* spp., o genoma de *H. chlorophenolicum* CPW301 apresenta os genes para a produção de poli(3-hidroxi-alcanoatos). Foram também encontrados dois genes que codificam para RuBisCOs-*like*, aparentemente, não fixadoras de carbono, que também poderiam estar envolvidas no recuperação de metionina (TABITA *et al*., 2007).

Embora não tenha sido encontrada associada com plantas, essa bactéria apresenta em sua sequência genômica o T3SS, que até então só havia sido encontrado em *Herbaspirillum* spp. associadas com plantas (STRAUB *et al*., 2013). Isso pode indicar que *H. chlorophenolicum* CPW301 descende de uma bactéria associativa, que pode ter se adaptado a um novo modo de vida, provavelmente pela modificação ambiental causada pelo homem. No entanto, não é descartada a hipótese dessa bactéria colonizar pequenas plantas resistentes à poluição, que também funcionariam como potenciais biorremiadores, como por exemplo, algumas gramíneas (SICILIANO & GERMIDA, 1997). Além disso, bactérias e fungos podem

realizar a rizodegradação, que consiste em agir na rizosfera, de forma a eliminar substâncias nocivas à planta e permitir seu desenvolvimento, enquanto aproveitam as condições geoquímicas proporcionadas pela rizosfera da planta (SUSARLA *et al*., 2002).

Na anotação genômica de *H. rhizosphaerae* UMS-37 (FIGURA 6.22) também não foi encontrado o gene que codifica para a PFK-1 (EC 2.7.1.11). Porém, assim como em *H. chlorophenolicum* CPW31 e *H. seropedicae* SmR1, a rota para o metabolismo de D-glucose e D-frutose pode desviar para a via das pentoses fosfato. De fato, a espécie é descrita como capaz de metabolizar esses açúcares (JUNG *et al.*, 2007). Essa bactéria também é descrita como capaz de metabolizar D-manose e D-galactose (JUNG *et al.*, 2007), mas genes relacionados com a conversão de D-manose a D-manose-6P e o gene que codifica para a D-galactose 1-desidrogenase (EC 1.1.1.48) não foram encontrados. A bactéria ainda é descrita como incapaz de crescer utilizando L-ramnose como única fonte de carbono e energia (JUNG *et al.*, 2007) e genes para a degradação de L-ramnose não foram encontrados. Além disso, ela é descrita como capaz de crescer unicamente na presença de sorbitol (JUNG *et al.*, 2007), de modo que o gene que codifica para a L-iditol 2-desidrogenase, relacionada com a degradação desse composto, foi encontrado (EC 1.1.1.14).

Para a produção de trealose foram encontrados tanto o gene *ostB* quanto os genes *treXYZ*, assim como observado na maioria das *Herbaspirillum* spp. (STRAUB *et al*., 2013). Também, como descrito, a bactéria não apresenta o gene relacionado com a beta-galactosidase (EC 3.2.1.23) (JUNG *et al.*, 2007).

Embora apresente os genes que codificam para as enzimas NADH desidrogenase, succinato desidrogenase, citrocromo C redutase e oxidase, complexo bd e ATPase, não foram encontrados genes que codificam para o complexo cbb3. Esses genes estão presentes nas demais estirpes de *Herbaspirillum* spp., o que torna *H. rhizosphaerae* UMS-37 diferente de todo o grupo com relação à cadeia oxidativa.

As análises *in silico* levam a supôr que essa bactéria degrade um conjunto mínimo de aminoácidos (L-alanina, L-aspartato, L-asparagina, L-glutamato, L-glutamina, L-glicina, L-serina, L-cisteína, e L-arginina) que parecem comuns às *Herbaspirillum* spp. A bactéria poderia efetuar o ciclo da ureia, de maneira que genes relacionados com esse ciclo e com a enzima urease foram encontrados. Do

mesmo modo que *H. seropedicae* SmR1 e *H. clorophenolicum* CPW301, *H. rhizosphaerae* UMS-37 poderia também degradar arginina via agmatinase (EC 3.5.3.11).

Como já havia sido descrito (JUNG *et al.*, 2007), os genes que codificam para a nitrogenase não foram encontrados no genoma parcial dessa bactéria. Da mesma forma que os demais *Herbaspirillum* não fixadores de nitrogênio, o genoma dessa bactéria apresenta os genes *nasAB*, *nirBD*, mas não os genes *nar*.

Assim como em *H. chlorophenolicum* CPW301, foram encontrados genes relacionados com a degradação de 4-clorofenol. A primeira reação, que converte 4-clorofenol para 4-clorocatecol, poderia ser realizada por alguma monoxigenase de função genérica (EC 1.14.13.-), como proposto para *H. chlorophenolicum* CPW301. Dentre elas estão as enzimas: 2-octaprenil-6-metoxifenol hidroxilase (EC 1.14.13.-); 2-octaprenil-3-metil-6-metoxi-1,4-benzoquinol hidroxilase (EC 1.14.13.-); nitrilotriacetato monoxigenase componentes A e B (EC 1.14.13.-); vanilato O-demetilase oxidoredutase (EC 1.14.13.-); e provável proteína vanilato O-demetilase oxigenase subunidade oxidoredutase (EC 1.14.13.-).

Diferentemente do que foi observado em *H. chlorophenolicum* CPW301, a partir de 4-clorocatecol a via seguiria pela clivagem *orto* através do gene que codifica a enzima catecol-1,2-dioxigenase (EC 1.13.11.1) (ARORA & BAE, 2014). Porém, na etapa seguinte da via, ao invés do 3-cloromuconato ser convertido a cis-dienelactona, conforme a literatura (ARORA & BAE, 2014), é proposto que ele seja convertido a uma toxina chamada protoanemonina pela enzima muconato cicloisomerase (EC 5.5.1.1). Essa toxina é produzida por plantas da família *Ranunculaceae* e é responsável por causar ardor na pele e irritação nas membranas da camada inferior da pele em humanos, mas também foi descrita como capaz de ser produzida por *Pseudomonas* sp. B13 e *Pseudomonas reinekei* MT1, além de funcionar como uma substância antimicrobiana e antimicótica (BOBADILLA FAZZINI *et al*., 2013; UÇMAK *et al*., 2014). Provavelmente *H. rhizospharae* UMS-37 utilize esses genes para produzir protoanemonina e então utilizá-la para inibir o crescimento de outros microrganismos. As rizobactérias costumam utilizar antibióticos para colonizar esse ambiente altamente competitivo (BERG *et al*., 2005).

Também foi verificado que *H. seropedicae* estirpes SmR1, Os34, Os45, AU14040, *H. rubrisubalbicans* M1, *H. huttiense* subsp. *putei* 7-2, *H. frisingenese* GSF30 e as estirpes GW103, YR522 e CF444 apresentam o gene que codifica para

muconato cicloisomerase, que são homólogos entre si. No entanto, a maioria deles (com exceção do próprio *H. rhizosphaerae* UMS-37, das estirpes Os34 e Os45 de *H. seropedicae* e da estirpe YR522) apresenta uma segunda muconato cicloisomerase evolutivamente divergente da primeira (FIGURA 6.21). Isso é diferente do que foi observado para outros organismos (por exemplo, *Azohydromonas australica*, *Collimonas arenae* e *Burkholderia terrae*), que apresentam duas muconato cicloisomerases muito parecidas entre si. A maior identidade (92%) entre a muconato cicloisomerase de *H. rhizosphaerae* UMS-37 com uma das homólogas de *Herbaspirillum sp* CF444 pode ser entendida pela proximidade entre esses dois membros do gênero. A presença do gene que codifica para essa enzima em grupos distintos de *Herbaspirillum* também leva a acreditar que, pelo menos, uma das cópias do gene seja inerente ao ancestral comum do gênero.

H. rhizospharae UMS-37, assim como *H. chlorophenolicum* CPW301, poderia converter salicilato a catecol. As hipóteses levantadas para *H. chlorophenolicum* CPW301, para a presença do gene que codifica a enzima responsável por essa reação, são também válidas para *H. rhizosphaerae* UMS-37, mas faz mais sentido acreditar na hipótese que a bactéria degrade salicilato por esse aumentar a suscetibilidade a antibióticos produzidos por outros organismos. Porém, diferentemente de *H. chlorophenolicum* CPW301, o restante da via, que vai da degradação de catecol a succinil-CoA, é comum a *H. seropedicae* SmR1.

O gene que codifica para a penicilina amidase (EC 3.5.1.11), que converte penicilina em ácido 6-aminopenicilânico (6-APA), está ausente no genoma de *H. seropedicae* SmR1. Nesse último, é encontrado apenas o gene que codifica para a beta-lactamase classe A (EC 3.5.2.6), enquanto em *H. rhizosphaerae* UMS-37 apresenta ambos. É importante ressaltar que o 6-APA pode ser usado para a produção de vários antibióticos análogos à penicilina (DESHPANDE *et al*., 2004).

H. rhizosphaerae UMS-37 também apresenta os sistemas de exportação de proteínas Sec e Tat, embora não apresente sistemas de secreção Sec e Tat-dependentes. O único sistema de secreção encontrado foi o T1SS, que é comum a todos os *Herbaspirillum*.

7.3 Comparação genômica

A variação do número de genes em 19 genomas de *Herbaspirillum* (entre 4.031 e 5.587 genes – FIGURA 6.23) é tão grande quanto o observado para 63 genomas de *Escherichia* e *Shigella* (entre 4.061 e 5.803 genes – LUKJANCENKO *et al*., 2010). Essa diferença corresponde, em grande parte, a genes únicos de cada genoma analisado, principalmente de algumas espécies como *H. autotrophicum* e *H. massiliense* que se mostraram mais divergentes em relação às demais. Esses genes, que correspondem a 15% da média de genes por genoma de *Herbaspirillum*, seriam responsáveis pela adaptação a ambientes específicos que esses organismos colonizam.

Em comparação com à espécie *E. coli*, a qual apresenta 100 novos genes a cada genoma sequenciado (LAND *et al*., 2015), *H. seropedicae* mostra maior divergência entre suas estirpes (203 genes novos a cada genoma sequenciado). Isso se deve provavelmente ao fato das estirpes Os34 e Os45 serem distintas das demais, enquanto as estirpes SmR1, BR11417 e BR11305 apresentam maior homologia.

A matriz BLAST (FIGURA 6.24) reforçou, de certa forma, o que é observado na filogenia do gene 16S rRNA de *Herbaspirillum*: as estirpes de *H. seropedicae* não se relacionam com estirpes Os34 e Os45, que se relacionam com *H. rubrisubalbicans*; *H. huttiense* subsp. *putei* e *Herbaspirillum* sp. GW103 são relacionados entre si; a estirpe CF444 e os membros do filogrupo 2 apresentam alguma relação; e *H. massiliense* JC206 é divergente dos demais *Herbaspirillum* (MONTEIRO *et al*., 2014).

Isso levou a acreditar que, de fato, as estirpes Os34 e Os45 pertenceriam à espécie *H. rubrisubalbicans*, que *Herbaspirillum* sp. GW103 poderia ser uma estirpe de *H. huttiense*, que *H. massiliense* JC206 poderia não pertencer ao gênero *Herbaspirillum* e que as estirpes YR522 e CF444 fariam parte de duas novas espécies.

De maneira geral, os *Herbaspirillum* se parecem entre si quanto à homologia dos conjuntos proteômicos, pois compartilham mais de 50% das proteínas (com exceção de *H. massiliense* JC206). Isso pode ser comparado, por exemplo, ao gênero *Lactobacillus*, no qual as diferentes espécies compartilham, em geral, menos que 20% do conteúdo proteômico, e *Bifidobacterium* que, embora não seja tão

divergente quanto *Lactobacillus*, em geral não apresenta mais de 50% do conjunto proteômico compartilhado entre as espécies (LUKJANCENKO *et al*., 2012).

O *core* genoma de *Herbaspirillum,* composto por 1.412 genes (FIGURA 6.25), é pouco menor que o *core* genoma obtido para 53 genomas de *E.coli* (1.472 genes) e representa uma porcentagem menor de genes compartilhados entre os organismos (29% para *Herbaspirillum* contra 36% para *E. coli* – calculado com base nas informações de LUKJANCENKO *et al*., 2010). Isso é esperado, pois a relação entre espécies de um gênero deve ser menor que a relação entre as estirpes de uma mesma espécie. Lukjancenko e colaboradores (2010) também afirmam que o *core* genoma de *Escherichia* decresce para 993 genes (20% da média de genes, calculado com base nas informações citadas pelos autores) quando adicionados genomas de *Shigella*, o que demonstra que *Herbaspirillum* é um grupo mais coeso que *Escherichia/Shigella*. Além disso, o *core* genoma de *Herbaspirillum* é proporcionalmente similar ao obtido para 21 genomas de *Lactobacillus* (~500 genes, ~25% da média de genes) e 26 genomas de *Streptococcus* (~600 genes, ~25% da média de genes segundo os autores e 30% segundo recálculo – LEFÉBURE & STANHOPE, 2007).

O pangenoma de *Herbaspirillum*, que corresponde a 19.945 genes (~4 vezes o número médio de genes por genoma – FIGURA 6.25), embora pareça grande, é proporcionalmente menor que o obtido para os 21 genomas de *Lactobacillus* (~13.000 genes; ~6,5 vezes a média de genes), 22 genomas de *Streptococcus* (~10.000 genes, ~5 vezes a média de genes), 19 genomas de *Bifidobacterium* (~7.000 genes, ~4,5 vezes a média de genes) (LUKJANCENKO *et al.*, 2012). Porém é proporcionalmente maior do que o obtido para 63 genomas de *Escherichia* e *Shigella* (13.279 genes, 2,7 vezes a média de genes) (LUKJANCENKO *et al.*, 2010). No entanto, um trabalho anterior, realizado com apenas 20 genomas de *E. coli*, demonstrou que o pangenoma de *E. coli* corresponde a 17.831 genes e o *core* genoma a 1.976 genes (TOUCHON *et al*., 2009). Um trabalho posterior, realizado com 186 genomas de *E. coli*, chegou a um *core* genoma de 3.000 genes. Dessa forma, a variação entre os pan e o *core* genomas pode estar relacionada com diferenças metodológicas.

A dinâmica do pangenoma em *Herbaspirillum* pode ser explicada universalmente: genomas similares, quando adicionados à análise, pouco acrescentam ao pangenoma, enquanto genomas divergentes promovem o aumento

do pangenoma (LUKJANCENKO *et al.*, 2010). Assim, o aumento do pangenoma de *Herbaspirillum* se deve em grande parte à adição dos conjuntos proteômicos de *H. autotrophicum* IAM 14942 e *H. massiliense* JC206.

Em *Herbaspirillum*, foi observado o aumento da proporção de genes nas categorias COG para metabolismo de carboidratos, metabolismo de íons inorgânicos e transcrição no pangenoma em relação ao *core* genoma (FIGURA 6.26). O aumento das classes metabolismo de carboidratos e transcrição também foi observado nos pangenomas de *Lactobacillus* e *Bifidobacterium* (LUKJANCENKO *et al.*, 2012), o que pode indicar que essas classes de genes são as principais responsáveis pela adaptação das bactérias aos ambientes que colonizam. O aumento da classe metabolismo de carboidratos pode estar relacionado com o fato de os açúcares serem a principal fonte de energia para a célula, bem como o aumento da classe para transcrição pode estar relacionada com a regulação transcricional dos genes que necessitam ser expressos em diferentes ambientes. Em *Pseudomonas aeruginosa*, essas categorias de genes estão enriquecidas no *core* genoma*,* ao invés de fazerem parte do genoma acessório, mas nesse caso a análise foi realizada apenas com estirpes de uma mesma espécie (VALOT *et al*., 2015).

A árvore do pangenoma (FIGURA 6.27) representa de uma maneira mais clara a relação entre os organismos mostrada na matriz BLAST. Essa abordagem foi capaz de separar as espécies de *Herbaspirillum* em dois filogrupos distintos, com a ressalva que *Herbaspirillum* sp. YR522 ficou posicionado no filogrupo 1, ao contrário do que é observado na análise realizada com o gene 16S rRNA (MONTEIRO *et al*., 2014). Porém, nas duas análises as estirpes Os34 e Os45 de *H. seropedicae* foram agrupadas com *H. rubrisubalbicans* (MONTEIRO *et al*., 2014). Da mesma forma, estirpes de *E. coli* foram separadas de espécies de *Shigella* usando a mesma abordagem, embora *E. coli* tenha apresentado maior relação com as espécies de *Shigella* do que com as demais *Escherichia* (LUKJANCENKO *et al.*, 2010).

Pouco mais de 100 genes diferenciam os *core* genomas dos dois filogrupos de *Herbaspirillum*, o que demonstra forte relação entre eles. Dentre os genes exclusivos de cada filogrupo, os genes para a biossíntese de EPS, encontrados no filogrupo 1 (TABELA 6.13), chamam a atenção por desempenhar um importante papel na colonização da planta hospedeira (BALSANELLI *et al*., 2014).

Os genes relacionados com o sistema *pili* tipo IV (TABELA 6.13), embora presentes em ambos os filogrupos, não aparecem no *core* genoma, pois apresentam identidade inferior a 50% entre organismos de filogrupos distintos. Com isso, é difícil dizer se eles apresentam uma origem comum. O sistema *pili* tipo IV está relacionado com a aderência a células eucarióticas durante a patogênese, formação de biofilme, motilidade, secreção de proteínas e captação de DNA (FRIEDRICH *et al*., 2014). Straub e colaboradores (2013) já haviam descrito a presença de um pequeno conjunto de genes que codificam para o *pili* tipo IV nos genomas de *Herbaspirillum*, quando comparados a *H. seropedicae* SmR1. Os autores também descreveram a semelhança entre o agrupamento de genes *pili* tipo IV de *H. seropedicae* SmR1 e os encontrados em *Herbaspirilum* sp. CF444 (que pertence ao filogrupo 2), mas não mencionaram a presença de *pili* tipo IV em *H. lusitanum* P6-12 (também pertencente ao filogrupo 2).

A classificação de *Herbaspirillum* baseada no *core* genoma mostrou que é possível separar esse gênero de outros gêneros relacionados (FIGURA 6.29), visto que os gêneros *Collimonas* e *Oxalobacter* aparecem posicionados junto aos *Herbaspirillum* em árvores do gene 16S rRNA (JUNG *et al*., 2007; ANANDHAM *et al*., 2013). Com relação a *Oxalobacter*, os genomas das duas estirpes analisadas apresentam menos genes que a intersecção dos *core* genomas dos dois filogrupos de *Herbaspirillum*, de modo que seria impossível posicioná-las dentro desse gênero.

H. massiliense JC206 também foi separado dos demais *Herbaspirillum* a ponto de sugerir sua reclassificação (FIGURA 6.29). Monteiro e colaboradores (2014) já haviam observado a relação de *H. massiliense* com *Noviherbaspirillum maltae*, o que poderia indicar que essa espécie pertence ao gênero *Noviherbaspirillum*. Um aspecto interessante desse organismo é ele apresentar o gene que codifica para a PFK-1, ausente em todas as estirpes de *Herbaspirillim* spp. analisadas.

Na análise de classificação, os *core* genomas não foram equilibrados entre os dois filogrupos, de modo que a linha que os separa (se estendida para todo o gráfico) não passaria pelo ponto zero em ambos os eixos no gráfico. Entretanto, é possível observar que organismos de gêneros diferentes de *Herbaspirillum* estão posicionados em uma linha reta, a qual cortaria o ponto zero do gráfico (FIGURA 6.29). Isso indica que organismos diferentes de *Herbaspirillum* podem também ser detectados pelo seu posicionamento nessa linha (não apresentam tendência a pertencer a nenhum dos dois filogrupos).

O funcionamento dessa abordagem depende da estabilização do *core* genoma, pois a adição de novos genomas poderia reduzir demais o ponto mínimo de delimitação dos *Herbaspirillum*. Também existe dúvida se somente a área que delimita os *Herbaspirillum* e a linha que passa pelo ponto zero seriam preenchidas, ou se seria possível que organismos pontuassem outras áreas do gráfico. A adição de novos genomas, por exemplo, genomas de *Noviherbaspirillum* e *Paraherbaspirillum*, que são gêneros mais próximos de *Herbaspirillum*, poderiam dizer se esse método de classificação é viável.

O BLAST atlas (FIGURA 6.30) proporciona uma nova perspectiva do que é observado na linha 1 da matriz BLAST, onde o genoma de *H. seropedicae* SmR1 é comparado a todos os outros. É possível ver graficamente que essa referência apresenta maior número de regiões homólogas às estirpes BR11335 e BR11417, mas que também se relaciona com as estirpes de *H. rubrisubalbicans*, *H. huttiense* subsp. *putei*, *Herbaspirillum* sp. GW103 e *H. frisingense* GSF30. É notável também a queda da quantidade de regiões homólogas quando se salta desses organismos para *H. chlorophenolicum* CPW301, *Herbaspirillum* sp. YR522 e para as estirpes de *Herbaspirillum* spp. do filogrupo 2.

Muitas das regiões únicas de *H. seropedicae* SmR1, ou compartilhadas apenas com as estirpes BR11335 e BR11417, estão acompanhadas de queda de conteúdo G+C. Dentre elas, Pedrosa e colaboradores (2011) já haviam descrito uma região próxima à origem de replicação, posterior aos genes *dnaA*, *dnaN* e *gyrB*, com conteúdo G+C de 52% e 16,6 Kb, que seria provavelmente uma região de transferência horizontal de genes e que foi também evidenciada com o BLAST atlas. *H. rubrisubalbicans* M1 apresenta uma região de transferência horizontal no mesmo ponto, mas os genes são diferentes dos adquiridos por *H. seropedicae* SmR1 (SOUZA *et al.*, *in prep*), o que poderia indicar uma região de recombinação para os *Herbaspirillum*.

Além disso, foram descritas 18 prováveis regiões de transferência horizontal para *H. seropedicae* SmR1 (PEDROSA *et al*., 2011), mas a simples análise visual mostra que *H. seropedicae* SmR1 apresenta maior número de regiões candidatas a terem sido obtidas horizontalmente (em torno de 32). Algumas delas não apresentam variação de conteúdo G+C e, portanto, podem não ter sido detectadas por ferramentas automáticas de busca por regiões transferidas horizontalmente.

Com relação a agrupamentos gênicos específicos, Straub e colaboradores (2013), em sua comparação genômica de *Herbaspirillum*, já haviam mencionado a presença de genes *nif* nas estirpes SmR1, Os34, Os45 de *H. seropedicae*, e GSF30 de *H. frisingense*, organizados de forma idêntica e com identidade de 96% em nível de proteínas. As estirpes BR11335 e BR11417 de *H. seropedicae*, bem como as estirpes M1 e BR11502 de *H. rubrisubalbicans*, podem ser incluídas nesse conjunto (FIGURA 6.31B). Também já havia sido observado que *H. seropedicae* AU14040 não possuía esse agrupamento de genes e que esse é um dos pontos que diferencia essa estirpe das demais (FAORO, Informação verbal).

Straub e colaboradores (2013) também já haviam comparado os sistemas T3SS, presente em *H. seropedicae* estirpes SmR1, Os34 e Os45, *H. rubrisubalbicans* M1, estirpes YR522 e CF444, e o sistema T6SS, presente na maioria dos genomas analisados (ausente somente em *H. massiliense* JC206 e *Herbaspirillum* sp. CF444). As análises realizadas com maior número de genomas sugerem que o T6SS seja exclusivo de espécies do filogrupo 1 (FIGURA 6.31C e D). Nesse contexto, a presença desse agrupamento de genes reforça o fato de *Herbaspirillum* sp. YR522 pertencer a esse filogrupo. A manutenção desses genes por *H. seropedicae* AU14040, que também já havia sido observada (FAORO, Informação verbal), pode sugerir que ele desempenhe uma função importante na interação com o hospedeiro humano. Porém, *H. huttiense* subsp. *putei*, isolado de amostras de água, também apresenta esse agrupamento.

O T3SS, embora esteja envolvido na secreção de proteínas efetoras responsáveis pela interação da bactéria com a planta (SCHMIDT *et al*., 2012), está presente em *H. chlorophenolicum* CPW301 (obtido de amostra ambiental), mas ausente em *H. frisingense* GSF30 (endofítico – FIGURA 6.31C). Da mesma forma, foi observada a ausência desses genes em outras bactérias endofíticas, tais como *Azoarcus sp*. BH72, *Klebsiella pneumoniae* 342, *Azospirillum sp*. B510 e *Glucanocetobacter diazotrophicus* PAI5 (STRAUB *et al*., 2013). Por outro lado, a presença desses genes em *H. hiltneri* N3 e *Herbaspirillum* sp. CF444, isolados de rizosfera, pode indicar que existam *Herbaspirillum* endofíticos presentes também no filogrupo 2.

Outro aspecto importante do T3SS é a sua conservação entre a estirpe M1 de *H. rubrisubalbicans* e as estirpes Os34 e Os45 de *H. seropedicae*, ao contrário do que é observado entre essas estirpes e *H. seropedicae* SmR1 (FIGURA 6.33). Isso é

mais um indício de que as estirpes Os34 e Os45 sejam estirpes de *H. rubrisubalbicans*.

A região do fago I de *H. seropedicae* SmR1 (FIGURA 6.31A) apresenta 25,3 Kb e conteúdo G+C de 66,4%, enquanto a região do fago II (FIGURA 6.31C) apresenta 38,5 Kb e 58,1% de conteúdo G+C (PEDROSA *et al*., 2011). No entanto, embora menor e com maior conteúdo G+C, a região do fago I parece ter sido uma aquisição recente por parte de *H. seropedicae* SmR1, enquanto a região do fago II parece ser uma aquisição antiga. Em *H. rubrisubalbicans* M1, a análise da região do fago II através da ferramenta PHAST ("*PHAge Search Tool"* - ZHOU *et al*., 2011) revelou similaridade com uma região de fagos de *Burkholderia*. Essa bactéria apresenta mais duas regiões relacionadas com fago, mas que não possuem homologia com as demais estirpes de *Herbaspirillum* spp. (SOUZA *et al*., *in prep*).

Além da análise desses agrupamentos, a presença de genes responsáveis pela biossíntese de celulose foi descrita como um dos elementos diferenciais entre os genomas de *H. seropedicae* SmR1 e *H. rubrisubalbicans* M1 (MONTEIRO *et al*., 2012). É interessante que esses genes estejam ausentes em *H. seropedicae* (FIGURA 6.36), com exceção das estirpes Os34 e Os45, pois a produção de celulose foi descrita como importante para o processo de colonização da planta por parte de *H. rubrisubalbicans* (MONTEIRO *et al*., 2012). Por outro lado, esses genes estão presentes em *H. huttiense* subsp. *putei*, isolado de amostras de água, de forma que é difícil estabelecer uma relação entre o conteúdo gênico e o hábito endofítico.

Pedrosa e colaboradores (2011) sugeriram que *H. seropedicae* houvesse adquirido o agrupamento de genes *nif* por transferência horizontal de genes, fato sustentado pela presença de uma transposase remanescente na região à montante desse agrupamento, e pela sua homologia com o agrupamento encontrado em *Burkholderia vietnamiensis* G4. Dentro da análise genômica comparativa, é possível supôr que o ancestral comum das três espécies de *Herbaspirillum* fixadoras de nitrogênio atmosférico o tenha recebido (FIGURA 6.37).

É importante ressaltar que a hipótese do ancestral comum para esse agrupamento de genes se baseia na topologia da árvore da ausência/presença de genes. Isso significa que ela está dentro do contexto da dinâmica de todos os genes ao longo da evolução de *Herbaspirillum* e não apenas do agrupamento observado, dando ênfase à herança vertical de genes e minimizando a HGT.

Essa última pode ser usada para explicar a origem dos genes quando se tem outro tipo de informação além do que se observa na árvore. Por exemplo, ela pode explicar a origem da região do fago II de *H. frisingense* GSF30, pois mesmo que a informação da árvore leve a concluir que ela tenha um ancestral comum com as regiões de fago II de outras estirpes de *Herbaspirillum* spp., uma análise individual da estrutura desse agrupamento mostra que não faz sentido acreditar que ele tenha, de fato, a mesma origem que os demais (FIGURA 6.35).

Pedrosa e colaboradores (2011) sugeriram também que o T3SS não tenha sido adquirido recentemente, ou talvez tivesse sido adquirido de uma bactéria relacionada evolutivamente. A ampla distribuição desse agrupamento de genes em grupos distintos de *Herbaspirillum* leva a acreditar que a primeira hipótese esteja correta. Dessa forma, o ancestral de *Herbaspirillum* já teria esse agrupamento de genes (FIGURA 6.37) e talvez já estivesse de alguma forma associado a plantas. Posteriormente, esses organismos teriam se diversificado e colonizado outros ambientes.

Análises de vias metabólicas específicas mostraram que a degradação de compostos fenólicos por *Herbaspirillum* spp. (FIGURA 6.38), seguindo pela via da clivagem *meta*, parece universal o suficiente para afirmar que é tão antiga quanto o ancestral do gênero. A segunda via, pela clivagem *orto,* surgiu dentro do filogrupo 1, no ancestral comum a esse filogrupo, e foi perdida na maior parte dos organismos. É interessante destacar que, se levado em consideração que a degradação pela via da clivagem *orto* tem como ancestral comum o ancestral de todo o filogrupo 1, isso significa que os genes para as duas vias coexistiram, mas apenas uma delas se manteve para a maioria dos organismos do filogrupo 1.

Outras análises envolvendo proteínas específicas, mostraram que as proteínas do tipo RLP (RuBisCO-*like protein*) aparentemente apresentam origens distintas. A primeira delas, por estar distribuída em ambos os filogrupos (FIGURA 6.40), parece ter uma origem mais antiga e também faria parte do conjunto proteômico ancestral. A segunda delas ocorre somente no filogrupo 2 e pode ter sido adquirida somente por esse filogrupo, mas também poderia ter sido perdida pelo ancestral do filogrupo 1. Os organismos que apresentam genes que codificam essas proteínas não apresentam o gene que codifica para a fosforribuloquinase, o que impede a fixação de carbono, e talvez essas proteínas estejam envolvidas no metabolismo de enxofre (STRAUB *et al*., 2013). Esse tipo de proteína também ocorre em *Bacillus* como parte

da via de recuperação da metionina, de modo que foi sugerido que ele possa ter sido adquirido horizontalmente a partir do genoma de planta (SEKOWSKA & DANCHIN, 2002).

A RuBisCO verdadeira, responsável pela fixação de carbono provavelmente foi adquirida por transferência horizontal e foi também encontrada na estirpe TSA66 de *Herbaspirillum*. Embora essa estirpe pudesse abrir o leque de organismos relacionados com *H. autotrophicum*, o qual parece ter características metabólicas distintas das demais *Herbaspirillum* spp., uma análise nas informações presentes no NCBI mostra que ela provavelmete será classificada em um novo gênero chamado *Denitrispirillum*, espécie *Denitrispirillum autotrophicum* (ISHII *et al*., não publicado).

As nuâncias entre os genomas analisados, tais como a ausência do T3SS em *H. frisingense* GSF30 (endofítico) e sua presença em *H. chlorophenolicum* CPW301 (ambiental), bem como a ausência dos genes *wss* em *H. seropedicae* SmR1 (endofítico), mas sua presença em *H. huttiense* subsp. *putei* (ambiental), levam a acreditar que espécies ambientais de *Herbaspirillum* sejam potenciais organismos associativos, ao mesmo tempo que *Herbaspirillum* associativos tenham potencial para colonizar outros ambientes.

Olivares e colaboradores (1996) verificaram que estirpes de *Herbaspirillum* apresentam baixa sobrevivência no solo, sem contato com a planta. Por outro lado, esses organismos poderiam ser pioneiros na colonização de depósitos vulcânicos devido às suas características metabólicas, com alguns isolados apresentando o gene *nifH* (LU *et al*., 2008). Um dos fatores que poderia auxiliar os *Herbaspirillum* a colonizar novos ambientes seriam os genes *ppk* e *ppx*, associados com resistência a estresse (SEUFFERHELD *et al*., 2008)**.**

Com relação aos genes que, de certa forma, separam os *Herbaspirillum* associativos daqueles que são ambientais, estão genes relacionados a transporte, amidase e liases, além de reguladores de transcrição. É difícil colocar esses genes num contexto associativo, porém o gene que codifica a proteína LrgA efetora de mureína hidrolase parece de suma importância. As mureína hidrolases desempenham uma série de funções em bactéria, entre elas, são capazes de alargar a parede celular para que sejam montados sistemas *pili*, flagelos e sistemas de secreção (VOLLMER *et al*., 2008). A proteína LrgA, especificamente, desempenha função importante na dinâmica de aderência do biofilme (RANJIT *et al*.,

2011) e, portanto, poderia ser fundamental no processo de colonização da planta hospedeira.

Com relação às análises taxonômicas, algumas sugestões de classificação de *Herbaspirillum* spp., baseadas na análise MLSA (*Multilocus Sequence Analysis*) do *core* genoma, haviam sido propostas. Nelas, a estirpe GW103 seria classificada como *H. huttiense* e a estirpe CF444 seria classificada como *H. lusitanum* (WEISS, 2014). Entretanto, as análises taxonômicas baseadas na sequência genômica, ANI e GGDH, não sustentam essas classificações, de forma que as estirpes GW130, YR522 e CF444 são novas espécies de *Herbaspirillum* (FIGURA 6.41).

Para *Herbaspirillum*, tanto a análise ANI quanto a análise GGDH concordaram em todas as classificações realizadas, pois sempre que a ANI foi maior que 95%, a GGDH foi maior que 70%. Essa relação foi observada em *Aeromonas*, nas quais houve apenas dois casos de controvérsia, nos quais os autores foram rígidos e estabeleceram a separação dos indivíduos (COLSTON *et al*., 2014).

Análises prévias realizadas com a montagem genômica parcial da estirpe tipo Z67^{T} de *H. seropedicae*, cedida pelo NFN durante a execução deste trabalho, mostraram ANIs de 88,6% e 88,5% e GGDHs de 30% e 30% em relação às estirpes Os34 e Os45 de *Herbaspirillum*, respectivamente (FIGURA 6.41). Com isso, essas duas estirpes não podem ser classificadas como pertencentes a tal espécie. Diante disso, foi sugerido classificar as estirpes Os34 e Os45 como pertencentes à espécie *H. rubrisubalbicans* ou, pelo menos, como subespécie dessa. Enquanto este trabalho estava sendo escrito, no NCBI essas estirpes foram, de fato, reclassificadas como *H*. *rubrisubalbicans* Os34 e *H*. *rubrisubalbicans* Os45.

Outra análise prévia, realizada com o genoma da estirpe RV1423, publicado durante a execução deste trabalho, mostra que ela também não pode ser classificada em nenhuma espécie já descrita, embora possa ser posicionada entre os organismos do filogrupo 2, com os quais apresenta ANIs em torno de 87% (contra ~83% em relação aos organismos do filogrupo 1), e GGDH em torno de 28% (contra ~23% em relação aos organismos do filogrupo 1 – FIGURA 6.41).

A análise genômica comparativa de *Herbaspirillum* forneceu informações relevantes sobre a evolução desses organismos. Ela também mostrou que eles podem ser divididos em dois grupos distintos filogeneticamente, os quais poderiam ser identificados por conteúdos gênicos específicos, mas que não estão relacionados com o estilo de vida desses organismos.

8 CONCLUSÕES

As montagens híbridas realizadas com *reads* provenientes da plataforma Illumina e *contigs* de montagens Illumina, previamente obtidos, foram capazes de gerar *drafts* genômicos de alta qualidade.

A anotação genômica de *H. autotrophicum* IAM 14942 revelou a presença de genes relacionados com a fixação de carbono (RuBisCO) e obtenção de energia a partir de hidrogênio (hidrogenase). Os *Herbaspirillum* apresentam genes que codificam para RLP, mas essas proteínas aparentemente não tem relação com a RuBisCO fixadora de carbono.

A anotação genômica de *H. chlorophenolicum* CPW301 revelou a presença de genes relacionados com a degradação de 4-clorofenol, embora não tenham sido obtidos todos os genes relacionados com a degradação desse composto. A via utilizada é também encontrada em *Herbaspirillum* sp. GW103.

A anotação genômica de *H. rhizosphaerae* UMS-37 revelou genes relacionados com a degradação de fenol e produção de protoanemonina. Esses genes estão presentes na maioria das estirpes de *Herbaspirillum* spp. analisadas.

A comparação genômica de estirpes de *Herbaspirillum* spp. mostrou que o *core* genoma do gênero compreende 1.412 genes e o pangenoma compreende 19.945 genes. Essa comparação também mostrou que os *Herbaspirillum* podem ser divididos em dois filogrupos distintos.

O *core* genoma pode ser usado para separar as estirpes de *Herbaspirillum* spp. de estirpes de gêneros relacionados.

A análise de agrupamentos de genes mostrou que o ancestral dos *Herbaspirillum* fixadores de nitrogênio recebeu, além dos genes *nif*, genes para a biossíntese de celulose, os quais foram perdidos por algumas das estirpes. O T6SS pode ter sido adquirido pelo ancestral do filogrupo 1, e o T3SS provavelmente já estava presente no ancestral do gênero.

As análises de ANI e GGDH mostram que as estirpes GW103, YR522 e CF444 de *Herbaspirillum* spp. são novas espécies ainda não descritas.

REFERÊNCIAS BIBLIOGRÁFICAS

454 SEQUENCING. Disponível em: <http://www.454.com>. Acesso em: 10/06/2015.

ALIKHAN, N.-F.; PETTY, N. K.; BEN ZAKOUR, N. L.; BEATSON, S. A. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. **BMC Genomics**, v. 12, n. 1, p. 402, 2011.

ANANDHAM, R.; KIM, S.-J.; MOON, J. Y.; WEON, H.-Y.; KWON, S.-W. *Paraherbaspirillum soli* gen. nov., sp. nov. isolated from soil. **Journal of Microbiology**, v. 51, n. 2, p. 262–267, 2013.

ANDREWS, S. FastQC: a quality control tool for high throughput sequence data. 2010. Disponível em : <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.

ANI CALCULATOR. Disponível em: <http://enve-omics.ce.gatech.edu/ani/> Acesso em: 10/06/2015.

ALTSCHUL, S.F.; MADDEN, T.L.; SCHÄFFER, A. A.; *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic acids Research**, v. 25, p. 3389-3402, 1997.

APWEILER, R. UniProt: the Universal Protein knowledgebase. **Nucleic Acids Research**, v. 32, n. 90001, p. 115D–119, 2004.

ARAGNO, M.; SCHLEGEL, H.G. *Aquaspirillum autotrophicum*, a New Species of Hidrogen-Oxidazing, Facultatively Autotrophic Bacteria. **International Journal of Systematic Bacteriology**. v. 28, p. 112-116, 1978.

ARORA, P.; BAE, H. Bacterial degradation of chlorophenols and their derivatives. **Microbial Cell Factories**, v. 13, n. 1, p. 31, 2014.

AZIZ, R. K.; BARTELS, D.; BEST, A. A.; *et al*. The RAST Server: Rapid Annotations using Subsystems Technology. **BMC Genomics**, v. 9, n. 1, p. 75, 2008.

BAJERSKI, F.; GANZERT, L.; MANGELSDORF, K.; *et al*. *Herbaspirillum psychrotolerans* sp. nov., a member of the family *Oxalobacteraceae* from a glacier forefield. **International Journal of Systematic and Evolutionary Microbiology**, v. 63, n. Pt 9, p. 3197–3203, 2013.

BALDANI, J. I.; BALDANI, V. L. D.; SELDIN, L.; DÖBEREINER, J. Characterization of *Herbaspirillum seropedicae* gen. nov. sp. nov., a root associated nitrogen-fixing bacterium. **International Journal of Systematic Bacteriology**, v. 36, p. 86-93, 1986.

BALDANI, J. I.; POT, B.; KIRCHHOF, G.; *et al*. Emended description of *Herbaspirillum*; inclusion of [*Pseudomonas*] *rubrisubalbicans*, a mild plant pathogen, as *Herbaspirrillum rubrisubalbicans* comb. nov.; and classification of a group of

clinical isolates (EF Group 1) as species 3. **International Journal of Systematic Bacteriology**, v. 46, p. 802-810, 1996.

BALDANI, J. I.; BALDANI, V. L. D. History on the biological nitrogen fixation research in graminaceous plants: special emphasis on the Brazilian experience. **Anais da Academia Brasileira de Ciências**, v. 77, p. 549-579, 2005.

BALDANI, J.I.; ROUWS, L.; CRUZ, L. M.; *et al*. The Family *Oxalobacteraceae*. In: **The Prokaryotes**. USA: Springer, p. 919-974, 2014.

BALSANELLI, E.; BAURA, V. A.; PEDROSA, F. DE O.; SOUZA, E. M.; MONTEIRO, R. A. Exopolysaccharide Biosynthesis Enables Mature Biofilm Formation on Abiotic Surfaces by Herbaspirillum seropedicae. **PLoS ONE**, v. 9, n. 10, p. e110392, 2014.

BENEDUZI, A.; AMBROSINI, A.; PASSAGLIA, L. M. P. Plant growth-promoting rhizobacteria (PGPR): Their potential as antagonists and biocontrol agents. **Genetics and Molecular Biology**, v. 35, n. 4, p. 1044-1051. 2012.

BERG, G.; EBERL, L.; HARTMANN, A. The rhizosphere as a reservoir for opportunistic human pathogenic bacteria. **Environmental Microbiology**, v. 7, p. 1673-1685, 2005.

BERG, G.; SMALLA, K. Plant species and soil type cooperatively shape the structure and function of microbial communities in the rhizosphere: Plant species, soil type and rhizosphere communities. **FEMS Microbiology Ecology**, v. 68, n. 1, p. 1–13, 2009.

BERGLUND, E. C.; KIIALAINEN, A.; SYVÄNEN, A.-C. Next-generation sequencing technologies and applications for human genetic history and forensics. **Investigative Genetics**, v. 2, n. 1, p. 23, 2011.

BHARDWAJ, D.; ANSARI, M.; SAHOO, R.; TUTEJA, N. Biofertilizers function as key player in sustainable agriculture by improving soil fertility, plant tolerance and crop productivity. **Microbial Cell Factories**, v. 13, n. 1, p. 66, 2014.

BINNEWIES, T. T.; MOTRO, Y.; HALLIN, P. F.; *et al*. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. **Functional & Integrative Genomics**, v. 6, n. 3, p. 165–185, 2006.

BLAST: Basic Local Alignment Search Tool. Disponível em: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>. Acesso em: 16/09/15.

BLÜMEL, S.; BUSSE, H. J.; STOLZ, A.; KÄMPFER, P. *Xenophilus azovorans* gen. nov., sp. nov., a soil bacterium that is able to degrade azo dyes of the Orange II type. **International Journal of Systematic and Evolutionary Microbiology**, v. 51, n. Pt 5, p. 1831–1837, 2001.

BOBADILLA FAZZINI, R. A.; SKINDERSOE, M. E.; BIELECKI, P.; *et al*. Protoanemonin: a natural quorum sensing inhibitor that selectively activates iron starvation response. **Environmental Microbiology**, v. 15, n. 1, p. 111–120, 2013.

BOWIEN, B.; SCHLEGEL, H. G. Physiology and Biochemistry of Aerobic Hydrogen-Oxidizing Bacteria. **Annual Review of Microbiology**, v. 35, n. 1, p. 405–452, 1981.

BOYD D. A., CVITKOVITCH D. G., HAMILTON I. R. Sequence, expression, and function of the gene for the nonphosphorylating, NADP-dependent glyceraldehyde-3-phosphate dehydrogenase of *Streptococcus mutans*. **Journal of Bacteriology** V. 177, p. 2622–2627, 1995.

BRODER, S.; VENTER, J. C. Sequencing the Entire Genomes of Free-Living Organisms: The Foundation of Pharmacology in the New Millennium. **Annual Review of Pharmacology and Toxicology**, v. 40, n. 1, p. 97–132, 2000.

BROWN, S. D.; UTTURKAR, S. M.; KLINGEMAN, D. M.; *et al*. Twenty-One Genome Sequences from *Pseudomonas* Species and 19 Genome Sequences from Diverse Bacteria Isolated from the Rhizosphere and Endosphere of *Populus deltoides*. **Journal of Bacteriology**, v. 194, n. 21, p. 5991–5993, 2012.

CALIZ, J.; VILA, X.; MARTÍ, E.; *et al*. Impact of chlorophenols on microbiota of an unpolluted acidic soil: microbial resistence and biodegradation. **FEMS Microbiology Ecology**, v. 78, p. 150-164, 2011.

CAMPANA, S.; TACCETTI, G.; RAVENNI, N.; *et al*. Transmission of *Burkholderia cepacia* Complex: Evidence for New Epidemic Clones Infecting Cystic Fibrosis Patients in Italy. **Journal of Clinical Microbiology**, v. 43, p. 5136-5142, 2005.

CANELLAS, L. P.; BALMORI, D. M.; MÉDICI, L. O.; *et al*. A combination of humic substances and *Herbaspirillum seropedicae* inoculation enhances the growth of maize (*Zea mays* L.). **Plant and Soil**, v. 366, n. 1-2, p. 119–132, 2013.

CARDOSO, R.L.A**. Montagem Genômica da Bactéria Endofítica Diazotrófica *Herbaspirillum rubrisubalbicans* M1**. 109 f. Dissertação (Mestrado em Bioinformática) – Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, Curitiba, 2011.

CARRO, L.; RIVAS, R.; LÉON-BARRIOS, M.; *et al*. *Herbaspirillum canariense* sp. nov., *Herbaspirillum aurantiacum* sp. nov. and *Herbaspirillum soli* sp. nov., isolated from volcanic mountain soil, and emended description of the genus *Herbaspirillum*. **International Journal of Systematic and Evolutionary Microbiology,** v. 61, p. 1300-1306, 2011.

CARVER, T.; HARRIS, S. R.; BERRIMAN, M.; PARKHILL, J.; MCQUILLAN, J. A. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. **Bioinformatics**, v. 28, n. 4, p. 464–469, 2012.

CHAO, H.; ZHOU, N.-Y. Involvement of the Global Regulator GlxR in 3-Hydroxybenzoate and Gentisate Utilization by Corynebacterium glutamicum. **Applied and Environmental Microbiology**, v. 80, n. 14, p. 4215–4225, 2014.

CHEMALY, R. F.; DANTES, R.; SHAH, D. P.; *et al*. Cluster and Sporadic Cases of *Herbaspirillum* Species Infections in Patients With Cancer. **Clinical Infectious Diseases**, v. 60, n. 1, p. 48–54, 2015.

CHEN, G. **DNA Sequencing and Short Reads Assembly**. Department of Science, Systems and Models, Roskilde University, Denmark, 2008.

CHEN, J.; SU, Z.; LIU, Y.; *et al*. *Herbaspirillum* Species: A Potencial Pathogenic Bacteria Isolated from Acute Lymphoblastic Leukemia Patient. **Current Microbiology**, v. 62, p. 331-333, 2011.

CLC BIO. Disponível em: <http://www.clcbio.com>. Acesso em: 10/06/2015.

COENYE, T.; GORIS, J.; SPILKER, T.; VANDAMME, P.; LIPUMA, J.J. Characterization of Unusual Bacteria Isolated from Respiratory Secretions of Cystic Fibrosis Patients and Description of *Inquilinus limonus* gen. nov., sp. nov. **Journal of Clinical Microbiology**, v. 40, p. 2062-2069, 2002.

COGs Phylogenetic classification of proteins encoded in complete genomes. Disponível em: <http://www.ncbi.nlm.nih.gov/COG>. Acesso em: 01/09/2015.

COLSTON, S. M.; FULLMER, M. S.; BEKA, L.; *et al*. Bioinformatic Genome Comparisons for Taxonomic and Phylogenetic Assignments Using *Aeromonas* as a Test Case. **mBio**, v. 5, n. 6, p. e02136–14, 2014.

CRUZ, L. M.; SOUZA, E. M.; WEBER, O. B.; *et al*. 16S Ribosomal DNA Characterization of Nitrogen-Fixing Bacteria Isolated from Banana (*Musa* spp.) and Pineapple (*Ananas comosus* (L.) Merril). **Applied and Environmental Microbiology**, v. 67, n. 5, p. 2375–2379, 2001.

CRUZ, L. M.; SOUZA, E.M.; MONTEIRO, R.A.; *et al*. Comparative analysis of complete genome sequence of *Herbaspirillum seropedicae* and a draft genome sequence of *Herbaspirillum rubrisubalbicans*. In: 9th European Nitrogen Fixation Conference, 2010, Geneva – Suíça. **The abstract book of 9th European Nitrogen Fixation Conference**. p. 221, 2010.

CRUZ, L. M. ; WEISS, V. ; FAORO, H. ; CARDOSO, R. L. A. ; *et al*. Comparative Genomics of *Herbaspirillum* Genus. In: XLI Reunião Anual da Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq, Foz do Iguaçu. **XLI Reunião Anual da Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq**, 2012.

DE SOUZA, V.; PIRO, V. C.; FAORO, H.; *et al*. Draft Genome Sequence of *Herbaspirillum huttiense* subsp. *putei* IAM 15032, a Strain Isolated from Well Water. **Genome Announcements**, v. 1, n. 1, p. e00252–12–e00252–12, 2013.

DELCHER, A. L.; HARMON, D.; KASIF, S.; WHITE, O.; SALZBERG, S. L. Improved microbial gene identification with GLIMMER. **Nucleic Acids Research**, v. 27, p. 4636-4641, 1999.

DESHPANDE, A. D.; BAHETI, K. G.; CHATTERJEE, N. R. Degradation of β-lactam

antibiotics. **Current Science**, v. 87, n. 12, p. 1684–1695, 2004.

DING, L.; YOKOTA, A. Proposals of *Curvibacter gracilis* gen. nov., sp. nov. and *Herbaspirillum putei* sp. nov. for bacterial strains isolated from well water and reclassification of [*Pseudomonas*] *huttiensis*, [*Pseudomonas*] *lanceolata*, [*Aquaspirillum*] *delicatum* and [*Aquaspirillum*] *autotrophicum* as *Herbaspirillum huttiense* comb. nov., *Curvibacter lanceolatus* comb. nov., *Curvibacter delicatus* comb. nov. and *Herbaspirillum autotrophicum* comb. nov. **International Journal of Systematic and Evolutionary Microbiology**, v. 54, n. 6, p. 2223–2230, 2004.

DOBRITSA, A.P.; REDDY, M.C.S.; SAMADPOUR, M. Reclassification of *Herbaspirillum putei* as a later heterotypic synonym of *Herbaspirillum huttiense*, with the description of *H. huttiense* subsp. *huttiense* subsp. nov. and *H. huttiense* subsp. *putei* subsp. nov., and description of *Herbaspirillum aquaticum* sp. nov. **International Journal of Systematic and Evolutionary Microbiology**, v. 60, p. 1418-1426, 2010.

EBERZ, G.; HOGREFE, C.; KORTLÜKE, C.; *et al*. Molecular cloning of structural and regulatory hydrogenase (*hox*) genes of *Alcaligenes eutrophus* H16. **Journal of Bacteriology**, v. 168, n. 2, p. 636–641, 1986.

EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Research**, v. 32, n. 5, p. 1792–1797, 2004.

EDWARDS, D. J.; HOLT, K. E. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. **Microbial Informatics and Experimentation**, v. 3, n. 1, p. 2, 2013.

ELBELTAGY, A.; NISHIOKA, K.; SATO, T.; *et al*. Endophytic Colonization and In Planta Nitrogen Fixation by a *Herbaspirillum* sp. Isolated from Wild Rice Species. **Applied and Environmental Microbiology**, v. 67, n. 11, p. 5285–5293, 2001.

FILLINGER, S. Two Glyceraldehyde-3-phosphate Dehydrogenases with Opposite Physiological Roles in a Nonphotosynthetic Bacterium. **Journal of Biological Chemistry**, v. 275, n. 19, p. 14031–14037, 2000.

FREIRE, R. S.; PELEGRINI, R.; KUBOTA, L. T.; DURÁN, N. Novas tendências para o tratamento de resíduos industriais contendo espécies organocloradas. **Química Nova**, v. 23, p. 504-511, 2000.

FRIEDRICH, C.; BULYHA, I.; SOGAARD-ANDERSEN, L. Outside-In Assembly Pathway of the Type IV Pilus System in *Myxococcus xanthus*. **Journal of Bacteriology**, v. 196, p. 378-390, 2014.

GGDC 2.0. Disponível em: <http://ggdc.dsmz.de/distcalc2.php>. Acesso em: 10/06/2015.

GORDON, D.; ABAJIAN, C.; GREEN, P. Consed: A Graphical Tool for Sequence Finishing. **Genome Research**, v. 8, n. 3, p. 195–202, 1998.

GORIS, J.; KONSTANTINIDIS, K. T.; KLAPPENBACH, J. A.; *et al*. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. **International Journal of Systematic and Evolutionary Microbiology**, v. 57, n. 1, p. 81–91, 2007.

GULATI, A.; SOOD, S.; RAHI, P.; *et al*. Diversity Analysis of Diazotrophic Bacteria Associated with the roots of Tea (*Camellia sinensis* (L.) O. Kuntze). **Journal of Microbiology Biotechnology,** v. 21, p. 545-555, 2011.

GUZMÁN, E.; ROMEU, A.; GARCIA-VALLVE, S. Completely sequenced genomes of pathogenic bacteria: A review. **Enfermedades Infecciosas y Microbiología Clínica**, v. 26, n. 2, p. 88–98, 2008.

HAMMER, Ø., HARPER, D. A. T., RYAN, P. D. PAST: Paleontological statistics software package for education and data analysis. **Palaeontologia Electronica,** v. 4(1): 9pp, 2001.

ILLUMINA. Disponível em: <http://www.illumina.com>. Acesso em: 10/06/2015.

IM, W.; BAE, H.; YOKOTA, A.; LEE, S.T. *Herbaspirillum chlorophenolicum* sp. nov., a 4-chlorophenol-degrading bacterium. **International Journal of Systematic and Evolutionary Microbiology**, v. 54, p. 851-855, 2004.

ISHII, S.; YAMAMOTO, M.; KIKUCHI, M.; *et al*. Microbial Populations Responsive to Denitrification-Inducing Conditions in Rice Paddy Soil, as Revealed by Comparative 16S rRNA Gene Analysis. **Applied and Environmental Microbiology**, v. 75, n. 22, p. 7070–7078, 2009.

JAUREGUI, R.; RODELAS, B.; GEFFERS, R.; *et al*. Draft Genome Sequence of the Naphthalene Degrader *Herbaspirillum* sp. Strain RV1423. **Genome Announcements**, v. 2, n. 2, p. e00188–14–e00188–14, 2014.

JUNG, S.; LEE, M.; OH, T.; YOON, J. *Herbaspirillum rhizophaerae* sp. nov., isolated from rhizosphere soil of *Allium victorialis* var. *platyphyllum*. **International Journal of Systematic and Evolutionary Microbiology**, v. 57, p. 2284-2288, 2007.

JÚNIOR, P. I.; PEREIRA, G. M.; PERIN, L.; *et al*. Diazotrophic bacteria isolated from wild rice *Oryza glumaepatula* (*Poacea*) in the Brazilian Amazon. **Revista de Biologia Tropical**, v 61 (2), p. 991-9, 2013.

KANEHISA, M.; GOTO, S.; SATO, Y.; FURUMICHI, M.; TANABE, M. KEGG for integration and interpretation of large-scale molecular data sets. **Nucleic Acids Research**, v. 40, n. D1, p. D109–D114, 2012.

KANG, Y.; GU, C.; YUAN, L.; *et al*. Flexibility and Symmetry of Prokaryotic Genome Rearrangement Reveal Lineage-Associated Core-Gene-Defined Genome Organizational Frameworks. **mBio**, v. 5, n. 6, p. e01867–14, 2014.

KAWAICHI, S.; ITO, N.; YOSHIDA, T.; SAKO, Y. Bacterial and Archaeal Diversity in an Iron-Rich Coastal Hydrothermal Field in Yamagawa, Kagoshima, Japan. **Microbes and Environments**, v. 28, n. 4, p. 405–413, 2013.

KELLY, L. C.; COCKELL, C. S.; THORSTEINSSON, T.; *et al*. Pioneer Microbial Communities of the Fimmvörðuháls Lava Flow, Eyjafjallajökull, Iceland. **Microbial Ecology**, v. 68, n. 3, p. 504–518, 2014.

KIRCHHOF, G.; ECKERT, B.; STOFFELS, M.; *et al*. *Herbaspirillum frisingense* sp. nov., a new nitrogen-fixing bacterial species that occurs in C4-fibre plants. **International Journal of Systematic and Evolutionary Microbiology**, v. 51, p. 157-168, 2001.

KUHN, E.; ICHIMURA, A. S.; PENG, V.; *et al*. Brine Assemblages of Ultrasmall Microbial Cells within the Ice Cover of Lake Vida, Antarctica. **Applied and Environmental Microbiology**, v. 80, n. 12, p. 3687–3698, 2014.

KURTZ, S., PHILLIPPY, A., DELCHER, A. L., *et al*. Versatile and open software for comparing large genomes. **Genome Biology**, v. 5:R12, 2004.

LAGESEN, K.; HALLIN, P.; RODLAND, E. A.; *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. **Nucleic Acids Research**, v. 35, n. 9, p. 3100–3108, 2007.

LAGIER, J.-C.; GIMENEZ, G.; ROBERT, C.; RAOULT, D.; FOURNIER, P.-E. Non-contiguous finished genome sequence and description of *Herbaspirillum massiliense* sp. nov. **Standards in Genomic Sciences**, v. 7, n. 2, p. 1–14, 2012.

LAING, C.; BUCHANAN, C.; TABOADA, E. N.; *et al*. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. **BMC Bioinformatics**, v. 11, n. 1, p. 461, 2010.

LAND, M.; HAUSER, L.; JUN, S.-R.; *et al*. Insights from 20 years of bacterial genome sequencing. **Functional & Integrative Genomics**, v. 15, n. 2, p. 141–161, 2015.

LAPIERRE, P.; GOGARTEN, J. P. Estimating the size of the bacterial pan-genome. **Trends in Genetics**, v. 25, n. 3, p. 107–110, 2009.

LEE, G. W.; LEE, K.-J.; CHAE, J.-C. Genome Sequence of *Herbaspirillum* sp. Strain GW103, a Plant Growth-Promoting Bacterium. **Journal of Bacteriology**, v. 194, n. 15, p. 4150–4150, 2012.

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

LEPIDI, A.; CASELLA, S.; TOFFANIN, A.; PETRASSI, S.; NUTI, M. Hydrogen-oxidizing bacteria for biomass production. **International Journal of Hydrogen Energy**, v. 15, n. 7, p. 485–489, 1990.

LEVEAU, J. H. J.; UROZ, S.; DE BOER, W. The bacterial genus *Collimonas* : mycophagy, weathering and other adaptive solutions to life in oligotrophic soil environments. **Environmental Microbiology**, v. 12, n. 2, p. 281–292, 2010.

LIFE TECHNOLOGIES. Disponível em: <https://www.lifetechnologies.com/br/en/home.html>. Acesso em: 10/06/2015.

LIN, S.-Y.; HAMEED, A.; ARUN, A. B.; *et al*. Description of *Noviherbaspirillum malthae* gen. nov., sp. nov., isolated from an oil-contaminated soil, and proposal to reclassify *Herbaspirillum soli*, *Herbaspirillum aurantiacum*, *Herbaspirillum canariense* and *Herbaspirillum psychrotolerans* as *Noviherbaspirillum soli* comb. nov., *Noviherbaspirillum aurantiacum* comb. nov., *Noviherbaspirillum canariense* comb. nov. and *Noviherbaspirillum psychrotolerans* comb. nov. based on polyphasic analysis. **International Journal of Systematic and Evolutionary Microbiology**, v. 63, n. Pt 11, p. 4100–4107, 2013.

LOWE, T.M.; EDDY, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. **Nucleic Acids Research**, v. 25, p. 955-964, 1997.

LU, H.; FUJIMURA, R.; SATO, Y.; *et al*. Characterization of *Herbaspirillum*- and *Limnobacter*-Related Strains Isolated from Young Volcanic Deposits in Miyake-Jima Island, Japan. **Microbes and Environments.** v. 23, p. 66-72, 2008.

LUKJANCENKO, O.; WASSENAAR, T. M.; USSERY, D. W. Comparison of 61 Sequenced *Escherichia coli* Genomes. **Microbial Ecology**, v. 60, n. 4, p. 708–720, 2010.

LUKJANCENKO, O.; USSERY, D. W.; WASSENAAR, T. M. Comparative Genomics of *Bifidobacterium*, *Lactobacillus* and Related Probiotic Genera. **Microbial Ecology**, v. 63, n. 3, p. 651–673, 2012.

MARQUES, A. C. Q.; PALUDO, K. S.; DALLAGASSA, C. B.; *et al*. Biochemical Characteristics, Adhesion, and Cytotoxicity of Environmental and Clinical Isolates of Herbaspirillum spp. **Journal of Clinical Microbiology**, v. 53, n. 1, p. 302–308, 2015.

MARTIN, W.; SCHNARRENBERGER, C. The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. **Current Genetics**, v. 32, n. 1, p. 1–18, 1997.

MATASSA, S.; BOON, N.; VERSTRAETE, W. Resource recovery from used water: The manufacturing abilities of hydrogen-oxidizing bacteria. **Water Research**, v. 68, p. 467–478, 2015.

MACINTYRE, D. L.; MIYATA, S. T.; KITAOKA, M.; PUKATZKI, S. The *Vibrio cholera* type VI secretion system displays antimicrobial properties. **PNAS**, v. 107, p. 19520-19524.

MEIER-KOLTHOFF, J. P.; AUCH, A. F.; KLENK, H.-P.; GÖKER, M. Genome

sequence-based species delimitation with confidence intervals and improved distance functions. **BMC Bioinformatics**, v. 14, n. 1, p. 60, 2013.

MÉDIGE, C.; MOSZER, I. Annotation, comparison and databases for hundreds of bacterial genomes. **Research in microbiology**, v. 158, p. 724-736, 2007.

MEDINI, D.; SERRUTO, D.; PARKHILL, J.; *et al*.. Microbiology in the post-genomic era. **Nature Reviews Microbiology**, v.6, p. 419-430, 2008.

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

MILLER, J. R.; KOREN, S.; SUTTON, G. Assembly algorithms for next-generation sequencing data. **Genomics**, v. 95, n. 6, p. 315–327, 2010.

MIYAGAWA, Y.; TAMOI, M.; SHIGEOKA, S. Overexpression of a cyanobacterial fructose-1,6-/sedoheptulose-1,7-bisphosphatase in tobacco enhances photosynthesis and growth. **Nature Biotechnology**, v. 19, n. 10, p. 965–969, 2001.

MONTEIRO, R. A.; BALSANELLI, E.; TULESKI, T.; *et al*. 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. 441–451, 2012.

MONTEIRO, R. A.; CRUZ, L. M.; WASSEM, R.; *et al*. Comparative Genomics of *Herbaspirillum* Species. In: **Plasticity in Plant-Growth and Phytopathogenic Bacteria**. USA: Springer, 2014. 171-198.

MORIYA, Y.; ITOH, M.; OKUDA, S.; YOSHIZAWA, A. C.; KANEHISA, M. KAAS: an automatic genome annotation and pathway reconstruction server. **Nucleic Acids Research**, v. 35, n. Web Server, p. W182–W185, 2007.

NARZISI, G.; MISHRA, B. Comparing De Novo Genome Assembly: The Long and Short of It. **PLoS ONE**, v. 6, n. 4, p. e19175, 2011.

NCBI. Disponível em: <http://www.ncbi.nlm.nih.gov/>. Acesso em: 10/06/2015.

OH, H.-M.; LEE, K.; JANG, Y.; *et al*. Genome Sequence of Strain IMCC9480, a Xanthorhodopsin-Bearing Betaproteobacterium Isolated from the Arctic Ocean. **Journal of Bacteriology**, v. 193, n. 13, p. 3421–3421, 2011.

OLIVARES, F.L.; BALDANI, V. L. D.; REIS, V., M.; BALDANI, J. I.; DÖBEREINER, J. Occurrence of the endophytic diazotrophs *Herbaspirillum* spp. in roots, stems and leaves predominantly of *Gramineae*. **Biology and Fertility of Soils,** v. 21, p. 197-200, 1996.

OVERBEEK, R.; OLSON, R.; PUSCH, G. D.; *et al*. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). **Nucleic Acids Research**, v. 42, n. D1, p. D206–D214, 2014.

PALLEN, M.J.; WREN, B.W. Bacterial Pathogenomics. **Nature**, v. 449, p. 835-842, 2007.

PEDROSA, F.O.; MONTEIRO, R.A.; WASSEM, R.; CRUZ, L.M.; AYUB, R.A.; *et al.* Genome of *Herbaspirillum seropedicae* Strain SmR1, a Specialized Diazotrophic Endophyte of Tropical Grasses. **Plos Genetics**, v. 7, p. 1-10, 2011.

PEDROSA, F. O. ; CARDOSO, R. L. A. ; WEISS, V. A.; *et al*. Genomic comparison of *Herbaspirillum* species. In: 11th European Nitrogen Fixation Conference, Tenerife. **11th European Nitrogen Fixation Conference - Abstract Book**, 2014.

PERL. Disponível em: <https://www.perl.org>. Acesso em: 10/06/2015.

POHLMANN, A.; FRICKE, W. F.; REINECKE, F.; *et al*. Genome sequence of the bioplastic-producing "Knallgas" bacterium *Ralstonia eutropha* H16. **Nature Biotechnology**, v. 24, n. 10, p. 1257–1262, 2006.

PRICE, C. T. .; LEE, I. R.; GUSTAFSON, J. E. The effects of salicylate on bacteria. **The International Journal of Biochemistry & Cell Biology**, v. 32, n. 10, p. 1029–1043, 2000.

PUMPHREY, G. M.; RANCHOU-PEYRUSE, A.; SPAIN, J. C. Cultivation-Independent Detection of Autotrophic Hydrogen-Oxidizing Bacteria by DNA Stable-Isotope Probing. **Applied and Environmental Microbiology**, v. 77, n. 14, p. 4931–4938, 2011.

PYTHON. Disponível em: <https://www.python.org>. Acesso em: 10/06/2015.

RAAIJMAKERS, J. M.; PAULITZ, T. C.; STEINBERG, C.; *et al*. The rhizosphere: a playground and battlefield for soilborne pathogens and beneficial microorganisms. **Plant and Soil**, v. 321, n. 1-2, p. 341–361, 2009.

RAMOS, R. T.; CARNEIRO, A. R.; BAUMBACH, J.; *et al*. Analysis of quality raw data of second generation sequencers with Quality Assessment Software. **BMC Research Notes**, v. 4, n. 1, p. 130, 2011.

RANJIT, D. K.; ENDRES, J. L.; BAYLES, K. W. *Staphylococcus aureus* CidA and LrgA Proteins Exhibit Holin-Like Properties. **Journal of Bacteriology**, v. 193, n. 10, p. 2468–2476, 2011.

REIS, V. M.; URQUIAGA, S.; SILVA, M. F.; *et al*. Eficiência agronômica do inoculante de cana-de-açúcar aplicado em três ensaios conduzidos no Estado do Rio de Janeiro durante o primeiro ano de cultivo. **Embrapa Agrobiologia**. Boletim de Pesquisa & Desenvolvimento, 45. Seropédica, 2009.

RICHARDSON, E. J.; WATSON, M. The automatic annotation of bacterial genomes. **Briefings in Bioinformatics**, v. 14, n. 1, p. 1–12, 2013.

ROTHBALLER, M.; SCHMID, M.; KLEIN, I.; *et al*. *Herbaspirillum hiltneri* sp. Nov., isolated from surface-sterilized wheat roots. **International Journal of Systematic**

and Evolutionary Microbiology, v. 56, p. 1341-1348, 2006.

RUST, A. G.; MONGIN, E.; BIRNEY, E. Genome annotation techniques: new approaches and challenges. **Drug Discovery Today**, v. 7, n. 11, p. S70–S76, 2002.

SALZBERG, S. L.; PHILLIPPY, A. M.; ZIMIN, A.; *et al*. GAGE: A critical evaluation of genome assemblies and assembly algorithms. **Genome Research**, v. 22, n. 3, p. 557–567, 2012.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U.S.A., v. 74, n. 12, p. 5463–5467, 1977.

SANTOS, F.; BOELE, J.; TEUSINK, B. A Practical Guide to Genome-Scale Metabolic Models and Their Analysis. **Methods in Enzymology**, v. 500, p.509–532, 2011.

SCHIFFELS, J.; PINKENBURG, O.; SCHELDEN, M.; *et al*. An innovative cloning platform enables large-scale production and maturation of an oxygen-tolerant [NiFe]-hydrogenase from *Cupriavidus necator* in *Escherichia coli*. **PloS One**, v. 8, n. 7, p. e68812, 2013.

SCHLEIFER, K. H. Classification of Bacteria and Archaea: Past, present and future. **Systematic and Applied Microbiology**, v. 32, n. 8, p. 533–542, 2009.

SCHMIDT, M.; BALSANELLI, E.; FAORO, H.; *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, n. 1, p. 98, 2012.

SEKOWSKA, A.; DANCHIN, A. The methionine salvage pathway in *Bacillus subtilis*. **BMC Microbiol**. v. 2, p. 8, 2002.

SEUFFERHELD, M. J.; ALVAREZ, H. M.; FARIAS, M. E. Role of Polyphosphates in Microbial Adaptation to Extreme Environments. **Applied and Environmental Microbiology**, v. 74, n. 19, p. 5867–5874, 2008.

SICILIANO, S. D.; GERMIDA, J. J. Bacterial inoculants of forage grasses that enhance degradation of 2-chlorobenzoic acid in soil. **Environmental Toxicology and Chemistry**, v. 16, n. 6, p. 1098–1104, 1997.

SILVA, R.M.; CAUGANT, D.A.; ERIBE, E.R.K.; *et al*. Bacterial diversity in aortic aneurysms determined by 16S ribosomal RNA gene analysis. **Journal of Vascular Surgery**, v. 44, p. 1055-1060, 2006.

SNIPEN, L.; USSERY, D. W. Standard operating procedure for computing pangenome trees. **Standards in Genomic Sciences**, v. 2, n. 1, p. 135–141, 2010.

SPILKER, T.; ULUER, A. Z.; MARTY, F. M.; et al. Recovery of *Herbaspirillum* Species from Persons with Cystic Fibrosis. **Journal of Clinical Microbiology**, v. 46, n. 8, p. 2774–2777, 2008.

STEIN, L. Genome Annotation: From Sequence to Biology. **Nature Reviews Genetics**, v. 2, p. 493-503, 2001.

STOTHARD, P.; WISHART, D. S. Automated bacterial genome analysis and annotation. **Current Opinion in Microbiology**, v. 9, n. 5, p. 505–510, 2006.
STRAUB, D.; ROTHBALLER, M.; HARTMANN, A.; LUDEWIG, U. The genome of the endophytic bacterium *H. frisingense* GSF30T identifies diverse strategies in the *Herbaspirillum* genus to interact with plants. **Frontiers in Microbiology**, v. 4, 2013.

SUSARLA, S.; MEDINA, V. F.; MCCUTCHEON, S. C. Phytoremediation: An ecological solution to organic chemical contamination. **Ecological Engineering**, v. 18, n. 5, p. 647–658, 2002.

TABITA, F. R.; HANSON, T. E.; LI, H.; *et al*. Function, Structure, and Evolution of the RubisCO-Like Proteins and Their RubisCO Homologs. **Microbiology and Molecular Biology Reviews**, v. 71, n. 4, p. 576–599, 2007.

TAMURA, K.; STECHER, G.; PETERSON, D.; FILIPSKI, A.; KUMAR, S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. **Molecular Biology and Evolution**, v. 30, n. 12, p. 2725–2729, 2013.

TAN, M. J.; OEHLER, R. L.; Lower Extremity Cellulitis and Bacteremia With *Herbaspirillum seropedicae* Associated With Aquatic Exposure in a Patient With Cirrhosis. **Infectious Diseases in Clinical Pratice**, v. 13, p. 277-279, 2005.

TAN, Z.Q.; MEN, R.; ZHANG, R.Y.; HUANG, Z. First Report of *Herbaspirillum rubrisubalbicans* Causing Mottled Stripe Disease on Sugarcane in China. **The American Phytopathological Society**, v. 94, p. 379, 2010.

TATUSOV, R.L.; FEDOROVA, N.D.; JACKSON, J.D.; *et al*. The COG database: an updated version includes eukaryotes. **BMC Bioinformatics**, v. 4, p. 41, 2003.

THOMPSON, C. C.; CHIMETTO, L.; EDWARDS, R. A.; *et al*. Microbial genomic taxonomy. **BMC Genomics**, v. 14, n. 1, p. 913, 2013.

TIEPPO, E. **Montagem e Análise Preliminar do Genoma de *Bradyrhizobium elkanii* 587 Utilizando Leituras de Sequências de DNA Curtas**. 79 f. Dissertação (Mestrado em Bioinformática) – Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, Curitiba, 2011.

TOUCHON, M.; HOEDE, C.; TENAILLON, O.; *et al*. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. **PLoS genetics**, v. 5, n. 1, p. e1000344, 2009.

UÇMAK, D.; AYHAN, E.; MELTEM AKKURT, Z.; HAYDAR UÇAK. Presentation of three cases with phyto contact dermatitis caused by *Ranunculus* and *Anthemis* genera. **Journal of Dermatological Treatment**, v. 25, n. 6, p. 467–469, 2014.

VACHERON, J.; DESBROSSES, G.; BOUFFAUD, M.-L.; et al. Plant growth-promoting rhizobacteria and root system functioning. **Frontiers in Plant Science**, v. 4, 2013.

VALOT, B.; GUYEUX, C.; ROLLAND, J. Y.; *et al*. What It Takes to Be a *Pseudomonas aeruginosa*? The Core Genome of the Opportunistic Pathogen Updated. **PLOS ONE**, v. 10, n. 5, p. e0126468, 2015.

VALVERDE, A. *Herbaspirillum lusitanum* sp. nov., a novel nitrogen-fixing bacterium associated with root nodules of *Phaseolus vulgaris*. **International Journal Of Systematic And Evolutionary Microbiology**, v. 53, n. 6, p. 1979–1983, 2003.

VAN TONDER, A. J.; MISTRY, S.; BRAY, J. E.; *et al*. Defining the Estimated Core Genome of Bacterial Populations Using a Bayesian Decision Model. **PLoS Computational Biology**, v. 10, n. 8, p. e1003788, 2014.

VIEIRA, L. DO N.; FAORO, H.; ROGALSKI, M.; *et al*. The Complete Chloroplast Genome Sequence of *Podocarpus lambertii*: Genome Structure, Evolutionary Aspects, Gene Content and SSR Detection. **PLoS ONE**, v. 9, n. 3, p. e90618, 2014.

VOLLMER, W.; JORIS, B.; CHARLIER, P.; FOSTER, S. Bacterial peptidoglycan (murein) hydrolases. **FEMS Microbiology Reviews**, v. 32, n. 2, p. 259–286, 2008.

WEBER, O. B.; CRUZ, L. M.; BALDANI, J. I.; DÖBEREINER, J. *Herbaspirillum*-Like Bacteria in Banana Plants. **Brazilian Journal of Microbiology**, v. 32, n. 3, p. 201–205, 2001.

WEISS, V.A. **Estratégias de finalização da montagem do genoma da bactéria diazotrófica endofítica *Herbaspirillum seropedicae* SmR1**. Dissertação (Mestrado em Bioquímica) – Setor de Ciências Biológicas, Universidade Federal do Paraná, 2010.

WEISS, V. A.; FAORO, H.; SFEIR, M.Z.T.; RAITTZ, R.T.; SOUZA, E.M.; *et al*. Draft Genome Sequence of *Herbaspirillum lusitanum* P6-12, an Endophyte Isolated from Root Nodules of *Phaseolus vulgaris*. **Journal of Bacteriology** (Print), 2012.

WEISS, V. A.. **Montagem, Anotação e Análise Comparativa do Genoma da Bactéria *Herbaspirillum lusitanum* P6-12**. 93 f. Tese (Doutorado em Ciências-Bioquímica) – Setor de Ciências Biológicas, Universidade Federal do Paraná, Curitiba, 2014.

XU, H.X.; WU, H.Y.; QIU, Y.P.; SHI, X.Q.; HE, G.H.; ZHANG, J.F.; WU, J.C. Degradation of fluoranthene by a newly isolated strain of *Herbaspirillum chlorophenolicum* from activated sludge. **Biodegradation**, v. 22, p. 335-345, 2011.

YE, W.; YE, S.; LIU, J.; et al. Genome Sequence of the Pathogenic *Herbaspirillum seropedicae* Strain Os34, Isolated from Rice Roots. **Journal of Bacteriology**, v. 194, n. 24, p. 6993–6994, 2012.

ZERBINO, D. R. Using the Velvet *de novo* Assembler for Short-Read Sequencing Technologies. In: BAXEVANIS, A. D.; DAVISON, D. B.; PAGE, R. D. M.; *et al*. **Current Protocols in Bioinformatics**, 2010. Hoboken, NJ, USA: John Wiley & Sons, Inc.

ZHOU, Y.; LIANG, Y.; LYNCH, K. H.; DENNIS, J. J.; WISHART, D. S. PHAST: A Fast Phage Search Tool. **Nucleic Acids Research**, v. 39, n. suppl, p. W347–W352, 2011.

ZHU, B.; YE, S.; CHANG, S.; *et al*. Genome Sequence of the Pathogenic *Herbaspirillum seropedicae* Strain Os45, Isolated from Rice Roots. **Journal of Bacteriology**, v. 194, n. 24, p. 6995–6996, 2012.

ZIGA, E. D.; DRULEY, T.; BURNHAM, C.-A. D. *Herbaspirillum* Species Bacteremia in a Pediatric Oncology Patient. **Journal of Clinical Microbiology**, v. 48, n. 11, p. 4320–4321, 2010.

ZOUACHE, K.; VORONIN, D.; TRAN-VAN, V.; MAVINGUI, P. Composition of Bacterial Communities Associated with Natural and Laboratory Populations of *Asobara tabida* Infected with *Wolbachia*. **Applied and Environmental Microbiology**, v. 75, n. 11, p. 3755–3764, 2009.