UNIVERSIDADE FEDERAL DO PARANÁ
VINÍCIUS ALMIR WEISS

MONTAGEM, ANOTAÇÃO E ANÁLISE COMPARATIVA DO GENOMA DA BACTÉRIA *Herbaspirillum lusitanum* P6-12

CURITIBA
2014

VINÍCIUS ALMIR WEISS

MONTAGEM, ANOTAÇÃO E ANÁLISE COMPARATIVA DO GENOMA DA BACTÉRIA *Herbaspirillum lusitanum* P6-12

Tese apresentada ao Curso de Pós-Graduação em Ciências- Bioquímica da Universidade Federal do Paraná, como requisito parcial para a obtenção do título de Doutor em Ciências Bioquímica.

Orientador:
Prof. Dr. Leonardo Magalhães Cruz

Coorientador:
Prof. Dr. Roberto Tadeu Raittz

CURITIBA

2014

Universidade Federal do Paraná
Sistema de Bibliotecas

Weiss, Vinícius Almir
Montagem, anotação e análise comparativa do genoma da bactéria *Herbaspirillum lusitanum* P6-12. / Vinícius Almir Weiss. – Curitiba, 2014.
93 f.: il. ; 30cm.

Orientador: Leonardo Magalhães Cruz
Co-orientador: Roberto Tadeu Raittz

Tese (doutorado) - Universidade Federal do Paraná, Setor de Ciências Biológicas. Programa de Pós-Graduação em Bioquímica.

1. Herbaspirillum 2. Bioinformática 3. Genômica I. Título II. Cruz, Leonardo Magalhães III. Raittz, Roberto Tadeu IV. Universidade Federal do Paraná. Setor de Ciências Biológicas. Programa de Pós-Graduação em Bioquímica.

CDD (20. ed.) 589.9
574.192

TERMO DE APROVAÇÃO

VINÍCIUS ALMIR WEISS

MONTAGEM, ANOTAÇÃO E ANÁLISE COMPARATIVA DA BACTÉRIA *Herbaspirillum lusitanum* P6-12.

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

Prof. Dr. Leonardo Magalhães Cruz(Orientador)
Departamento de Bioquímica e Biologia Molecular – UFPR

Dr. Hélisson Faoro
Departamento de Bioquímica e Biologia Molecular – UFPR

Prof. Dr. Humberto Maciel França Madeira
Escola de Ciências Agrárias e Medicina Veterinária – PUCPR

Dr. Christian Macagnan Probst
Departamento de Bioinformática – ICC-FIOCRUZ.

Prof. Dr. Rommel Thiago Jucá Ramos
Laboratório de Polimorfismo de DNA - UFPA

Curitiba, 28 de março de 2014.

Dedico este trabalho à minha esposa Izabella,
aos meus pais Almir e Simome e à minha irmã Pâmela.

AGRADECIMENTOS

Muito aconteceu até este momento. Muitas mudanças, situação boas e outras não tão boas, mas o importante é que tudo correu bem. Gostaria de agradecer a Deus por permitir esta caminhada, sem Ele eu não estaria aqui.

Muitas pessoas também tiveram papel central a quem devo muitos agradecimentos.

Primeiramente à minha esposa Izabella, que é minha parceira para todas as horas, minha alma gêmea com quem divido tudo, as alegrias e as tristezas.

Aos meus pais, Almir e Simone, agradeço por todo o apoio. Vocês são responsáveis por muito do que sou e tenho nesta vida. À minha irmã Pâmela pela amizade que temos que é muito importante. Aos meus sogros Alceu e Eli Izabel por me receberem tão bem em sua família.

Agradeço à Universidade Federal do Paraná pela oportunidade e à CAPES pela concessão da bolsa de estudos.

Agradeço aos meus orientadores Leonardo Magalhães Cruz e Roberto Tadeu Raittz, por todos os ensinamentos, pelo companheirismo desde o meu estágio, há seis anos, e principalmente pelas amizades.

Aos Professores Emanuel Maltempi de Souza e Fábio de Oliveira Pedrosa pela oportunidade e por acreditarem em mim.

À Prof. Maria Berenice R. Steffens pela sua ajuda desde o início, ainda na minha preparação para o mestrado.

Aos meus colegas de laboratório Rodrigo Cardoso e Helisson Faoro pelas discussões e pela parceria.

Aos Professores Alain Denise e Olivier Lespinet por me receberem em seu laboratório durante o meu Doutorado Sanduíche na França e ao departamento de Bioquímica e Biologia Molecular por proporcionar esta oportunidade de ampliar minha formação e conhecimento.

Agradeço aos membros da banca examinadora pela disponibilidade em avaliar este trabalho bem como pelas críticas e sugestões.

“Agrada-te do Senhor, e Ele satisfará
os desejos do teu coração”.
Salmos 37:4

RESUMO

O *Herbaspirillum lusitanum* P6-12 foi isolado de nódulos de raiz da planta *Phaseolus vulgaris* (feijão) em Portugal. O genoma foi obtido através do programa *CLCWorkbench* utilizando a combinação de leituras de sequências SOLiD e Illumina. Utilizando o programa RAST e posterior revisão manual, a anotação do genoma obteve 4488 genes, cobrindo 87% do genoma, 51 tRNA e um operon ribosomal na ordem 16S rRNA-23S rRNA-5S rRNA. O *H. lusitanum* P6-12 possui as vias metabólicas *Entner-Doudoroff*, pentoses fosfato, ácidos tricarboxílicos *Embden Meyerhof-Parnas*, com exceção da enzima 6-fosfofrutoquinase (EC 2.7.1.11). Não foram encontrados os genes responsáveis pela fixação de nitrogênio, mas foi encontrado o gene que codifica a 1-aminociclopropano-1-carboxilato (ACC) deaminase, relacionada ao desenvolvimento da planta sob condições de estresse. Também foi encontrada a sequência parcial do gene que codifica para a proteína Ribulose-1,5-bisfosfato carboxilase/oxigenase (*RuBisCO*), ligada a fixação do carbono, embora os outros genes desta via não tenham sido identificados. Análises na diferença de cobertura global em relação à cobertura na região do operon, sinalizaram a presença de mais um operon ribosomal. Foram desenvolvidos métodos de tratamento das regiões de *gap* que permitiram a finalização do *draft* em 30 *super-contigs,* totalizando 4.919.496 *pb*. A análise do gene 16S rRNA confirmou que a espécie de *Herbaspirillum* sequenciada foi a de *H*. *lusitanum* estirpe P6-12. Foram utilizadas duas metodologias de agrupamento para as proteínas, a ligação simples e a ligação completa. Os grupos de genes ortólogos foram determinados como a intersecção entre os três métodos empregados nas análises *OrthoMCL*, *INPARANOID* e *BBH*, identificando 5218 grupos de genes ortólogos entre as 12 espécies estudas: *Herbaspirillum sp.* JC206, CF444, GW103, YR522, *H. seropedicae* SmR1, Os45, Os34, AU14040, *H. rubrisubalbicans* M1, *H. frisingense* GSF30 e *H. huttiense* subsp putei 7-2. Dentro dos 5218 grupos de ortólogos, 768 estiveram presentes em todos os genomas e foram definidos como *core* genoma*.* As análises filogenéticas baseadas nos grupos ortólogos sinalizaram uma possível classificação das estirpes *Herbaspirillum sp.* CF444 como *Herbaspirillum lusitanum* CF444 e *Herbaspirillum sp.* GW103 como *H. huttiense subsp. putei* GW103.

Palavras-chave: *Herbaspirillum lusitanum* P6-12, anotação, grupos ortólogos, bioinformática, genômica.

ABSTRACT

The *Herbaspirillum lusitanum* P6 -12 was isolated from root nodules of the *Phaseolus vulgaris* plant (beans) in Portugal. The genome was assembled in *CLCWorkbench* program using the combination of SOLiD and Illumina reads. Using the RAST program and subsequent manual review, the genome annotation obtained 4.488 genes, covering 87 % of the genome, 51 tRNA and one ribosomal operon 16S rRNA-23S rRNA-5S rRNA. *H. lusitanum* P6-12 has the Entner- Doudoroff, pentose phosphate, citric acid and Embden Meyerhof-Parnas metabolic pathways, with the exception of the enzyme 6-phosphofructokinase (EC 2.7.1.11). The presence of genes responsible for nitrogen fixation have not been found, but the gene encoding 1-aminocyclopropane-1-carboxylate (ACC) deaminase was present. This gene is related with plant development under stress conditions. Another finding was the partial sequence from the carbon fixation protein, Ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco). Analysis of the difference in overall genome coverage compared to the operon coverage, showed the presence of another ribosomal operon. Methods of treating regions of gap were developed and allowed the completion of the draft in 30 *super-contigs* and 4.919.496 bp. The analysis confirmed that the 16S rRNA gene of the species *Herbaspirillum lusitanum* was sequenced strain P6-12. Orthologous analysis showed the difference in prediction between three methods evaluated, *OrthoMCL*, *INPARANOID* and *BBH*. Two methods of clustering, simple and complet link were used. Groups of orthologous genes were determined as the intersection between the three methods identifying 5218 clusters of orthologous genes among the 12 species studied: *Herbaspirillum* sp. JC206, CF444, GW103, YR522, *H. seropedicae* SMR1, Os45, Os34, AU14040, *H. rubrisubalbicans* M1, *H. frisingense* GSF30 and *H. huttiense subsp. putei* 7-2. Within the groups of orthologous, 768 were present in all genomes being defined as core genome. Phylogenetic analysis based on orthologous groups showed a possible classification of *Herbaspirillum* CF444 as *Herbaspirillum lusitanum* CF444 and
Herbaspirillum GW103 as *H. huttiense subsp. putei* GW103.

Keywords: *Herbaspirillum lusitanum* P6-12, annotation, orthologous clustering, bioinformatics, genomics.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

16S	rRNA 16S *ribosomal Ribonucleic Acid* (ácido ribonucleico ribosomal)
16S-23S-5S	*Ribosomal operon* (operon ribosomal)
BACs	Bacteriófagos
BBH	*Best Blast Hits* (Melhores alinhamentos *BLAST*)
COG	*Cluster of Orthologous Groups* (Agrupamento de grupos ortólogos)
ddATP	2',3'-Dideoxyadenosine-5'-Triphosphate (2',3'-dideoxicitidina-5'- trifosfato)
DDBJ *DNA*	*Databank of Japan* (Banco de DNA do Japão)
ddCTP	2',3'-Dideoxycytidine-5'-Triphosphate (2',3'-dideoxicitidina-5'-trifosfato)
ddGTP	2',3'-Dideoxyguanosine-5'-Triphosphate (2',3'-dideoxiguanosina-5'-trifosfato)
ddTTP	2',3'-Dideoxythymidine-5'-Triphosphate (2',3'-dideoxitimidina-5'- trifosfato)
DNA	*Deoxyribonucleic acid* (Ácido desoxirribonucleico)
EMBL	*European Molecular Biology Laboratory* (Laboratório Europeu de Biologia Molecular)
GAAT	*Genome Automatic Annotation Tools* (Ferramenta de anotação automática de genomas)
Gb	Giga base
GenBank	*Genetic sequence database* (Banco de Dados de Sequência Genética)
HGT	*Horizontal gene transfer* (Transferência horizontal gênica)
INSDC	*International Nucleotide Sequence Database Collaboration* (Banco de Dados de Sequência de Colaboração Internacional)
IS	*Insersion Sequence* (Sequências de Inserção)
k-mers	Subsequência de tamanho k de uma sequência de DNA
Kb	Kilo base
KEGG	*Kyoto Encyclopedia of Genes and Genomes* (Enciclopédia de genes e genoma de Kyoto)

Mb	Mega bases
ML	*Maximum Likelihood* (Máxima verossimilhança)
Mpb	Mega pares de bases
MRP	*Matrix Representation using Parsimony* (Representação de matriz com parsimônia)
MSSA	*Most Similar Supertree* (Super-árvore mais similar)
NCBI	*National Center for Biotechnology Infromation* (Centro Nacional de Informação Biotecnológica)
NNIs	*Nearest Neighbor Interchanges* (Troca entre vizinhos mais próximos)
OLC	*Overlap Consensus* (Consenso de sobreposição)
Operon	Segmento de DNA que funciona como uma única unidade de transcrição
ORFs	*Open Read frames* (Fases abertas de leitura)
pb	Pares de bases
PCR	*Polimerase Chain Reaction* (Reação em cadeia da polimerase)
PDB	*Protein Databank* (Banco de dados de proteínas)
PFAM	*Protein Families* (Família de proteínas)
QFIT	*Maximum Quartet Fit* (Ajuste máximo de quarteto)
RNA	*Ribonucleic Acid* (Ácido ribonucleico)
SFIT	*Maximum Splits* Fit (Ajuste máximo de quebras)
SMRT	*Single molecule real time* (Molécula única em tempo real)
SPR	*Subtree Pruning and Regrafting* (Poda e reinserção da superárvore)
Tb	Tera bases
tRNA	*Transfer Ribonucleic acid (Ácido ribonucleico de transferência)*
UPGMA	*Unweighted Pair Group Method with Arithmetic Mean* (Método de grupos de pares não ponderados com média aritmética)
WGS	*Wole Genome Sequence* (Sequência total do genoma)

SUMÁRIO

1 REVISÃO BIBLIOGRÁFICA

1.1 GÊNERO *Herbaspirillum*

Taxonomicamente o gênero *Herbaspirillum* pertence ao filo *Proteobacteria*, classe *Betaproteobacteria*, ordem *Burkholderiales*, família *Oxalobacteraceae*. Foi originalmente descrito em bactérias isoladas de milho (*Zea mays*), arroz (*Oryza sativa*), sorgo (*Sorghum bicolor*), trigo (*Triticum aestivum*) e cana-de-açúcar (*Saccharum oficinarum*) (BALDANI *et al*., 1986). Todas estas culturas apresentam grande importância e impacto econômico.

As bactérias do gênero *Herbaspirillum* possuem grande interesse biotecnológico devido a sua capacidade de colonizar plantas bem como, em condições microaeróbias, fixar N_2 atmosférico. Esta capacidade permite que a bactéria seja utilizada como um promotor de crescimento vegetal, diminuindo assim os gastos com a adubação química (BALDANI *et al*., 1986.

Atualmente o gênero contém 11 espécies descritas, já que as espécies *Herbaspirillum soli, Herbaspirillum canariensis, Herbaspirillum aurantiacum e Herbaspirillum psychotolerans* foram reclassificadas como *Noviherbaspirillum (*LIN *et al.,* 2013*)* (Tabela 1). A caracterização destas espécies permitiu um maior esclarecimento da diversidade metabólica presente no gênero, demostrando que além de características diazotróficas endofíticas em plantas, existem patógenos específicos para algumas culturas de cana-de-açúcar, espécies não fixadoras de N_2, e isolados de humanos associados a doenças. Dentre as doenças descritas que acometem humanos estão aneurisma da aorta (MARQUES DA SILVA *et al*., 2006), fibrose cística (SPILKER *et al*., 2008) e leucemia linfoblástica aguda (CHEN *et al*., 2011; ZIGA; DRULEY e BURNHAM, 2010). O mecanismo pelo qual essas bactérias atuam nestas doenças ainda não está esclarecido, sendo classificadas como patógenos oportunistas (BERG; EBERL e HARTMANN, 2005).

TABELA 1 – ESPÉCIES DO GÊNERO *Herbaspirillum*

Espécie	Capacidade de Fixação de Nitrogênio	Lugar de Isolamento	Autor
Herbaspirillum seropedicae	+	Milho (*Zea mays*), arroz (*Oryza sativa*) sorgo (*Sorghum bicolor*)	BALDANI *et al.*, 1986
Herbapirillum rubrisubalbicans	+	*Saccharurn spp.*	BALDANI *et al.*, 1986
Hesbaspirillum frisingense	+	*Spartina pectinata /Miscanthus spp*	KIRCHHOF *et al.*, 2001
Herbaspirillum lusitanum	+	*Phaseolus vulgaris* (Feijão)	VALVERDE *et al.*, 2003
Herbaspirillum huttiense	-	Água de poço	DING *et al.*, 2004
Herbaspirillum hiltneri	+	Água de lago	ROTHBALLER *et al.*, 2006
Herbaspirillum autotroficum	-	Solo	DING *et al.*, 2004
Herbaspirillum rhizosphaerae	-	Rizosfera	JUNG *et al.*, 2007
Herbaspirillum aquaticum	-	Água destilada	DOBRITSA *et al.*, 2009
Herbaspirillum chlorophenolicum	-	Solo	IM *et al.*, 2004
*Herbaspirillum psychrotolerans**	-	Solo	BAJERSKI *et al.*, 2013
*Herbaspirillum canariense**	-	Solo	CARRO *et al.*, 2011
*Herbaspirillum aurantiacum**	-	Solo	CARRO *et al.*, 2011
*Herbaspirillum soli**	-	Solo	CARRO *et al.*, 2011
Herbaspirillum massiliense	-	Humano	LAGIER *et al.*, 2012

*Reclassificados como *Noviherbaspirillum* (LIN *et al.,* 2013).

1.2 *Herbaspirillum lusitanum* P6-12

O *Herbaspirillum lusitanum* P6-12*,* nome dado em referência à Lusitânia, o nome romano de Portugal, foi isolado de nódulos de raiz da planta *Phaseolus vulgaris* (feijão) (VALVERDE *et al*., 2003). A bactéria é gram negativa e apresenta a forma de bastão curvado com um ou dois flagelos. Foi observado que o *H. lusitanum* P6-12 tem capacidade de crescimento em temperaturas entre 20º e 40°C e pH ótimo de crescimento entre 5 e 8. Testes fisiológicos e bioquímicos demostraram que o *H. lusitanum* P6-12 é capaz de redução de nitrato, produção de β-galactosidase, e pode assimilar meso-inositol, meso-erithritol, L-ramnose e arabinose. Testes usando antibióticos demostraram resistência gentamicina, cefotaxima, ceftazidima, tobramicina, netilmicina e amicacina. O resultado dos testes bioquímicos estão apresentados na Tabela 2, em comparação com os outros membros do gênero (VALVERDE *et al*., 2003).

TABELA 2 – CARACTERÍSTICAS FENOTÍPICAS DA BACTÉRIA *Herbaspirillum lusitanum* P6-12.

Testes \ Organismos	*H. lusitanum*	*H. seropedicae*	*H. rubrisubalbicans*	*H. frisingense*	*H. huttiense subs. putei*	*H. hiltneri*	*H. rhizosphaerae*	H. autotroficum
Redução de nitratos a nitritos	-	+	+	+	-		-	-
β-galactosidase	-	+	+	+	-	-	-	-
Assimilação de:								
N-acetil glucosamina	+	+	-	+	-	+		-
Meso-inositol	-	+	-	-	+		-	-
L-ramnose	+	+	-	-	-	-	+	-
Meso-eritritol	-	-	+	-	-			-
Arabinose	+	+	+	-	+		+	+
Fenol								-
Clorofenol								-
Resistência à:								
Gentamicina	-	+	-	-			-	
Cefotaxima	+	-	+	+				
Ceftazidima	+	+	-	-				
Tobramicina	-	+	+	-				
Netilmicina	-	+	-	-				

O sinal "+" indica que o resultado do teste é positivo/resistente ao antibiótico analisado. O sinal "-" indica que o resultado foi negativo/sensível ao antibiótico analisado. O sinal "x" indica que o teste não foi realizado.
FONTE: Adaptado de VALVERDE *et al.,* 2003; BALDANI *et al.,* 1986; KIRCHHOF *et al.,* 2001; DING *et al.*, 2004; ROTHBALLER *et al*., 2006; JUNG *et al*., 2007; DOBRITSA *et al.,* 2010; CARRO *et al*., 2011; BAJERSKI *et al*., 2013.

Hebaspirillum lusitanum P6-12 foi descrito como uma bactéria fixadora de nitrogênio através do crescimento em meio livre de nitrogênio e da amplificação do gene *nifD,* embora a banda correspondente ao produto amplificado em gel de agarose não tenha sido sequenciada. Seu DNA foi estimado com um conteúdo de G+C de 57,9 mol% (VALVERDE *et al*., 2003).

1.3 Tecnologias de sequenciamento de DNA

O processo de sequenciamento genômico teve seu início quando Sanger e colaboradores apresentaram ao mundo o método de sequenciamento de DNA utilizando terminadores de cadeias (SANGER; NICKLEN e COULSON, 1977). Este método lhe rendeu seu segundo Prêmio Nobel de Química em 1980 e permitiu projetos de sequenciamento de DNA nos últimos 30 anos. Esta técnica foi utilizada desde o primeiro genoma de organismo não viral sequenciado, a bactéria *Haemophilus influenzae* (FLEISCHMANN *et al*., 1995) até o genoma humano (LANDER *et al*., 2001). Porém, devido ao alto custo e tempo, o sequenciamento de larga escala ficou restrito apenas a grandes centros genômicos e a alguns laboratórios (HALL, 2007).

Contudo, este cenário mudou com a introdução de novas técnicas de sequenciamento de DNA, chamadas de sequenciamento de próxima geração (*next generation sequencing*) ou sequenciamento de segunda geração (SHENDURE e JI, 2008). Estas tecnologias não só mudaram a abordagem do sequenciamento genômico, dispensando a necessidade de clonagem *in vivo,* como também o tempo e o custo envolvido no processo. Com isso, o sequenciamento de segunda geração forneceu a base para criar linhas de pesquisa e impulsionar as chamadas "*omics*", incluindo a genômica, metagenômica, transcriptômica, metabolômica, interactômica, proteômica e a identificação de polimorfismos (MARDIS, 2008). Dentre os sequenciadores de segunda geração destacam-se o 454-ROCHE, ILLUMINA HiSeq/MiSeq e o SOLiD System (Life Technology). O sequenciador 454-ROCHE foi o primeiro a ser comercializado e o primeiro a introduzir o sistema de clonagem por PCR em emulsão, dispensando a clonagem *in vivo.* O método de sequenciamento também foi inovador, chamado de pirossequenciamento, no qual a incorporação de um nucleotídeo pela DNA polimerase libera pirofosfato, iniciando uma série de reações em cadeia que resultam na produção de luz pela enzima luciferase (MARGULIES *et al*., 2005).

Em 2006 o sistema SOLiD® foi disponibilizado e introduziu a tecnologia de sequenciamento de duas bases, baseada na hibridização-ligação, na qual a reação de sequenciamento é catalisada por uma DNA ligase, e não por uma DNA polimerase (MARDIS, 2008; MCKERNAN *et al*., 2006). A preparação das

bibliotecas e a clonagem dos fragmentos de DNA genômico por PCR em emulsão utiliza microesferas e é similar à utilizada pelo sequenciador 454 ROCHE® (HUTCHISON, 2007). A diferença entre estas duas plataformas está no tamanho das microesferas: elas apresentam um tamanho menor no SOLiD, 1 µm de diâmetro quando comparadas aos 26 µm de diâmetro daquelas utilizadas no 454 ROCHE®. Outra diferença é a disposição das microesferas na superfície, aleatória no SOLiD e ordenada em poços no 454. Estas diferenças permitiram ao SOLID uma maior densidade de deposição de microesferas, aumentando o número de sequências geradas pelo sequenciador e promovendo maior flexibilidade no sequenciamento de mais de uma biblioteca (MCKERNAN *et al*., 2006). As bibliotecas são sequenciadas com uso de sondas de 8 bases ligadas a terminadores fluorescentes. A fluorescência é emitida quando há a hibridização da sonda no DNA alvo, e desaparece com a clivagem das 3 últimas bases da sonda. A sequência do fragmento é deduzida a partir de 5 rodadas de sequenciamento através do deslocamento da posição de hibridização do *primer* no adaptador em uma posição. Este método permite que cada base seja lida duas vezes aumentando a fidelidade do sequenciamento (SHENDURE e JI, 2008). Assim, cada sinal de fluorescência codifica para um dinucleotídeo em vez de uma única base, e a decodificação é realizada analisando a transição das cores de cada ligação. Como a sequência do adaptador P1 é conhecida, a primeira base do fragmento é identificada, direcionando a transição de cores para a identificação das demais bases (MARDIS, 2008). Atualmente o SOLiD está disponível na versão 5500xl.

No mesmo ano, a empresa Solexa lançou o *Genome Analyzer* (GA) e em 2007 a companhia foi comprada pela Illumina. Este sequenciador introduziu a tecnologia de sequenciamento por síntese (LIU *et al*., 2012). Na preparação das bibliotecas o DNA ligado a adaptadores é desnaturado em fita simples e depositado na lâmina de sequenciamento. A clonagem dos fragmentos é realizada pela amplificação em ponte gerando os grupos de fragmentos clonados que posteriormente são clivados em fita simples novamente com a ajuda da enzima de linearização. São usados 4 tipos de nucleotídeos (ddATP, ddCTP, ddTTP, ddGTP), que contém diferentes terminadores fluorescentes cliváveis e grupos bloqueadores removíveis que complementam o fragmento, uma base por vez. O sinal é então capturado por uma câmera de alta resolução

CCD (LIU *et al*., 2012). Atualmente a Illumima disponibiliza no mercado a versão HiSeq 2500 para sequenciamento genômico. A comparação entre as tecnologias de sequenciamento de segunda geração e o método de Sanger estão resumidas na Tabela 3.

TABELA 3 – COMPARAÇÃO ENTRE AS TECNOLOGIAS DE SEQUENCIAMENTO.

Sequenciador	**454 GS FLX**	**HiSeq 2500**	**SOLiD 5500xl**	**Sanger 3730 xl**
Técnica	Pirossequencia-mento	Sequenciamento por síntese	Sequenciamento por ligação	Terminadores de cadeia
Tamanho das leituras	700 pb	1x 36 pb, 2x50 pb, 2x100 pb 2x 125 pb	75 pb, 75 pb + 35 pb 60 + 60 pb	400~900 pb
Fidelidade	99,99%	98,00%	99,94	99,99%
Dados gerados	0.7 Gb	128 Mb - 1 Tb	300 Gb	1,9~84 Kb
Tempo	24 horas	29 horas a 6 dias	1 a 6 dias	20 min a 3 horas
Vantagem	Tamanho das leituras, rápidez	Alta capacidade	Fidelidade	Alta qualidade e tamanho das leituras
Desvantagem	Taxa de erro com homopolímeros maiores que 6, alto custo, baixa capacidade	Leituras curtas	Leituras curtas	Alto custo, baixa capacidade
Preço do sequenciador	$500.000	$740,000	$595,000	$95.000
Custo por milhão de bases em dólar	$10	$0.07	$0.07	$240

SE: do inglês *Single end.* Representa sequenciamento de ponta simples, PE: do inglês *Pair end.* Representa biblioteca de pares, Kb: Quilobases, Gb: Giga bases.
FONTE: Adaptado de Liu *et al.,* 2012, http://www.lifetechnologies.com/, http://www.illumina.com/

Embora as tecnologias de sequenciamento de segunda geração tenham proporcionado uma nova era no sequenciamento genômico, um novo capítulo já está sendo escrito com o desenvolvimento dos sequenciadores de terceira geração. Duas características definem estes novos sequenciadores, 1- a etapa de PCR anterior ao sequenciamento torna-se desnecessária, o que diminui o tempo de preparo das bibliotecas e o mais importante, 2- o sinal do

sequenciamento é captado em tempo real, independente da origem, através do monitoramento em tempo real da reação enzimática (LIU *et al*., 2012). Destacam-se dois métodos: o sequenciamento de molécula única em tempo real (SMRT: do inglês, *Single Molecule Real Time*) (NIEDRINGHAUS et al., 2011).

1.4 BANCOS DE DADOS BIOLÓGICOS

A pesquisa genômica tem avançado significativamente nos últimos anos impulsionada pelas novas tecnologias de sequenciamento. Este crescimento proporcionou a criação do *INSDC* (*International Nucleotide Sequence Database Collaboration*) em 1987 (COCHRANE; KARSCH-MIZRACHI e NAKAMURA, 2011). O *INSDC* consiste na colaboração de três grandes bancos de dados sincronizados, o *GenBank* (*NIH, National Institute of Health database*) dos Estados Unidos da América, o *DDBJ* (*DNA Databank of Japan*) do Japão e o *EMBL* (*European Molecular Biology Laboratory*), situado na Inglaterra. A finalidade destes bancos é coletar, armazenar e disseminar dados de sequências de DNA e RNA. Este serviço gratuito tem papel central na manutenção e no desenvolvimento de diversas áreas de pesquisa genômica (COCHRANE; KARSCH-MIZRACHI e NAKAMURA, 2011). O NCBI *(National Center for Biotechnology Information*) que pertence ao *NIH,* por exemplo, possui 2 bancos de sequências, o *GenBank* e o *WGS* (*Whole genome shotgun*). O *GenBank* contém os depósitos de sequências desde 1982, enquanto o banco *WGS* foi iniciado em 2002 e engloba sequências de projetos de sequenciamento de genomas. Em dezembro de 2013 estavam catalogadas mais de $1X10^{12}$ bases no *GenBank*. Este número só é superado pelo banco *WGS*, que contém $6X10^{12}$ bases, demostrando o aumento no número de projetos de sequenciamento de genomas completos a partir de 2002 (Figura 1).

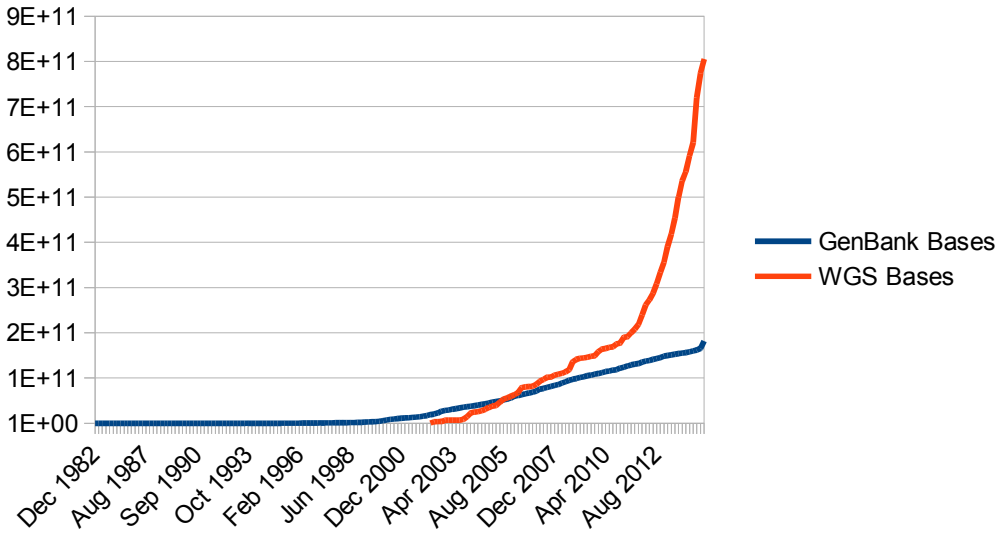


FIGURA 1- CRESCIMENTO DOS BANCOS DE DADOS DO *NCBI*.
FONTE: Adaptado do *NCBI*. http://www.ncbi.nlm.nih.gov/genbank/statistics

Além destes grandes bancos de dados existem diversos bancos menores relacionados a assuntos específicos. Dentre estes se destacam o *PDB* (*Protein Database Bank*), banco de dados de proteínas onde pode-se encontrar as sequências e estruturas cristalográficas das proteínas armazenadas (WESTBROOK *et al*., 2003); o *PFAM*, outro banco de dados de proteínas, que agrupa as sequências de proteínas em famílias representadas por alinhamentos múltiplos (PUNTA *et al*., 2012); o *RDP*, banco de sequências de genes *rRNA* ribosomais de bactérias (COLE *et al*., 2014); e o *KEGG* (*Kyoto Encyclopedia of Genes and Genomes*) que tem como objetivo a análise sistemática das funções gênicas classificando as proteínas de acordo com os processos celulares nos quais participam, metabolismo, transporte de membrana, transdução de sinal e ciclo celular (KANEHISA e GOTO, 2000).

Estes bancos, junto com outros não citados, formam uma rede de informações essenciais para o desenvolvimento da pesquisa em biologia molecular, permitindo o intercâmbio de conhecimento entre os laboratórios pelo mundo.

1.5 MONTAGEM DE GENOMAS

Genoma é todo material genético contido em determinado organismo, incluindo toda a informação necessária para manutenção da vida. Genomas de organismos são sequenciados e depositados nos bancos de dados em todo o mundo, fornecendo informações valiosas dos mais diversos organismos e sistemas. Porém, antes de agregar valor biológico para estas sequências, o processo de montagem das sequências dos genomas é uma etapa fundamental em um projeto de sequenciamento (MILLER; KOREN e SUTTON, 2010).

Não existe hoje tecnologia capaz de sequenciar grandes moléculas de DNA em um único passo, embora iniciativas estejam sendo desenvolvidas, como o caso do sequenciamento de molécula única em tempo real (*SMRT*: do inglês, *Single Molecule Real Time*) (NIEDRINGHAUS et al., 2011; FLUSBERG *et al*., 2010). Todas as técnicas de sequenciamento utilizam a fragmentação do DNA genômico, produzindo fragmentos de DNA pequenos que são sequenciados diretamente e então sobrepostos para reconstruir o genoma. Este procedimento é utilizado desde o início do sequenciamento pelo método de Sanger (SANGER; NICKLEN e COULSON, 1977), até os dias de hoje, mesmo com a introdução de modernos métodos de sequenciamento de DNA de terceira geração. Fragmentação química ou física são largamente utilizadas para diminuir o tamanho da sequência a ser lida pelo sequenciador, chegando ao tamanho máximo de sequências obtidas próximo a 1.000 pb, muito pequeno quando comparado ao tamanho do genoma de um eucarioto, bactéria e até vírus (SHENDURE e JI, 2008).

Devido a esta limitação, para o sequenciamento do genoma é necessária a construção de bibliotecas de DNA genômico. O tamanho do fragmento de DNA da biblioteca varia de acordo com a estratégia, de 0,1 a 100 kb. Estes fragmentos de DNA podem ser clonados *in vivo* utilizando como vetores plasmídeos, cosmídeos, fosmídeos, bacteriófagos e BACs, ou clonados *in vitro,* nas plataformas de sequenciamento de segunda geração*.* Cada sequenciador tem o seu protocolo individual de construção das bibliotecas de DNA, porém algumas características são comuns a todos e serão descritas a seguir.

As bibliotecas de fragmento *(fragment* ou *single-end*), permitem ao sequenciador produzir apenas uma leitura parcial do fragmento de DNA molde. Já as bibliotecas de pares de leituras (*mate-pair* ou *paired-end*) possuem a vantagem de produzir duas leituras para cada fragmento de DNA, uma em cada extremidade, gerando mais informação e auxiliando o processo de montagem do genoma (BERGLUND; KIIALAINEN e SYVANEN, 2011). As bibliotecas de pares possuem nomenclaturas distintas na literatura de acordo com o método de construção utilizado. Na biblioteca *paired-end* são ligados adaptadores nas extremidades do fragmento de DNA, adaptador P1 na região 5' e P2 na região 3'. O primeiro passo do sequenciamento é realizado através do adaptador P1 para gerar a leitura 5' do par, e P2 para gerar a leitura 3'. Já na biblioteca *mate-pair,* o DNA molde é circularizado utilizando adaptadores biotinilados para depois ser lido. Este método permite o sequenciamento de fragmentos longos de DNA, por exemplo 10 kb, importante para resolver regiões de repetição no genoma (BERGLUND; KIIALAINEN e SYVANEN, 2011) (Figura 2). Com isso, o tamanho do DNA molde varia, bem como o sentido do sequenciamento. Estas informações são importantes pois direcionam o processo de montagem. Como resultado final, o sequenciamento produz milhares de fragmentos que são sobrepostos para gerar a sequência completa do genoma. Este processo é conhecido como montagem.

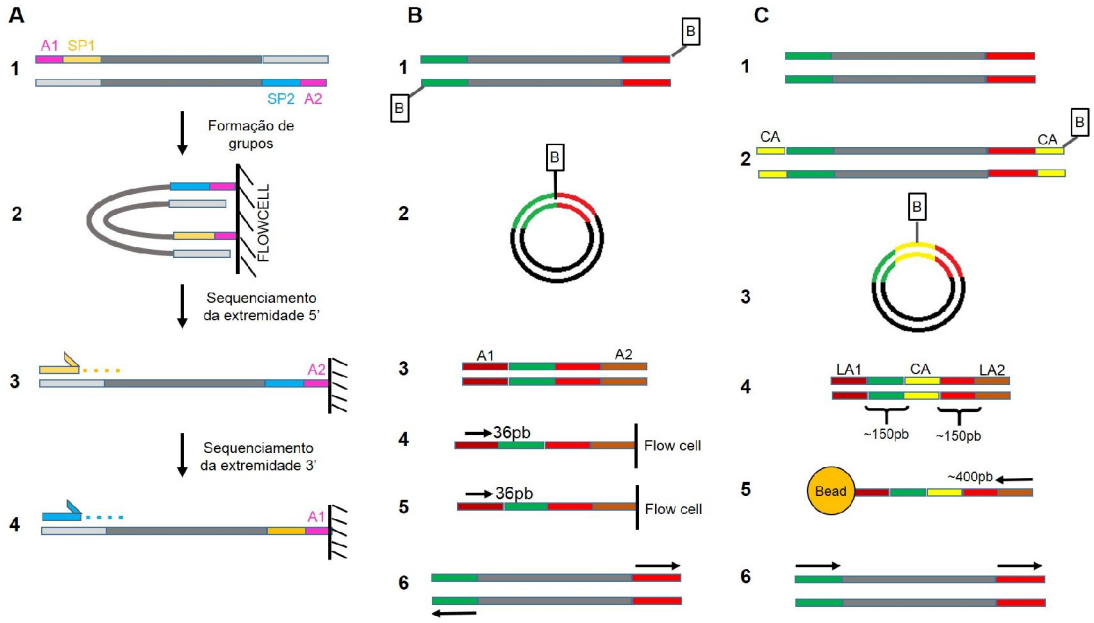


FIGURA 2- CONSTRUÇÃO DAS BIBLIOTECAS DE FRAGMENTO, *PAIRED-END E MATE-PAIR* PARA AS PLATAFORMAS ILLUMINA E SOLID.

A – Preparação da biblioteca *paired-end* Illumina. (1) São ligados adaptadores nas extremidades do fragmento. (2) Ocorre a formação dos grupos. (3) O sequenciamento do par 5' é realizado a partir do *primer* P1. (4) O sequenciamento da extremidade 3' é realizado a partir do *primer* P2.

B – Preparação de bibliotecas *mate-pair* Illumina. Os fragmentos são pareados usando nucleótidos biotinilados (1). Após a circularização, as duas extremidades do fragmento (verde e vermelho) ficam localizados adjacentes uns aos outros (2). O DNA circular é fragmentado, e os fragmentos biotinilados são purificados e capturados por afinidade. Adaptadores de sequenciamento (A1 e A2) são ligados nas extremidades dos fragmentos obtidos (3), e os fragmentos são hibridizados em uma célula de fluxo (*flow cell*), na qual eles são amplificados em ponte. A primeira sequência lida é obtida com o adaptador A2 ligado à célula de fluxo (4). A cadeia complementar é sintetizada e linearizado com adaptador A1 ligado à célula de fluxo, e a segunda sequência é obtida (5). Por fim a sequência de duas leituras (setas) é dirigida para o exterior a partir do fragmento original (6).

C– Preparação de bibliotecas *mate-pair* SOLiD. Os fragmentos originais (1) são pareados com os nucleotídeos não marcados e ligados a um adaptador interno (AI) construído com biotina (2). Depois da circularização (3), da fragmentação e da purificação por afinidade, adaptadores (P1 e P2) são ligados às novas extremidades do fragmento (4). O sequenciamento é realizado com dois iniciadores diferentes, complementares ao adaptador P1 e ao adaptador interno, respectivamente (5). Como resultado as leituras terão a mesma orientação (6).

FONTE: Adaptado de BERGLUND, 2011.

Como o sequenciamento é aleatório, não existe garantia de que todas as regiões do DNA serão representadas, por isso, normalmente, é necessário um número de bases muito maior do que o tamanho do genoma sequenciado. Existem projetos de sequenciamento genômico em que a sequência do genoma foi obtida com 10x de cobertura, mas o mais comum é 30 ou mais vezes de cobertura (SCHATZ; DELCHER e SALZBERG, 2010). A quantidade de dados gerados está ligada a complexidade do genoma estudado, e suas características específicas.

O processo de montagem consiste na união das leituras de sequências de DNA (*reads)* gerados pelo sequenciador. Esta união é obtida através do alinhamento das sequências e sobreposição das bases de DNA, sendo chamado de *contig*. O consenso é construído com a base mais frequente no alinhamento para cada posição do *contig.*

Muitos dos projetos de genomas não almejam a representação total do cromossomo, mas um mapa global *(draft)* contendo a maior parte da sequência do genoma, sequenciando as regiões faltantes de acordo com a necessidade. Estes mapas globais são formados por vários *contigs* ordenados chamados *super-contigs* como apresentado na Figura 3.

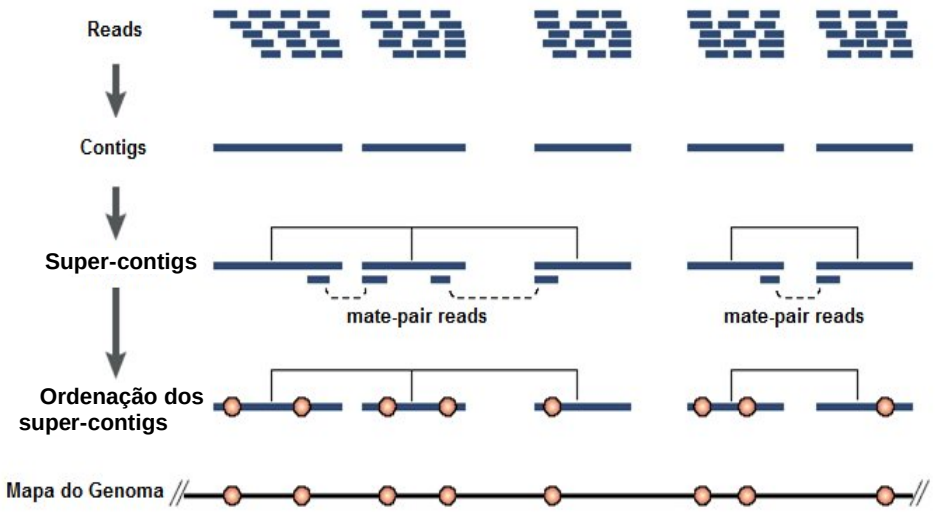


FIGURA 3- FLUXOGRAMA DA MONTAGEM DE GENOMAS.

A partir da fragmentação e sequenciamento do DNA as leituras (*reads*) são empilhadas pelo processo de montagem gerando sequências contíguas (*contigs*). As leituras pareadas (*mate-pair)* são utilizadas para criar e orientar os *super-contigs* para a geração do mapa genômico (*draft*).
FONTE: Adaptado de http://www.exploreable.wordpress.com/2011/05/03

Os algoritmos mais utilizados pelos montadores são *OLC* (*Overlap Layout Consensus*) e grafos de *Bruijin*. O algoritmo *OLC* compara cada leitura com todas as outras, independente da orientação. Diferentes algoritmos *OLC* possuem diferentes critérios para avaliar as sobreposições. No caso do montador Celera, um dos primeiros a implementar o algoritmo *OLC*, as sobreposições de alta qualidade eram definidas com 40 nucleotídeos apresentando mais de 94% de similaridade (MYERS *et al*., 2000).

O grafo de montagem é construído usando os dois tipos de sobreposições onde os vértices representam as leituras e os ramos representam as sobreposições de alta qualidade. Assim, a montagem do genoma consiste em encontrar um caminho através do grafo que passe por cada vértice exatamente uma vez (caminho Hamiltoniano). A fim de diminuir o tamanho do grafo, o grafo de montagem *OLC* é simplificado na fase de *layout*, em que os segmentos do grafo de montagem são comprimidos em *contigs (*MILLER; KOREN e SUTTON, 2010). Assim, no grafo de sobreposição, um *contig* seria um subgrafo ou um grupo de vértices com várias ligações entre si, e as sobreposições são representadas por uma sequência (Figura 4A e 4B). Uma vez identificado um subgrafo, esses vértices e arestas são compactados em um só vértice, ou um *contig* (Figura 4C). Apenas os vértices de início e fim ligam com nós de outros subgrafos (POP, 2009).

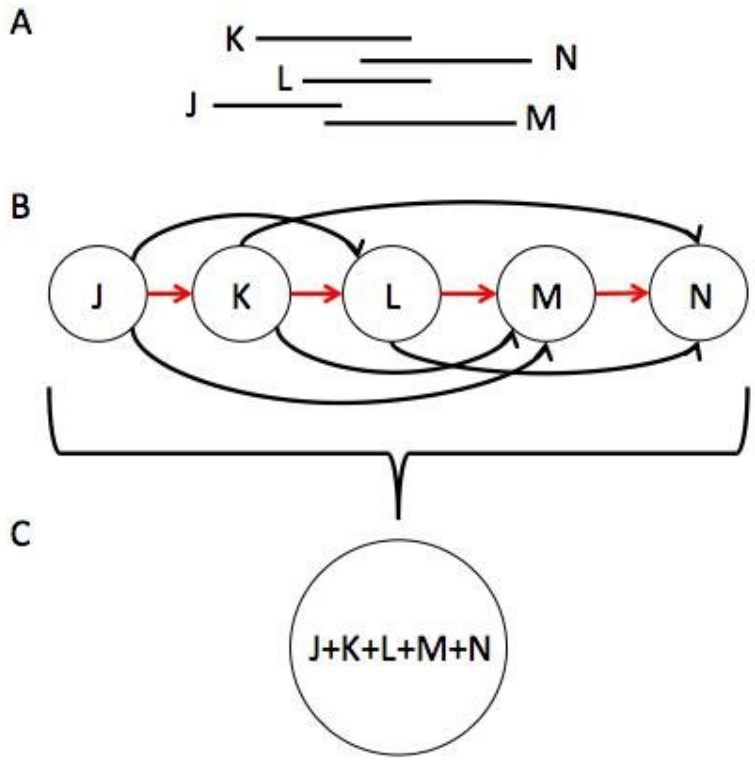


FIGURA 4 - CRIAÇÃO DE CONTIGS PELO ALGORITMO OLC.

Em A cada letra representa a sequência consenso da sobreposição das leituras (*contigs*). Em B cada *contig* possui múltiplas ligações que são simplificadas em apenas uma entrada e uma saída, representada pela seta vermelha. Em C é observada a ordenação dos *contigs.*
FONTE: http://www.gcat.davidson.edu/phast/olc.html

O grafo de Brujin divide a sequência em janelas com tamanho definidos (*k-mers)* que representam a sequência original. Em seguida, um grafo orientado é construído ligando pares de *k-mers* com sobreposição entre os primeiros k-1 nucleotídeos e os últimos k-1 nucleotídeos (Figura 5C). A simplificação do grafo é realizada quando vértices possuem apenas uma entrada e saída entre si, desta forma as ligações transitivas são eliminadas. Esta simplificação proporciona uma redução da complexidade do grafo e é clara quando comparada com o grafo de sobreposição 5B, proporcionando menor demanda de processamento e consequente aumento de velocidade, por isso é muito utilizado para grandes quantidades de dados. O sentido da seta é orientado do *k-mer* cujos últimos k-1 nucleotídeos estão sobrepostos, para o *k-*

mer cujos primeiros k-1 nucleotídeos estão sobrepostos (ZERBINO e BIRNEY, 2008).

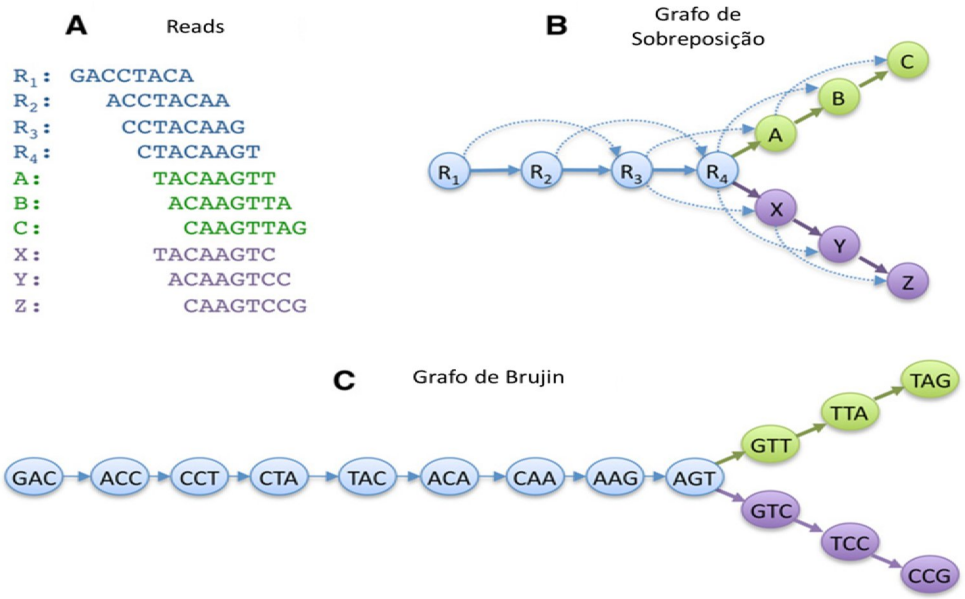


FIGURA 5- GRAFO DE BRUJIN.

A- *Reads* sobrepostos.
B- Cada *leitura* representa um nó. As sobreposições diretas >5 bp são mostradas com setas sólidas. As sobreposições transitivas, implicando em sobreposições em nós distantes, são mostradas com setas pontilhadas.
C- Para cada sequência é criado um nó, porém com um *k-mer* de 3 bp. A sobreposição é contabilizada a cada (*k-mer* -1) bases. Podemos observar que em ambas as abordagens a sequência repetida possibilitou dois caminhos distintos, porém o grafo de Brujin obteve um melhor direcionamento no crescimento das sequências.
FONTE: Adaptado de Schatz *et al.,* 2010.

A quantidade de dados gerados pelos sequenciadores, além da necessidade de implementação e desenvolvimento de algoritmos robustos para montagem, exige a disponibilidade de uma infraestrutura que permita a análise destes dados, com computadores de grande capacidade de processamento e armazenamento. A metodologia de sequenciamento empregada interfere no tamanho das sequências geradas e na sua qualidade, e pode proporcionar erros de leitura que dificultam o processo de montagem. A característica do genoma escolhido, como a presença de transposases, regiões repetidas, plasmídeos, ou genomas que possuem mais de um cromossomo, como o genoma humano (diplóide) ou do trigo *Triticum aestivum* (hexaplóide), são exemplos dos muitos desafios encontrados no processo de montagem

(FRANCKI e APPELS, 2002). Uma abordagem que auxilia na resolução destas dificuldades é a comparação do DNA genômico com genomas completos de organismos próximos. Esta comparação permite que o genoma seja estudado com base em informações biológicas já disponíveis, confirmando as diferenças observadas e identificando os possíveis problemas oriundos do processo de montagem.

1.6 GENÔMICA COMPARATIVA E HOMOLOGIA ENTRE SEQUÊNCIAS

A genômica comparativa é a análise e comparação de genomas com o propósito de obter um melhor entendimento de como as espécies evoluíram bem como de determinar a função de genes e de regiões não codificantes do genoma (KARLIN; CAMPBELL e MRÁZEK, 1998; ABBY e DAUBIN, 2007).

Estudos comparativos de genomas de procariotos têm revelado a sua complexa estrutura e organização, e a enorme diversidade genética entre estes organismos, mesmo entre isolados de uma mesma espécie. A diversidade genética é identificada através da presença de elementos de transferência lateral (*HGT)*, elementos de inserção (*IS)*, ilhas de patogenicidade, plasmídeos e transposons; e levanta questionamentos importantes sobre os mecanismos pelos quais estes microrganismos evoluem e sua taxonomia (COENYE *et al*., 2005; BINNEWIES *et al*., 2006).

Estudos de contexto filogenético, como a reconstrução de árvores e redes filogenéticas, assim como a determinação de grupos de genes ortólogos entre espécies, revelaram padrões de evolução de características bioquímicas e morfológicas importantes, como a fixação de nitrogênio e mecanismos de defesas químicas (DUTILH *et al*., 2007; DELSUC; BRINKMANN e PHILIPPE, 2005). Também foi observado que o número de genes compartilhados entre as espécies é inferior ao total de genes identificados. Estes genes presentes em todas os organismos de uma mesma espécie ou gênero representam o *core genome*, composto por genes considerados essenciais (*housekeeping)*, normalmente responsáveis pelos processos de tradução, transcrição e replicação/reparo do DNA. Já o total de genes identificados para cada espécie ou gênero *(pan-genome)* cresce cada vez que um genoma é disponibilizado.

Esta característica indica que regiões podem ter sido obtidas ou perdidas durante o processo de especiação (LEFEBURE e STANHOPE, 2007).

Os avanços para elucidar muitos questionamentos relacionados a aspectos fundamentais da genética, da bioquímica e da evolução de um grande número de espécies são resultado da crescente quantidade de genomas completos depositados nos bancos de dados, aliados às análises comparativas entre suas sequências de DNA (ABBY e DAUBIN, 2007; HUYNEN; GABALDÓN e SNEL, 2005).

O estudo de homologia entre sequências precede a era genômica e foi difundida por Fitch (1970), que analisou o relacionamento evolutivo de convergência e divergência entre sequências de proteínas (FITCH, 1970). Posteriormente, ele definiu a homologia como a relação entre dois indivíduos que descenderam com divergência de um ancestral comum (FITCH, 2000). Existem diferentes relações de homologia entre as proteínas, cujas definições são importantes e serão descritas a seguir.

Sequências ortólogas são sequências homólogas, derivadas de um evento de especiação, a partir da sequência do último ancestral comum das espécies que estão sendo comparadas. Ortólogos normalmente executam funções equivalentes em espécies estreitamente relacionadas, sendo utilizados na anotação do genoma de novas espécies (KOONIN, 2005). Parálogos são sequências derivadas de um evento de duplicação de uma sequência. Ocorrem tanto em um mesmo genoma como entre diferentes genomas e eles podem evoluir para exercer novas funções, mecanicamente distintas porém biologicamente relacionadas (FITCH, 1970). In-parálogos são parálogos que resultam de duplicação linhagem específica após um processo de especiação, mantendo funções parecidas dentro da espécie. Out-parálogos são parálogos resultantes de uma duplicação precedendo um evento de especiação e estas proteínas podem exercer diferentes funções (SONNHAMMER e KOONIN, 2002). Xenólogos são sequências homólogas cuja história, desde seu ancestral comum, envolve uma transferência do material genético interespécies (horizontal) para, pelo menos, uma das espécies (KOONIN, 2005). O evento de transferência horizontal de genes *HGT* (*Horizontal gene transfer*) está inserido neste contexto, onde o material genético é transferido entre espécies, e não segue linhagem vertical de sua descendência (KUZNIAR *et al*., 2008). A Figura

6 apresenta os diferentes tipos de homologia entre sequências.

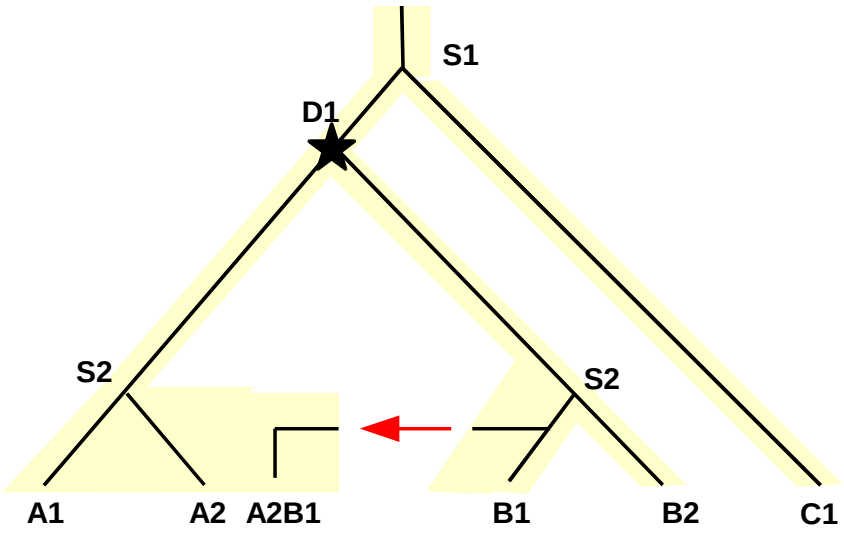


Figura 6 - TIPOS DE HOMOLOGIA ENTRE PROTEÍNAS.
S1 e S2 são eventos de especiação. D1 é uma duplicação. (A1,A2) ou (B1,C1) representam ortólogos. (A1,B1) e (A2,B2) representam parálogos. (A1,B2) são in-parálogos com respeito à especiação S1. (A1,B2) são out-paralogos comrespeito à especiação S2. A seta vermelha representa a transferência do gene B1 para a espécie A2. O gene A2B1 é xenólogo dos outros 5 genes.
FONTE: Adaptado de (FITCH, 2000).

Um exemplo de aplicação utilizando grupos de proteínas ortólogas é o banco de grupos ortólogos *COG* (*Cluster of orthologous groups*) que possui proteínas de diferentes espécies agrupadas por função. Este conjunto de dados tem a capacidade de elucidar padrões entre as sequências, anotar genomas realizando a caracterização funcional de novas proteínas e avaliar as relações filogenéticas entre as proteínas bem como os organismos (TATUSOV; KOONIN e LIPMAN, 1997). Outra aplicação é a Enciclopédia de genes e genomas de Kyotto (*KEGG*) que utiliza o agrupamento de proteínas ortólogas para realizar a caracterização funcional das proteínas visando elucidar e comparar os sistemas biológicos de cada organismo. O banco também oferece anotação automática de genomas com base nos grupos de genes ortólogos (KANEHISA e GOTO, 2000).

1.7 FILOGENIA

A classificação é uma das mais fundamentais preocupações da ciência. Fatos e objetos devem ser organizados, a fim de seus princípios serem descobertos e usados como base de extrapolação (SNEATH e SOKAL, 1973). A filogenia é a ciência de inferência do passado evolucionário de um indivíduo ou grupo. A filogenia teve inicialmente como base o estudo de diferenças morfológicas, fisiológicas e de comportamento. O entomólogo alemão Willi Henning (1913-1976) foi fundamental no avanço da filogenia sistemática, redefinindo os objetivos da área além da inserir novas ferramentas de estudo, como cladogramas, dendogramas (HENNIG e DAVIS, 1999). Posteriormente Sneath e Sokal apresentaram pela primeira vez métodos quantitativos para a classificação dos organismos (SNEATH e SOKAL, 1973).

Na filogenia os relacionamentos entre os organismos são representados por árvores filogenéticas compostas por nós e ramos. Ramos conectam nós e o nó é um ponto a partir do qual dois ou mais ramos divergem. Um nó interno corresponde a um último ancestral comum, a origem. Nós terminais correspondem às sequências a partir das quais a árvore foi gerada. Uma árvore filogenética pode ser construída utilizando famílias de genes diferentes (árvores de genes), por um único gene de várias espécies (árvores de espécies) ou pela combinação de ambos. No primeiro caso, os nós internos correspondem à duplicação gênica e no segundo, a eventos de especiação. Um nó e todos os ramos subsequentes a este nó são chamados de grupo monofilético ou holofilético, derivados de um único ancestral, compartilhando suas características únicas (Figura 7A). Um grupo, excluindo alguns de seus descendentes, é chamado de grupo parafilético (ex: excluindo a e b do grupo, Figura 7B) e grupos descendentes de mais de um ancestral são chamados de polifiléticos (Figura 7C). O modelo de árvore filogenética está representado na Figura 6 (BALDAUF, 2003).

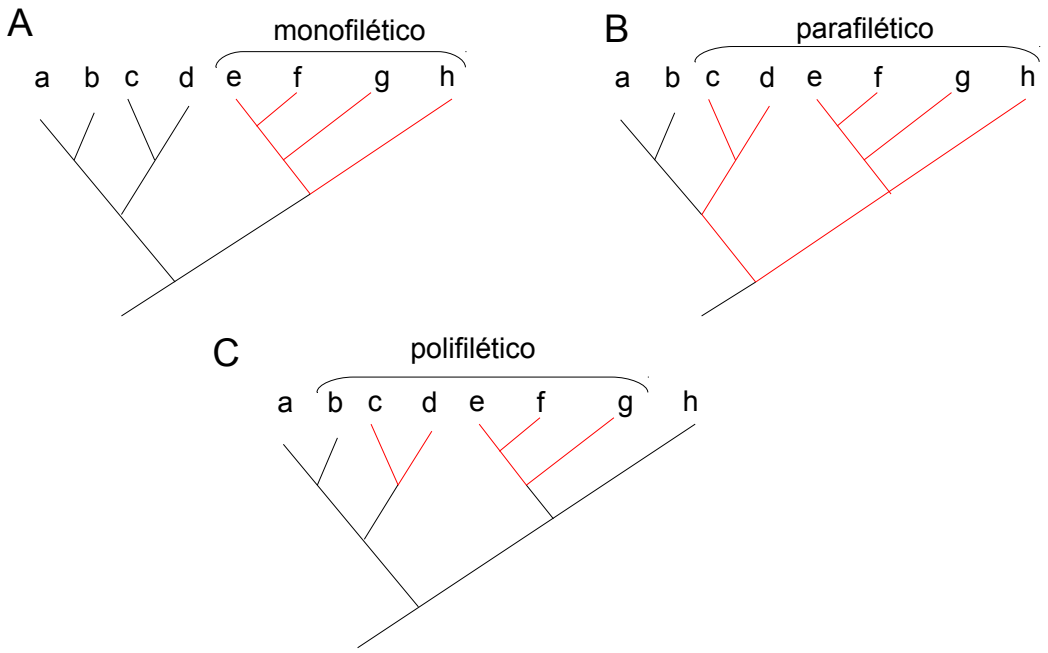


FIGURA 7 – MODELO DE ÁRVORE FILOGENÉTICA.
A- Representa um grupo monofilético.
B- É demostrado um grupo parafilético, devido a exclusão de um ramo do grupo.
C- Está representado um grupo polifilético que descendem de mais de um ancestral.
FONTE: Adaptado de BALDAUF, 2003.

Os métodos de construção de árvores filogenéticas estão divididos em duas categorias: matriz de distância e dados discretos, também conhecidos como métodos de procura de árvore. Como exemplos de métodos de matriz de distância destacam-se *UPGMA* (*Unweighted Pair Group Method with Arithmetic Mean*), (SNEATH e SOKAL, 1973) *neighbour-joining* (SAITOU e NEI, 1987), *Fitch–Margoliash* (FITCH e MARGOLIASH, 1967); e de dados discretos destacam-se, *maximum parcimony* (DAY, 1987), *maximum likelihood* (FELSENSTEIN, 1981) e o método bayesiano (RANNALA e YANG, 1996). Enquanto os métodos de matrizes utilizam cálculos estatísticos para determinar a distância entre cada combinação de pares de nós para então agrupá-los em uma árvore, os métodos discretos examinam separadamente cada coluna lida de alinhamento buscando a árvore que melhor acomode toda informação. Apesar dos métodos de distância serem mais rápidos, os métodos discretos são mais ricos em informação gerando resultados mais confiáveis (BALDAUF, 2003).

O teste mais comum para definir limites de confiança para árvores filogenéticas é o valor de *bootstrap*. A ideia envolve inferir a variabilidade em uma distribuição não conhecida na qual o dado é analisado através de subconjuntos aleatórios (FELSENSTEIN, 1985). Assim o cálculo do valor de *bootstrap* é realizado em três etapas: primeiramente o conjunto de dados é aleatoriamente amostrado e substituído, visando a criação de múltiplos conjuntos de dados aleatórios com o mesmo tamanho do original (Figura 8a). Em sequência, são calculadas árvores individuais para cada subconjunto junto com um valor de *score* dependendo de qual nó é obtido com mais frequência (Figura 8b). Com isso, o resultado é o consenso destas árvores aleatórias junto com o valor de *bootstrap*, calculado através dos *scores* individuais (Figura 8c) (BALDAUF, 2003). Por exemplo, se o padrão achado em todas as árvores dos subconjuntos o valor de *bootstrap* é 100%, em 2/3 o valor é de 67% (Figura 9c). O valor mínimo de *bootstrap* para garantir a confiabilidade da árvore é de 70% (HILLIS e BULL, 1993).

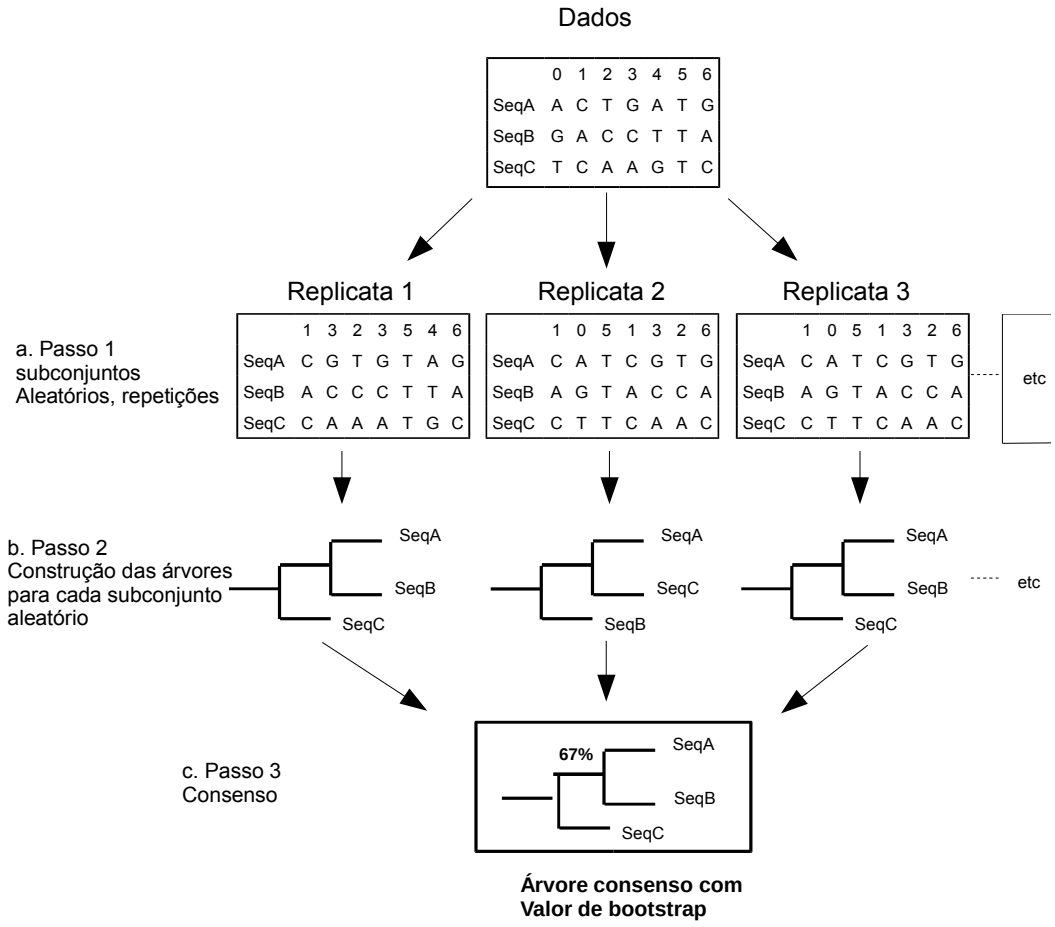


FIGURA 8 – DETERMINAÇÃO DO VALOR DE *BOOTSTRAP*.

a – Criação de múltiplos conjuntos de dados aleatórios com o mesmo tamanho do original.
b- São calculadas árvores individuais para cada subconjunto junto com um valor de *score* dependendo de qual nó é obtido com mais frequência.
c – O resultado é o consenso destas árvores aleatórias junto com o valor de *bootstrap*, calculado através dos valores de *bootstrap* individuais.
FONTE: Adaptado de (BALDAUF, 2003).

Com o advento do sequenciamento de DNA, a filogenia molecular, baseada na comparação de DNA ou sequências de proteínas, passou a ter grande papel na taxonomia (BALDAUF, 2003). Com isso, o estudo de diversidade e a determinação de relações filogenéticas passou a utilizar, além do gene 16S rRNA, todas as proteínas do genoma, mais especificamente os genes que estão presentes em todas as espécies do grupo analisado (*core* genes) (KAAS *et al*., 2012).

2 JUSTIFICATIVA

As bactérias do gênero *Herbaspirillum* possuem grande interesse biotecnológico como promotores de crescimento em plantas. Atualmente o gênero contém 11 espécies descritas e isoladas de diversos habitats como plantas, solo, água e humanos. A fim de facilitar sua utilização é muito importante conhecer as características destes organismos. Além de contribuir para a geração de mais informação a respeito do gênero, o sequenciamento do genoma da bactéria *H. lusitanum* P6-12 é relevante pelo fato desta bactéria ter sido isolada de nódulos de raiz de planta, indicando a presença de genes relacionados com a fixação de nitrogênio. A comparação do genoma de *H. lusitanum* P6-12 com as outras bactérias do gênero é essencial para o conhecimento das características do organismo. A anotação e classificação das proteínas destas espécies através de estudos de homologia é fundamental para analisar seu potencial biotecnológico. O estudo comparativo das proteínas também permite identificar as características adquiridas e perdidas através do tempo por estas bactérias, bem como determinar o seu relacionamento filogenético.

3 OBJETIVOS

3.1 Objetivo geral

O presente trabalho tem como objetivo a montagem anotação e análise comparativa da bactéria *H. lusitanum* P6-12 bem como o desenvolvimento de uma ferramenta de comparação genômica funcional e filogenética.

3.2 Objetivos específicos

- Sequenciar, montar e anotar do genoma de *H. lusitanum* P6-12;
- Estudar a homologia de sequências de nucleotídeos e aminoácidos entre *H. lusitanum* e as outras espécies de *Herbaspirillum*;
- Identificar grupos de genes ortólogos entre os genomas de *Herbaspirillum*;
- Determinar grupos de genes essenciais e acessórios para o gênero *Herbaspirillum*;
- Classificação funcional dos genes do *H. lusitanum* P6-12;
- Determinar as vias metabólicas do genoma de *H. lusitanum* P6-12;
- Determinar a filogenia das espécies de *Herbaspirillum*;
- Aplicar perfis de alinhamento de grupos ortólogos para a anotação dos genes de *Herbaspirillum*.

4 MATERIAL E MÉTODOS

4.1 Artemis

Ferramenta de navegação e anotação de genoma desenvolvida pelo Instituto Sanger (http://www.sanger.ac.uk/Software/Artemis/). Foi utilizada para revisar as anotações automáticas. O programa Artemis é escrito em Java, e está disponível para diversas plataformas, como UNIX, Macintosh e Windows. O programa utiliza sequências armazenadas em texto nos formatos EMBL e GENBANK ou FASTA (RUTHERFORD *et al*., 2000). Foi utilizado para realizar a revisão da anotação.

4.2 *Blast*

BLAST (do inglês, *Basic Local Alignment Search Tool*) identifica regiões de similaridade local entre as sequências. O programa *BLAST* pode ser usado para inferir relações evolutivas e funcionais entre as sequências, assim como para ajudar a identificar os membros de famílias proteicas (ALTSCHUL *et al*., 1990. O *BLAST* foi utilizado para comparar as sequências de nucleotídeos ou de proteínas entre os genomas do gênero, fornecendo a significância estatística dos alinhamentos tanto para anotação quanto para a determinação dos genes ortólogos. Foram considerados os alinhamentos com valor de E superior a 1.10^{10} e cobertura acima de 70% para garantir alta identidade entre as sequências.

4.3 *GAAT*

O programa *GAAT* é uma ferramenta de anotação de genomas. Ele utiliza um conjunto de programas específicos para a anotação, como *GLIMMER* (SALZBERG *et al*., 1998) e *tRNAscan* (LOWE e EDDY, 1997), fornecendo uma interface *web* para visualização e edição de genomas. Foi desenvolvido inicialmente para atender a rede GENOPAR de sequenciamento e anotação do genoma da bactéria *Herbaspirillum seropedicae* SmR1. Neste

estudo foi utilizado para anotar o genoma da bactéria *Herbaspirillum lusitanum* P6-12*.*

4.4 *Gapkiller*

O *script gapkiller* é um algoritmo escrito em *perl* (*Practical Extraction and Report Language*) e utiliza bibliotecas do pacote *bioperl*. O *pipeline* foi desenvolvido para auxiliar o fechamento de *gaps* dentro de sequências de *ORFs* (*Open read frames*) e utiliza o programa *BLAST, fastacmd* e *phredPhrap/Consed* como programas acessórios. Ele tem como finalidade descobrir se existem sequências dentro do banco de leituras do organismo, capazes de preencher a região faltante na *ORF*. O *pipeline* funciona da seguinte forma: a sequência da *ORF* é comparada ao banco de sequências do *NCBI* a fim de recuperar a sequência da proteína com maior identidade. Esta proteína é comparada ao banco de leituras do organismo utilizando o programa *BLAST* visando identificar as leituras capazes de reproduzir a região faltante. As leituras selecionadas são submetidas ao processo de montagem utilizando o pacote phred/phrap/Consed gerando uma sequência consenso. O programa Phrap é utilizado com os parâmetros padrões. Por fim esta sequência consenso é comparada novamente com o banco de dados NCBI para confirmar a obtenção da *ORF completa.* A estratégia pode ser visualizada na Figura 9.

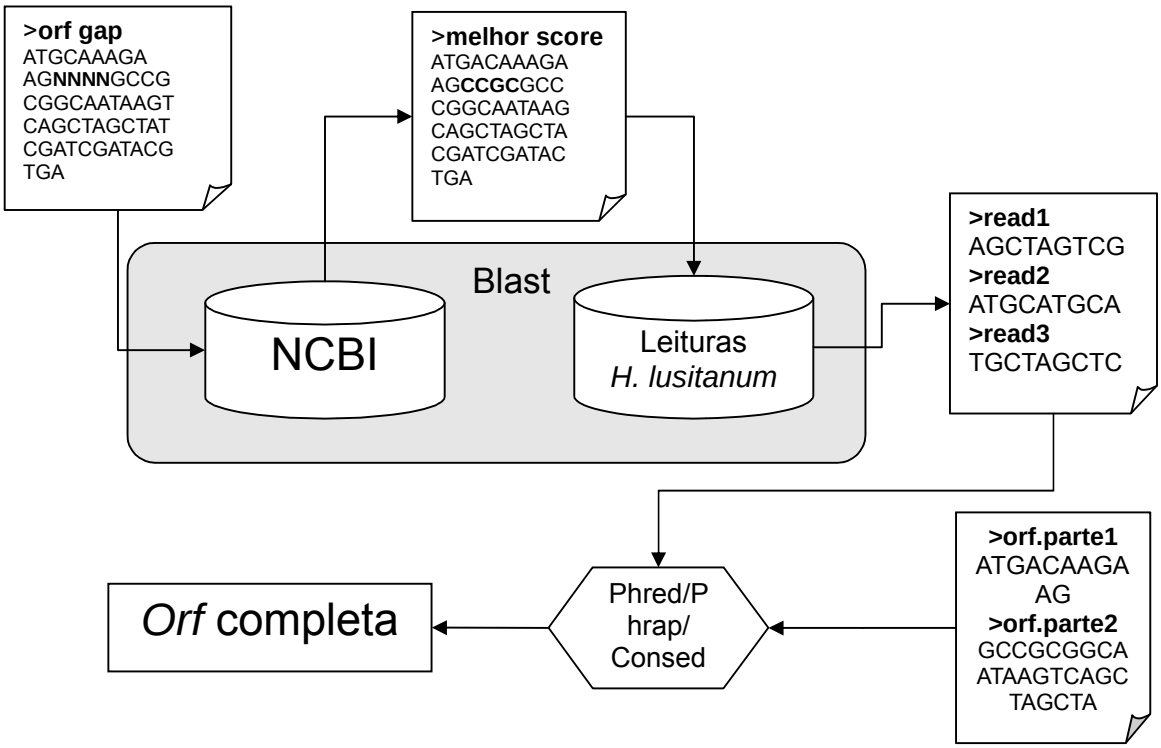


FIGURA 9- FLUXOGRAMA DE EXECUÇÃO DO *SCRIPT Gapkiller.*

4.5 Determinação do GCskew

O GCskew é uma medida que utiliza a característica dos genomas bacterianos de variar a quantidade de guanina e citosina no genoma (LOBRY *et al.,* 1996). Este fenômeno promove uma segregação do genoma em duas regiões, uma com excesso de guanina sobre citosina, caracterizando a fita líder; e a outra com excesso de citosina sobre guanina, caracterizando a fita atrasada. A região onde ocorre esta mudança de frequência, é relacionada com a origem de replicação (GRIGORIEV, 1998; FRANK e LOBRY, 1999). O GCskew é definido como o excesso normalizado de C sobre G, (C − G)/(C + G) calculado através de uma janela deslizante pelo genoma (LOBRY, 1996). O gráfico de GCskew cumulativo foi construído utilizando o programa Matlab, percorrendo janelas de 100 pares de bases pela sequência do genoma.

4.6 *Fgap*

O programa tem o objetivo de finalizar *drafts* de genomas através da fusão de diferentes versões de montagens ou através da incorporação de novas sequências (PIRO et al., 2014). O programa tem a capacidade de analisar a região de *gap* e as pontas de *contigs,* e indicar a melhor sequência para fechá-los. O *Fgap* utiliza o programa *BLAST* como programa acessório para identificar as possíveis sobreposições nas pontas de *contigs*. O algoritmo funciona da seguinte forma: os *gaps*, presentes na montagem na forma de um "N", são identificados e então comparados a um banco de dados disponibilizado pelo usuário. O banco pode ser de leituras ou *contigs*. Os alinhamentos entre a região do *gap* e a sequência candidata são analisados. Os parâmetros do alinhamento são ajustáveis pelo usuário e regulam a rigorosidade do alinhamento. Os parâmetros são três: 1- valor de corte mínimo de *score* do alinhamento, 2- Maior valor de *E-value*, e 3- mínimo valor de identidade. O *Fgap* utiliza estes parâmetros nesta ordem para restringir os alinhamentos apresentados, escolhendo a melhor combinação. Também pode ser definido a distância a montante quanto a jusante ao *gap* para ser analisada, e o resto é ignorado. O valor empregado para esta distância foi de 300pb quando utilizado o conjunto de *contigs* e 50pb quanto utilizado o conjunto de *ORFs*. Os parâmetros de alinhamento foram os valores padrões. A sequência obtida é substituída no *scaffold* no ponto que existe o *gap*, como mostra a figura 10.

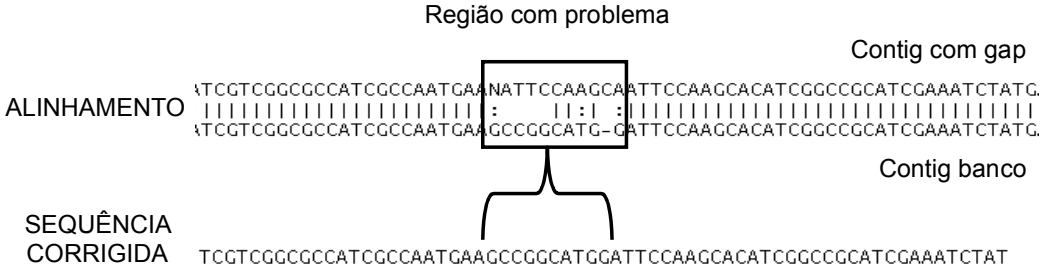


FIGURA 10- EXEMPLO DE CORREÇÃO DA REGIÃO DE *GAP.*

A figura demonstra o alinhamento entre a sequência com o *gap* e a sequência candidata. A região com o problema é substituída e o *gap* é resolvido.

4.7 *Phred/Phrap/Consed*

Phred/Phrap/Consed é um pacote de programas que utiliza arquivos de sequenciamento de DNA (eletroforetograma) para realizar a montagem do genoma. O programa *Phred* é responsável por fazer a chamada de bases e estimar a probabilidade de erro de cada base da sequência. O programa *Phrap* é um montador para sequências de DNA e utiliza parâmetros de alinhamento e as qualidades atribuídas pelo programa *Phred* para realizar a montagem do genoma em *contigs*, a partir das sobreposições entre as leituras (EWING *et al*., 1998). Junto com as sequências consenso, o programa fornece informações sobre a montagem ajudando no tratamento de eventuais problemas encontrados. A ferramenta *Consed/Autofinish* foi utilizada para visualizar a montagem criada pelo montador Velvet e do montador *Phrap*. Além da visualização foi utilizado para edição e fechamento dos *gaps*. Entre os recursos de tratamento de *gaps* destacam-se: a capacidade de escolha e sugestão de oligonucleotídeos iniciadores; a identificação de regiões com problemas de montagem, como baixa qualidade nas sequências; fornece sugestões de sequenciamentos adicionais e agrupa os *contigs* em *super-contigs* (*contigs* ordenados por evidências de ligação) (GORDON; ABAJIAN e GREEN, 1998). Foi utilizado junto ao *script gapkiller.pl* para a construção das *ORFs* com problema na montagem SOLiD.

4.8 Quality assessment (QA)

O programa *Quality assessment* é um programa de análise de qualidade para leituras do sequenciador SOLiD. O programa tem a capacidade de exibir graficamente estatísticas de qualidade como qualidade média e qualidade acumulada para cada base das leituras (RAMOS *et al*., 2011). O programa foi utilizado para definir valores de corte na seleção dos *reads* utilizados para a montagem SOLiD.

4.9 *SOLiD™ de novo accessory tools 2.0*

SOLiD™ de novo accessory tools 2.0 é um *pipeline* desenvolvido pela Life Technologies para a montagem automática do genoma exigindo a mínima interação com o usuário. Ele foi desenvolvido para a otimização do uso das leituras de sequências em *color-space* gerados pelo sequenciador SOLiD. Ele utiliza a alta capacidade de sequenciamento, o valor de qualidade das leituras e a vantagem do sistema de dupla leitura de cada base para obter o melhor resultado de montagem. O *pipeline* é executado em cinco passos integrados descritos a seguir e apresentados na Figura 11.

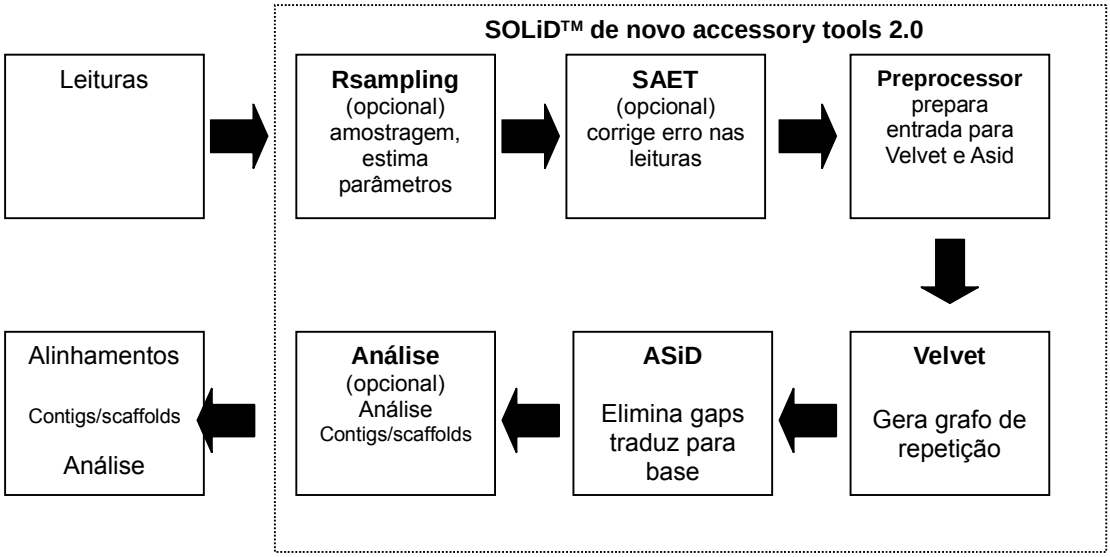


FIGURA 11- FLUXOGRAMA DE EXECUÇÃO DO PIPELINE *SOLID™ DE NOVO ACCESSORY TOOLS* 2.0.
FONTE: Adaptado de:
http://gsaf.cssb.utexas.edu/wiki/images/7/71/DeNovo_Assembly_Pipeline_2.0.pdf

O *pipeline* exige apenas que o usuário forneça a informação do tipo de biblioteca, o tamanho das leituras e o tamanho estimado do genoma estudado para a execução com os parâmetros padrões. Os programas que constituem o pipeline são:

1- *Rsampling:* é capaz de determinar precisamente a cobertura da montagem em leituras de sequências. Ele utiliza esta informação para otimizar o uso das leituras criando um subconjunto de 300x de cobertura quando o este

supera 600x. Isto é realizado para otimizar o consumo de memória do computador. Esta etapa é opcional.

2- *SAET* (*SOLiD™ Accuracy Enhancement Tool)*: é um corretor de erros de sequenciamento. Ele aproveita a alta cobertura do sequenciamento, os valores de qualidade e o sistema de dupla leitura de cada base para corrigir falhas de leitura quando a base não é lida pelo sequenciador, e erros de leituras, quando a base lida está errada. O algoritmo cria uma lista de todos os *k-mers* presentes nas leituras de sequências. Os *k-mers* com uma frequência maior que o valor de corte, ajustável pelo usuário, são considerados confiáveis. Assim as leituras de sequência são corrigidas para que existam apenas *k-mers* confiáveis. Em média, o SAET reduz a taxa de erro bruto por um fator de 5x, otimizando o processo de montagem.

3- *Preprocessor*: faz a conversão das leituras *color-space* para *double-encoded,* formato de entrada requisitado pelo montador *Velvet*. O formato *double-encoded* é a representação das leituras de sequências em *color-space* em bases sem a perda da dupla interrogação.

4- *Velvet:* é o montador usado para gerar a montagem. Como resultado ele gera os *contigs* ou *super-contigs* da montagem no formato *double-encoded*.

5- *AsiD* (*Assembly Assistant for SOLiD™ System tool*): é responsável pela finalização do processo de montagem. Ele possui a função de fechamento de *gaps* entre *super-contigs* e a conversão da montagem de *color-space* para *bases.* O fechamento de *gap* é realizado através de uma mini-montagem. O conjunto de *leituras* usado é selecionado se cumprir a condição de ter seu par dentro da região de vizinhança do *contig*. Se o *contig* gerado com estas leituras tiver sobreposição ao par de *contigs* que flanqueiam o *gap*, então ele é considerado fechado e a região é incorporada.

6- Analisys: *script* responsável pela análise comparativa e estatística da montagem.

Este programa foi utilizado para construir as montagens *SOLiD.* Os parâmetros para a criação das montagens estão apresentados na Tabela 4.

TABELA 4 – PARÂMETROS PARA A CRIAÇÃO DAS MONTAGENS *SOLiD™ DE NOVO ACCESSORY TOOLS* 2.0.

Nome da Montagem		**k17**	**k19**	**k21/Frag21**
Parâmetro	**Função**	**Valor**		
-hsize	Tamanho do *k-mer* usado na matriz de alinhamento	17 pb[1]	19 pb	21 pb
-ins_length	Estimativa de tamanho do inserto	1000 pb	1000 pb	1000 pb
-ins_length_sd	Estimativa de variação to tamanho do inserto	500 pb	500 pb	500 pb
-tamanho do genoma	Tamanho estimado que o genoma deve ter	5.6 Mpb[2]	5.6 Mpb	5.6 M
-min_contig_lgth	Tamanho mínimo para um *contig* ser considerado	100 pb	70 pb	100
-exp_cov	Cobertura esperada	300x	300x	250x
-cov_cutoff	Tamanho mínimo de cobertura esperada para formar *contig*	20 pb	20 pb	17 pb
-min_pair_count	Número de *mate-pair leituras* para considerar uma ligação de super-contigs confiáveis	240	240	240

[1]pb=pares de bases.
[2]Mpb= Milhões pares de bases.

4.10 *Velvet* 1.2.10

O programa *Velvet* é uma solução para montagem de genomas utilizando sequências curtas (ZERBINO e BIRNEY, 2008). Ele utiliza grafos de Brujin para resolver as sobreposições e repetições no processo de montagem e é uma solução capaz de manipular uma enorme quantidade de dados gerados por sequenciadores como SOLiD, ILLUMINA e 454 (ZERBINO e BIRNEY, 2008). *Velvet* foi o montador utilizado para a construir as montagens SOLiD através do programa *SOLiD™ de novo accessory tools 2.0*.

4.11 Genomas utilizados

As sequências de aminoácidos foram extraídas dos genomas de 13 estirpes do gênero *Herbaspirillum,* descritas na Tabela 5.

TABELA 5 – CARACTERÍSTICAS DOS GENOMAS DAS ESTIRPES DE *Herbaspirillum.*

Organismo	Estirpe	Genes	Scaffolds	Genoma (Mb)	Código Genbank
Herbaspirillum sp.	JC206	4122	30	4,17	CAHF00000000
	CF444	5214	125	5,59	NZ_AJVC00000000
	GW103	4784	6	5,05	NZ_AKJW00000000.1
	YR522	5650	168	5,11	AKJA00000000.1
H. seropedicae	SmR1	4804	1	5,51	NC_014323
	Os34	5941	253	6,15	AMSB00000000
	Os45	5504	145	5,63	AMSA00000000
	14040	4804	45	5,43	Não Depositado
H. rubrisubalbicans	M1	4673	1	5,6	Não Depositado
H. frisingense	GSF30	5655	1545	4,74	NZ_AEEC00000000.1
H. lusitanum	P6-12	4488	30	4,9	Não Depositado
H. huttiense putei	IAM 15032	5684	29	5,77	ANJR00000000

Os tópicos 3.16 ao 3.39 foram realizados durante o período de Doutorado Sanduíche na França, em colaboração com os professores doutores Alain Denise e Olivier Lespinet do grupo de Bioinformática do Instituto de Biologia molecular (IGM) da *Université Paris-Sud XI,* Orsay, França.

4.12 Identificação dos grupos de genes ortólogos

Existem diferentes metodologias para a identificação de grupos ortólogos baseados em árvores filogenéticas, grafos ou híbridos, porém nenhuma delas aparenta ser completamente infalível (KUZNIAR *et al*., 2008). Por esta razão foi utilizada a intersecção de três métodos conhecidos, (*BRH*, *INPARANOID* e *OrthoMCL*) para a obtenção dos grupos ortólogos confiáveis (GROSSETÊTE; LABEDAN e LESPINET, 2010). Após a intersecção foi realizado um passo adicional visando comparar o perfil de cada grupo com cada proteína que foi excluída previamente como resultado da intersecção.

Esta etapa serviu para verificar a inserção e exclusão das proteínas nos grupos (Figura 12). Cada método será explicado detalhadamente a seguir.

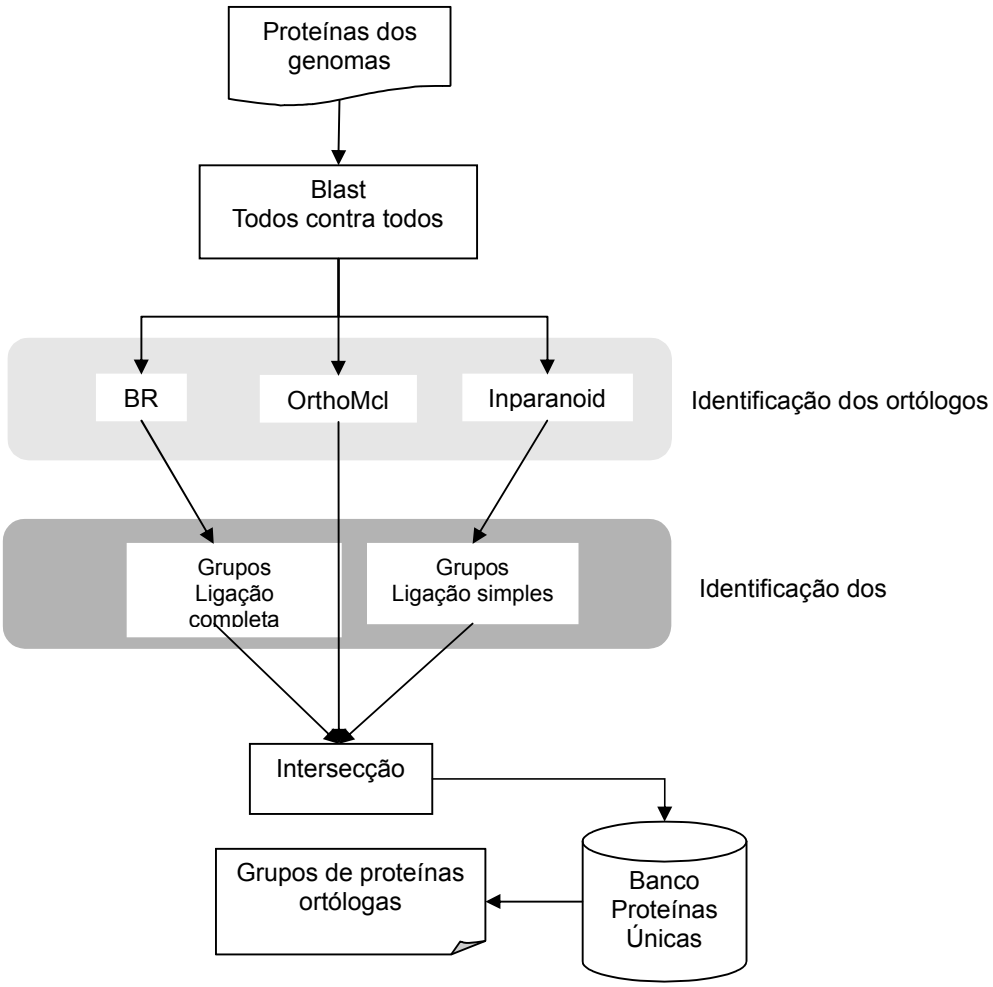


FIGURA 12- FLUXOGRAMA DE IDENTIFICAÇÃO DOS GRUPOS DE GENES ORTÓLOGOS.

As proteínas dos genomas são comparadas entre si através de similaridade de sequência utilizando o programa Blast. O resultado é utilizado pelos métodos de determinação de ortólogos BRH, OrthoMcl e Inparanoid. A criação dos grupos é obtida utilizando as metodologias de agrupamento de ligação simples e ligação completa. Os grupos iniciais são obtidos através da intersecção dos métodos. É criado um perfil Hmm de cada grupo obtido pela intersecção. As proteínas não contidas na intersecção foram reinseridas aos grupos através da comparação com os perfís Hmm utilizando o valor de corte de E de 1e-10.

4.13 Busca de similaridade entre sequências usando o BLAST

As sequências de aminoácidos das proteínas foram armazenadas em arquivos multifasta distintos, a fim de separar as proteínas de cada genoma. A busca de similaridade entre as sequências foi conduzida utilizado o programa BLAST+ 2.2.25. Não foi aplicado nenhum parâmetro de corte como *score* ou *E-value,* visando obter todos os alinhamentos encontrados. Os resultados foram armazenados em formato tabular e utilizados como informação de entrada para os programas *INPARANOID* e *OrthoMCL* e *BRH*.

4.14 Melhores alinhamentos recíprocos *BLAST* (*BRH*)

Para a determinação dos alinhamentos bidirecionais *BLAST* (*BRH*, do inglês *BLAST Reciprocal Hits*) entre as proteínas, cada proteína de cada genoma foi comparada com todas as outras e com ela mesma (todos contra todos) (Figura 14). O objetivo de comparar a sequência com ela mesma visa identificar o valor ideal de *score* para o alinhamento. Foram preservados os pares de proteínas que tiveram reciprocidade no alinhamento e obedeceram a 4 condições: 1- Obter o mesmo melhor par, independente de estar na situação de sequência de entrada ou banco; 2- Ser seu melhor par quando analisada contra ela mesma; 3- Razão acima de 60% entre o tamanho do alinhamento e o tamanho total da proteína identificada no banco, com o objetivo de minimizar a conservação local; 4- Foram excluídos alinhamentos onde a razão entre o valor do *score* do alinhamento e o valor de *score* total for inferior a 20%.

		Entrada	
		A	B
Banco	A	AA	AB
	B	BA	BB

FIGURA 13- EXEMPLO DE MATRIZ DE ALINHAMENTO *BLAST* TODOS CONTRA TODOS.
A é a proteína do genoma A e B é a proteína do genoma B. AA é o resultado do alinhamento *BLAST* da proteína A com ela mesma. AB é o resultado do alinhamento *BLAST* da proteína A com B. BA é o resultado do alinhamento *BLAST* da proteína B com A e BB é o resultado do alinhamento *BLAST* da proteína B com ela mesma.

4.15 *INPARANOID*

O programa *INPARANOID* 4.1 permite a identificação de pares de ortólogos entre dois conjuntos de dados. O programa foi executado com os parâmetros padrões (REMM; STORM e SONNHAMMER, 2001).

Para determinação dos grupos o programa utiliza o valor de similaridade entre as sequências (*score*) do alinhamento gerado pelo programa *BLAST* e um valor de confiança calculado sobre o alinhamento (REMM; STORM e SONNHAMMER, 2001). O alinhamento com o maior *score* determina o par de ortólogos, porém outros genes podem ser agrupados sendo identificados como parálogos. O valor de confiança representa o nível de relacionamento entre uma proteína e seu principal ortólogo, em uma escala de 0% a 100%, calculado com a seguinte fórmula.

Valor de confiança para A = 100%

$$x\frac{(scoreAAp - scoreAB)}{(scoreAA - scoreAB)}$$

Valor de confiança para B = 100%

$$x\frac{(scoreBBp - scoreAB)}{(scoreAA - scoreAB)}$$

Onde Ap é um in-parálogo do conjunto A e Bp é um in-parálogo do conjunto B. A é o principal ortólogo do conjunto A e B é o principal ortólogo do conjunto B. *Score* é o valor de similaridade entre as proteínas.

Os resultados obtidos pelo programa *BLAST* foram processados pelo programa *INPARANOID* 4.1.

4.16 *OrthoMCL*

OrthoMCL 2.0 é um programa para agrupamento de genes ortólogos entre genomas (LI; STOECKERT e ROOS, 2003). Ele utiliza o algoritmo de criação de grupos de *Markov* (*MCL*), que é baseado no modelo probabilístico de *Markov* e na teoria de grafo. Ele permite classificação simultânea de

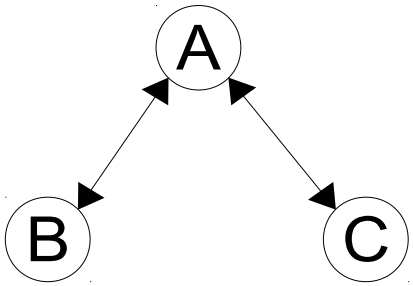
relacionamentos globais em um espaço de similaridade. *MCL* simula caminhos aleatórios pelo grafo usando matrizes de *Markov* para determinar as possibilidades ou probabilidades de transições entre os nós do grafo. A matriz de similaridade utilizada para a determinação dos grupos é criada utilizando os valores de *score* do programa BLAST (LI; STOECKERT e ROOS, 2003). Não é necessário utilizar nenhum processamento adicional de agrupamento, pois o programa já fornece os grupos formados. Os resultados obtidos pelo programa *BLAST* foram concatenados em um único arquivo para servir de entrada ao programa que permite a identificação de grupos de ortólogos entre mais de 2 genomas. As tabelas de resultados foram carregadas em um banco de dados MySQL 5.0. O programa foi executado com os parâmetros padrões.

4.17 Agrupamento dos alinhamentos bidirecionais *BRH* e *INPARANOID*

Os grupos de genes ortólogos foram obtidos utilizando dois modelos de agrupamento Ligação Simples e Ligação Completa que serão descritos a seguir.

4.17.1 Ligação simples

O método de ligação simples é um método permissivo de agrupamento, que utiliza a característica de transitividade entre as proteínas ortólogas, isto é, se uma proteína A é próxima a B, e B é próxima a C, então A também será próxima a C. O método permite o agrupamento dos pares de proteínas exigindo apenas a ligação de uma única proteína (Figura 14). Esta abordagem foi utilizada para identificar o maior número de grupos e adotado no resultado do programa *INPARANOID*.

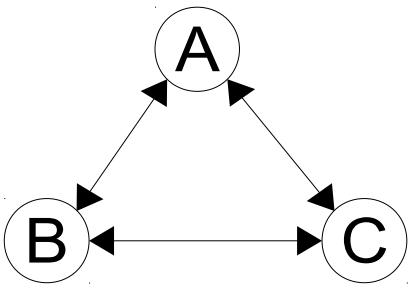


	A	B	C
A	1	1	1
B	1	0	0
C	1	0	0

FIGURA 14- REGRA DE AGRUPAMENTO DE LIGAÇÃO SIMPLES ENTRE PROTEÍNAS COM BASE EM SIMILARIDADE DE SEQUÊNCIA.
A proteína A é ortóloga às proteínas B e C. As proteínas B e C são ortólogas apenas à proteína A e não entre si. O grupo ABC foi formado apenas com base em uma única ligação, a proteína A. Visualização da matriz de comparação. Valor 1 na existência da ligação e 0 para a ausência.

4.17.2 Ligação completa

O método de ligação completa permite o agrupamento dos pares apenas se existir reciprocidade entre todas as proteínas do grupo, levando em consideração a propriedade transitiva dos genes ortólogos. Para um par de homólogos ser adicionado ao grupo, cada proteína deve ser ortóloga de cada proteína do grupo como pode ser observado na Figura 15.



	A	B	C
A	1	1	1
B	1	1	1
C	1	1	1

FIGURA 15- REGRA DE AGRUPAMENTO DE LIGAÇÃO COMPLETA ENTRE PROTEÍNAS COM BASE EM SIMILARIDADE DE SEQUÊNCIA.
A proteína A é ortóloga às proteínas B e C. As proteínas B e C são ortólogas apenas à proteína A. As proteínas B e C são ortólogas entre si. O grupo foi formado respeitando a reciprocidade de todos os pares. Visualização da matriz de comparação. Valor 1 na existência da ligação e 0 para a ausência.

4.18 Intersecção dos resultados *BRH*, *Iparanoid* e *OrthoMCL*

O arquivo de intersecção contendo os grupos de ortólogos foi criado através da intersecção dos 3 métodos.

Esta etapa adequou os grupos à intersecção diminuindo o seu tamanho. Isto permitiu a criação de um conjunto de ortólogos padronizados com maior confiabilidade, pelo fato de serem obtidos pelos 3 métodos.

4.19 Recuperação das proteínas discrepantes entre os métodos *BRH*, *INPARANOID* e *OrthoMCL*

Cada abordagem empregada gerou um resultado distinto. O consenso dos resultados foi obtido utilizando a intersecção dos três métodos, *BRH*, *INPARANOID* e *OrthoMCL*. Devido a esta etapa, alguns grupos perderam proteínas. Como estas proteínas foram identificadas pelo menos por um método, surgiu o questionamento se elas seriam realmente ortólogas e algum método falhou em identificá-las. Para responder à dúvida, as proteínas foram comparadas com os grupos novamente seguindo a estratégia descrita na Figura 16.

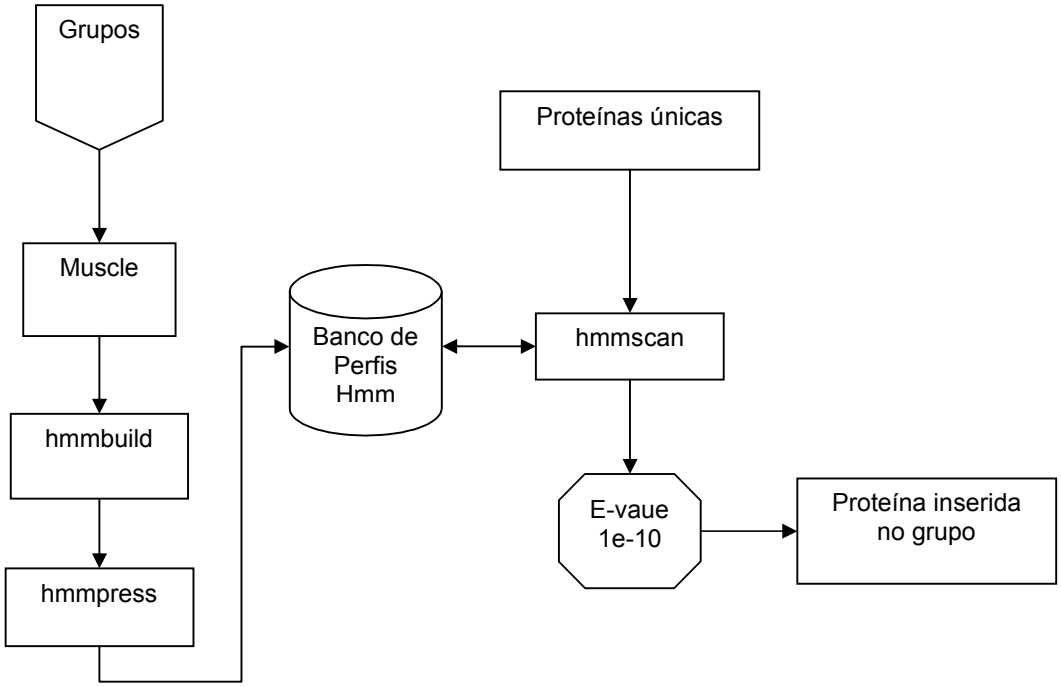


FIGURA 16- RECUPERAÇÃO DE PROTEÍNAS ORTÓLOGAS APÓS INTERSECÇÃO DOS MÉTODOS *BRH, IMPARANOID* E *ORTHOMCL*.

As proteínas de cada grupo foram alinhadas entre si usando o programa Muscle3.8.31. *Hmmbuild* foi utilizado para construir o perfil do alinhamento múltiplo. *Hmmpress* transforma o banco de perfis em formato binário para a busca de similaridade utilizando o programa *Hmmscan*. O parâmetro *E-value de* 1E-10 foi utilizado como limiar de corte.

4.20 Alinhamento múltiplo dos grupos ortólogos utilizando do software MUSCLE 3.8.31

O software Muscle 3.8.31 é um programa de alinhamento múltiplo de sequência de DNA, RNA ou proteína (EDGAR, 2004). Ele foi utilizado para obter o alinhamento múltiplo das proteínas de cada grupo.

4.21 Construção dos perfis *HMMER* e inserção de proteínas

O programa *HMMER* 3.1 foi utilizado para a criação dos perfis de alinhamentos múltiplos. Estes perfis são a representação do alinhamento em um sistema de score posição-específico adequado para identificar sequências homólogas distantes, com baixo grau de similaridade. (EDDY, 1998). *Hmmbuid*

foi utilizado para construir o perfil do alinhamento múltiplo criado pelo programa *MUSCLE* 3.8.31. *Hmmpress* transformou o banco de perfis em formato binário. Após esta etapa cada proteína foi testada contra todos os grupos. Foi utilizado o parâmetro de corte *E-value* de 1e-10 para decidir se a proteína seria ou não inserida.

4.22 Anotação funcional das proteínas

O conjunto de dados utilizado contém genomas completos e fragmentados de diferentes origens disponíveis no *GenBank*. Portanto, a anotação destes genomas varia quanto a qualidade, sendo revisada ou não. Partindo da premissa de que os genes que formam o grupo são ortólogos, o esperado seria a mesma função para todos. Por isso os grupos foram anotados novamente a fim de verificar discrepâncias entre a anotação, enriquecê-la, bem como confirmar os grupos encontrados. Os genomas foram comparados contra o banco de dados *KEGG*, *Refseq* e S*wissprot*.

4.23 Determinação dos genes centrais e acessórios de cada gênero

Os grupos de genes centrais (*core*) foram identificados dentro dos grupos de genes ortólogos. Os grupos com genes presentes em todas as espécies foi denominado como *core*. O restante foi classificado como genes acessórios.

4.24 Criação das árvores filogenéticas

As árvores filogenéticas do gênero *Herbaspirillum* foram construídas utilizando duas abordagens distintas. A árvore utilizando somente os grupos de genes *core* foi construída através da concatenação dos alinhamentos dos grupos. Foi utilizado o método *ML* (*Maximum likelihood*) implementado no programa *PhyMl* (GUINDON *et al*., 2010). A segunda árvore utilizou o método de super árvore filogenética. Foi construída uma árvore para cada grupo de ortólogos e então as árvores foram combinadas em uma super árvore através da metodologia descrita na Figura 17.

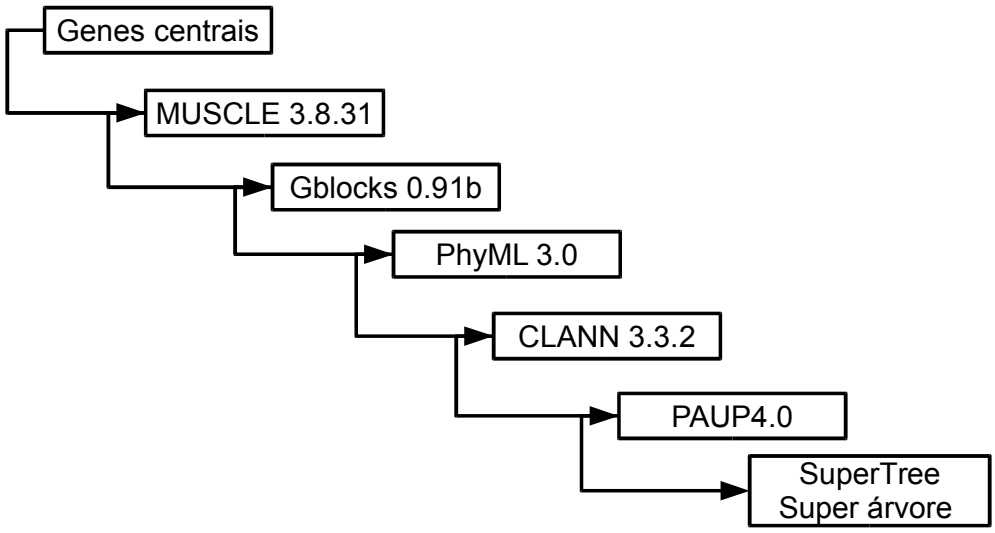


FIGURA 17- FLUXOGRAMA DE CRIAÇÃO DAS ÁRVORES.
Os grupos ortólogos foram alinhados utilizando o programa *MUSCLE* 3.8.31. Os alinhamentos foram filtrados através do programa *Gblocks* 0.91b. A árvore filogenética de cada grupo foi construída utilizando o programa *PhyML* 3.0. O programa *CLANN* 3.3.2 criou a matriz MRP. O programa *PAUP* 4.0 construiu a árvore.

4.25 Filtragem dos alinhamentos utilizando *GBLOCKS* 0.91b

A qualidade do alinhamento tem um grande impacto nas análises filogenéticas. Não apenas o alinhamento, mas também o método usado para tratá-las tem um papel decisivo na árvore final (TALAVERA e CASTRESANA, 2007). Deste modo o programa *Gblocks* foi utilizado para remover regiões divergentes e de baixa similaridade entre os grupos de genes ortólogos. Os blocos de alinhamento são escolhidos utilizando 5 valores de corte, IS, FS, CP, BL1, e BL2.

1- O grau de conservação de todas as posições do alinhamento múltiplo é avaliado. O valor padrão para IS e FS são ajustados para 50% do número de sequências + 1 e a 85 % do número de sequências, respectivamente.

2- Todos os trechos de posições contíguas não conservadas menores que CP são rejeitados. O valor padrão para CP é de 8 posições.

3- No restante dos blocos, as extremidades são examinadas e posições são removidas até que os blocos sejam rodeados por regiões de alta conservação. Dessa forma, blocos selecionados estão ancorados por posições que podem

ser alinhadas com confiança.

4- Somente blocos com comprimentos maiores ou igual a BL1 são mantidos a fim de evitar pequenas regiões em que a qualidade do alinhamento torna-se difícil de acessar. O valor padrão de BL1 é de 15 posições.

5- Todas as posições com lacunas (*gaps*) são removidas. Também são removidas posições não conservadas adjacentes às lacunas até que uma posição conservada é encontrada. Por fim, blocos pequenos remanescentes após remoção das lacunas também são removidos. O valor padrão BL2 é de 10 posições (TALAVERA e CASTRESANA, 2007). Os alinhamentos após o tratamento com o *Gblocks* foram utilizados para a construção das árvores filogenéticas.

4.26 Construção das árvores filogenéticas por grupo de genes ortólogos utilizando PHYML 3.0

PhyML é um software de filogenia baseado no princípio de máxima verossimilhança (do inglês; *maximum-likelihood*). Ele permite a verificação das mudanças entre os alinhamentos próximos (NNIs do inglês, *Nearest Neighboor Interchanges*) visando aumentar a confiabilidade do resultado. Permite o ajuste da intensidade do espaço da árvore (SPR, do inglês *Subtree Pruning and Regrafting*) e utiliza critérios de parsimônia para filtrar modificações topológicas com respeito a função de máxima semelhança (GUINDON *et al*., 2010). PhyMl 3.0 foi utilizado para a obtenção das árvores filogenéticas para cada grupo de genes ortólogos.

4.27 Obtenção da matriz de parsimônia resultante do agrupamento de todas as árvores utilizando o programa *CLANN* 3.32

A construção de super-árvores filogenéticas depende da combinação da informação contida nas árvores bases. O método pode combinar a informação desde que estas possuam um ramo comum (BAUM, 1992). *CLANN* 3.3.2 é um programa capaz de construir super-árvores utilizando diferentes abordagens tais como: Matriz de representação usando parsimônia *MRP (Matrix Representation using Parsimony*), super-árvore mais similar (*MSSA*, do inglês,

Most Similar Supertree) ajuste máximo do quarteto *QFIT (Maximum Quartet Fit*) e ajuste de separação máxima *SFIT* (*Maximum Splits Fit*) (CREEVEY e MCINERNEY, 2005).

A matriz de representação das árvores foi construída através do método *MRP* que utiliza como entrada uma coleção de árvores bases e as transforma em uma matriz binária com o objetivo de obter a árvore mais próxima às árvores bases, usando critério de parsimônia (LAPOINTE; WILKINSON e BRYANT, 2003). O resultado é chamado de super-árvore.

4.28 Criação da super-árvore filogenética utilizando o programa *PAUP**4.0

A super-árvore foi obtida utilizando o programa *PAUP**4.0. *PAUP**4.0 é um programa de filogenia que utiliza o princípio de máxima parsimônia para construção de árvores filogenéticas. O método de parsimônia procura por árvores que minimizem o número de mudanças evolucionárias para explicar um determinado dado sobre um conjunto pré especificado de mudanças de caráter admissíveis (SWOFFORD, 1991). O programa *PAUP**4.0 construiu a super-árvore utilizando como base a matriz de parsimônia construída pelo programa *CLANN* 3.3.2.

4.29 *MEGA* 5

MEGA é uma ferramenta integrada para a realização de alinhamento de sequências, inferindo árvores filogenéticas, estimando tempos de divergência, extraindo bases de dados online, estimando taxas de evolução molecular, inferindo sequências ancestrais, e testando hipóteses evolutivas (TAMURA *et al*., 2011). O programa foi utilizado para a edição das árvores filogenéticas.

5 RESULTADOS E DISCUSSÃO

5.1 Análise das sequências de *Herbaspirillum lusitanum* P6-12

A extração do DNA genômico e análise do gene 16S rRNA para confirmação da estirpe de *Herbaspirillum lusitanum* foi realizada pela Dra. Michelle Zibeti Tadra-Sfeir, do Departamento de Bioquímica e Biologia Molecular da Universidade Federal do Paraná. As sequências foram obtidas através de três sequenciamentos, dois utilizando o sequenciador SOLiD 4 (LifeTecnologies®) e um utilizando o sequenciador MiSeq (Illumina®). No primeiro sequenciamento, com o sequenciador SOLiD, foram obtidas 107.890.594 leituras de 50 pares de bases de comprimento, a partir de uma biblioteca de pares (*mate-pair*) (Tabela 6). No segundo sequenciamento (MiSeq Illumina®), foram obtidas 4.890.413 leituras a partir de uma biblioteca de fragmento *(single-end).* Uma terceira biblioteca de pares interligados *(paired-end)* foi construída e sequenciada no sequenciador MiSeq e foram obtidas 4.263.275 leituras. A preparação das bibiotecas e operação do sequenciador foi conduzida pelo Dr. Helisson Faoro, do Departamento de Bioquímica e Biologia Molecular da Universidade Federal do Paraná.

TABELA 6 – DADOS DE SEQUENCIAMENTO DO GENOMA DO *Herbaspirillum lusitanum* P6-12.

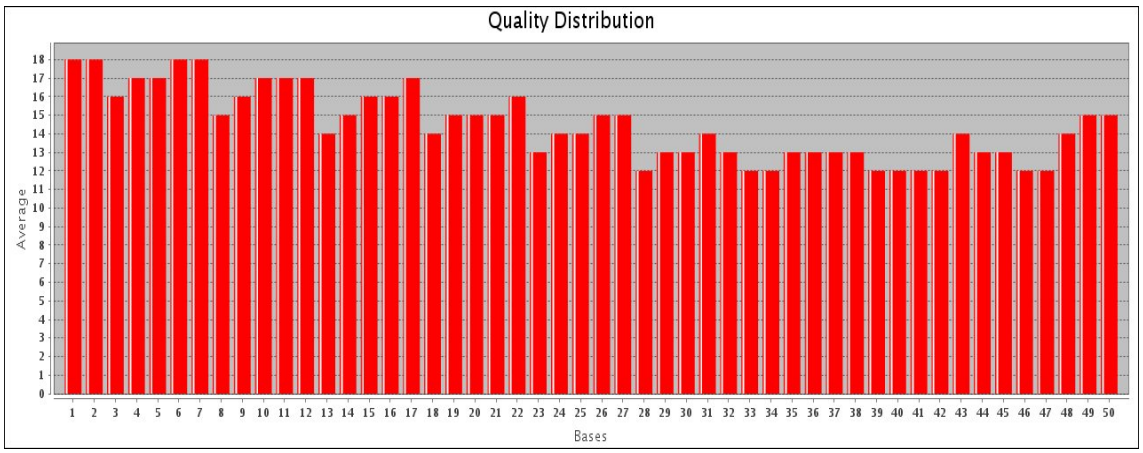
	SOLiD		**MiSeq**
	Single-end	*Mate-pair*	*Paired-end*
Número de leituras	4.890.413	107.890.594	4.263.275
Tamanho das leituas	50	50x2	250x2
Distância dos pares	-	2500	2500

5.1.1 Leituras SOLiD

O sequenciamento do genoma de *H. lusitanum* P6-12 com uso da plataforma SOLiD foi realizado a partir de duas bibliotecas diferentes. A biblioteca de *mate pair*, preparada a partir de fragmentos de 1000 pb gerou um total de 107.890.594 leituras de 50 pb de comprimento, totalizando

5.394.529.700 bases e proporcionando cerca de 900x de cobertura do tamanho previsto para o genoma. A biblioteca de fragmento gerou um total de 4.890.413 leituras, totalizando 227.490.799 bases e uma cobertura de cerca de 46x. O total de leituras geradas pela plataforma SOLiD foi de 112.781.007. A análise de distribuição de qualidade revelou um índice de qualidade médio que ficou abaixo de phred20 para todas as posições (Figura 18). De forma geral, a Figura 18A mostra uma distribuição média de qualidade superior à Figura 18B, que tem maior parte das posições com valores de qualidade abaixo de phred10. A baixa qualidade pode ter influenciado de forma negativa no resultado da montagem do genoma.

A



B

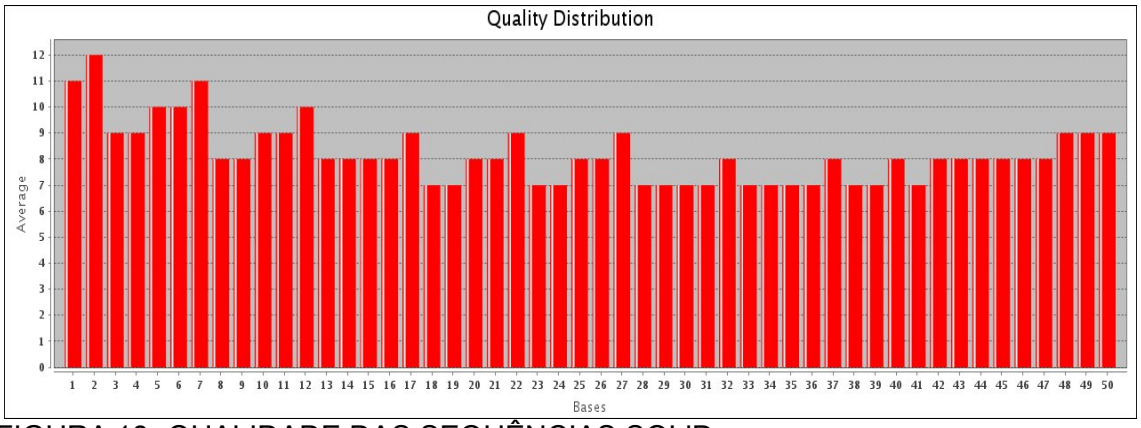


FIGURA 18- QUALIDADE DAS SEQUÊNCIAS SOLiD.

Sequências obtidas no sequenciador automático SOLiD. A média de qualidade foi inferior a phred20.
A- biblioteca de pares. Refere-se as duas sequências
B- biblioteca de fragmento.
Figura obtida através do programa *Quality assessment.*

As bases que não são lidas de forma confiável pelo sequenciador são substituídas por símbolos (“N”). No caso do conjunto de dados em *color-space*, essas bases indeterminadas são identificadas por um ponto (“.”). No SOLiD este é o resultado da falha de uma ligação no sequenciamento, impossibilitando a identificação da cor e consequentemente, impedindo a leitura. Como esta plataforma utiliza a informação da ligação anterior para determinar a leitura da próxima base, a penalização é maior quando comparada a outros sequenciadores. Isso acontece porque quando as leituras são decodificadas para bases, a informação que sucede o ponto é perdida.O número de leituras com bases indeterminadas nas duas bibliotecas foi de 2.907.538 (2,58%). Na construção da biblioteca são ligados adaptadores aos fragmentos de DNA para a PCR em emulsão. Estes adaptadores, se sequenciados, podem causar erros no processo de montagem. Foram identificadas 2.606.772 leituras com sequência de adaptadores. A fim de diminuir o erro nas leituras, foram eliminadas todas as leituras que apresentaram bases indeterminadas ou presença de sequência de adaptador. O total de leituras remanescentes foi de 107.266.697 (95%).

5.1.2 Leituras MiSeq

O sequenciamento na plataforma MiSeq com a biblioteca *paired-end* obteve 4.263.275 leituras e 871.464.565 bases permitindo 177x de cobertura. As leituras apresentaram uma variação de tamanho entre 35 a 251 como observado na Figura 19.

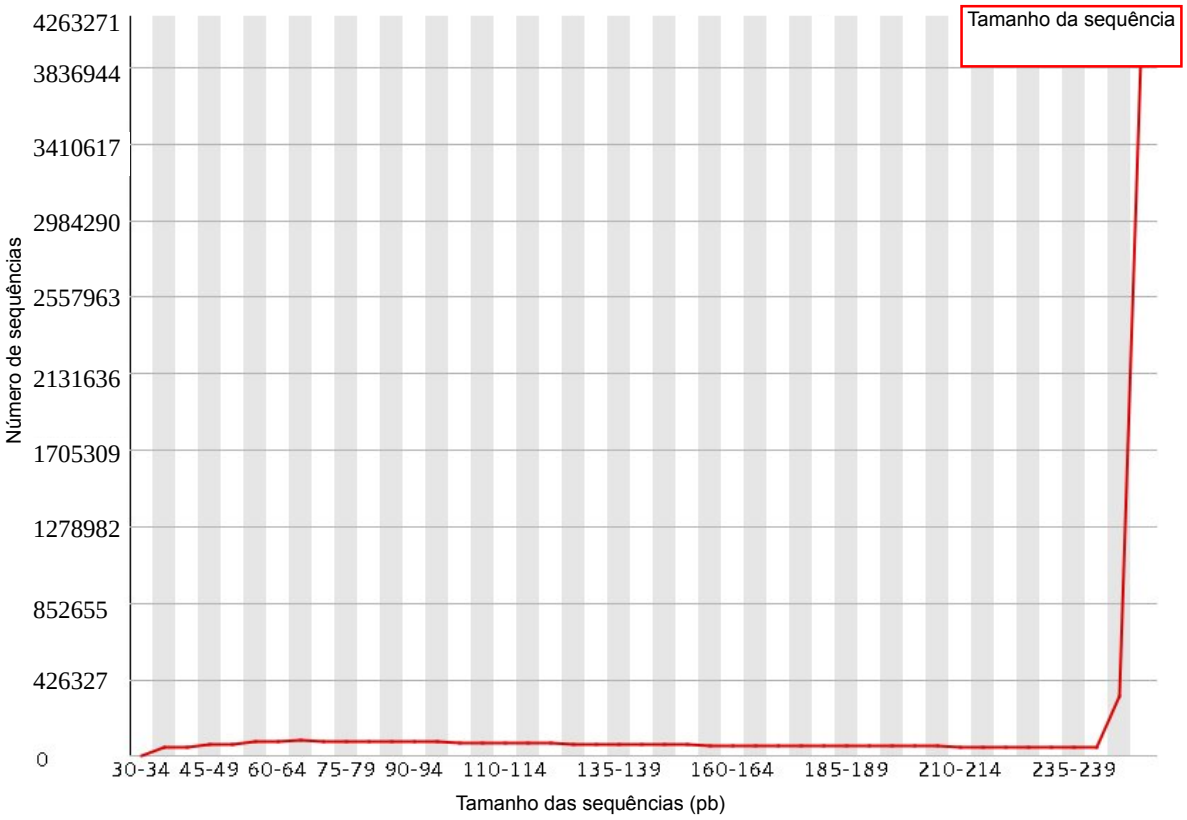


FIGURA 19- DISTRIBUIÇÃO DO TAMANHO DAS LEITURAS MiSeq.
O gráfico mostra a predominância de leituras longas como resultado do sequenciamento.
Gráfico obtido com o programa FASTQC.

A qualidade média das leituras MiSeq observadas oscilou entre phred25 a phred40 (Figura 20). A presença de bases indeterminadas foi baixa, menor que 1%.

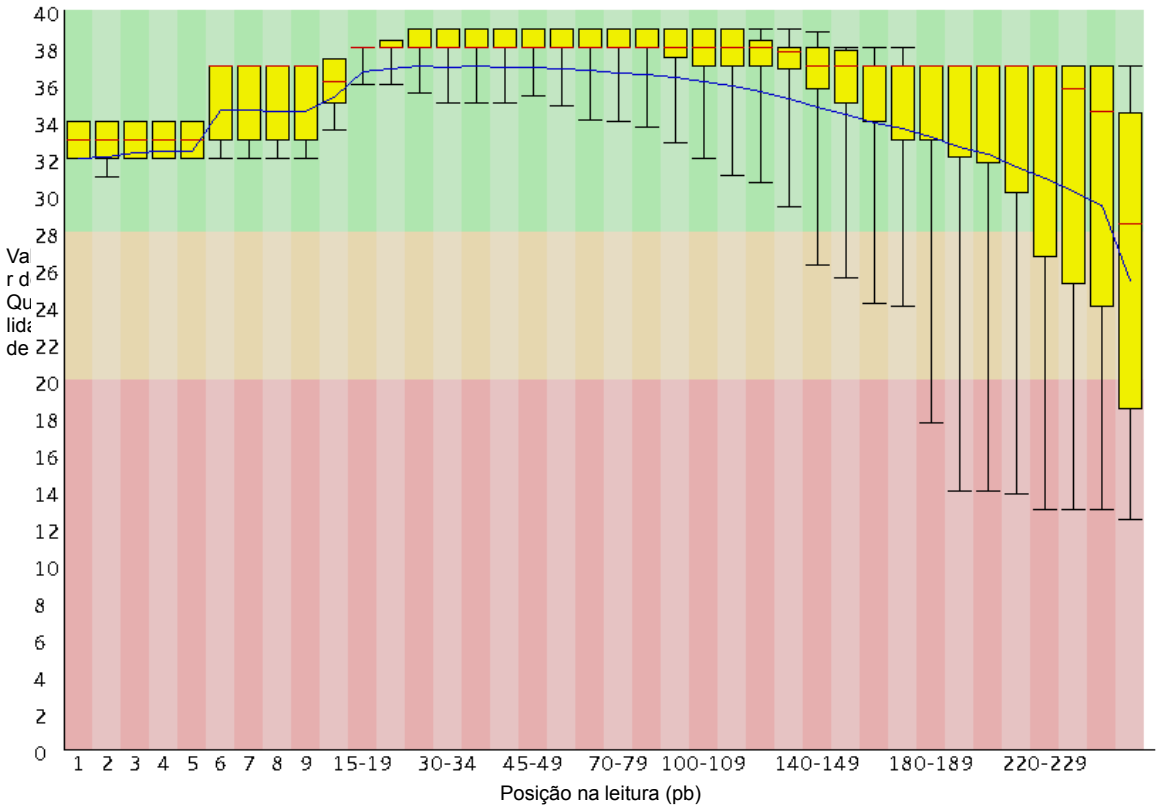


FIGURA 20- MÉDIA DE QUALIDADE POR BASE DAS LEITURAS MiSeq.
O gráfico mostra a média da qualidade das leituras por base. A barra mostra a variação de qualidade para cada base. Os valores médios variaram entre phred18 a phred39.
Gráfico obtido com o programa FASQC.

5.2 Montagens do genoma da bactéria *Herbaspirillum lusitanum* P6-12

5.2.1 Montagens a partir do sequenciamento com a plataforma SOLiD

A montagem do genoma do *Herbaspirillum lusitanum* P6-12 utilizou o programa Velvet, versão 1.1.04. Ele foi utilizado junto com o pipeline SOLiD™ *System de novo Accessory Tools* 2.0 desenvolvido pela *LifeTecnologies*®. Foram realizadas diferentes montagens visando otimizar os parâmetros do *pipeline* (Tabela 5). Os critérios utilizados para avaliação foram o maior valor de N50, o menor número de *contigs* e o tamanho final do genoma. Com isso a montagem com *k-mer* de 21, que apresentou o maior valor de N50 (250 Kb) foi escolhida como a base da montagem do genoma de *H. lusitanum* e chamada

de montagem k21 (Tabela 7).

TABELA 7 – RESULTADO OBTIDO NO TESTE DE *k-mers.*

	k-mer		
	k17	**k19**	**k21**
Contigs (pb)	7046	6513	6028
Super-contigs	2168	2500	1927
Média do Tamanho dos *Scaffods* (*pb*)	2277	1389	2595
N50 (pb)	104611	206273	252584
Maior *Scaffold* (pb)	249953	508975	563402
Maior *Contig* (pb)	8317	10637	16476
Tamanho estimado do Genoma (pb)	4936725	5234000	4807499

*pb=pares de bases.

Os c*ontigs* da montagem k21 com tamanho inferior à 1.000 pb foram retirados da montagem para serem utilizados no fechamento dos *gaps* e reinseridos na montagem. As leituras da biblioteca de fragmentos foram montadas separadamente, gerando 4.996 *contigs* e a montagem foi chamada de Frag21. Todos os *contigs* da montagem Frag21 mais os *contigs* inferiores a 1.000 pb da montagem K21 foram utilizados para o fechamento de *gaps* com o programa *Fgap* (Tabela 8).

TABELA 8 – COMPARAÇÃO DAS MONTAGENS ANTES DO FECHAMENTO DOS *GAPS.*

	Montagens		
	Frag21	**K21**	**K21up**
Contigs	4.996	6.028	2.156
Super-contigs	2.500	1.927	37
Média do Tamanho dos Scaffolds (pb)	303	2.595	2.595
N50	315	252.584	252.584
Maior *Scaffold* (pb)	-	563.402	563.402
Maior *Contig* (pb)	3.519	16.476	16.476
Contigs inferiores a 1000 (pb)	2.791	3872	0
Tamanho estimado do Genoma (pb)	1.516.969	4.807.499	4.846.950

*pb=pares de bases.

Antes do início da etapa de fechamento dos *gaps*, foi realizada a anotação automática da montagem k21up. Após a anotação, os *gaps* foram divididos em 3 categorias: G1 – *gaps* no interior de *ORFs* anotadas, G2 – *gaps* em regiões intergênicas, G3 – *gaps* de pontas de *super-contigs;* como apresentado na Figura 21.

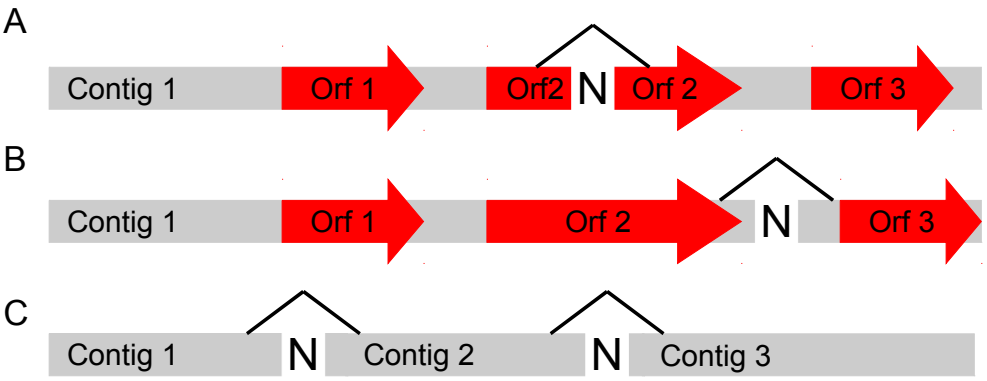


FIGURA 21- CATEGORIAS DE GAPS NO GENOMA DE *H. lusitanum.*
A figura apresenta o esquema gráfico dos *gaps* no genoma.
A – Gaps G1 no interior de *orfs* anotadas.
B – Gaps G2 em regiões intergênicas.
C– Gaps G3 de pontas de *super-contigs.* Os *gaps* estão representados pela letra N.

Os *gaps* G1 foram resolvidos utilizando o *script gapkiller.pl*. Foram identificadas 1.449 regiões G1 e a abordagem permitiu resolver 666 delas. Os *gaps* G2 de regiões intergênicas foram resolvidos utilizando o programa *Fgap* e nesta etapa foram resolvidos 670 *gaps*. Os *gaps* de pontas de *super-contigs* G3 foram analisados e apresentam indícios de ligação, porém nenhum foi fechado (Figura 22).

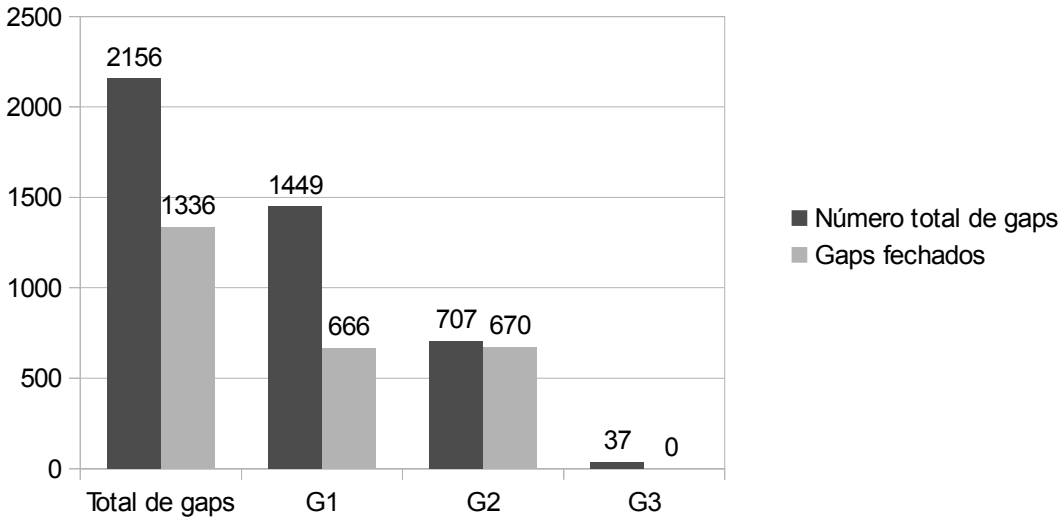


FIGURA 22- FECHAMENTO DOS *GAPS* DA MONTAGEM K21up DO GENOMA DE *H. lusitanum*.
Em cinza escuro está representado o total de *gap* de cada categoria e em cinza claro o número de *gaps* fechados após a aplicação das estratégias de fechamento de *gaps*. G1 – *gaps* em *orfs* anotadas G2– *gaps* em regiões intergênicas G3– *gaps* de pontas de *contigs*.

Ao todo foram resolvidos 1.336 *gaps*, representando 62 % to total de 2156. O número de *contigs* diminuiu para 820 considerando os *contigs* maiores ou iguais a 1000. A montagem foi anotada automaticamente utilizando o programa RAST. Foram identificados 5.240 *ORFs* do cromossomo cobrindo 84%, 38 tRNAs e um operon ribosomal 16S-23S-5S. O tamanho do genoma obtido foi de ~4.9Mb (WEISS *et al*., 2012). Os dados desta montagem estão resumidos na Tabela 9 abaixo.

TABELA 9 – ESTATÍSTICAS DA MONTAGEM SOLiD K21up APÓS FECHAMENTO DOS *GAPS.*

Tamanho (pb*)	4.919.393
Contigs	820
Super-contigs	37
Genes	5.240
tRNAs	38
G+C	60.2%
rRNA 16S-23-5S	1
Cobertura (pb*)	214x

*pb=pares de bases. Fonte: Adaptado de WEISS *et al*.,2012.

5.2.2 Montagem híbrida utilizando dados da montagem SOLiD e MiSeq

Uma montagem híbrida (denominada HI01) foi obtida utilizando a união dos conjuntos de leituras *SOLiD* e *Illunima MiSeq*. Esta montagem foi construída com o pacote *CLC Wokbench 5* e as estatísticas estão resumidas na Tabela 10.

TABELA 10 – CARACTERÍSTICAS DA MONTAGEM HI01.

Contigs	55
Super-contigs	31
Média do Tamanho dos Scaffolds (pb*)	11.201
N50 (pb*)	1.004.417
Maior *Scaffold* (pb*)	1.767.941
Maior *Contig* (pb*)	197.000
Cobertura em bases (pb*)	851x
Tamanho estimado do Genoma (pb*)	4.930.839

*pb=pares de bases.

Os *contigs* da montagem k21up e os contigs inferiores à 1.000 pb da montagem HI01 foram utilizados no programa *Fgap* para resolver os *gaps*. Só foi possível resolver os *gaps* internos, G1 e G2, diminuindo o número de *contigs*, mas não de *super-contigs*. A montagem HI01 foi definida como a segunda versão da montagem no genoma de *H. lusitanum* P6-12 e suas características estão descritas na Tabela 11.

TABELA 11 – CARACTERÍSTICAS FINAIS DA MONTAGEM HI01.

	Pré-fechamento dos *gaps*	Pós-fechamento dos *gaps*
Contigs	55	31
Super-contigs	31	31
Média do Tamanho dos *super-contigs* (pb*)	11.201	11.201
N50 (pb*)	1.004.417	1.004.417
Maior *super-contig* (pb*)	1.767.941	1.767.941
Maior *contig* (pb*)	197.000	197.000
Cobertura em bases	851x	851x
Tamanho do Genoma	4.930.839	4.919.496

*pb=pares de bases.

A correlação entre as duas montagens pode ser observada no gráfico *dotplot* da Figura 23. Ele mostra que as montagens possuem sintenia mas que a montagem k21up apresentou algumas incongruências de posicionamento na ordenação dos *contigs.* Esta dificuldade, provavelmente, é consequência do alto número de *contigs* presente na montagem K21up.

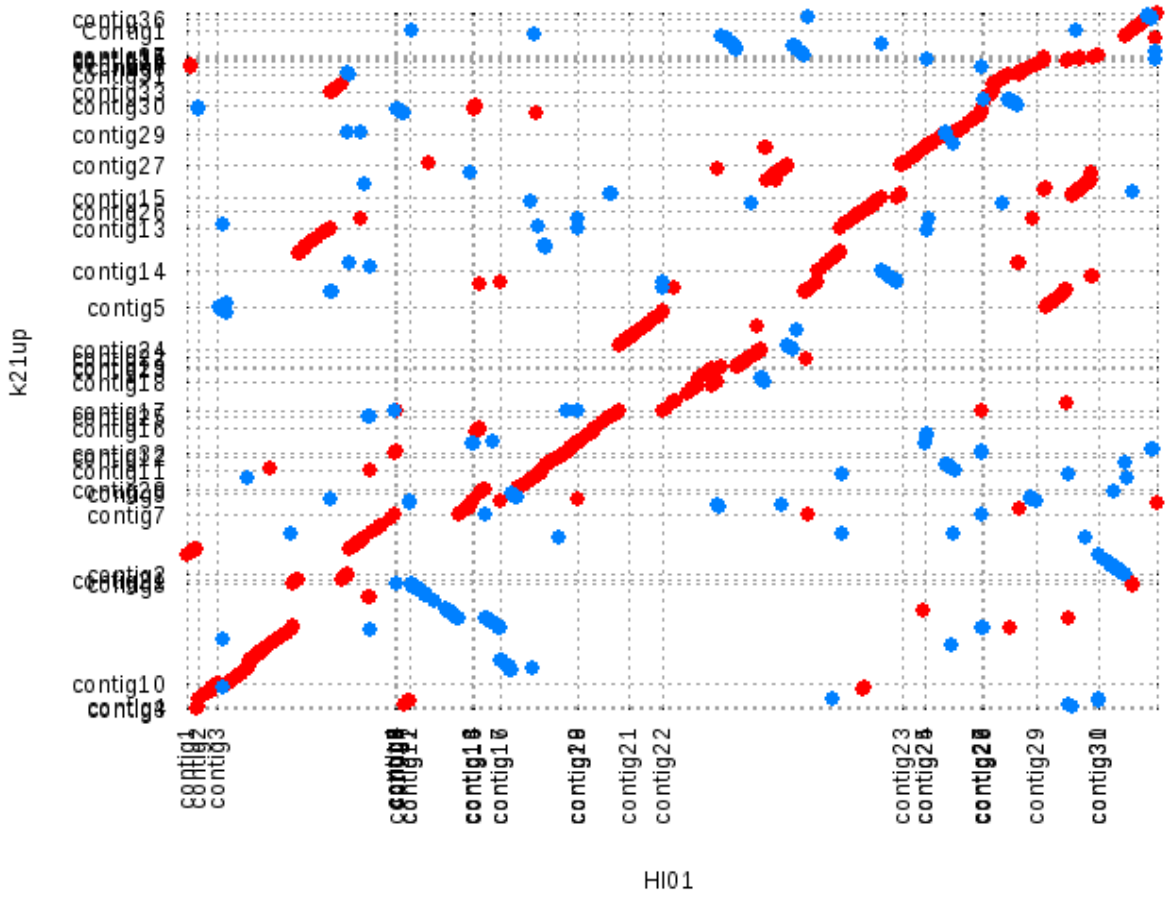


FIGURA 23- *Dotplot* ENTRE AS MONTAGENS k21up E HI01.
Em vermelho estão representadas as regiões de similaridade entre as duas montagens.
Figura construída com o programa Mummer 3.32.

Outra análise que comprovou a melhoria de qualidade obtida na montagem HI01 foi o perfil *GCskew.* Ele avalia a frequência de G sobre C no genoma. Ambas as montagens apresentaram um perfil de acordo com a literatura (EPPINGER *et al*., 2004), demonstrando as características dos genomas bacterianos de apresentar uma maior quantidade de GC sobre AT na fita líder na replicação. Porém a montagem HI01 apresentou um perfil mais suave, que representa um perfil crescente e contínuo até a inversão, demostrando que a ordenação dos *super-contigs* foi melhorada em relação a montagem K21up Figura 24.

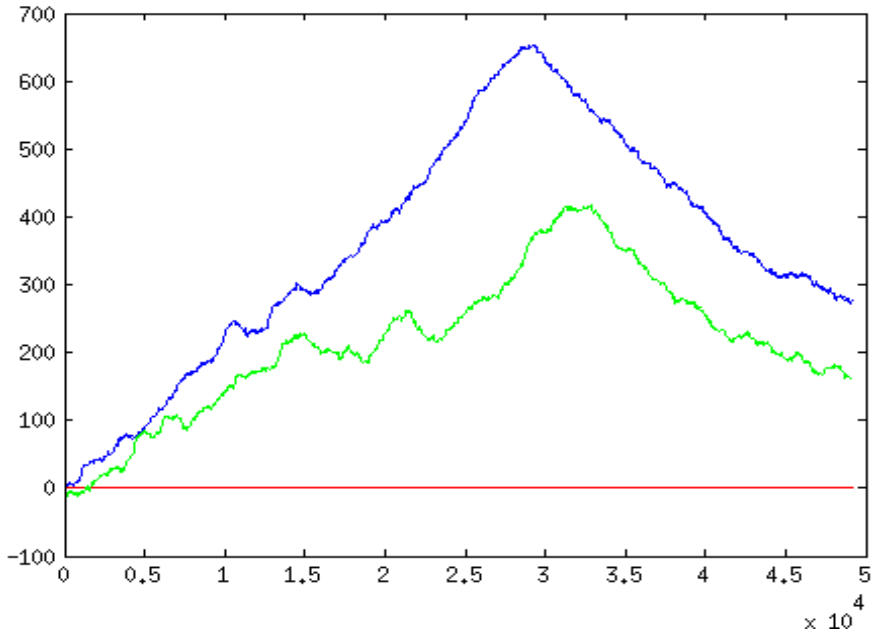


FIGURA 24- COMPARAÇÃO DO GC*skew* ENTRE AS MONTAGENS k21up e HI01.

O gráfico mostra o GCskew cumulativo nas montagens. A janela utilizada para o cálculo foi de 100 pares de bases. A montagem HI01 está representada em Azul. Em verde a montagem k21up. A montagem Hi01 possui um perfil mais suave e continuo, com menos inversões na frequência de GC.

A figura foi construída utilizando o programa Matlab.

5.3 Operon de rRNA do *H. lusitanum* P6-12

A anotação automática conseguiu identificar um operon ribosomal na montagem HI01. O operon tem a ordem 16S rRNA – 23S rRNA – 5S rRNA. Análise comparativa entre a sequência do gene 16S rRNA da montagem Hi01 e a sequência depositada no NCBI demostrou identidade de 100%, confirmando a espécie *H. lusitanum* estirpe P6-12. A análise filogenética do gene 16S rRNA do *H. lusitanum* mostra uma proximidade evolutiva com o *Herbaspirillum sp.* estirpe CF444 como demonstrado na Figura 25.

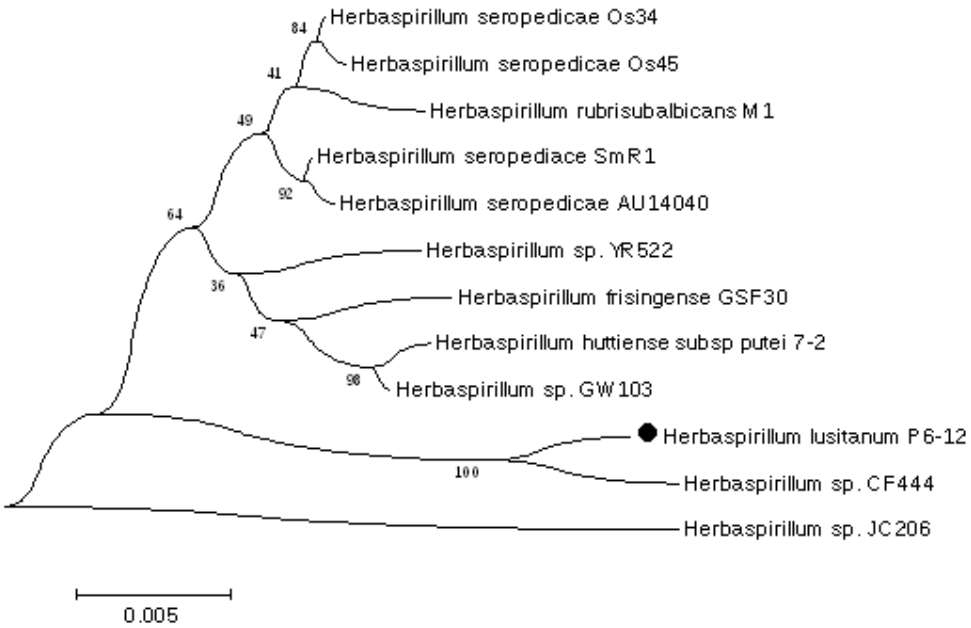


FIGURA 25- ÁRVORE FILOGENÉTICA DAS ESPÉCIES DE *Herbaspirillum* COM BASE NO GENE 16S rRNA.

A árvore sem raiz foi construída utilizando a metodologia de *Neighbor-Joining*. O valor da soma do comprimento dos ramos foi 0,08404273. O valor de *bootstrap* foi determinado através de 10.000 replicatas e está mostrado nas bifurcações dos ramos. A distância evolutiva foi determinada utilizando o método de *Maximum Composite Likelihood*. Todas as posições contendo *gaps* ou erros de alinhamento foram eliminadas totalizando 1481 sítios. Sequências foram retiradas dos genomas depositados no Genbank (Tabela 6).

Os operons de rRNA são sequências de alta similaridade e formam um único *contig. A*s diferenças que separam as cópias são tratadas como erros pontuais pelo algoritmo de montagem. A comparação da cobertura da região do operon com a cobertura global do genoma pode dar indícios da possível presença de mais de uma cópia. A cobertura global observada para a montagem Hi01 foi de 814x, já a região que representa o operon ribosomal possui uma cobertura de 1.600x, ou seja, 2x a média do genoma, indicando a possível presença de 2 cópias deste operon (Figura 26). Este dado difere do que foi encontrado em *Herbaspirillum seropedicae* SmR1, onde foi observada a presença de 3 operons (PEDROSA *et al*., 2011).

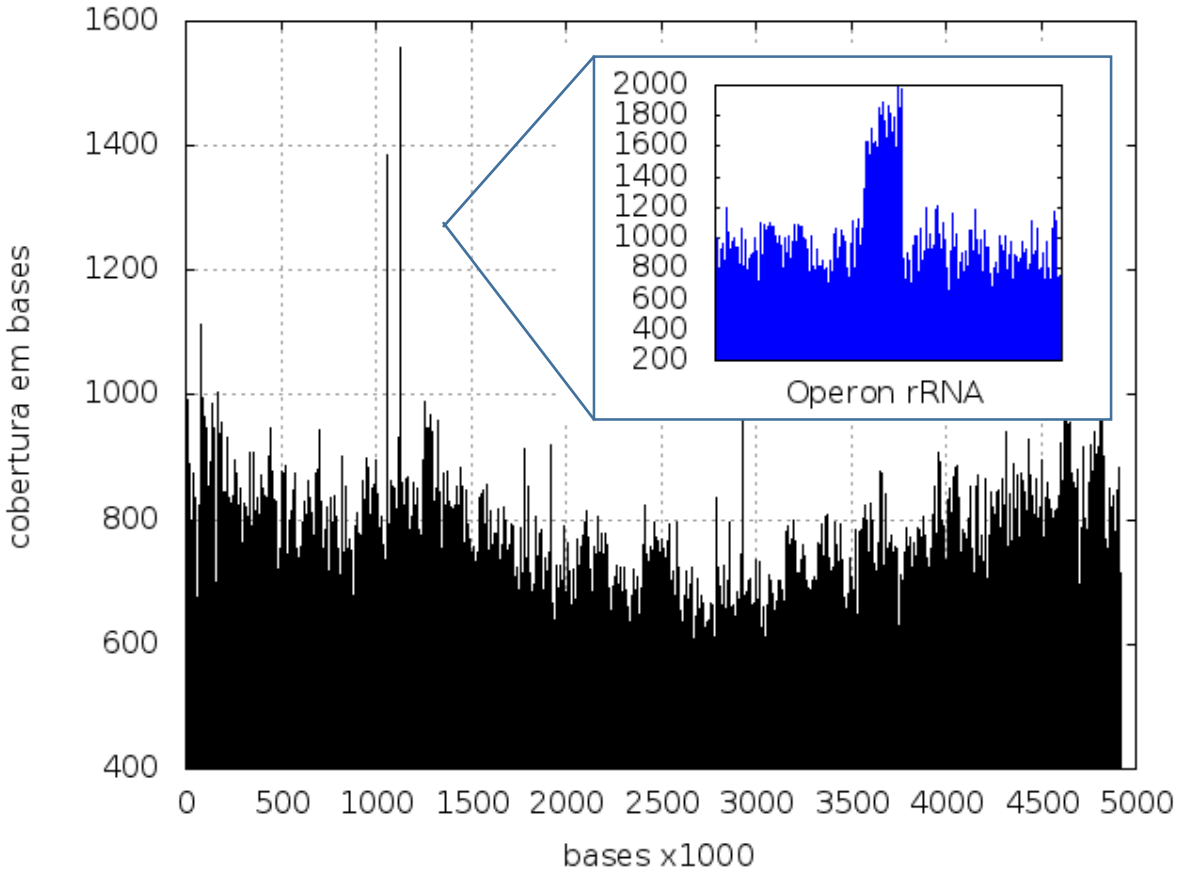


FIGURA 26- IDENTIFICAÇÃO DO NÚMERO DE CÓPIAS DO OPERON RIBOSOMAL NO GENOMA DO *H. lusitanum*.

A cobertura em bases do genoma está em preto. Em azul está ampliada a região onde se encontra o operon 16S rRNA – 23S rRNA – 5SrRNA. Pode ser observada uma duplicação da cobertura na região.

5.4 Anotação da montagem HI01

A anotação realizada com o programa RAST revelou 4488 genes e 51 tRNA, correspondendo ao total de 87% das bases do genoma, embora o *H. lusitanum* P6-12 tenha sido descrito como fixador de nitrogênio através do experimento de redução de acetileno e posterior amplificação do gene *nifD*, neste estudo não foi observada a presença de genes relacionados com a fixação de nitrogênio no genoma (WEISS *et al*., 2012). O *H. lusitanum* P6-12 foi isolado de nódulo de feijão, porém ele não possui nenhum gene *nod* relacionado com a genes de nodulação como os presentes em *rizhobium* (SCHULTZE *et al*., 1994). Foram encontrados genes *HlyB, HlyD, TolC,* codificadores de proteínas relacionadas com o sistema de secreção do tipo I, que é responsável pela secreção de diversas moléculas como íons, drogas e proteínas de diversos tamanhos (DELEPELAIRE, 2004). Os genes *pil* e f*irm* que codificam para os sistemas de secreção do tipo II e IV também foram encontrados. Estes sistemas possuem um relacionamento evolutivo próximo com proteínas compartilhando funções (PEABODY *et al*., 2003), porém não há informação de como estes sistemas agem na bactéria *H. lusitanum*. O sistema de secreção tipo III responsável pela interação planta bactéria e presente no *H. rubrisubalbicans* e *H. seropedicae* SmR1 não foi encontrado (MONTEIRO *et al*., 2012). *H. lusitanum* P6-12 possui as vias metabólicas, *Entner-Doudoroff*, das pentoses fosfato e dos ácidos tricarboxílicos. Os genes que codificam para a via de *Embden Meyerhof-Parnas* também foram encontrados, com exceção da enzima 6-fosfofrutoquinase (EC 2.7.1.11) tal como em *H. seropedicae* SmR1.

Foi observada a presença de um gene que codifica uma proteína semelhante (83% de similaridade) à ribulose-1,5-bisfosfato carboxilase/oxigenase (RuBisCO), enzima chave envolvida na fotossíntese e na fixação de carbono (F ROBERT TABITA, 2008). Proteínas *RuBisCO* semelhantes e com participação na via de recuperação da metionina, foram descritas em outras bactérias heterotróficas como por exemplo, no *Bacillus subtilis* (ASHIDA *et al*., 2008). Os genomas das bactérias *Herbaspirillum sp*. GW103, *Herbaspirillum sp*. YR522 e *H. frisingense* também possuem proteínas homólogas a *RuBisCO* presente em *H. lusitanum* porém classificadas como

proteínas *RuBisCO-like* (Figura 27).

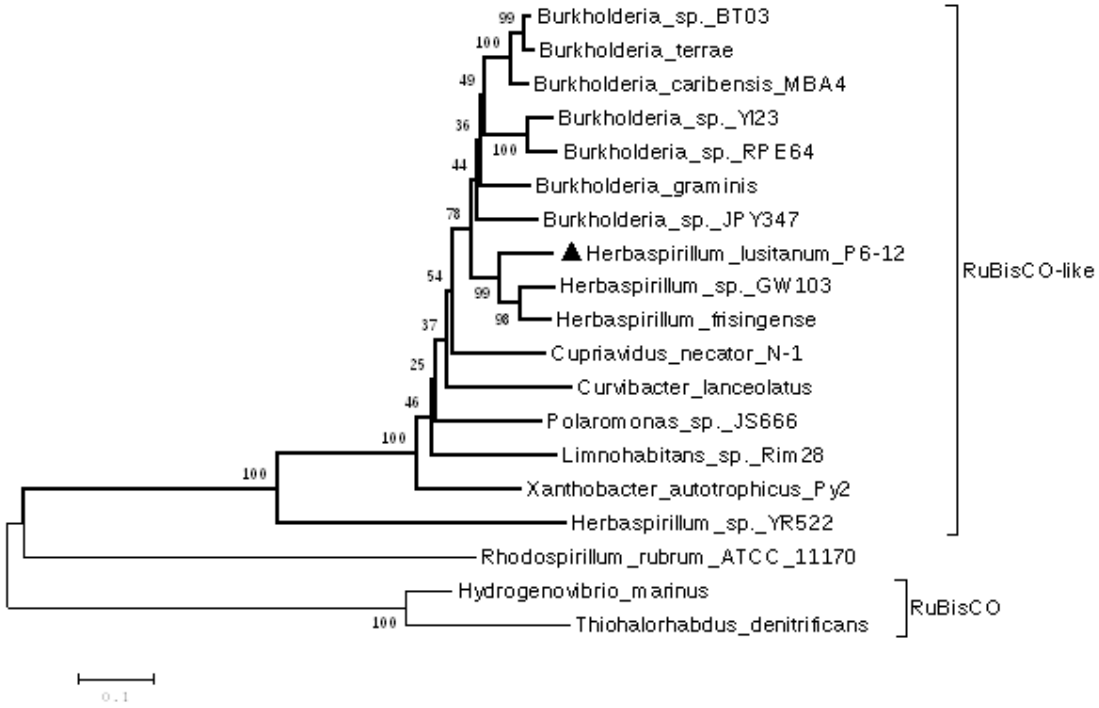


FIGURA 27- ÁRVORE FILOGENÉTICA DA PROTEÍNA RuBisCO DE *H. lusitanum* P6-12.

A árvore foi construída utilizando a metodologia de Bio-*Neighbor-Joining*. O valor da soma do comprimento dos ramos foi 3.70560105. O valor de *bootstrap* foi determinado através de 1000 *replicatas* e está demostrado perto dos ramos. A distância evolucionária foi determinada utilizando o método de correção de Poisson. Todas as posições contendo *gaps* ou erros de alinhamento foram eliminados totalizando 290 sítios.

Também foi encontrado o gene que codifica a 1-aminociclopropano-1-carboxilato (ACC) deaminase, como no *Herbaspirillum seropedicae* SMR1, indicando uma provável contribuição para o desenvolvimento da planta sob condições de estresse. A anotação com base nas categorias funcionais *COG* sendo que para 20% das proteínas foi atribuída uma função geral, e para apenas 4% não foi possível atribuir uma função. Isso revela que o genoma não apresenta grandes diferenças em seu conteúdo gênico dos que já foram descritos por Pedrosa e colaboradores (2011) para *H. seropedicae*. As funções dos genes presentes no genoma de *H. lusitanum* estão resumidas na Figura 28.

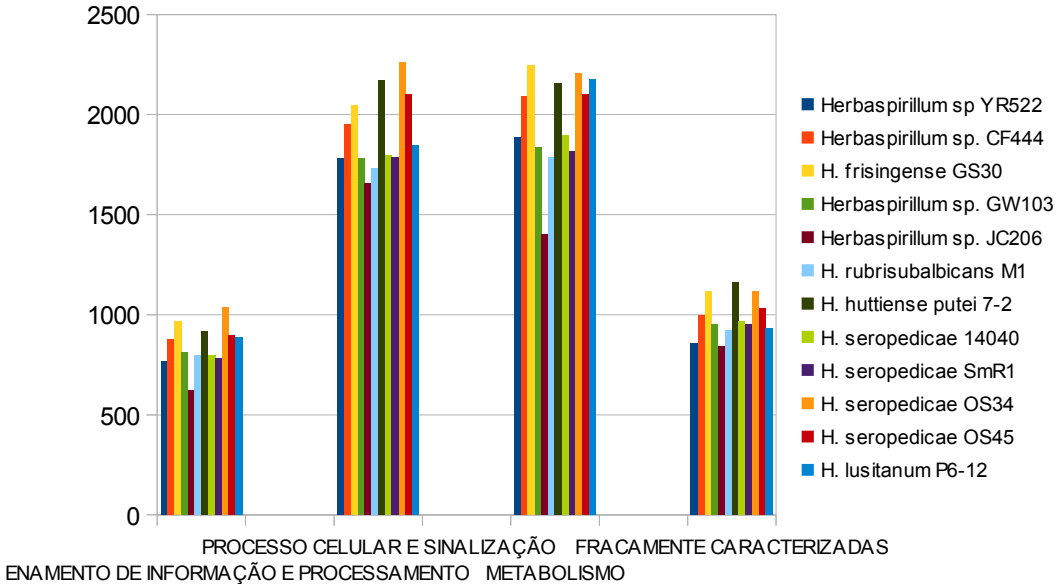


FIGURA 28- CLASSIFICAÇÃO DOS GRUPOS NAS CATEGORIAS FUNCIONAIS COG.

A figura apresenta o número de genes ortólogos classificados nas categorias funcionais COG para cada organismo: Armazenamento e processamento de informação; processo celular e sinalização; metabolismo e fracamente caracterizadas.

5.5 Clusterização dos genes ortólogos do gênero *Herbaspirillum*

5.5.1 Determinação de genes ortólogos

Os métodos de determinação de grupos ortólogos apresentaram resultados distintos em relação ao número e tamanho dos grupos como observado por Grossetête e colaboradores (GROSSETÊTE; LABEDAN e LESPINET, 2010). O programa *OrthoMCL* identificou o maior número de grupos (7.601), enquanto o *BRH* combinado com a metodologia de agrupamento ligação completa obteve a menor soma (5.812). Os valores ficaram próximos ao número médio de genes presentes nos genomas analisados, que é de 5.200 genes. Os tamanhos dos grupos foram próximos entre si, sendo que o *OrthoMCL* contabilizou 48 proteínas no maior grupo e *BRH* 23 proteínas. Já o programa *INPARANOID* obteve o maior grupo com 141 proteínas. Isto foi

possivelmente foi devido a utilização da metodologia ligação simples, mais permissiva na criação dos grupos (Tabela 12).

TABELA 12- NÚMERO DE GRUPOS FORMADOS PELOS MÉTODOS DE DETERMINAÇÃO DE ORTÓLOGOS.

	BRH ligação completa	*INPARANOID* ligação simples	*OrthoMCL*
Número de grupos	5812	6516	7601
Maior grupo	23	141	48

A distribuição da frequência dos grupos otólogos nos diferentes métodos se apresentou na forma de um gráfico bi-modal (Figura 29). O primeiro pico de frequência mostrou uma predominância de grupos pequenos com até 4 proteínas, o segundo pico revelou grupos contendo o mesmo número de proteínas que o de genomas utilizados para a análise, determinados como grupos contendo os genes do *core* genoma. O core genoma é o conjunto de genes que estão presentes em todas as espécies do gênero.

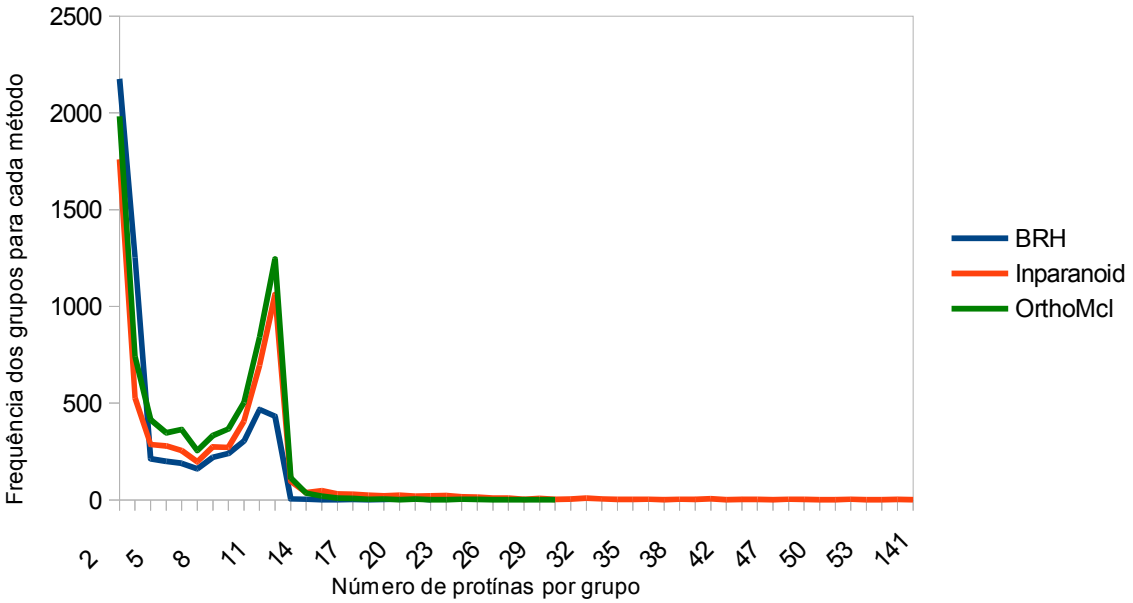


FIGURA 29- FREQUÊNCIA RELATIVA DOS GRUPOS PARA OS MÉTODOS *OrthoMCL*2.0 *INPARANOID* E *BRH*.
O eixo y apresenta a frequência dos grupos para cada método. O eixo x mostra o número de proteínas por grupo. O segundo pico representa os grupos contendo os genes core.

5.5.2 Resultado da intersecção entre os métodos de agrupamento *OrthoMCL*, *INPARANOID* e *BRH* e recuperação das proteínas divergentes

Os resultados divergentes obtidos pelos métodos mais comuns de predição de genes ortólogos *OrthoMCL*, *INPARANOID* e *BRH* revelaram a dificuldade em se obter grupos confiáveis. O programa *OrthoMCL* obteve o maior número de grupos únicos (1.670), número maior que o do programa *INPARANOID* (611) que utilizou a metodologia de agrupamento mais permissiva (ligação simples). Visando contornar este problema, foi realizada a intersecção dos resultados dos três métodos a fim de extrair apenas os grupos compartilhados entre os três. No total foram obtidos 5.218 grupos que foram considerados com alta confiança e foram utilizados para os restantes das análises (Figura 30).

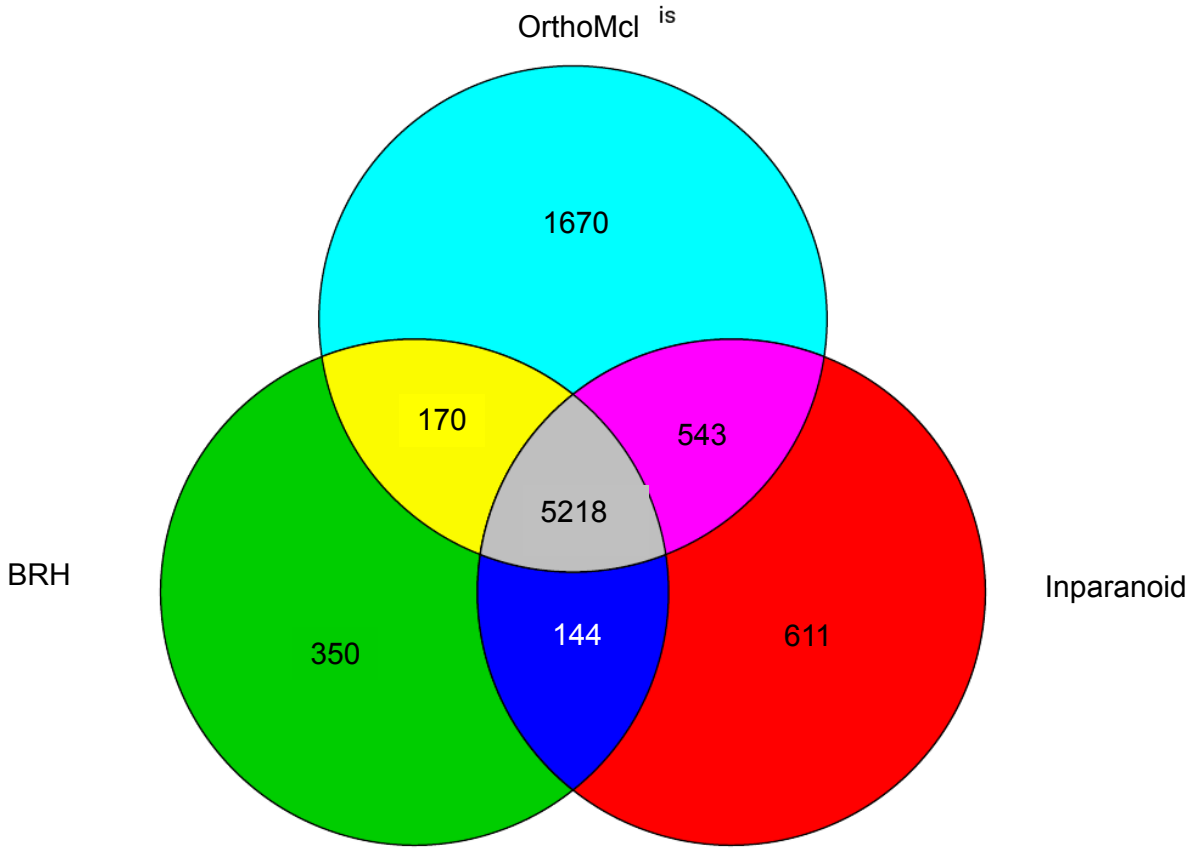


FIGURA 30- GRUPOS DE GENES ORTÓLOGOS IDENTIFICADOS PELOS MÉTODOS *OrthoMCL, INPARANOID* E *BRH*.
Em azul-claro está representado os grupos de ortólogos obtidos pelo método *OrthoMCL*. Em verde pelo do método *BRH* e em vermelho pelo método *INPARANOID*. Em amarelo a intersecção apenas dos métodos *BRH* e *OrthoMCL*. Em azul a interseção apenas dos métodos *BRH* e *INPARANOID*. Em rosa apenas a intersecção dos métodos *INPARANOID* e *OrthoMCL*. Em cinza a interseção dos três métodos.

A utilização da intersecção preservou apenas 45% do número total de proteínas (Tabela 13). Para contornar este efeito, as proteínas que não foram comuns aos 3 métodos foram reinseridas nos 5.218 grupos de alta confiança. Para reinserção das proteínas foram criados consensos das sequências de cada um dos 5.218 grupos através de um alinhamento múltiplo seguido da construção de um perfil e *HMM*. Cada proteína foi individualmente comparada contra os grupos utilizando a ferramenta *Hmmscan* do pacote *Hmmer.* O valor menor ou igual a 1E-10 foi utilizado como limiar de decisão de inserção. Esta etapa permitiu a recuperação de 92,5% das proteínas, aumentando o uso para 64% do conjunto total (Tabela 14).

TABELA 13- NÚMERO DE PROTEÍNAS INSERIDAS COM O *Hmmscan.*

	Proteínas fora da intersecção	Proteínas recuperadas	Proteínas não recuperadas
Número de proteínas	12.428 (100%)	11.499 (92,5%)	929 (7,4%)

TABELA 14- NÚMERO FINAL DE PROTEÍNAS UTILIZADAS NA FORMAÇÃO DOS GRUPOS ORTÓLOGOS.

	Intersecção	Intersecção + Hmmer	Total de proteínas
Número de proteínas	27.505 (45%)	39.004 (64%)	60.941 (100%)

Por fim foram obtidos 5.218 grupos de genes ortólogos dentre os quais 768 grupos foram classificados como *core* e 4.450 grupos como acessórios (Tabela 15). Os grupos core foram determinados como grupos contendo um gene de cada organismo, os demais foram classificados como acessórios.

TABELA 15- CLASSIFICAÇÃO DOS GRUPOS ORTÓLOGOS.

	Grupos Genes core	Grupos de Genes acessórios	Total de grupos
Número de grupos	768	4.450	5.218
Número de proteínas	9.228	29.776	39.005

5.5.3 Anotação funcional dos grupos

Os grupos de genes *core e* acessórios foram anotados utilizando o banco de dados COG (*Cluster of orthologous groups*). Foi possível observar uma grande diferença na distribuição do número de genes entre as categorias. Também foi possível observar que a representatividade entre as categorias é diferente nos genes acessório e *core* (Figura 31). Os genes acessórios possuem a maior quantidade de genes como um todo e as funções mais representadas estão envolvidas com o Metabolismo e transporte de Aminoácidos (E), Transcrição (K), Mecanismos de sinal e transdução (T), Metabolismo e transporte de íons inorgânicos (P). O número de genes que

integra o *core* genoma é inferior ao grupo de genes acessórios porém pode-se observar que algumas categorias são pouco representadas no grupo acessório e seu número é próximo ao número de genes observado para a categoria no grupo *core*. Isso mostra uma mudança na representatividade destas funções direcionada para o grupo *core*. Dentre elas destacam-se ciclo celular, divisão celular e particionamento do cromossomo (D), modificação pós-traducional e chaperonas (O), replicação, recombinação e reparo (L) e metabolismo e transporte de lipídeos (I) (Figura 31).

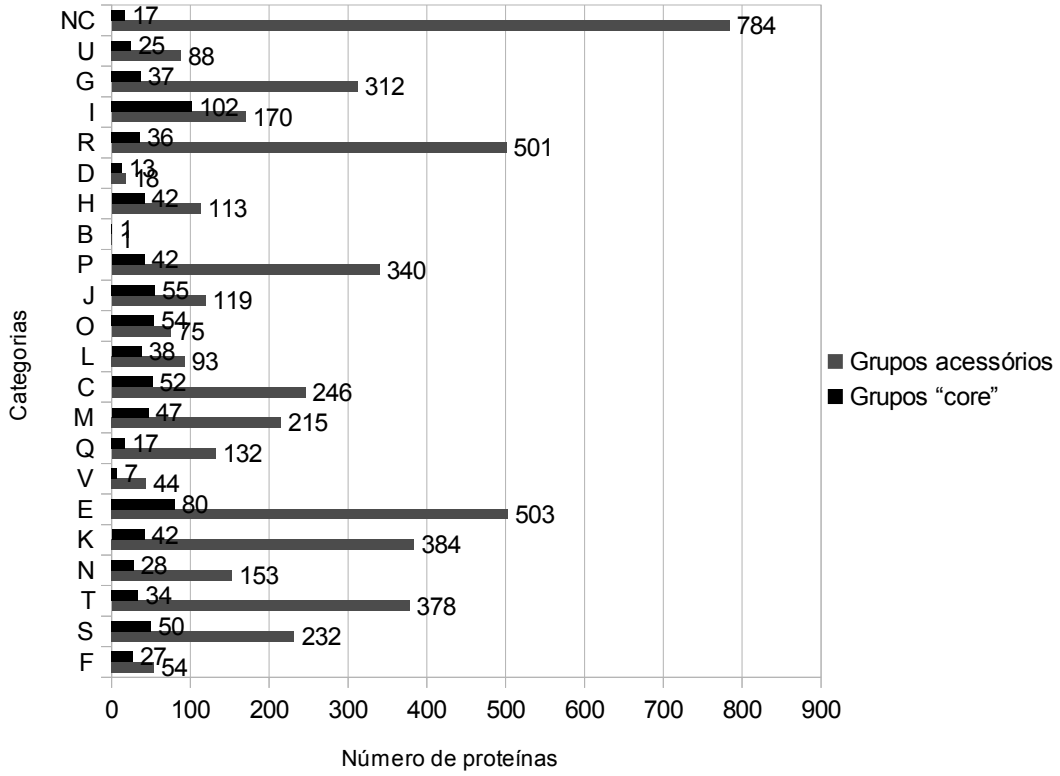


FIGURA 31- CLASSIFICAÇÃO DOS GRUPOS ORTÓLOGOS EM CATEGORIAS FUNCIONAIS COG.

Armazenamento e processamento de informação: [J] Tradução, estrutura e biogênese ribossomal, [K] Transcrição, [L] Replicação, recombinação e reparo, [B] Dinâmica e estrutura da cromatina. Processo celular e sinalização:[D] Ciclo celular, divisão celular, particionamento do cromossomo, [V] Mecanismo de defesa, [T] Mecanismos de sinal e transdução, [M] Parede celular/Membrana/biogênese do envelope, [N] Motilidade Celular, [U] Tráfego intracelular, secreção e transporte vesicular, [O] Modificação pós-traducional, chaperonas. Metabolismo: [G] Transporte de carboidratos e metabolismo, [E] Metabolismo e transporte de aminoácidos, [F] Metabolismo e transporte de nucleotídeos, [H] Metabolismo e transporte de coenzimas, [I] Metabolismo e transporte de lipídeos, [P] Metabolismo e transporte de íon inorgânico, [Q] Biossíntese, transporte e catabolismo de metabólitos secundários, Fracamente caracterizadas: [R] Função geral, [S] Função desconhecida, [NC] Não Caracterizadas.

5.5.4 Árvore filogenética do Gênero *Herbaspirillum*

A árvore filogenética foi construída utilizando os 768 genes *core.* Os genes foram concatenados em um único alinhamento com o total de 216.610 sítios (Figura 32). A árvore filogenética dos genes *core* revelou uma proximidade evolutiva entre os genes dos organismos *H. lusitanum* e *Herbaspirillum sp.* CF444 além da já confirmada pela análise do gene 16S rRNA.

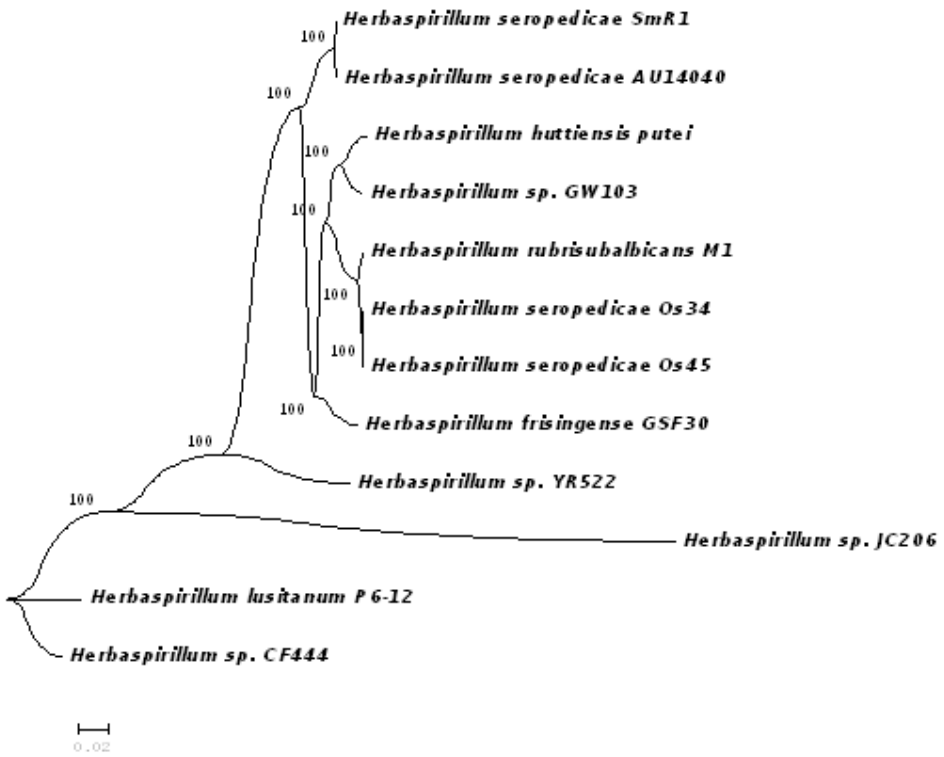


FIGURA 32- ÁRVORES FILOGENÉTICAS COM BASE NA CONCATENAÇÃO DOS GENES ORTÓLOGOS CORE.

A árvore foi construída utilizando a metodologia de Bio-*Neighbor-Joining*. O valor da soma do comprimento dos ramos foi 0.87210. O valor de *bootstrap* foi determinado através de 100 replicatas e está demostrado perto dos ramos. A distância evolucionária foi determinada utilizando o método de *Maximun Composite Likelihood*. Todas as posições contendo *gaps* ou erros de alinhamento foram eliminados totalizando 216610 sítios.

A super-árvore filogenética foi construída visando confirmar o resultado obtido com a árvore dos genes *core*. Para isso foram utilizados todos os 5.218 genes ortólogos com a finalidade de observar se os relacionamentos observados na árvore mudariam. Contudo o relacionamento entre *H. lusitanum* e *Herbaspirillum sp.* CF444 foi mantido (Figura 33).

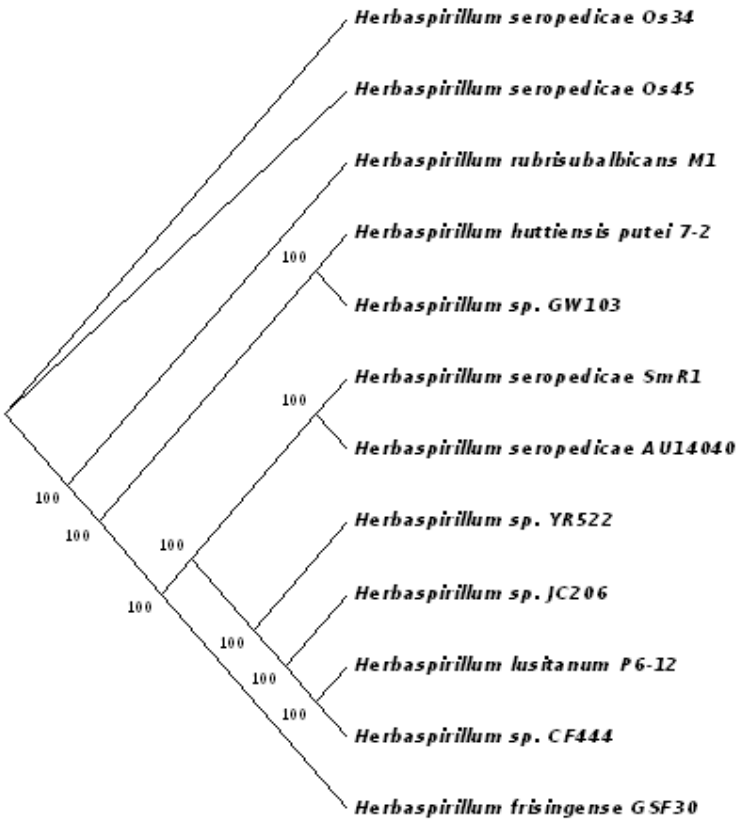


FIGURA 33- SUPER-ÁRVORE CONSTRUÍDA USANDO TODAS AS SEQUÊNCIAS DOS GRUPOS DE GENES ORTÓLOGOS DE *Herbaspirillum spp.*

A super-árvore foi construída através da matriz de parsimônia do resultado de árvores individuais para cada grupo. A árvore foi gerada a partir de 5218 árvores individuais. As árvores individuais foram construídas utilizando a metodologia de Bio-*Neighbor-Joining*. A distância evolutiva foi determinada com o método de *Maximun Composite Likelihood*

Estes resultados sugerem que *Herbaspirillum sp.* CF444 pertence a espécie *H. lusitanum*. Além disso a estirpe *Herbaspirillum* GW103 aparentemente é relacionada com *H. huttiense* subsp. *Putei* 7-2, embora neste caso a sequência de 16S rRNA apresente 4% de dissimilaridade.

6 CONCLUSÃO

A metodologia de sequenciamento Illumina MiSeq obteve um conjunto de leituras de sequências com maior qualidade do que a plataforma SOLiD.

A combinação de mais de uma metodologia de sequenciamento, no caso MiSeq e SOLiD, promoveu melhora significativa no processo de montagem do genoma de *H. lusitanum*. P6-12.

As leituras MiSeq e SOLiD conseguiram boa complementariedade quando submetidas em conjunto ao montador.

As estratégias de fechamento de *gap*, com o uso combinado do *script gapkill.pl e Fgap* foram eficazes, sendo uma ferramenta útil no processo de finalização da sequência genômica.

Uma sequência parcial do genoma de *H. lusitanum* P6-12 foi obtido gerando 31 *super-contigs*.

A metodologia de determinação dos genes ortólogos através da combinação dos métodos *BRH*, *ORTHOMCL* e *INPARANOID,* foi eficaz para a determinação de ortólogos de alta confiança, podendo ser utilizada para a identificação de ortólogos em outras espécies.

A filogenia, utilizando os grupos ortólogos *core,* corroborou a proximidade evolutiva das estirpes *Herbaspirillum lusitanum* e *Herbaspirillum* CF444 sugerindo que *Herbaspirillum sp.* CF444 seja reclassificado como *Herbaspirillum lusitanum* CF444.

A filogenia baseada nas proteínas ortólogas do *core* genoma *também* sugere relação evolutiva entre *Herbaspirillum huttiense subsp. putei* 7-2 e *Herbaspirillum sp* GW103.

REFERÊNCIAS BIBLIOGRÁFICAS

ABBY, S.; DAUBIN, V. Comparative genomics and the evolution of prokaryotes. **Trends Microbiol**, v. 15, n. 3, p. 135–141, 2007.

ALTSCHUL, S. F. *et al*. Basic local alignment search tool. **J Mol Biol**, v. 215, n. 3, p. 403–410, 1990.

ASHIDA, H. *et al*. RuBisCO-like proteins as the enolase enzyme in the methionine salvage pathway: functional and evolutionary relationships between RuBisCO-like proteins and photosynthetic RuBisCO. **J Exp Bot**, v. 59, n. 7, p. 1543–1554, 2008.

BAJERSKI, F. *et al. Herbaspirillum psychrotolerans* sp. nov., a member of the family Oxalobacteraceae from a glacier forefield.**Int J Syst Evol Microbiol**, v. 63, n. Pt 9, p. 3197–3203, 2013.

BALDANI, J. I. *et al*. Characterization of Herbaspirillum seropedicae gen. nov., sp. nov., a Root-Associated Nitrogen-Fixing Bacterium. **Int J Syst Bacteriol**, v. 36, n. 1, p. 86–93, 1986.

BALDAUF, S. L. Phylogeny for the faint of heart: a tutorial. **Trends Genet.**, v. 19, n. 6, p. 345–351, 2003.

BANERJEE, R.; MUKHOPADHYAY, S. Niche specific amino acid features within the core genes of the genus Shewanella. **Bioinformation**, v. 8, n. 19, p. 938–942, 2012.

BAUM, B. R. Combining Trees as a Way of Combining Data Sets for Phylogenetic Inference, and the Desirability of Combining Gene Trees. **Taxon**, v. 41, n. 1, p. 3, 1992.

BERG, G.; EBERL, L.; HARTMANN, A. The rhizosphere as a reservoir for opportunistic human pathogenic bacteria. **Environ Microbiol**, v. 7, n. 11, p. 1673–1685, 2005.

BERGLUND, E. C.; KIIALAINEN, A.; SYVANEN, A.C. Next-generation sequencing technologies and applications for human genetic history and forensics. **Investig Genet**, v. 2, p. 23, 2011.

BINNEWIES, T. T. *et al*. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. **Funct Integr Genomics**, v. 6, n. 3, p. 165–185, 2006.

CARRO, L. *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.**Int J Syst Evol Microbiol**, v. 62, n. Pt 6, p. 1300–1306, 2011.

CHEN, J. *et al*. Herbaspirillum Species: A Potential Pathogenic Bacteria Isolated from Acute Lymphoblastic Leukemia Patient. **Curr Microbiol**, v. 62, n. 1, p. 331–333, 2011.

COCHRANE, G.; KARSCH-MIZRACHI, I.; NAKAMURA, Y. The International Nucleotide Sequence Database Collaboration. **Nucleic Acids Res**, v. 39, p. D15–D18, 2011.

COENYE, T. *et al*. Towards a prokaryotic genomic taxonomy. **FEMS Microbiol Rev**, v. 29, n. 2, p. 147–167, 2005.

CREEVEY, C. J.; MCINERNEY, J. O. Clann: investigating phylogenetic information through supertree analyses. **Bioinformatics**, v. 21, n. 3, p. 390–392, 2005.

DAY, W. H. E. Computational complexity of inferring phylogenies from dissimilarity matrices. **Bull Math Biol**, v. 49, n. 4, p. 461–467, 1987.

DELEPELAIRE, P. Type I secretion in gram-negative bacteria. **Biochim Biophys Acta**, v. 1694, n. 1–3, p. 149–161, 2004.

DELSUC, F.; BRINKMANN, H.; PHILIPPE, H. Phylogenomics and the reconstruction of the tree of life. **Nat Rev Genet**, v. 6, n. 5, p. 361–375, 2005.

DING, L. e 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. **Int J Syst Evol Microbiol**, v. 54, 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., comb. nov., and description of Herbaspirillum aquaticum sp. Nov. **Int J Syst Evol Microbiol**, v. 60, n. 6, p. 1418–1426, 2009.

DUTILH, B. E. *et al*. Assessment of phylogenomic and orthology approaches for phylogenetic inference. **Bioinformatics**, v. 23, n. 7, p. 815–824, 2007.

EDDY, S. R. Profile hidden Markov models. **Bioinformatics**, v. 14, n. 9, p. 755–763, 1998.

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

EPPINGER, M. *et al*. Comparative analysis of four Campylobacterales. **Nat Rev**

Microbiol, v. 2, n. 11, p. 872–885, 2004.

EWING, B. *et al*. Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. **Genome Res**, v. 8, n. 3, p. 175–185, 1998.

FELSENSTEIN, J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. **Evolution**, v. 39, n. 4, p. 783, 1985.

FINN, R. D. *et al*. The Pfam protein families database. **Nucleic Acids Res**, v. 36, n. suppl 1, p. D281–D288, 2008.

FITCH, W. M.; MARGOLIASH, E. Construction of Phylogenetic Trees. **Science**, v. 155, n. 3760, p. 279–284, 1967.

FITCH, W. M. Distinguishing Homologous from Analogous Proteins. **Syst Biol**, v. 19, n. 2, p. 99–113, 1970.

FITCH, W. M. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. **Syst Zool**, v. 20, n. 4, p. 406, 1971.

FITCH, W. M. Homology. **Trends Genet**, v. 16, n. 5, p. 227–231, 2000.

FLEISCHMANN, R. D. *et al*. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. **Science (New York, N.Y.)**, v. 269, n. 5223, p. 496–512, 1995.

FLUSBERG, B. A. *et al*. Direct detection of DNA methylation during single-molecule, real-time sequencing. **Nature Methods**, v. 7, n. 6, p. 461–465, 2010.

FRANCKI, M.; APPELS, R. Wheat functional genomics and engineering crop improvement. **Genome Biology**, v. 3, n. 5, p. reviews1013, 2002.

FRANCK, A.C.; LOBRY, J. R. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. **Gene**, v. 238, n. 1, p. 65–77, 1999.

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

GRIGORIEV, A. Analyzing genomes with cumulative skew diagrams. **Nucleic Acids Res**, v. 26, n. 10, p. 2286–2290, 1998.

GROSSETÊTE, S.; LABEDAN, B.; LESPINET, O. FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology. **BMC Genomics**, v. 11, n. 1, p. 81, 2010.

GUINDON, S. *et al*. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. **Syst Biol**, v. 59, n. 3, p. 307–321, 2010.

HALL, N. Advanced sequencing technologies and their wider impact in microbiology. **J Exp Biol**, v. 210, n. 9, p. 1518–1525, 1 maio 2007.

HENNIG, W.; DAVIS, D. D., ZANGERL, RAINER. **Phylogenetic systematics**. Urbana: University of Illinois Press, 1999.

HILLIS, D. M.; BULL, J. J. An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. **Syst Biol**, v. 42, n. 2, p. 182–192, 1993.

HUTCHISON, C. A. DNA sequencing: bench to bedside and beyond†. **Nucleic Acids Res**, v. 35, n. 18, p. 6227–6237, 2007.

HUYNEN, M. A.; GABALDÓN, T.; SNEL, B. Variation and evolution of biomolecular systems: Searching for functional relevance. **FEBS Letters**, v. 579, n. 8, p. 1839–1845, 2005.

IM, W.-T. *et al. Herbaspirillum chlorophenolicum* sp. nov., a 4-chlorophenol-degrading bacterium. **Int J Syst Evol Microbiol**, v. 54, n. 3, p. 851–855, 2004.

JUNG, S.-Y. *et al.* Herbaspirillum rhizosphaerae sp. nov., isolated from rhizosphere soil of Allium victorialis var. platyphyllum.**Int JSyst Evol Microbiol**, v. 57, n. 10, p. 2284–2288, 2007.

KAAS, R. S. *et al*. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. **BMC Genomics**, v. 13, n. 1, p. 577, 2012.

KANEHISA, M.; GOTO, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. **Nucleic Acids Res**, v. 28, n. 1, p. 27–30, 2000.

KARLIN, S.; CAMPBELL, A. M.; MRÁZEK, J. Comparative Dna Analysis Across Diverse Genomes. **Annu Rev Genet**, v. 32, n. 1, p. 185–225, 1998.

KIRCHHOF, G. *et al. Herbaspirillum frisingense* sp. nov., a new nitrogen-fixing bacterial species that occurs in C4-fibre plants. **Int J Syst Evol Microbiol**, v. 51, n. 1, p. 157–168, 2001.

KOONIN, E. V. Orthologs, Paralogs, and Evolutionary Genomics1. **Annu Rev Genet**, v. 39, n. 1, p. 309–338, 2005.

KUZNIAR, A. *et al*. The quest for orthologs: finding the corresponding gene across genomes. **Trends Genet**, v. 24, n. 11, p. 539–551, 2008.

LANDER, E. S. *et al*. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860–921, 2001.

LAGIER, J.C. *et al.* Non-contiguous finished genome sequence and description of *Herbaspirillum masiliense* sp. nov. **Stand Genomic Sci**, v. 7, p. 200-209, 2012.

LAPOINTE, F.-J.; WILKINSON, M.; BRYANT, D. Matrix Representations with Parsimony or with Distances:Two Sides of the Same Coin? **Syst Biol**, v. 52, n. 6, p. 865–868, 2003.

LEEKITCHAROENPHON, P. *et al*. Genomic variation in Salmonella enterica core genes for epidemiological typing. **BMC Genomics**, v. 13, n. 1, p. 88, 2012.

LEFEBURE, 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.

LI, L.; STOECKERT, C. J.; ROOS, D. S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. **Genome Res**, v. 13, n. 9, p. 2178–2189, 2003.

LIN, S.-Y. *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. **Int J Syst Evol Microbiol**, v. 63, n. Pt 11, p. 4100–4107, 2013.

LIU, L. *et al*. Comparison of Next-Generation Sequencing Systems. **Biomed Res Int**, v. 2012, 2012.

LOBRY, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. **Mol Biol Evol**, v. 13, n. 5, p. 660–665, 1996.

LOWE, T. M.; EDDY, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. **Nucleic Acids Res**, v. 25, n. 5, p. 955–964, 1997.

MARDIS, E. R. Next-Generation DNA Sequencing Methods. **Annu Rev Genomics Hum Genet**, v. 9, n. 1, p. 387–402, 2008.

MARQUES DA SILVA, R. *et al*. Bacterial diversity in aortic aneurysms determined by 16S ribosomal RNA gene analysis. **J Vasc Surg**, v. 44, n. 5, p. 1055–1060, 2006.

MCKERNAN, D. P. *et al*. Age-Dependent Susceptibility of the Retinal Ganglion Cell Layer to Cell Death. **Invest Ophthalmol Vis Sci**, v. 47, n. 3, p. 807–814, 1 2006.

MILLER, J. R.; KOREN, S.; SUTTON, G. Assembly Algorithms for Next-Generation Sequencing Data. **Genomics**, v. 95, n. 6, p. 315–327, 2010.

MONTEIRO, R. A. *et al*. Genomic comparison of the endophyte Herbaspirillum seropedicaeSmR1 and the phytopathogen Herbaspirillum rubrisubalbicans M1 by suppressive subtractive hybridization and partial genome sequencing. **FEMS**

Microbiol Ecol, v. 80, n. 2, p. 441–451, 2012.

MOSS, W. W.; HENDRICKSON, J. A. Numerical Taxonomy. **Annu Rev Entomol**, v. 18, n. 1, p. 227–258, 1973.

MYERS, E. W. *et al*. A whole-genome assembly of Drosophila. **Science (New York, N.Y.)**, v. 287, n. 5461, p. 2196–2204, 2000.

NIEDRINGHAUS, T. P. *et al*. Landscape of Next-Generation Sequencing Technologies. **Anal chem**, v. 83, n. 12, p. 4327–4341, 2011.

PEABODY, C. R. *et al*. Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. **Microbiology**, v. 149, n. 11, p. 3051–3072, 2003.

PEDROSA, F. O. *et al*. Genome of Herbaspirillum seropedicae Strain SmR1, a Specialized Diazotrophic Endophyte of Tropical Grasses. **PLoS Genet**, v. 7, n. 5, p. e1002064, 2011.

PIRO, V. C. *et al*. FGAP: an automated gap closing tool. **BMC Research Notes**, v. 7, n. 1, p. 371, 18 jun. 2014.

POP, M. Genome assembly reborn: recent computational challenges. **Brief Bioinform** v. 10, n. 4, p. 354–366, 2009.

PUNTA, M. *et al*. The Pfam protein families database. **Nucleic Acids Res**, v. 40, n. D1, p. D290–D301, 2012.

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

RANNALA, B.; YANG, Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. **J Mol Evol**, v. 43, n. 3, p. 304–311, 1996.

REMM, M.; STORM, C. E. V.; SONNHAMMER, E. L. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. **J Mol Biol**, v. 314, n. 5, p. 1041–1052, 2001.

ROTHBALLER, M. *et al. Herbaspirillum hiltneri* sp. nov., isolated from surface-sterilized wheat roots. **Int J Syst Evol Microbiol**, v. 56, n. 6, p. 1341–1348, 2006.

RUTHERFORD, K. *et al*. Artemis: sequence visualization and annotation. **Bioinformatics**, v. 16, n. 10, p. 944–945, 2000.

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Mol Biol Evol**, v. 4, n. 4, p. 406–425, 1987.

SALZBERG, S. L. *et al*. Microbial gene identification using interpolated Markov models. **Nucleic Acids Res**, v. 26, n. 2, p. 544–548, 1998.

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.

SCHATZ, M. C.; DELCHER, A. L.; SALZBERG, S. L. Assembly of large genomes using second-generation sequencing. **Genome Res**, v. 20, n. 9, p. 1165–1173, 2010.

SCHULTZE, M. *et al*. Cell and Molecular Biology of Rhizobium-Plant. In: KWANG W. JEON AND JONATHAN JARVIK (Ed.). **Int Rev Cytol** [s.l.] Academic Press, v. 156, p. 1–75, 1994.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. **Nat Biotechnol**, v. 26, n. 10, p. 1135–1145, 2008.

SNEATH, P. H. A.; SOKAL, R. R. **Numerical taxonomy**. San Francisco: W.H. Freeman, 1973.

SONNHAMMER, E. L. L.; KOONIN, E. V. Orthology, paralogy and proposed classification for paralog subtypes. **Trends Genet.**, v. 18, n. 12, p. 619–620, 2002.

SPILKER, T. *et al*. Recovery of Herbaspirillum Species from Persons with Cystic Fibrosis. **J Clin Microbiol**, v. 46, n. 8, p. 2774–2777, ago. 2008.

STACEY, G.; BURRIS, R. H.; EVANS, H. J. **Biological Nitrogen Fixation**. [s.l.] Springer, 1992.

SWOFFORD, D. L. PAUP: Phylogenetic analysis using parsimony, version 3.1, March 1993. [s.l.] **Center for Biodiversity, Illinois Natural History Survey**, [s.d.].

TABITA, F.R. *et al.* Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships. **J Exp Bot**, v. 59, n. 7, p. 1515–24, 2008.

TALAVERA, G.; CASTRESANA, J. Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. **Syst Biol**, v. 56, n. 4, p. 564–577, 2007.

TAMURA, K. *et al*. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. **Mol Biol Evol**, v. 28, n. 10, p. 2731–2739, 2011.

TATUSOV, R. L.; KOONIN, E. V.; LIPMAN, D. J. A genomic perspective on protein families. **Science**, v. 278, n. 5338, p. 631–637, 1997.

VALVERDE, A. *et al*. Herbaspirillum lusitanum sp. nov., a novel nitrogen-fixing bacterium associated with root nodules of Phaseolus vulgaris. **Int J Syst Evol Microbiol**, v. 53, n. 6, p. 1979–1983, 2003.

WEISS, V. A. *et al*. Draft Genome Sequence of Herbaspirillum lusitanum P6-12, an Endophyte Isolated from Root Nodules of Phaseolus vulgaris. **J Bacteriol**, v. 194, n. 15, p. 4136–4137, 2012.

WESTBROOK, J. *et al*. The Protein Data Bank and structural genomics. **Nucleic Acids Res**, v. 31, n. 1, p. 489–491, 2003.

YANG, Z.; RANNALA, B. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. **Mol Biol Evol**, v. 14, n. 7, p. 717–724, 1997.

ZERBINO, D. R.; BIRNEY, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. **Genome Res**, v. 18, n. 5, p. 821–829, 2008.

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