UNIVERSIDADE FEDERAL DO PARANÁ

ROXANA BEATRIZ RIBEIRO CHAVES

MONTAGEM E ANOTAÇÃO DO GENOMA DA BACTÉRIA ENDOFÍTICA DIAZOTRÓFICA *Herbaspirillum rubrisubalbicans* M4

CURITIBA

2016

ROXANA BEATRIZ RIBEIRO CHAVES

MONTAGEM E ANOTAÇÃO DO GENOMA DA BACTÉRIA ENDOFÍTICA DIAZOTRÓFICA *Herbaspirillum rubrisubalbicans* M4

Dissertação de Mestrado apresentado ao programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica da Universidade Federal do Paraná, como parte das exigências para a obtenção do título de Mestre em Bioinformática.

Orientadora: Profa. Dra. Rose Adele Monteiro
Co-Orientador: Prof.Dr. Leonardo Magalhães Cruz

CURITIBA

2016

Chaves, Roxana Beatriz Ribeiro
C512 Montagem e anotação do genoma da bacteria endofitica diazotrofica *Herbaspirillum rubrisubalbicans* M4 / Roxana Beatriz Ribeiro Chaves / / Curitiba, 2016 -
55 f il , tabs, grafs

Orientadora Rose Adele Monteiro
Co-orientador Leonardo Magalhães Cruz
Dissertação (Mestrado) – Universidade Federal do Parana, Setor de Educação Profissional e Tecnologica, Curso de Pos-Graduação em Bioinformatica
Inclui Bibliografia

1 Genoma 2 *Herbaspirillum* 3 *Herbaspirillum rubrisubalbicans* 4 Bioinformatica I Monteiro, Rose Adele II Cruz, Leonardo Magalhães III Titulo IV Universidade Federal do Parana

CDD 575 113

TERMO DE APROVAÇÃO

ROXANA BEATRIZ RIBEIRO CHAVES

"MONTAGEM E ANOTAÇÃO DO GENOMA DA BACTÉRIA ENDOFÍTICA DIAZOTRÓFICA *Herbaspirillum rubrisubalbicans M4*"

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora

Drª Rose Adele Monteiro
Universidade Federal do Parana - UFPR

Dr Marcelo Muller dos Santos
Universidade Federal do Paraná - UFPR

Dr Eduardo Balsanelli
Universidade Federal do Paraná - UFPR

Curitiba, 31 de agosto de 2016

AGRADECIMENTOS

Gostaria de agradecer aos meus orientadores, Rose Adele Monteiro, a quem admiro profundamente, pela professora, cientista e mulher que é. Obrigada pela paciência, carinho, sorrisos, confiança e incentivo, por tudo que me ensinou nesses dois anos, e mais ainda nesses últimos 6 meses. Pela presença tão marcante e tão importante em muitos momentos que se mostraram difíceis. Leonardo Magalhães Cruz, pela confiança e ensinamentos. Maria Isabel Stets, que sempre teve tempo para tirar as dúvidas que surgiam no decorrer desses dois anos. Mesmo sem conhecer pessoalmente, tenho certeza de ser uma pessoa muito querida e atenciosa.

Valter Baura, Michele Zibetti e Eduardo Balsanelli, pelas explicações sobre assuntos quando ainda eram novos para mim.

Rodrigo Cardoso, que mesmo tão ocupado, sempre tirava um tempinho, as vezes um tempão, para me ajudar, me explicar tantas vezes quanto eu perguntava (haja paciência! Haha). Muito obrigada mesmo!!!!

Helba Cirino (Nega), que foi fundamental e esteve comigo nas horas boas e não tão boas. Obrigada pelas conversas, por me ouvir, por permitir que eu te ouvisse. Pelas trocas de conhecimento, e de histórias de vida. Amo-te!!

Estevan Tomazini, (Lehninger ambulante). Que tentou me ensinar tantas coisas com palavras tão difíceis, mas que para ele são tão óbvias!! Mas eu aprendi, viu!!?

Aos meus amigos da bioinfo, especialmente ao Bruno Nichio e ao Alexandre Lejambre, irmãos de alma que foram fundamentais em vários momentos! Obrigada pela paciência, pelas risadas, pelas conversas de bar (no bar), pela confiança e carinho comigo!! Amo vocês imensamente!!

À secretaria da Bioinformática, e aos professores, que contribuíram para meu crescimento e amadurecimento acadêmico.

À Suzana, secretária da Bioinformática, pelas risadas, conversas, palavras carinhosas que tanto me colocaram para cima.

A todos da bioquímica que me acolheram com tanto carinho.

À minha família, que sempre esteve comigo. Pelo apoio e amor, compreensão e tolerância que sempre tiveram comigo!!

Ao meu namorido, que me ajudou a perceber e a ser mais forte por todas as coisas que passamos, que não foram poucas, mas que foram muito boas, mesmo quando eram ruins!

A todos citados ou não, mas que de alguma forma estiveram comigo por todos esses dois anos e alguns meses, hoje sou uma pessoa melhor, tanto

no pessoal quanto no profissional, graças à contribuição que cada um me deu. Muito obrigada!!

RESUMO

Herbaspirillum rubrisubalbicans M4 é uma bactéria endofítica diazotrófica que provoca a doença da estria mosqueada em algumas variedades de cana-de-açúcar e sorgo. No presente estudo, um total de 4.101.514 *reads* oriundos do Illumina MiSeq (cobertura de 289x) e 4.112.524 *reads* provenientes da plataforma ION Próton (cobertura de 30x) foram obtidos e utilizados para montagem e análise do genoma de *H. rubrisubalbicans* M4. A montagem *de novo* com os *reads* Illumina geraram 3.157 *contigs* de 3.183 pb e um GC de 60,1%; a montagem *de novo* com os *reads* ION geraram 2.100 *contigs* de 5.412.692 bp e um GC de 61,5%. As análises de cada corrida seguiram caminhos distintos, sendo analisadas em separado e em grau de comparação. A montagem ION ao ser finalizada pelo programa GFinisher, gerou 163 *contigs,* um GC de 62,3% e um genoma de 4.784.841bp. Os *scaffolds* de ambas as montagens foram ordenados usando o genoma de *H. rubrisubalbicans* M1 como referência com o programa MUMmer 3.0, o que demonstrou similaridades entre os genomas e constatou-se um alto grau de repetições de sequências, caracterizado pelos *reads* que não foram alinhados, característica já observada na estirpe *H. rubrisubalbicans* M1. A anotação automática desse genoma foi feita utilizando o programa RAST, sendo anotados 4.929 genes. Dentre esses genes foram encontrados genes potencialmente envolvidos no processo de interação entre a bactéria e a planta.

Palavras chave: *Herbaspirillum*, *Herbaspirillum rubrisubalbicans*, Diazotrófico, Genoma

ABSTRACT

Herbaspirillum rubrisubalbicans M4 is a diazotrophic endophytic bacterium, as well as the M1 strain of the same species causes the mottled stripe disease in some varieties of sugarcane. In this study, a total of 4,101,514 *reads* coming from the Illumina Genome Analyser (covering 215x) and 4,112,524 reads from the Proton ION platform (30x coverage), were obtained and used for assembly and genome analysis of *Herbaspirillum rubrisubalbicans* M4. The new assembly with Illumina *reads* generated 3,157 *contigs* with 3,183 bases in each and a GC of 60.1%; the new assembly with ION *reads* generated 2,100 contigs with 5,412,692 bp in each *contig*, and a GC 61.5%. The analysis of each race followed different paths, and analyzed separately and in degree of comparison. The assembly ION to be finalized by GFinisher program generated 163 *contigs*, a GC of 62.3% and a genome 4.784.841bp. The scaffolds of both assemblies were sorted using the genome *Herbaspirillum rubrisubalbicans* M1 with reference to the MUMmer 3.0 program, which showed similarities between genomes and found a high degree of repeat sequences, characterized by a read left outside the alignment feature already known species *Herbaspirillum rubrisubalbicans*. Were recorded 4,929 genes, including genes involved in plant bacteria interaction.

Key words: *Herbaspirillum*, *Herbaspirillum rubrisubalbicans*, Diazotrophic, Genome

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O gênero *Herbaspirillum* pertence a classe Beta das proteobactérias. A maioria das espécies desse gênero fixam nitrogênio atmosférico sob condições de microaerofilia e crescem utilizando N_2 como única fonte de nitrogênio (BALDANI *et al*., 1986; BALDANI *et al*., 1992). Essas espécies fixadoras de nitrogênio associam-se com várias plantas de interesse econômico, como milho (*Zea mays*), arroz (*Oryza sativa*), sorgo (*Sorghum bicolor*), trigo (*Triticum aestivum*), cana-de-açúcar (*Saccharum officinarum*), bananeiras (*Musa* sp.), palmeiras e abacaxizeiros (*Ananas comosus)* (BALDANI *et al*., 1986; BALDANI *et al*., 1992; CRUZ *et al*., 2001).

H. seropedicae e *H. rubrisubalbicans* são as espécies mais estudadas desse gênero em relação a bioquímica e a fisiologia (MONTEIRO *et al.,* 2012). Eles têm características fisiológicas muito similares e são filogeneticamente mais próximos do que com outras espécies de *Herbaspirillum* (MONTEIRO *et al.,* 2013). *H. seropedicae* e *H. rubrisubalbicans* promovem o crescimento vegetal e são componentes do inoculante para cana de açúcar formulado pela EMBRAPA. Na cana de açúcar B-4362, tanto *H. seropedicae* como *H. rubrisubalbicans* agem como bactérias endofiticas, entretanto, o *H. rubrisubalbicans* age, também como um fitopatógeno.

A fim de abranger os conhecimentos da espécie *H. rubrisubalbicans*, bem como conhecer as características particulares dentre as estirpes pertencentes a essa espécie, este projeto tem como objetivo sequenciar, montar e anotar o genoma da estirpe tipo M4 de *H. rubrisubalbicans*, bem como, comparar suas características genômicas com a sequência genômica completa de *H. rubrisubalbicans* M1.

2. REVISÃO BIBLIOGRÁFICA

2.2 GÊNERO Herbaspirillum

A espécie *Herbaspirillum seropedicae*, isolada de raízes e da rizosfera de alguns cereais, foi originalmente descrita como uma nova

espécie do gênero *Azospirillum*, por crescer em meio semi-sólido livre de nitrogênio. Entretanto, Baldani *et al.* e Falk *et al*. descobriram que essa nova espécie pertencia à classe das Betaproteobactérias e criaram um novo gênero, *Herbaspirillum*, com uma única espécie, *H. seropedicae*. Mais tarde mostrou-se que esta espécie pertencia a um agrupamento de rRNA também contendo [*Pseudomonas*] *rubrisubalbicans*, e um grupo de estirpes de origem clínica (BALDANI *et al*., 1996).

Bactérias pertencentes ao gênero *Herbaspirillum* são bacilos gram-negativos, curvos, podendo também ser espiralados. Taxonomicamente pertencem ao filo Proteobacteria, classe Betaproteobacteria, ordem Burkholderiales, família Oxalobacteriaceae. As células individuais têm cerca de 0,6µm a 0,7µm de diâmetro e 1,5µm a 5µm de comprimento; são flageladas, com dois flagelos polares, ocasionalmante três (BALDANI *et* al, 1986).

Este gênero atualmente tem 11 espécies descritas, pois *Herbaspirillum soli*, *Herbaspirillum canariensis*, *Herbaspirillum aurantiacum* e *Herbaspirillum psychotolerans*, foram reclassificadas como *Noviherbaspirillum* (LIN et *al*., 2013). Além de 11 genomas depositados no NCBI (*Herbaspirillum*, *H. autotrophicum*, *H. chlorophenolicum*, *H. frisingense*, *H. hiltneri*, *H. huttiense*, *H. lusitanum*, *H. mossiliense*, *H. rhizosphaerae*, *H. rubrisubalbicans*, e *H. seropedicae*). Originalmente o gênero *Herbaspirillum* foi encontrado em isolados de milho (*Zea mays*), sorgo (*Sorghum bicolor*), arroz (*Oryza sativa*), cana de açúcar (*Saccharum oficinarum*) e trigo (*Triticum aestivum*) (BALDANI *et al.*, 1986). Associam-se a raízes, caules e folhas de plantas, geralmente de gramíneas (sorgo, arroz, cana-de-açúcar) de valor econômico, e menos comumente encontradas associadas a plantas forrageiras e frutas tropicais (banana, abacaxi) (JAMES *et al*, 1997).

2.3 *Herbaspirillum rubrisubalbicans*

No fim da década de 1920 e começo da década de 1930, os pesquisadores do Louisiana Agricultural Experiment Station, descreveram uma doença, cujos sintomas eram semelhantes aos de certas doenças

bacterianas tropicais, que estavam ocorrendo em cultivos de cana-de-açúcar do estado da Louisiana, no sul dos Estados Unidos, acometendo folhas e bainhas foliares desta planta. Denominaram-na estria mosqueada, devido às características que imprimia nas plantas (CHRISTOPHER & EDGERTON, 1930). Segundo pesquisas feitas na década de 70 por Hale & Wilkie, a estria mosqueada acometeu apenas cultivos de cana-de-açúcar no estado de Louisiana, ao sul dos Estados Unidos, e mais tarde a estria vermelha foi encontrada em cultivos de sorgo no estado de Queensland, Austrália, mas nenhum caso foi encontrado nos cultivos de cana-de-açúcar no Brasil. Segundo pesquisas feitas por Galli *et al*. em 1980, apenas um cultivo em Barbados apresentou susceptibilidade à doença, ao passo que todos os outros cultivos, agronomicamente importantes se mostraram resistentes mesmo após inoculação artificial da espécie *H. rubrisubalbicans*. Nos anos 90 a espécie *Herbaspirillum rubrisubalbicans* foi isolada e classificada como *Pseudomonas rubrisubalbicans,* por conter um grupamento de rRNA pertencente à esse gênero, e descrita como agente causal da estria mosqueada descrita por Christopher e Edgerton (FIGURA 01) (OLIVARES *et al*., 1996).

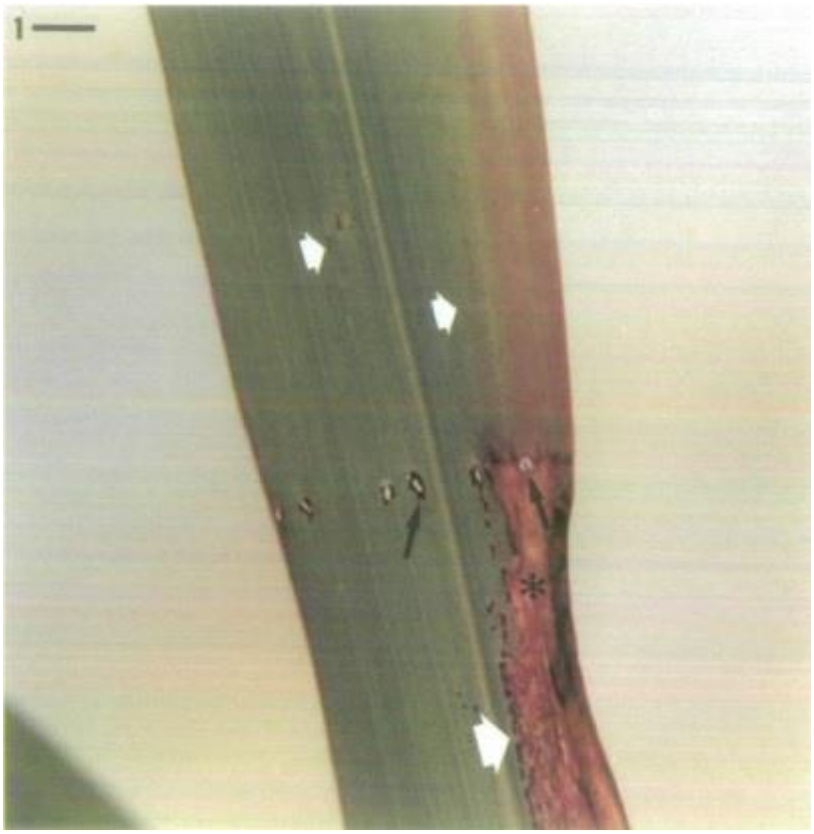


FIGURA 01. FOLHA DE SORGO (*Sorghum bicolor*) MOSTRANDO OS SINTOMAS DA DOENÇA DA ESTRIA VERMELHA.

Os pontos de inoculação (setas negras) são cercados por tecido necrótico. Os sintomas da doença podem ser vistos nas setas brancas. O asterisco (*) mostra uma extensiva área de necrose. A escala da barra é 5mm. FONTE: JAMES *et al.* (1997)

Pimentel *et. al.* (1991), testaram a patogenicidade de diversas estirpes de *H. rubrisubalbicans* e *H. seropedicae* em cana-de-açúcar, sorgo e capim Napier, concluindo que a inoculação artificial dos isolados de *H. rubrisubalbicans* causava sintomas nas três plantas usadas no ensaio. Posteriormente, James *et. al.* e Olivares *et. al.*, (1997) testaram a patogenicidade das estirpes de *H. rubrisubalbicans* e compararam com estirpes de *H. seropedicae* através de injeções de suspensão de células diretamente no limbo foliar. E espécie *H. rubrisubalbicans* desenvolveu sintomas de estria mosqueada, já a *H. seropedicae* não desenvolveu sintomas da doença, além de pequenas estrias com cerca de < 3mm de largura.

Apesar do potencial fitopatogênico de algumas estirpes *H. rubrisubalbicans*, eles também são capazes de corrigir o N_2 em caso de ausência deste componente no meio, como demonstrado pela redução de

acetileno e incorporação do $^{15}N_2$ em meio semi-sólido com ausência de nitrogênio (BALDANI *et al*., 1992), além de serem encontradas em cana-de-açúcar cultivadas no campo em condições assintomáticas (BALDANI *et al*., 1996; OLIVARES *et al*., 1996).

A estirpe *Herbaspirillum rubrisubalbicans* M4 (ATCC 19308) é a estirpe tipo e foi originalmente isolada a partir das folhas e do caule da cana-de-açúcar dos Estados Unidos (URETA *et al*., 1994). A estirpe M1 de *H. rubrisubalbicans* é a mais agressiva em causar a doença e já teve o seu genoma sequenciado pelo Núcleo de Fixação de Nitrogênio da UFPR.

2.4 GENOMA DE *Herbaspirillum rubrisubalbicans* M1

A montagem do genoma da estirpe M1 foi feito utilizando sequências obtidas do 454 FLX Titanium (560.138 mate-paired reads) e também do MegaBace 1000 (22.985 reads). O genoma de *H. rubrisubalbicans* M1 é composto de um único cromossomo contendo 5.611.261 pares de bases com 61.5% G+C. Foram anotadas 4.758 ORFs, 3.941 codificam pra proteínas de função conhecida, 439 codificam para proteínas conservadas hipotéticas e 312 para proteínas hipotéticas (BALSANELLI et al., 2016).

O genoma de *H. rubrisubalbicans* M1 contém genes que codificam para as seguintes vias do metabolismo geral: Entner-Doudoroff, pentose fosfato, TCA e cadeia de transporte de elétrons. De acordo com os genes encontrados no genoma o *H. rubrisubalbicans* M1 pode crescer utilizando amônio, nitrato, nitrito, ureia e nitrogênio atmosférico (BALSANELLI et al., 2016). Também foram encontrados genes que codificam para proteínas envolvidas com 5 sistemas de secreção, TSI, TSII, TSIII, TSV, TSVI e o pili tipo IV; genes envolvidos com a adesão a superfícies celulares; síntese de celulose; produção de fitohormônios, entre outros genes que podem estar envolvidos na interação entre a planta e a bactéria (BALSANELLI *et al*., 2016).

3 OBJETIVOS

Montar e anotar o genoma da bactéria *Herbaspirillum rubrisubalbicans* M4, bem como comparar as características encontradas nos genomas das estirpes M1 e M4 de *H.rubrisubalbicans*

4 MATERIAIS E MÉTODOS

4.1 SEQUENCIAMENTO DO GENOMA DE *Herbaspirillum rubrisubalbicans* M4 (*H.r.*M4)

Os dados do *H.r.*M4 foram obtidos a partir de bibliotecas construídas no laboratório de Fixação Biológica de Nitrogênio-UFPR pela Dra. Michele Zibetti Tadra e pelo Dr. Eduardo Balsanelli. O genoma da estirpe *H.r.*M4 foi sequenciado pelas plataformas Illumina MiSEQ e ION PROTON, em que foram gerados *reads* pela técnica de DNA polimerase e pela detecção de H^+ liberados na síntese, respectivamente.

Após o sequenciamento foi realizada a montagem dos fragmentos obtidos com a ferramenta CLC Genomics Workbench v.7.5.

4.1.1 Procedimento do sequenciamento Illumina MiSEQ

O preparo da biblioteca foi feito com kit NEXTERA XT, que é ideal para pequenos genomas, amplicons e plasmídeos (NGS). Foram feitas duas corridas, MISEQ1 e MISEQ2, em Paired-End do genoma de *H.r.*M4 pela plataforma Illumina MiSEQ usando kit V3 com 600 pb, gerando 1.136.658 *reads* e 2.964.856 *reads*, respectivamente, pelo método da DNA polimerase.

4.1.2 Procedimento do sequenciamento Ion Próton

O DNA da bactéria foi usado para a construção da biblioteca utilizando o Ion Plus Fragment Library Kit (Thermo Fisher Scientific™), seguindo as recomendações do fabricante.

4.2 ANÁLISE DE QUALIDADE DO CONJUNTO DE DADOS DAS CORRIDAS ILLUMINA E ION

O FASTQC é uma ferramenta que visa fornecer de maneira simples a execução de algumas verificações de controle de qualidade em dados de sequência de matérias provenientes de pipelines de sequenciamento de alto rendimento. Fornece um conjunto modular de análises que permite ao analista saber de quaisquer problemas que seu conjunto de dados tenha antes de qualquer análise posterior (FASTQC, 2015). Disponível em: http://www.bioinformatics.babraham.ac.uk/index.html

A ferramenta permite a importação de dados de arquivos BAM, SAM ou FastQ, gera relatórios em forma de gráficos e tabelas de forma resumida a fim de otimizar o trabalho do pesquisador (FASTQC, 2015). Disponível em: http://www.bioinformatics.babraham.ac.uk/index.html

4.3 MONTAGEM DOS *READS* MISEQ E ION

Após o sequenciamento foi realizada a montagem do conjunto de dados composto pelos *reads* MiSEQ e Ion com a ferramenta CLC Genomics Workbench v.7.5, cujos parâmetros usados foram *Word size* de 24 e tamanho mínimo para *contig* de 500. Para as opções de mapeamento foram adotados valor 2 para mismatch, fração de comprimento de 50% e fração de similaridade de 80%.

4.4 ALINHAMENTO DO GENOMA

A fim de avaliar a montagem feita com os *reads* Illumina e ION PROTON, os dados foram alinhados com os genomas de *Herbaspirillum seropedicae* (SmR1) e o *Herbaspirillum rubrisubalbicans* M1 (*H.r.*M1), os quais estão presentes no banco de dados público do NCBI, cujos números de acesso são CP002039.1 e CP013737.1, respectivamente. Os alinhamentos foram gerados através do software MUMmer 3.0, que alinha genomas inteiros ou incompletos de forma rápida usando pouca memória ram. O software é um pacote modular e versátil que se baseia em uma estrutura de dados de sufixos para correspondência de padrão. Os

alinhamentos são produzidos através de MUMs (Maximum Unique Match), que são fragmentos de dois genomas que possuem alta similaridade.

Os programas que compõe o pacote do software MUMmer que foram usados para análise do genoma da estirpe M1 foram:

- NUCmer – é um script em Pearl usado para alinhamento de sequências de nucleotídeos com alta similaridade;
- Run-mummer 1 – script em cshell que faz o alinhamento de duas sequências de DNA. O programa gera bons resultados quando alinha sequências de DNA muito semelhantes e identifica as diferenças entre elas. É recomendado para comparações do tipo um contra um, mas sem rearranjos (MUMmer 3 Manual – (http://mummer.sourceforge.net/manual/#description).

4.5 CONTIGS ILLUMINA

Os *contigs* resultantes das montagens MiSEQ 1 e 2 e ION, gerados pelo CLC Workbench v.7.5, foram analisados um a um pelo software UGENE a fim de analisar a distribuição dos reads que compõem os *contigs* de cada montagem. A fim de verificar a qualidade da distribuição dos *reads* em toda extensão do genoma, foram verificados a cobertura e presença ou ausência de empilhamento dos *reads*,

.

4.6 ANOTAÇÃO DO GENOMA

4.6.3 RAST

Para a pré-anotação do genoma da bactéria *H.r.*M4 referente aos dados das montagens Illumina (MISEQ1+MISEQ2), e ION, os arquivos contendo apenas as sequências dos *contigs* do genoma de *H.r.*M4 (.fasta), extraídos do programa CLC Workbench 7.5, foram submetidos ao anotador RAST. Foram gerados arquivos gbk, fasta e faa (contendo os aminoácidos oriundos da anotação), os quais foram usados para verificação de genes presentes no genoma no software Artemis, alinhamentos no Blast e COG, o qual faz inferência de função dos genes encontrados e construção de

vias metabólicas pelo KEGG. Os parâmetros usados foram os pré-estipulados pelo programa.

4.6.1 Blast (Basic Local Alignment Search Tool)

O BLAST (Basic Local Alignment Search Tool) é um programa que compara sequências de nucleotídeos ou base de dados de sequências de proteínas e calcula a significância estatística dos resultados.

O algoritmo Blast foi usado a fim de buscar similaridades e distinções entre a estirpe de estudo contra a referência, a partir dos resultados gerados pela anotação. Foram usados o blastn, que buscam o melhor alinhamento entre sequências compostas de nucleotídeos usando uma *query* composta, também, de nucleotídeos.

4.6.2 COG (Clusters of Orthologous Groups)

O banco de dados de COG (Clusters of Orthologous Groups of Protein) (CPV), que representa uma tentativa em uma classificação filogenética das proteínas codificadas em genomas completos, atualmente consiste de 2791 COGs incluindo 45 350 proteínas a partir de 30 genomas de bactérias, archaea e da levedura *Saccharomyces cerevisiae* (http://www.ncbi.nlm.nih.gov/COG) (TATUSOV *et. al.*, 2001).

A atualização de 2015 do COG foi usada para inferir função a partir dos dados gerados pela anotação do genoma feita pelo software RAST.

4.7 ANÁLISE DAS VIAS METABÓLICAS DO GENOMA DE *H. rubrisubalbicans* M4

A análise das vias metabólicas da espécie *H.r.*M4 foi feita pelo programa KEGG (*Kyoto Encyclopedia of Genes and Genomes*), que consiste em um recurso *on line* que contém uma coleção de banco de dados os quais possibilitam a compreensão das funções de sistemas biológicos como as células, organismos e ecossistemas a partir de

informações de nível molecular, especialmente conjuntos de dados moleculares de larga escala gerados pelos sequenciamentos de genomas e outros dados com alto rendimento.

Os dados gerados foram comparados com o metabolismo da espécie *H.r.*M1.

5 RESULTADOS E DISCUSSÃO

5.1 SEQUENCIAMENTO E ANÁLISE DA QUALIDADE DOS DADOS DE SEQUENCIAMENTO DO GENOMA DO *Herbaspirillum rubrisubalbicans* M4

O genoma da espécie *H. rubrisubalbicans* estirpe M4 foi sequenciado usando duas plataformas diferentes de sequenciamento, Illumina e ION. A primeira técnica usada gerou, em duas corridas, um total de 4.101.514 *reads*, já a segunda gerou 4.112.524 *reads* em apenas uma corrida. Os dados dos sequenciamentos Illumina e ION podem ser visualizados nas tabelas 01 e 02, respectivamente.

TABELA 01. DADOS PRODUZIDOS PELO SEQUENCIAMENTO ILLUMINA GENOME ANALYSER

Corridas	**MISEQ1**	**MISEQ2**
Nº *reads*	1,136,658	2,964,856
***Reads* pareados**	677,926	2,398,368
Nº de bases	163,623,718	721,856,390
Tamanho médio dos *reads*	143,95	251,62

FONTE: a autora (2016) com base em CLC Workbench e NGS.

TABELA 02. DADOS PRODUZIDOS PELO SEQUENCIAMENTO ION PROTON

Corridas	1
Nº *reads*	4.112.524
Nº de bases	313.474.028
Tamanho médio dos *reads*	76,22

FONTE: a autora (2016) com base em CLC Workbench e ION Próton.

As imagens de qualidade demonstradas abaixo mostram um gráfico ‘BoxWhisker’ ou ‘BoxPlot’, em que os elementos consistem em: linha vermelha central, que se refere ao valor da mediana; caixa amarela que representa o intervalo inter-quartil (25-75%); traços superiores e inferiores representam os pontos de 10% e 90%; e a linha azul, que refere-se à qualidade média (CIRINO, 2014).

Os gráficos seguintes demonstram uma análise de qualidade dos dados brutos dos *reads* ilumina e ION Próton. Em ambos observa-se uma diminuição na qualidade das bases a medida que se aproxima do fim. A análise da corrida Illumina, mostradas nas figuras 02 e 03, mostrou uma queda na qualidade de 50-60 últimas bases. Em contrapartida, a qualidade das bases Próton (FIGURA 04) demonstra superioridade às bases Illumina em toda extensão do genoma, apesar de apresentarem condição mediana, pode ser observada maior uniformidade entre as bases, ao contrário dos dados Illumina, que apesar de afigurarem melhor qualidade nas primeiras bases, não podemos deixar de observar a significativa queda na acurácea das mesmas entre as 50-60 últimas bases em ambas corridas geradas pela técnica de DNA polimerase. Os resultados verificados neste capítulo são vistos em análises posteriores feitas no CLC Workbench, descritos no próximo capítulo, em que a qualidade baixa dos *reads* MiSEQ comprometem bons resultados de montagem do genoma, a superioridade da qualidade dos *reads* ION em comparação aos *reads* MiSEQ vistas neste capítulo são demonstradas nos dados resultantes das montagens feitas no CLC Workbench, demonstrados a seguir.

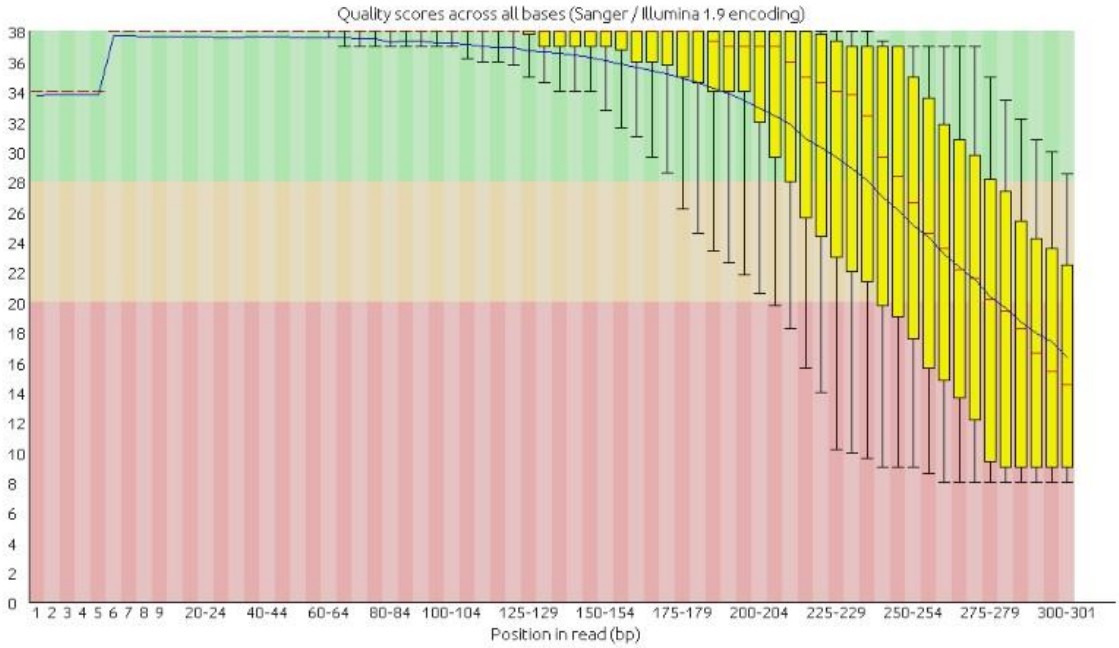


FIGURA 02. Gráfico tipo BoxPlot que demonstra a qualidade das bases dos dados do segundo sequenciamento (MISEQ2) da espécie H.r.M4. O eixo Y mostra os índices de qualidade. O eixo X representa a posição do *reads* de cada base.

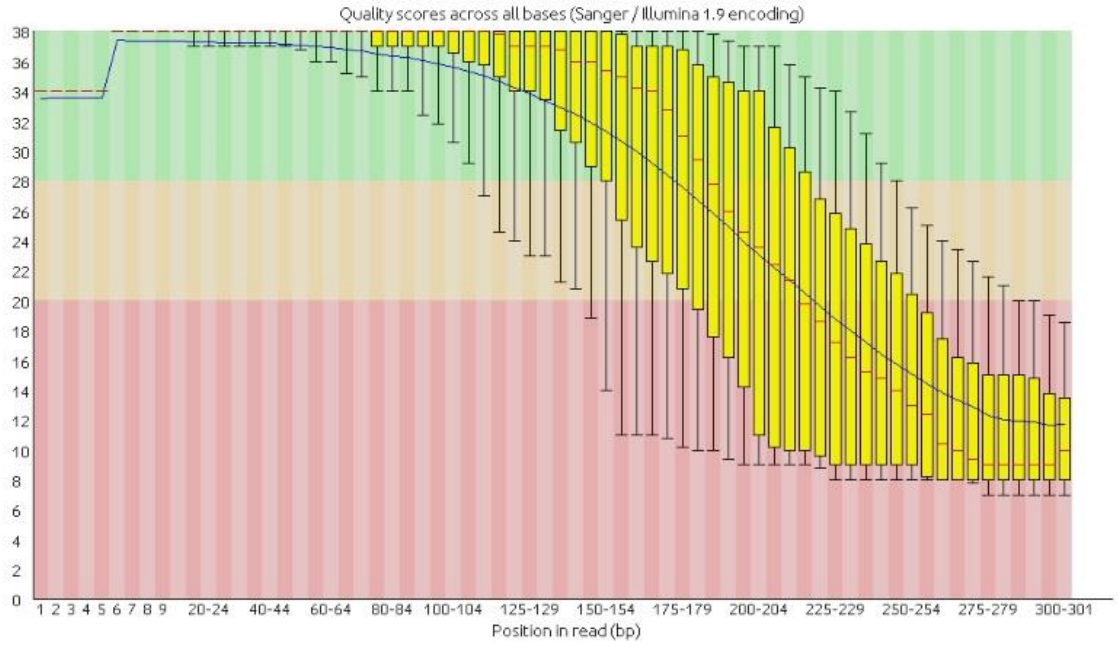


FIGURA 03. Gráfico tipo BoxPlot que demonstra a qualidade das bases dos dados do segundo sequenciamento (MISEQ1) da espécie H.r.M4. O eixo Y mostra os índices de qualidade. O eixo X representa a posição do *reads* de cada base.

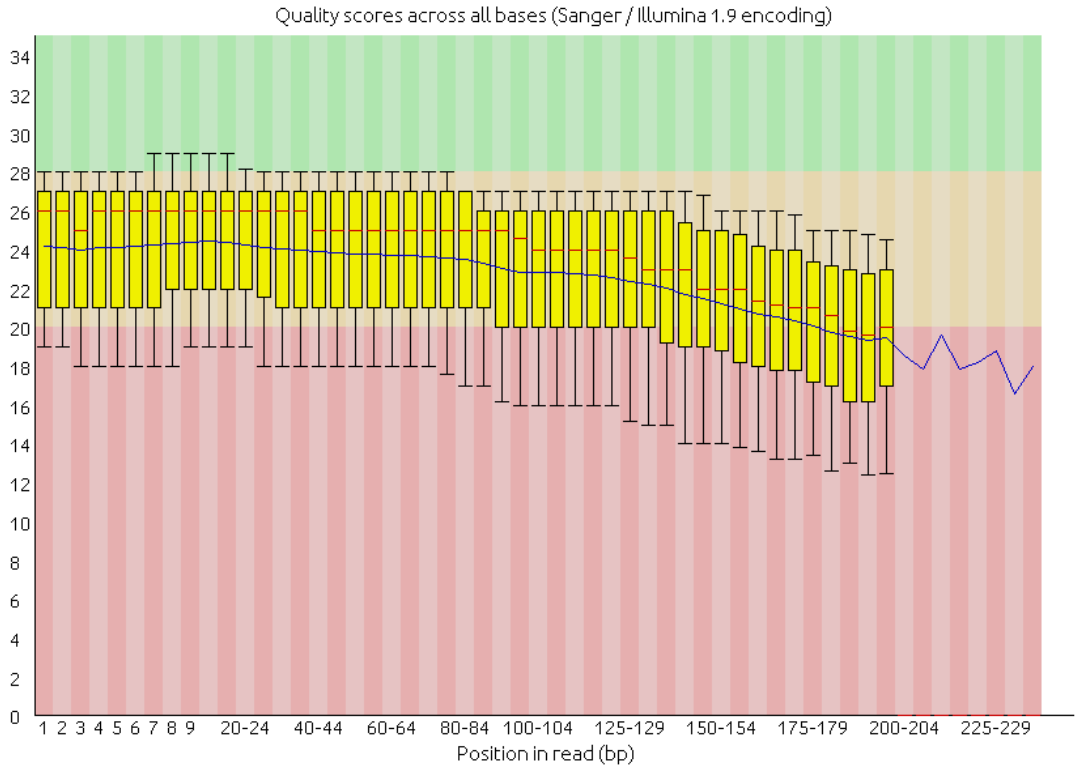


FIGURA 04. Gráfico tipo BoxPlot demonstrando a qualidade dos dados gerados pelo sequenciador ION Próton do genoma de *H.r.*M4. O eixo Y mostra os índices de qualidade. O eixo X representa a posição do *reads* de cada base.

5.2 MONTAGEM DOS DADOS DE SEQUENCIAMENTO DO GENOMA DE *Herbaspirillum rubrisubalbicans* M4 OBTIDOS UTILIZANDO AS PLATAFORMAS ILLUMINA E ION

As montagens Illumina, foram avaliadas a partir dos dados gerados pelo montador CLC Workbench, de acordo com a tabela 03, pode ser observada uma melhora geral entre MISEQ1 e MISEQ2, como já observado nos resultados de qualidade das corridas geradas pela plataforma Illumina. A montagem MISEQ2 apresentou um aumento da %GC, além do aumento da cobertura e tamanho estimado do genoma. Entretanto, foi observada grande falha na estrutura do genoma em ambas as montagens (ver capítulo 6.3). Uma terceira montagem (MISEQ3) foi feita utilizando os dados de MISEQ1 e MISEQ2, que resultou em um genoma de 4Mb de tamanho e cobertura de 204,25 vezes (TABELAS 04 e 05), em contra partida, devido à baixa qualidade das montagens que compõe o MISEQ3, ainda foi possível observar grande fragmentação dos *contigs*, o que dificulta a análise do genoma e sua comparação com outros genomas de *Herbaspirillum*.

Uma quarta montagem foi feita utilizando as leituras ION. Os dados da tabela 06 mostram que o genoma gerado tem 4,1Mb e uma cobertura de 30 vezes.

Na comparação entre as quatro montagens (MiSEQ1, MiSEQ2, MiSEQ3 e ION), pudemos verificar que a quarta montagem, gerada utilizando as leituras do ION Próton, foi a que apresentou os melhores resultados, com uma melhor qualidade, o que permitiu um um aumento do tamanho do genoma, melhor distribuição das leituras dentro dos *contigs,* o que confere a esta montagem uma cobertura mais bem distribuída em todo o genoma, propiciando uma melhor análise das características da estirpe.

A princípio acreditava-se que ao reunir os *reads* das três montagens haveria possibilidade conseguir cobrir uma maior extensão do genoma, fechando *gaps*, e aumentando a cobertura em áreas com baixa cobertura. Desta forma, foram feitas montagens com os *reads* MiSEQ1, MiSEQ2 e ION, mas os resultados encontrados foram piores do que os obtidos apenas com a montagem dos *reads* ION. Sugere-se que deve ter sido causado pela baixa qualidade dos *reads* MISEQ1, demonstrada desde as análises geradas pelo programa FASTQC, sendo desconsiderados das análises posteriores..

Desta forma, com base nos resultados já demonstrados, uma quinta montagem foi feita usando os *reads* da melhor montagem Illumina (MiSEQ2) com os *reads* ION, denomina-no-a de MiSEQ2_ION (TABELA 06). Podemos obervar que, de forma geral, houve melhora em alguns dados como aumento do valor de N50, no tamanho do genoma, bem como uma diminuição no número de *contigs*, se comparado com a montagem dos *reads* ION.

TABELA 03. MÉTRICAS DE MONTAGEM GERADOS COM O MONTADOR CLC7 WORKBENCH

	- MISEQ1	- MISEQ2
Reads alinhados	832,896	2,721,032
Bases alinhadas	134,804,431	684,668,665
Distância entre pares	40-60	40-60
Tamanho do genoma estimado	2,176,906	3,395,549
Cobertura do genoma	7,8	212,44

Número de bases nos *contigs*	17,845	23,179
Contig N50	917	1,131
Maior *contig*	17,845	23,179
Conteúdo G+C dos *scaffolds*	57,8	59,4
Número total de *contigs*	2,339	3,143

TABELA 04. DADOS PRODUZIDOS POR MONTAGEM DAS DUAS CORRIDAS (MISEQ1+MISEQ2)

Corridas	**MISEQ3**
Nº *reads*	4.101.514
Reads pareados	3.250.934
Nº de bases	885.480.108
Tamanho médio dos *reads*	215,89

FONTE: a autora, baseado em dados do CLC Workbench

TABELA 05. MÉTRICAS DAS MONTAGENS ILLUMINA REFERENTE ÀS DUAS CORRIDAS (MISEQ1+MISEQ2)

	MISEQ3
Reads alinhados	3.725.550
Bases alinhadas	834.239.040
Distância entre pares	50
Tamanho do genoma estimado	4.101.156
Número de bases nos *contigs*	3.183
Contig N50	1,373
Maior *contig*	44,277
Conteúdo G+C dos *scaffolds*	60,1%
Número total de *contigs*	3,157
Cobertura do genoma	204,25

FONTE: a autora, baseado em dados do CLC Workbench

TABELA 06. MÉTRICAS DA MONTAGEM ION

	ION	**MiSEQ_ION**
Reads alinhados	3.807.189	2.561.498
Bases alinhadas	291.107.314	
Distância entre pares	50	40-50
Tamanho do genoma estimado	5.412.692	5.630.759

Cobertura do genoma	30	
Número de bases nos *contigs*		
Contig N50	3.570	7.456
Maior *contig*	52.893	52.893
Conteúdo G+C dos *scaffolds*	61,5%	61,4%
Número total de *contigs*	2.100	1.208

FONTE: a autora, baseado em dados do CLC Workbench

5.3 *CONTIGS* DO *Herbaspirillum rubrisubalbicans* M4 OBTIDOS A PARTIR DE DADOS ILLUMINA

A análise dos *contigs* Illumina foi feita com o software UGene, em que foram escolhidos quatro *contigs* que demosntram como se apresenta toda a extensão do genoma, para essa montagem. Foi observada uma baixa qualidade dos *contigs* Illumina. As características observadas foram: quebra súbita de *reads*, como mostra a figura 06; falta de uniformidade na cobertura, caracterizada pelo empilhamento dos *reads* em uma parte do *contig* (FIGURA 05), e entre eles, como mostra as figuras 06 e 07.

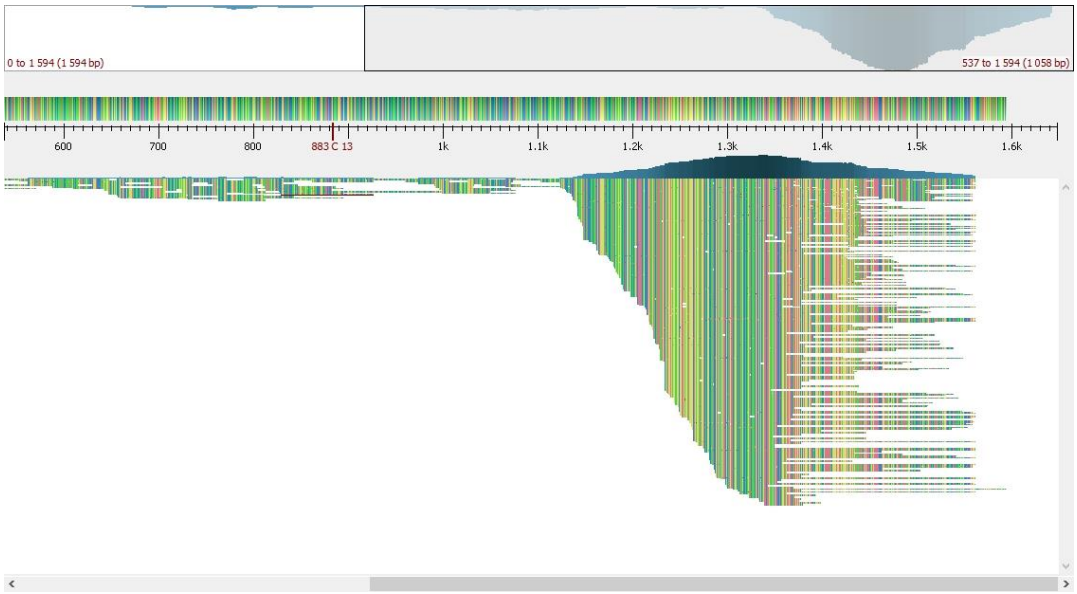


FIGURA 05. ESTRUTURA REFERENTE AO *CONTIG* 1010 DA bactéria *H.r.*M4. DEMONSTRA EMPILHAMENTO E COBERTURA IRREGULAR DE *READS.*
FONTE: UGENE

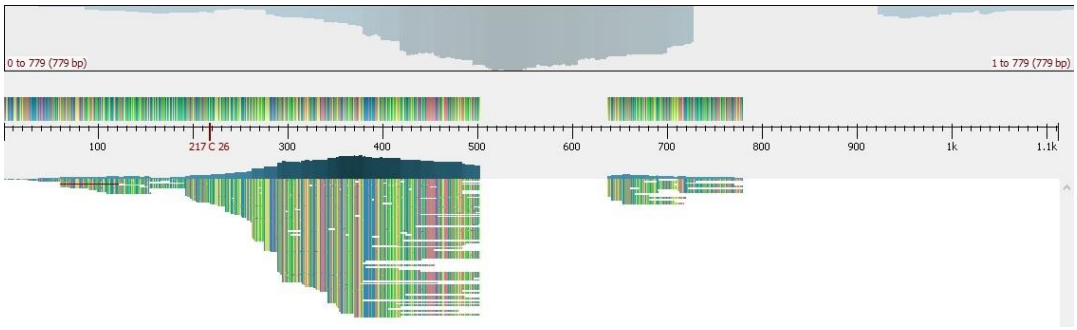


FIGURA 06. ESTRUTURA REFERENTE AO *CONTIG* 1014 DA bactéria *H.r.*M4. DEMONSTRA EMPILHAMENTO E COBERTURA IRREGULAR DE *READS,* ALÉM DE QUEBRA REPENTINA DO *CONTIG*.
FONTE: UGENE

FIGURA 07. ESTRUTURA REFERENTE AO *CONTIG* 102 DA *H.r.*M4. Demonstra empilhamento e cobertura irregular dos *contigs.*
FONTE: UGENE

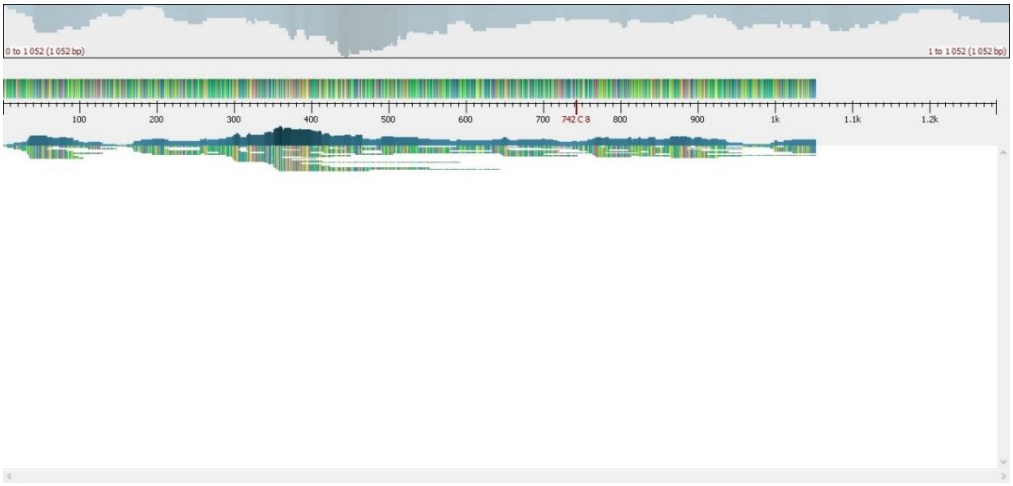


FIGURA 08. ESTRUTURA REFERENTE AO *CONTIG* 990 DA ESPÉCIE *H.r.*M4. Demonstra pouca cobertura, além de quebra do *contig.*
FONTE: UGENE

5.4 CONTIGS DO *Herbaspirillum rubrisubalbicans* M4 OBTIDOS A PARTIR DE DADOS ION

A análise dos *contigs* ION foram feitos com o software UGene, em que foram escolhidos três *contigs* (FIGURAS 09, 10 e 11) que demonstram como se apresenta toda a distribuição dos *reads*, para essa montagem. Foi observada uma cobertura bem distribuída em toda extensão do genoma da espécie *Hr*M4, não sendo encontrada quebra súbita, ou empilhamento dos *reads*.

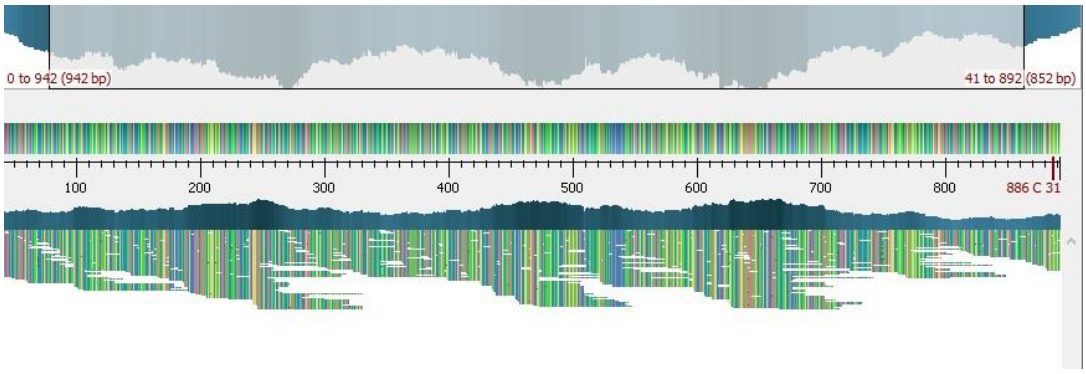


FIGURA 09. ESTRUTURA REFERENTE AO *CONTIG* 01 DA *H.r.*M4.
FONTE: UGENE

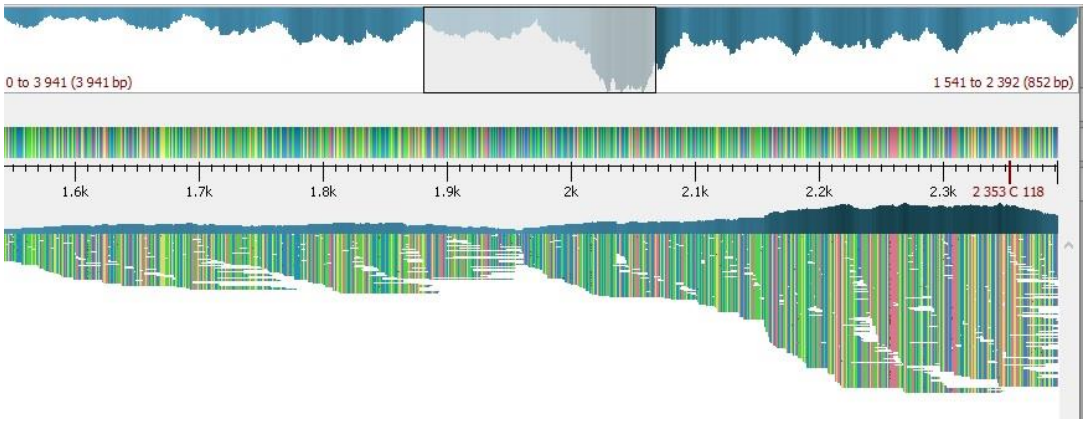


FIGURA 10. ESTRUTURA REFERENTE AO *CONTIG* 884 DA *H.r.*M4.
FONTE: UGENE

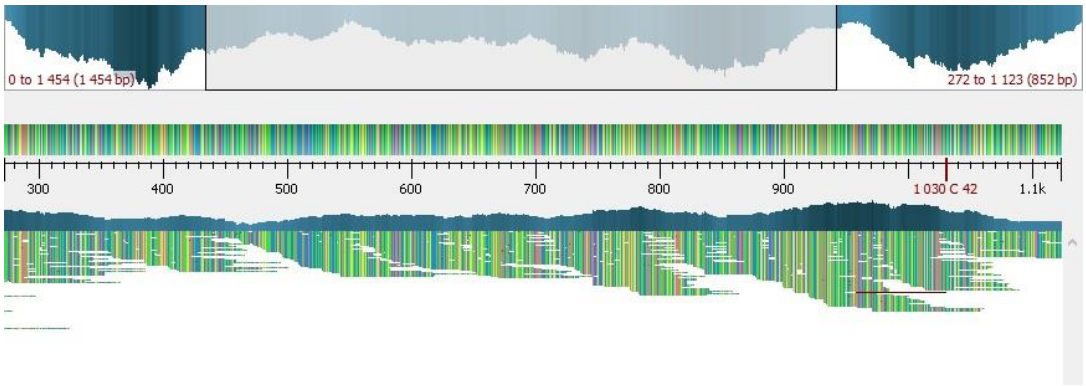


FIGURA 11. ESTRUTURA REFERENTE AO *CONTIG* 1830 DA *H.r.*M4.
FONTE: UGENE

Pode-se verificar, ao compararmos os *contigs* Illumina com os ION, representados pelas figuras 09, 10 e 11, uma grande diferença na distribuição da cobertura dos *contigs.* Com isso, pode-se constatar que o sequenciamento de *H. rubrisubalbicans* M4 teve maior cobertura de seu genoma pela técnica de sequenciamento ION Próton, o que pode ter sido causado pela qualidade ou quantidade de material utilizada da biblioteca. As observações feitas através da análise feita com o software UGENE, vêm confirmar a qualidade dos dados gerados pela plataforma ION Proton demonstrada em análises anteriormente feitas dos dados Illumina e ION.

A partir do demonstrado nas análises anteriores, submetemos as montagens com os melhores resultados, ION e MiSEQ2_ION, ao programa GFinisher. As verificações foram feitas separadamente para cada uma das duas montagens ION e MiSEQ_ION. Para a obtenção dos resultados foram usados três arquivos, os *contigs* da montagem ION e MiSEQ_ION, o mapeamento dos *reads* ION e o genoma de *H. rubrisubalbicans* M1 usado como referência. Os *contigs* das montagens foram usados para fechar o mapeamento, gerando os resultados da tabela 07, que demonstram que houve uma boa finalização dos genomas para ambas as montagens. Entretanto, foi obtido melhor resultado na montagem MiSEQ2_ION com aumento do genoma e na diminuição de *contigs*, consequentemente uma maior diminuição de gaps na mesma, que a observada da montagem ION.

TABELA 07. RESULTADO GERADO PELO PROGRAMA GFINISHER DAS MONTAGENS ION E MiSEQ_ION

	ION	**MiSEQ2_ION**
GAPs	162	114
Contigs	163	115
Tamanho	4.784.841	4.941.801
%GC	62,330	62,157

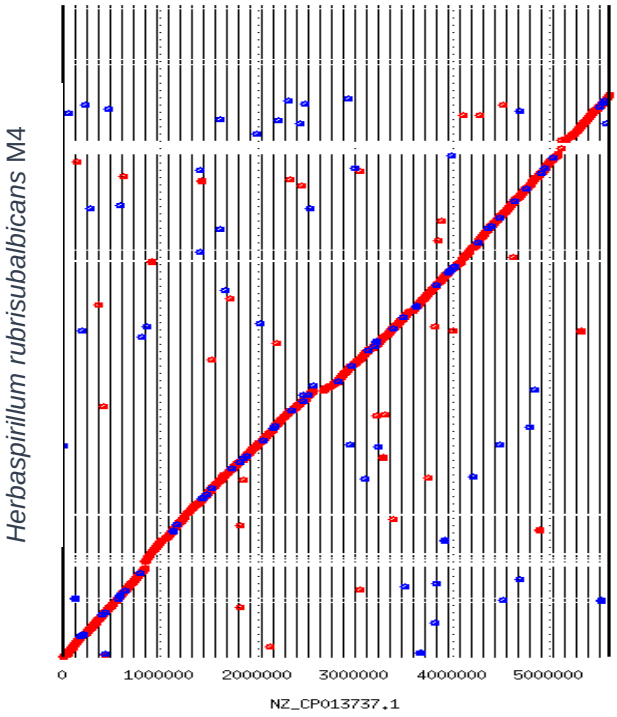
5.4 MAPEAMENTO E ORDENAÇÃO DOS SCAFFOLDS DO GENOMA DE *Herbaspirillum rubrisubalbicans* M4 OBTIDOS A PARTIR DAS MONTAGENS ILLUMINA E ION

Apesar de ter sido constatado que o melhor resultado encontrado fora o gerado pela montagem dos *reads* ION com os *reads* da montagem MiSEQ2 (MiSEQ2_ION), obtido anteriormente, os *scaffolds* das montagens MiSEQ3, ION, ION_GFinisher MiSEQ3 ION também foram ordenados e alinhados, a fim de fazer um comparativo entre eles. Podendo estabelecer quais foram os melhores alinhamentos, como também mensurar e quantificar quais sequências não foram alinhadas com a referência usada. O processo foi feito através do programa MUMmer versão 3.0, e os gráficos foram gerados com os comandos nucmer e mummerplot. O genoma da espécie *H.rubrisubalbicans* estirpe M1 foi utilizado como referência.

É possível encontrar similaridades nos três alinhamentos com a referência usada, embora sejam encontrados muitas sequências que não entraram no alinhamento, como demonstrado nas figuras 12, 13 e 14 é possível observar a relação taxonômica entre as duas espécies. O gráfico referente ao DOTPLOT dos dados ION (FIGURA 12) foi o que mais expressou coerência no alinhamento com a referência.

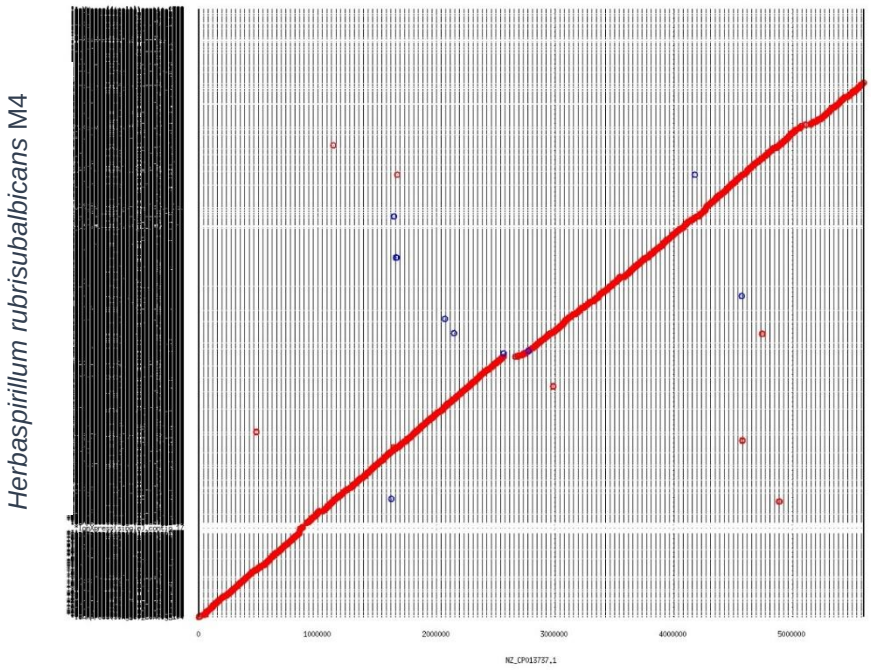
Os dados produzidos com o programa GFinisher, a partir da montagem ION, geraram um DOTPLOT (FIGURA 15) com cerca de 30 *contigs* que foram alinhados com a referência. Ainda é possível observar algumas sequências que não entraram no alinhamento, entretanto já é possível, a partir deste último resultado, ter uma melhor concepção da estrutura do genoma de *H. rubrisubalbicans* M4.

A FIGURA 16 mostra o alinhamento feito da montagem MiSEQ2_ION. Nele pode-se observar as mesmas características de análise encontradas nos outros alinhamentos, em que há o alinhamento de grande parte dos *contigs* com a referência, como também alguns *contigs* que não tiveram correlação com o genoma de comparação.



Herbaspirillum rubrisubalbicans M1

FIGURA 12. GRÁFICO DO ALINHAMENTO DOS *SCAFFOLDS* DA MOTAGEM ILLUMINA (MiSEQ 3) DE *H.r.*M4 CONTRA O GENOMA COMPLETO DE *H.r.*M1.
FONTE: Mummer



Herbaspirillum rubrisubalbicans M1

FIGURA 13. GRÁFICO DO ALINHAMENTO DOS *SCAFFOLDS* DA MONTAGEM ION DE *H.r.*M4 (ION) CONTRA O GENOMA COMPLETO DE *H.r.*M1, COM FILTRO DE ALINHAMENTOS
FONTE: Mummer.

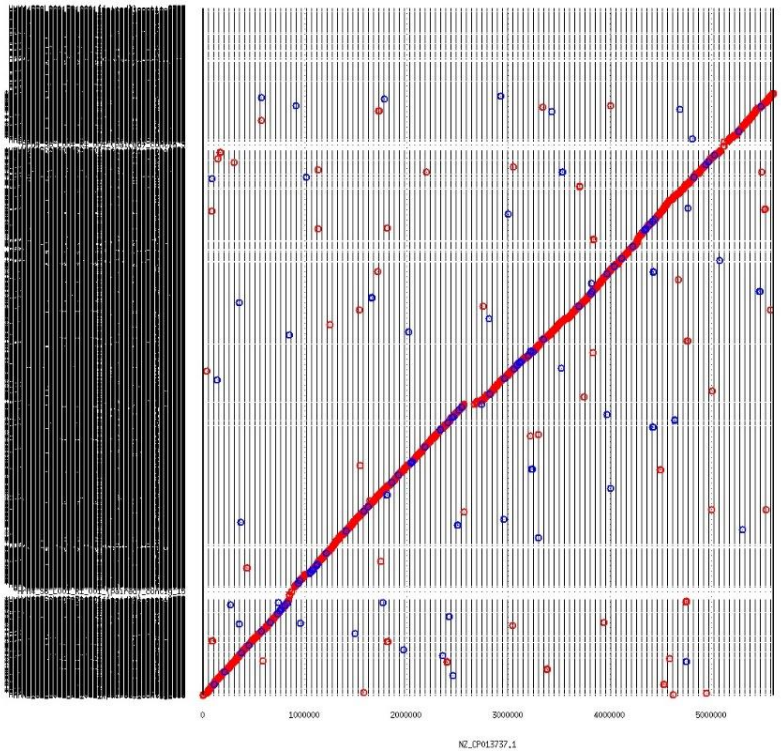


FIGURA 14. GRÁFICO *DO ALINHAMENTO* DOS *SCAFFOLDS* DA MONTAGEM MiSEQ3+ION DE *H.r.*M4 CONTRA O GENOMA COMPLETO DE *H.r.*M1, COM FILTRO DE ALINHAMENTOS
FONTE: Mummer.

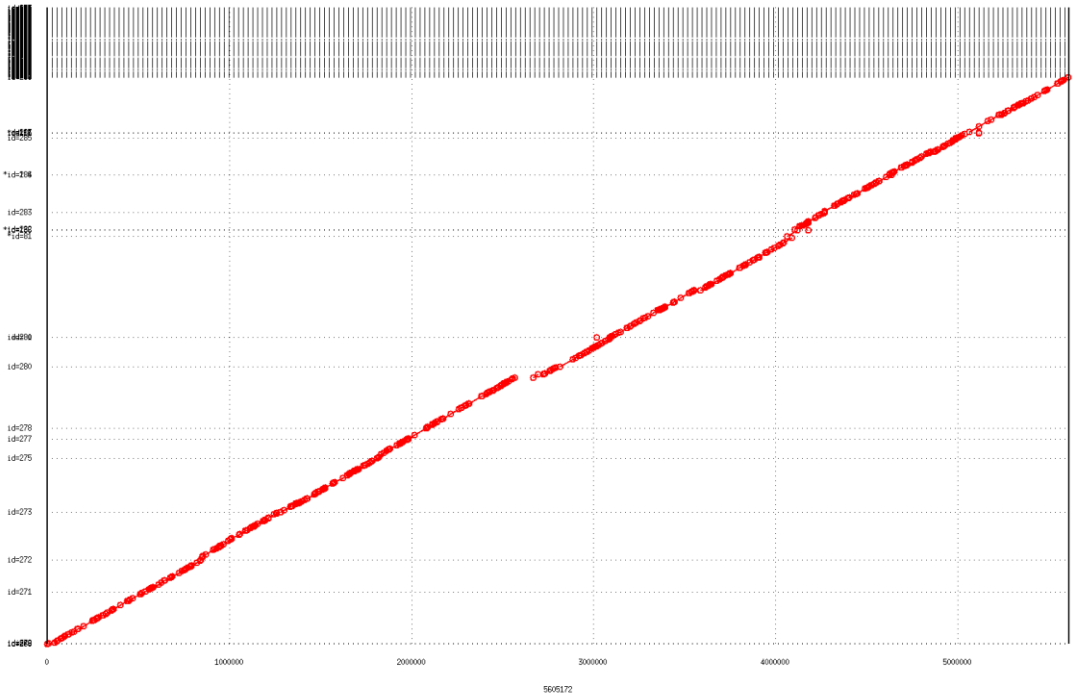


FIGURA 15. GRÁFICO DO ALINHAMENTO DOS *SCAFFOLDS* DA MONTAGEM ION DE *H.r.*M4 CONTRA O GENOMA COMPLETO DE *H.r.*M1.
FONTE: GFinisher.

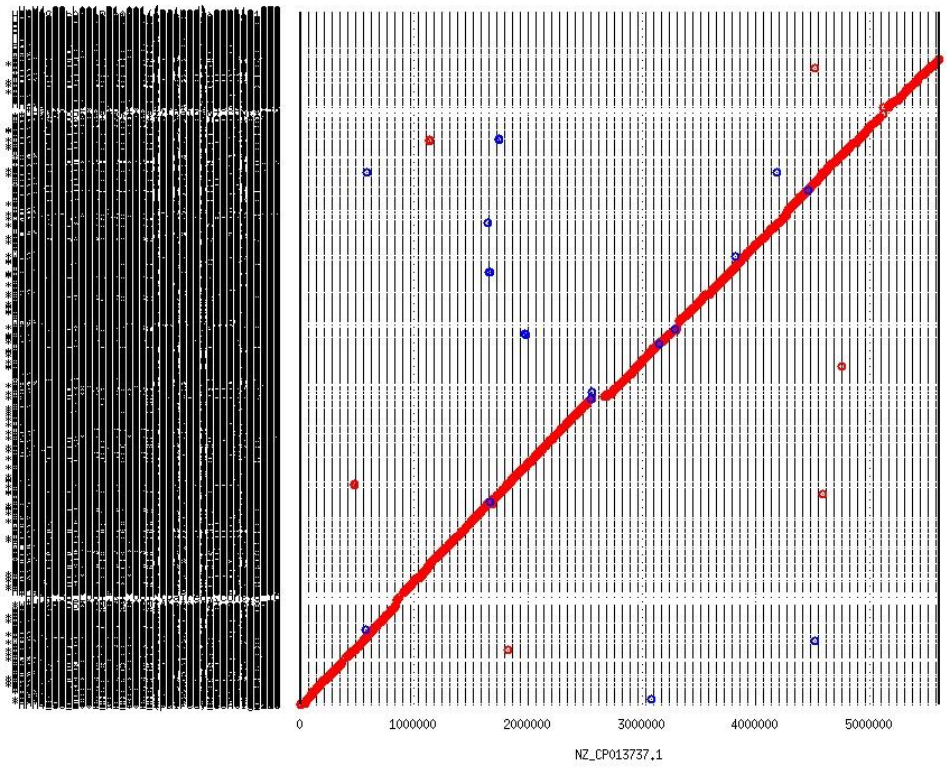


FIGURA 16. GRÁFICO DO ALINHAMENTO DOS *SCAFFOLDS* DA MONTAGEM MiSEQ2_ION DE *H.r.*M4 CONTRA O GENOMA COMPLETO DE *H.r.*M1.

Podemos perceber que em todos os gráficos gerados pelo programa MUMmer 3.0 foram encontradas sequências que não foram alinhadas com o genoma de *H.r.*M1. A fim de mensurar quantas não foram alinhadas para cada montagem, usamos o comando *show-coords*, o qual faz parte do pacote que compõe o software MUMmer. Os resultados mostraram (TABELA 08) que, o melhor alinhamento foi da montagem ION_GFinisher, apesar de ter menor número de sequências que a montagem MiSEQ2_ION, que teve 3.713 *contigs* não alinhados, dos 5.115 encontrados para essa montagem. Lembrando que, a montagem MiSEQ2_ION é resultado da junção de dois sequenciamentos, o que justifica o maior número de sequências, entretanto isso não contribuiu para um melhor alinhamento.

TABELA 08. ANÁLISE DE *CONTIGS* DAS MONTAGENS, QUE ESTÃO DESALINHADOS COM O GENOMA DE *H.R.*M1

Montagem	Número de sequências	Sequências alinhadas	Sequências não alinhadas
MiSEQ 3	3.885	3.459	426
ION	4.929	2.072	2.857
ION_GFinisher	4.419	4.285	134
MiSEQ3_ION	4.300	1.125	3.125
MiSEQ2_ION	5.115	1.402	3.713

5.5 IDENTIDADE ENTRE O GENOMA DE *H. rubrisubalbicans* M4 E O M1 CALCULADA PELO PROGRAMA ANI (AVERAGE NUCLEOTIDE IDENTITY)

A calculadora ANI estima a identidade média de nucleotídeos usando os dois melhores hits e melhores resultados entre dois conjuntos de dados genômicos. Normalmente, os valores ANI entre genomas da mesma espécie são acima de 95%, já valores abaixo de 75% não são considerados confiáveis. O genoma de *Hr*M4 foi comparado com o genoma de *Hr*M1 utilizando a plataforma ANI *Calculator*, e foi observada uma

identidade de 96,68% entre as duas estirpes. O resultado gerado pela ANI *Calculator* pode ser visualizado na figura 17.

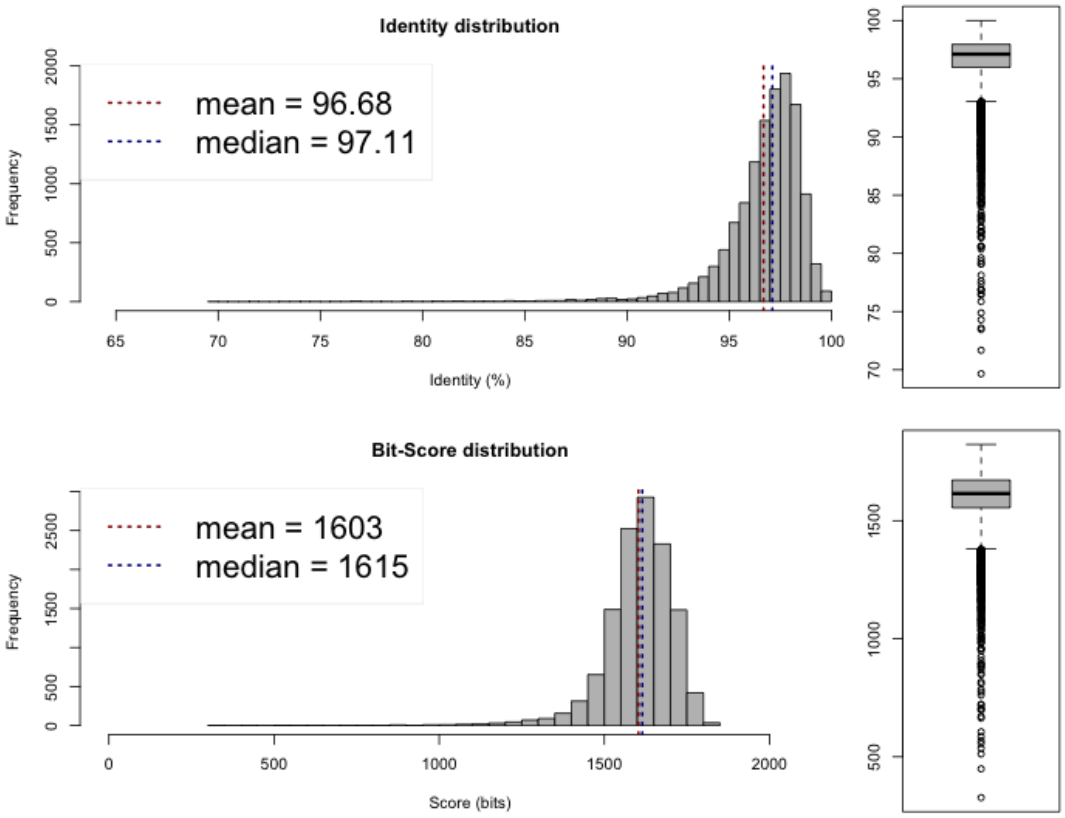


FIGURA 17. GRÁFICO DO RESULTADO GERADO POR ANI CALCULATOR NA COMPARAÇÃO DAS ESTIRPES M4 E M1 DA ESPÉCIE *Herbaspirillum rubrisubalbicans.*
FONTE: ANI *CALCULATOR*

5.6 DISTÂNCIA ENTRE OS GENOMAS DAS ESTIRPES M1 E M4 CALCULADA POR GGDC (GENOME-TO-GENOMA DISTANCE CALCULATOR)

O GGDC (*Genome-to-genoma Distance Calculator*) é um programa online que oferece métodos para inferir distâncias em todo genoma, capazes de imitar a hibridização DNA-DNA (DDH). O programa permite a obtenção de uma estimativa de semelhanças entre os genomas de duas estirpes. As funções de distância têm capacidade de trabalhar com genomas muito reduzidos e regiões de sequências repetitivas, como também mostram uma melhor correlação entre os rRNA 16S. O GGDC relata a diferença no teor G + C, o que pode ser usado para delineação de espécies (GGDC. Disponível em: (http://ggdc.dsmz.de/). O valor de DDH

mínimo para considerar dois microrganismos pertencentes à mesma espécie é de 70%.

O GGDC do genoma da espécie de *H. rubrisubalbicans* estirpe M4 foi calculado usando a estirpe M1 da mesma espécie como referência. O resultado gerado, de 72.70%, encontra-se acima da DDH mínima estipulada pelo cálculo (70%) para que dois microrganismos sejam considerados pertencentes à mesma espécie. Desta forma, pode-se afirmar que o genoma de estudo pertence à mesma espécie da referência utilizada.

5.7 PRÉ-ANOTAÇÃO GENOMA *H.r.*M4

O genoma de *Hr*M4, com tamanho aproximado de 5.6Mb, tem um GC de 61,4%, foi anotado pelo programa RAST (Rapid Annotation using Subsystem Technology). Na tabela 09 estão mostrados a classificação pelo COG dos 4.929 genes anotados na montagem do genoma de *H.r.*M4 obtida a partir dos dados gerados pela plataforma ION próton. Os genomas utilizados como referência foram do *H.r.*M1 e SmR1. Apesar de o genoma estar incompleto, comparando a distribuição dentro dessas categorias com os genes anotados de *H.r.*M1, ainda pode ser observada correspondência entre as duas espécies nas funções relacionadas a processos celulares e sinalização.

TABELA 09. GENES ANOTADOS NO GENOMA DE *H. rubrisubalbicans* (HrM4)

COG	FUNÇÃO	*H.r.*M4 MiSEQ2_ION	*H.r.*M1
	ARMAZENAMENTO E PROCESSAMENTO DE INFORMAÇÕES		
A	Modificação e processamento de RNA	1	-
J	Tradução, estrutura ribossomal e biogênese	243	114
K	Transcrição	339	307
L	Replicação, recombinação e reparo	128	135
B	Estrutura e dinâmica da cromatina	1	4
	PROCESSOS E SINALIZAÇÃO CELULAR		
D	Controle do ciclo celular, divisão celular, particionamento do cromossomo	54	25
Y	Estrutura do núcleo	-	-
V	Mecanismos de defesa	92	56
T	Mecanismos de tradução de sinal	252	211
M	Parede celular/membrana/envelope biogenesis	272	237
N	Motilidade celular	75	152

Z	Citoesqueleto	3	-
W	Estruturas extracelulares	2	-
U	Trafego intracelular, secreção, transporte de vesículas	75	50
O	Modificação pós-tradução, protein turnover, chaperones	162	115
	METABOLISMO		
C	Produção e conversão de energia	273	237
G	Transporte e metabolismo de carboidratos	255	252
E	Trasnporte e metabolismo de aminoácidos	411	422
F	Transporte e metabolismo de nucleotídeos	69	59
H	Transporte e metabolismo de coenzimas	161	115
I	Transporte e metabolismo de lipídios	206	163
P	Transporte e metabolismo de íons inorgânicos	263	230
Q	Biossintese de metabolitos, transporte e catabolismo secundário	102	84
	POORLY CHARACTERIZED		
R	Predição de funções gerais	347	342
S	Sem função conhecida	271	1633

Considerando que as estirpes M4 e M1 pertencem à mesma espécie, seria esperado que houvessem pequenas diferenças na quantidade de genes responsáveis por cada uma das funções. Entretanto, segundo os resultados demonstrados na tabela 09, pode-se observar algumas diferenças consideráveis na inferência de funções dos genes entre as duas estirpes. Algumas das funções representadas pelas letras na primeira coluna da tabela, tiveram um acúmulo de genes muito superior em M4, aos demonstrados em M1, como visto nas funções tradução, estrutura ribossomal e biogênese, motilidade celular, transporte e metabolismo de lipídios, entre outras, sendo a maior diferença encontrada nas proteínas sem função conhecida. Isso pode ser explicado pelo fato da alta fragmentação dos *contigs* em M4, podendo ocasionar que um gene, caso tiver sua sequência fragmentada, possa ser anotado mais de uma vez, como exemplificado na FIGURA 18, pois o anotador interpretará cada sequência como genes ‘diferentes” (Gene 1, Gene 2), mesmo que possuam sequências que condigam com o mesmo gene. Desta forma, poderá haver um aumento no número de genes determinados para cada função dentro do COG.

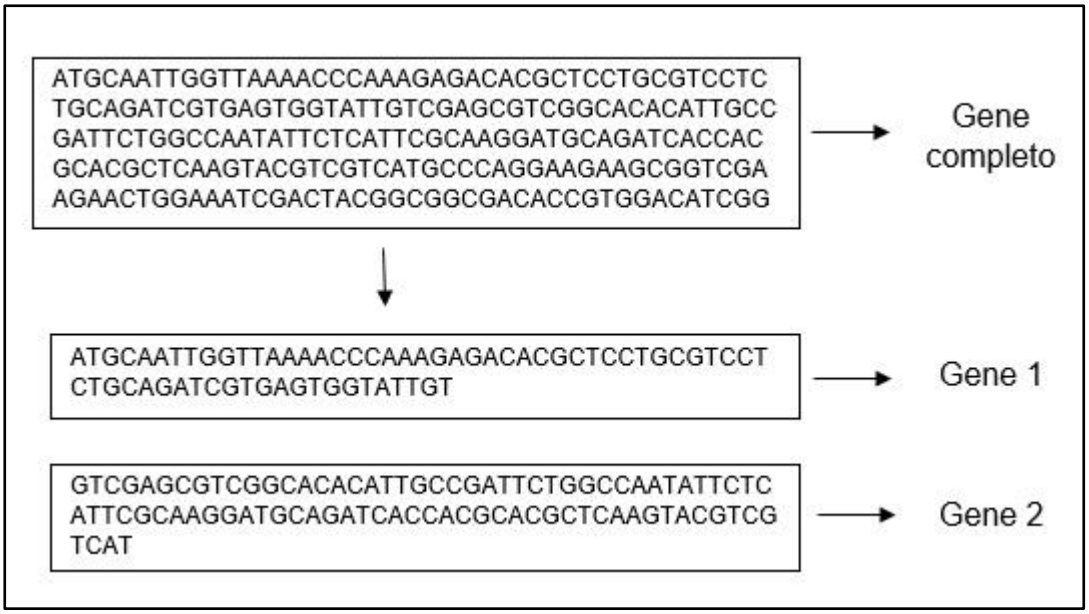


FIGURA 18. FIGURA ILUSTRATIVA DEMONSTRANDO A FRAGMENTAÇÃO DA SEQUÊNCIA DE UM GENE, PODENDO CAUSAR A ANOTAÇÃO DO MESMO GENE DUAS VEZES.

Fonte: a autora.

5.8 GENES DE *H. rubrisubalbicans* M4 ENVOLVIDOS NO PROCESSO DE INTERAÇÃO PLANTA BACTÉRIA.

O *H. rubrisubalbicans* é uma bactéria que pode se comportar como uma bactéria promotora do crescimento vegetal e como um fitopatogeno em algumas variedades de sorgo e cana de açúcar. Esse modo de vida contrastante dependente do hospedeiro está relacionado com o conjunto de genes que essa bactéria tem e também com o padrão de expressão destes genes durante a interação com o hospedeiro.

O padrão de expressão dos genes de *H. rubrisubalbicans* M1 durante a interação com o hospedeiro foi determinado por RNA-seq. Genes envolvidos com o metabolismo de carbono, fixação biológica de nitrogênio, síntese de peptidioglicano e LPS, secreção de proteínas e metabolismo de fitohormônios apresentaram um aumento de expressão durante a interação com o sorgo. Genes envolvidos nos processos acima foram identificados no genoma de M4 e estão descritos abaixo.

5.8.1 Metabolismo geral em *Herbaspirillum rubrisubalbicans* M4

O *Herbaspirillum rubrisubalbicans* M4 cresce em uma variedade de fontes de carbono, incluindo malato, glucose, arabinose, glicerol, manitol, xilose, frutose e meso-eritritol. A análise utilizando o programa KEGG mostra que o M4 tem a maquinaria enzimática que permite a utilização dessas vias de carbono. Em M4, assim como na estirpe M1 e em *H. seropedicae* SmrI, o gene que codifica para a enzima da glicólise, fosfofrutoquinase-1 esta ausente, isso indica que esse organismo deve utilizar as vias Entner-Doudoroff e Pentose fosfato pra metabolizar a glucose. Em M4 as vias das pentoses fosfatos, Entner-Doudoroff , síntese de novo de glucose, cadeia de transporte de életrons e o ciclo de Krebs estão completas, assim como em M1 (BALSANELLI et al., 2016).

5.8.2 Fixação Biológica de Nitrogênio em *Herbaspirillum rubrisubalbicans* M4

O *Herbaspirillum rubrisubalbicans* M4 fixa nitrogênio (Olivares, tese) e possui, os genes estruturais da fixação do nitrogênio, o *nifHDK*. O gene *nifH* codifica as Fe-proteínas, que consistem em dímeros (α_2) de subunidades idênticas, que contém o grupamento 4Fe-4S em cada dímero. O gene *nifD* codifica a subunidade α e o gene *nifK* codifica a subunidade β, ambas compõe a proteína Mo-Fe (molibdênio-ferro), um tetrâmero que corresponde a 4 subunidades ($\alpha_2\beta_2$) organizadas de forma heterodiméricas (REIS *et. al.*, 2006 e MARIN *et. al.*, 1998). Alem desses 3 genes, o M4 tem todos os genes necessários para a sintese e montagem da nitrogenase Fe-Mo (FIGURA 19). A organização dos genes *nif* é idêntica a da estirpe M1. Os genes *ntrBC* e os genes *glnB* e *glnK*, envolvidos na regulação do metabolismo de nitrogênio, também foram encontradps em M4.

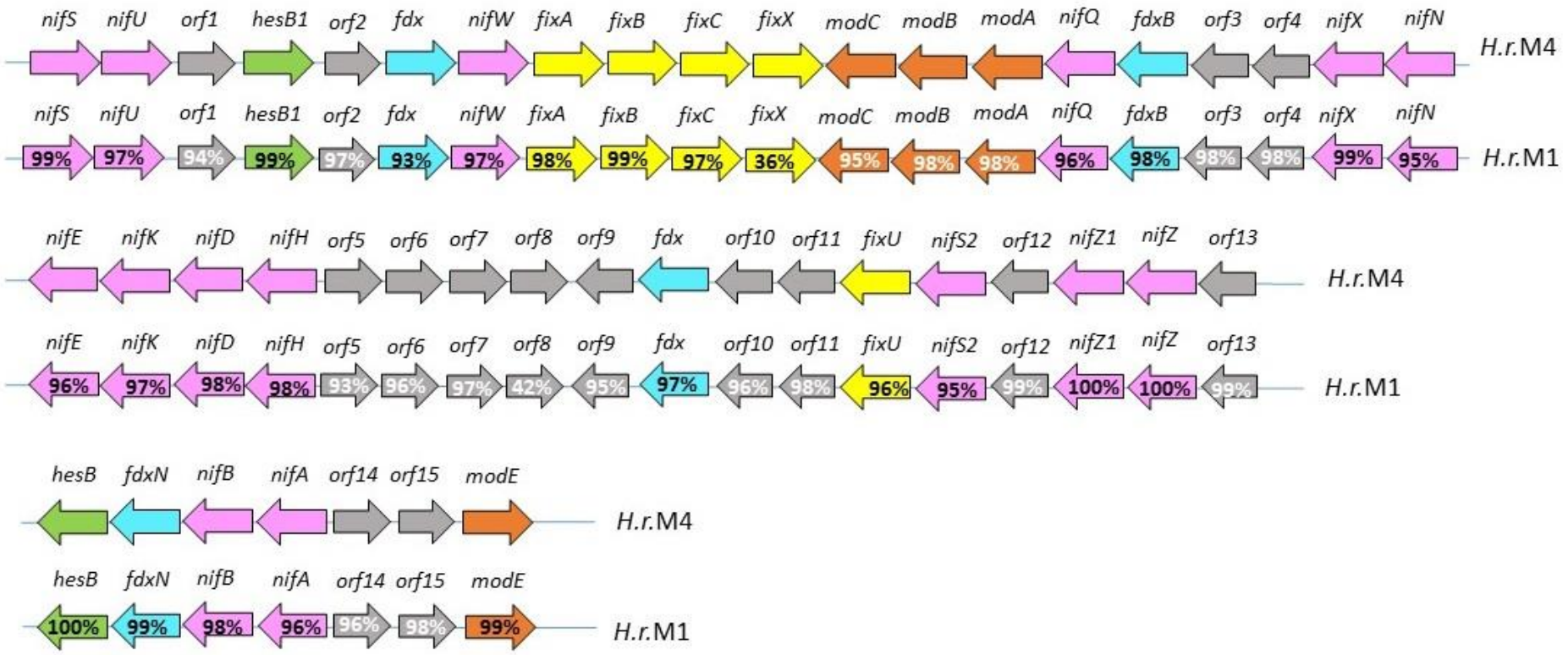


FIGURA 19. ESTRUTURA DO AGRUPAMENTO DE GENES *nif* DA ESPÉCIE *Herbaspirillum rubrisubalbicans* ESTIRPES M4 E M1.

As setas representam os genes na ordem e sentido que estão no genoma. A referência é *H. rubrisubalbicans* M4, e as porcentagens de identidade são todas em relação à referência. As setas em cinza correspondem aos genes que codificam proteínas hipotéticas.

5.8.3 Sintese de Hormônios de Plantas por *Herbaspirillum rubrisubalbicans* M4

Assim como em outras bactérias que interagem com plantas o *H. rubrisubalbicans* M4 parece ser capaz de sintetizar o ácido indol acético, envolvido no desenvolvimento da planta e degradar o 1- aminociclopropano 1-carboxilato (ACC) evitando que o mesmo seja convertido a etileno, envolvido com a resposta da planta ao estress. O M4 possui as enzimas triptofano aminotransferase e indol 3 piruvato monooxigenase que sintetizam o ácido indolacetico a partir do 2-oxoglutarato; e a enzima ACC deaminase que converte o ACC a amônia. A presença desses genes no genoma sugere que o M4 pode influenciar as respostas da planta.

5.8.4 Sistemas de secreção encontrados em *H.r.*M4

Durante a análise do genoma de *H. rubrisubalicans* M4 foram encontrados genes envolvidos com os sistemas de secreção II, VI e o pili tipo IV, que são sistemas de exportação Sec dependentes. Também foram encontrados genes de sistemas de secreção independentes de Sec, os sistemas I, III e VI. Os sistemas tipo III, tipo IV pili e tipo VI estão envolvidos com o processo de interação entre a planta e a bactéria, a estirpe M4 tem os mesmos genes que a estirpe M1 em relação a esses sistemas. Nas figuras 20 e 21 podemos obervar que a identidade entre os genes do Sistema de Secreção do Tipo III e do Pili Tipo IV tem uma identidade mais que 90% . Quanto as proteínas efetoras do sistema tipo III ainda não foi feita a análise na estirpe M4 para determinar se as duas estirpes tem as mesmas proteínas efetoras. Mutações nos genes do sistema do tipo III fazem com que a estirpe M1 seja incapaz de causar a doença da estria mosqueada na variedade de cana de açúcar B4362 (SCHMIDT et al., 2013)

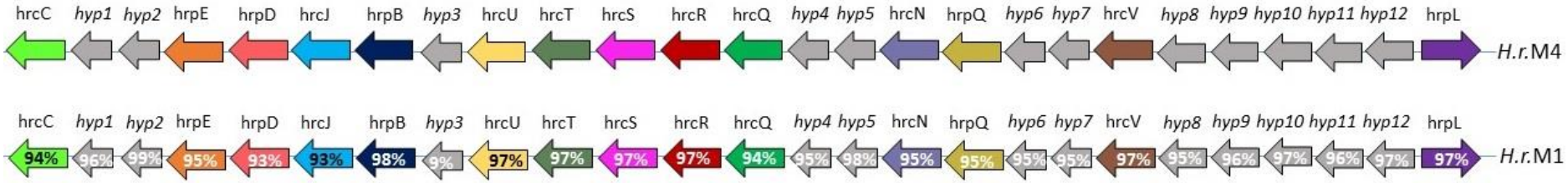


FIGURA 20. ESTRUTURA DO AGRUPAMENTO DE GENES DO SISTEMA DE SECREÇÃO TIPO III DA ESPÉCIE *Herbaspirillum rubrisubalbicans* ESTIRPES M4 E M1.

As setas representam os genes na ordem e sentido que estão no genoma. A referência é *H. rubrisubalbicans* M4, e as porcentagens de identidade são todas em relação à referência. As setas em cinza correspondem aos genes que codificam proteínas hipotéticas.

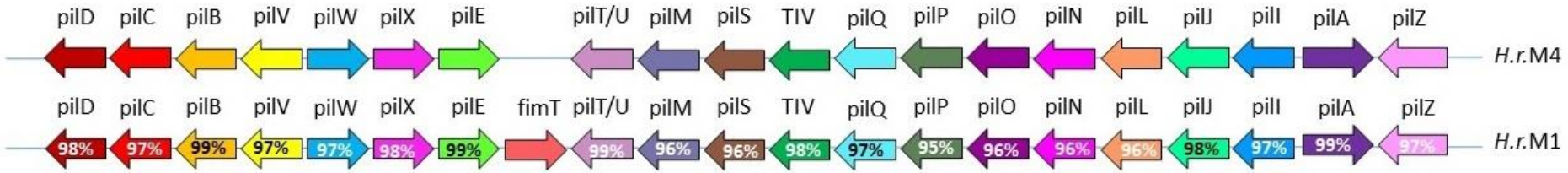


FIGURA 21. ESTRUTURA DO AGRUPAMENTO DE GENES DO SISTEMA DE SECREÇÃO TIPO IV PILIN DA ESPÉCIE *Herbaspirillum rubrisubalbicans* ESTIRPES M4 E M1.

As setas representam os genes na ordem e sentido que estão no genoma. A referência é *H. rubrisubalbicans* M4, e as porcentagens de identidade são todas em relação à referência.

5.8.5 Genes envolvidos com a síntese de LPS e celulose.

Os lipopolissacarídeos (LPS) são moléculas, cuja estrutura é altamente complexa, e são encontrados na monocamada externa da mambrana externa de bactérias gram-negativas (COLLINS & FERRIER, 1995). São formados por três regiões, distintas em sua estrutura, denominadas lipídio-A, núcleo oligossacarídico e antígeno-O (SUTHERLAND, 1985). A maioria dos genes envolvidos na síntese de LPS em *H. rubrisubalbicans* M1 foram também encontrados em M4, sendo alguns com alta similaridade de sequências, e a maioria com média similaridade de sequências. A identidade da maioria deles é superior a 50% (FIGURA 22 e 23), com exceção do gene *wecB*, que apresentou uma identidade baixa quando analisado pela sequência de aminoácidos, entretanto quando analisado pela sequência de nucleotídeos não foi encontrado no genoma de M4. Isto pode ter acontecido pelo genoma não estar completo, pois pode-se observar a presença do gene *gal* (FIGURA 22), que, assim como o gene *wec* é responsável pela diversidade estrutural do LPS. Desta forma, sugere-se que a estirpe de estudo provavelmente possui o gene *wecB*, pois se trata de um gene que está presente em todos os *Herbaspirillum*, e a ausência deste gene pode comprometer a quantidade e qualidade do LPS na membrana. O mesmo acontece com o gene *rfbB* e *rfbD*, também envolvido na síntese de LPS na região antígeno-O, que tiveram 68% e 48%, respectivamente, de similaridade com M1, mas quando analisado com a sequência de aminoácidos do gene, pois não foi encontrado no genoma de M4 ao ser analisado pela sequência de nucleotídeos. O que, assim como o gene *wecB*, não se pode afirmar que a estirpe M4 não possua o gene *rfbB* e *D*, mas que apenas a ausência de ambos seja em decorrência da parcialidade do genoma. Serrato, em 2010, fez estudos com estirpes da espécie de *Herbaspirillum*, a fim de analisar a composição química dos LPS desta. Os resultados demonstraram que a molécula de LPS de M4 é diferente do LPS de M1, tanto na composição, quanto na estrutura. Em *H. seropedicae* a molécula de LPS é importante para a interação entre essa bactéria e raízes de milho (BALSANELLI et al., 2013). A estirpe M4, assim como a estirpe M1 também possui genes

envolvidos com a síntese de celulose (FIGURA 23), em M1 a mutação nesses genes altera a formação de biofilme por essa estirpe e também a interação com raízes de milho (MONTEIRO et al., 2012). Esses genes não são encontrados em *H. seropedicae*, talvez estes genes estejam envolvidos no comportamento benéfico ou fitopatogênico do *H. rubrisubalbicans* quando este interage com as plantas.

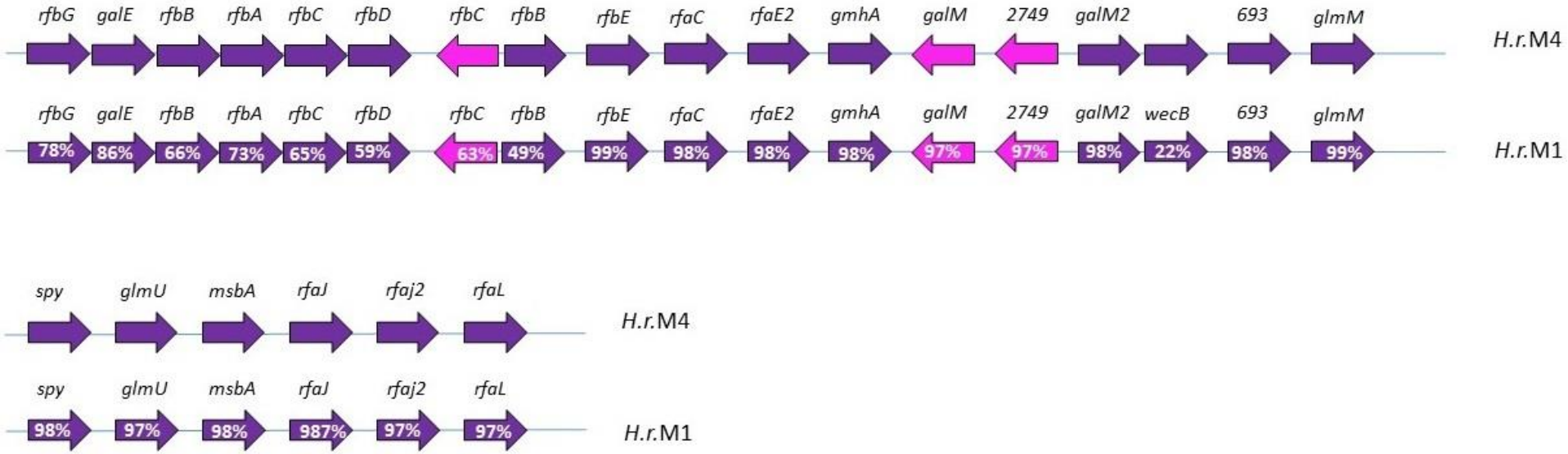


FIGURA 22. ESTRUTURA DE AGRUPAMENTO DOS GENES DA BIOSSÍNTESE DE OLIGOSSACARÍDEOS.

As setas representam os genes na ordem e sentido que estão no genoma. A referência é *H. rubrisubalbicans* M4, e as porcentagens de identidade são todas em relação à referência.

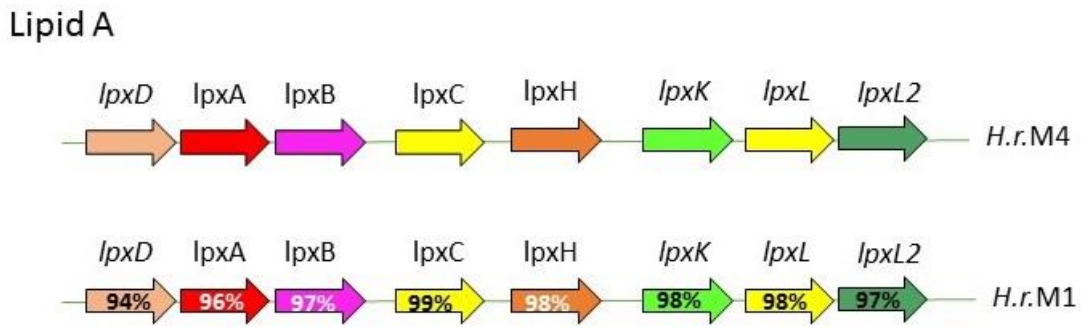


FIGURA 23. ESTRUTURA DE AGRUPAMENTO DOS GENES DA BIOSSÍNTESE DE LIPÍDIOS.

As setas representam os genes na ordem e sentido que estão no genoma. A referência é *H. rubrisubalbicans* M4, e as porcentagens de identidade são todas em relação à referência.

FIGURA 24. ESTRUTURA DE AGRUPAMENTO DOS GENES DA BIOSSÍNTESE DE CELULOSE.

As setas representam os genes na ordem e sentido que estão no genoma. A referência é *H. rubrisubalbicans* M4, e as porcentagens de identidade são todas em relação à referência.

6 CONCLUSÃO

A montagem prévia de *H. rubrisubalbicans* M4 com *reads* Illumina tem um tamanho de 4,1Mb, um conteúdo GC de 60,1% e 3.157 *contigs*. a montagem com os *reads* ION tem um tamanho de 4,1Mb, conteúdo GC de 61,5% e 2.100 *contigs*. A montagem Illumina continha grande fragmentação dos *contigs,* além de distribuição desigual da cobertura, causando empilhamento dos *reads.* Em contrapartida, a montagem ION, mostrou uma boa qualidade das bases, além de boa distribuição da cobertura. Foram feitas cinco montagens utilizando os dados dos sequenciamentos Illumina e ION, sendo constatado que as melhores resultados foram vistos nas denominadas ION e MiSEQ2_ION, com melhor distribuição dos *reads*, também foram as que mais se aproximaram em tamanho do genoma usado como referência.

As cinco montagens, compostas pelos dados Illumina e ION, foram ordenadas e mapeadas com o genoma de referência. A visualização do gráfico DOTPLOT, permitiu observar a relação taxonômica entre as duas espécies, apesar das sequências que não entraram no alinhamento. A montagem ION_GFinisher, teve menor número de *contigs* não alinhados com a referência. Por outro lado, a MiSEQ2_ION, apesar de se aproximar mais, em tamanho, da referência, foi a que teve maior número de *contigs* fora do alinhamento. A partir disso, levando em consideração de que o genoma da *estirpe* M4 não foi finalizado e que um genoma com alto grau de repetições de sequências seja uma característica já conhecida do *H. rubrisubalbicans,* sugere-se que os fragmentos não alinhadas possam ser sequências repetidas, ou ainda podem ser características diferentes das encontradas no genoma de referência.

As condições proporcionadas pela técnica Illumina, limitaram a análise real sobre o genoma da *estirpe* M4. Entretanto, com a alta cobertura obtida com os *reads* ION, e sua finalização pelo programa GFinisher, foi possível uma nova visão do genoma de *H. rubrisubalbicans* M4.

A melhor montagem foi submetida a uma pré-anotação, a qual foram encontrados 4.929 genes; 4.695 apresentam categorias funcionais no

COG. Sugere-se que as diferenças muito pronunciadas em referência à quantidade de genes para cada função inferida pelo COG, em relação ao genoma de *H. rubrisubalbicans* M1, sejam devido à alta fragmentação do genoma da estirpe M4. Durante estudos realizados com a estirpe M1 foi possível identificar genes que estão envolvidos na interação entre *H. rubrisubalbicans* e a planta, esses genes foram encontrados na estirpe M4, indicando que as duas estirpes podem ter o mesmo padrão de interação com as plantas.

REFERÊNCIAS

ANDREWS, J. H., HARRIS R. F. The Ecology and Biogeography of Microorganisms On Plant Surfaces. **Annu. Rev. Phytopathol. 38:145–80**, 2000.

ARTEMIS. Disponível em <http://www.sanger.ac.uk/science/tools/artemis>

BALDANI, J. I., BALDANI, V. L. D.; SELDIN, L.; DOBEREINER, J. Characterization of *Herbaspirillurn seropedicae* gen. nov. sp. nov. a Root-Associated Nitrogen-Fixing Bacterium. **International Journal of Systematic Bacteriology,** p. 86-93,1986.

BALDANI, J. I., POT, B**.,** KIRCHHOF, .G., FALSEN, E., BALDANI, V. I. D., OLIVARES, F. I., HOSTE, B., KERSTERS, K., HARTMANN, A., GILLIS, M., DÖBEREINER, J. Emended Description of *Herbaspirillum*; Inclusion of [*Pseudomonas*] *rubrisubalbicans*, a Mild Plant Pathogen, as *Herbaspirillum rubrisubalbicans* comb. nov.; and Classification of a Group of Clinical Isolates (EF Group 1) as *Herbaspirillum* Species 3. **International Journal of Systematic Bacteriology.** p. 802–810, 1996.

BALDANI, J. I.; BALDANI, V. I. D. History on the biological nitrogen fixation research in graminaceous plants: special emphasis on the Brazilian experience. **Anais da Academia Brasileira de Ciências** 77(3): 549-579, 2005.

BHATTACHARJEE S, LEE L-Y, OLTMANNS H, CAO H, VEENA, CUPERUS J, GELVIN. SB AtImpa-4, an Arabidopsis importin a isoform, is preferentially involved in Agrobacterium-mediated plant transformation. **Plant Cell** 20: 2661–2680, 2008.

CARVALHO, M.C.C.G. E SILVA, D.C.G. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas. **Ciência Rural,** Santa Maria, v.40, n.3, p.735-744, mar, 2010.

CHAN, E. Y. Advances in sequencing technology**. Mutation Research**, v. 573, p. 13-40, 2005.

CHRISTOPHER, W. N.; EDGERTON, C. W. Bacterial Stripe Diseases Of Sugarcane In Louisiana**. Journal of Agricultural Research,** Vol. 41, No. 3 Washington, D. C. Key No. La.-5. 1930.

CLC WORBENCH. Disponível em <http://www.clcbio.com/products/clc-main-workbench/>

FASTQC. Disponível em <www.bioinformatics.babraham.ac.uk/projects/fastaqc/>

GFINISHER. Disponível em < http://www.bioinfo.ufpr.br/gfinisher/>

HIRANO, S. S., UPPER, C. D. Ecology and Epidemiology of Foliar Bacterial Plant Pathogens. **Ann. Rev. Phytopathol**. 21:243-69, 1983.

JAMES, E.K. Nitrogen Fixation in endophytic and associative symbiosis. **Field Crops Research** 65 (2000) 197±209, 1999.

JAMES, E.K., OLIVARES, F. L. Infection and Colonization of Sugar Cane and Other Graminaceous Plants by Endophytic Diazotrophs. ***Critical Reviews in Plant Sciences,* 17(1):77–119 (1997)**, 1998.

KURTZ, S., PHILLIPPY, A., DELCHER, A.L., SMOOT, M., SHUMWAY, M., ANTONESCU C., SALZBERG S. L. Versatile and open software for comparing large genomes. **Genome Biol**., v. 5:R12, 2004.

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

MYLONA, P., PAWLOWSKI, K., AND BISSELING, T. Symbiotic nitrogen fixation. **Plant cell** 7, 869-885, 1995.

NELSON, D.L. Princípios de bioquímica de Lehninger. 5 ed.Ed. Artmed Pág. 292-294, 2011.

OKONECHNIKOV K., GOLOSOVA O., FURSOV M. Unipro UGENE. A unified bioinformatics toolkit. **Bioinformatics.** 28:1166-1167, 2012.

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

OLIVARES, F. L., JAMES, E. K., BALDANI, J. I., DÖBEREINER, J. Infection of mottled stripe disease-susceptible and resistant sugar cane varieties by the endophytic diazotroph *Herbaspirilium.* ***New Phytol.*** 135, 723-737, 1997.

PEDROSA F.O., MONTEIRO R.A., WASSEM R, CRUZ L.M., AYUB R.A., *et al*. Genome of Herbaspirillum seropedicae Strain SmR1, a Specialized Diazotrophic Endophyte of Tropical Grasses. **PLoS Genet** 7(5): e1002064. doi: 10.1371/journal.pgen.1002064, 2011.

PIMENTEL, D., HARVEY, C., RESOSUDARMO, P., SINCLAIR, K., KURZ, D., MCNAIR, M., CRIST, S., SPHPRITZ, L., FITTON, L., SAFFOURI, R.,

BLAIR, R. Environment and economic costs of soil erosin and conservation benefits. **Science** 267, 1117-1123, 1995.

ROESCH, L. F. W.; PASSAGLIA, L. M. P.; BENTO, F. M. B.; TRIPLETT, E. W.; CAMARGO, F. A. O. C. Diversidade De Bactérias Diazotróficas Endofíticas Associadas a Plantas de Milho. **R. Bras. Ci. Solo.** 31:1367-1380, 2007.

SABINO, D.C.C., FERREIRA, J.S.F., GUIMARÃES, S.L. E BALDANI, V.L.D. Bactérias diazotróficas como promotoras do desenvolvimento inicial de plântulas de arroz. **Enciclopédia Biosfera**, Centro Científico Conhecer, Goiânia, v.8, n.15, 2012

SHENDURE, J. AND JI, H. Next-generation DNA sequencing. **Nature Biotechnology**. doi:10.1038/nbt1486, 2008.

SPRENT, J. AND SPRENT, P. Nitrogen fixing organsims. Pure and applied aspects. **Chapman and Hall, Camdbrige**. Great Britain. pp. 256, 1990.

TATUSOV RL, NATALE DA, GARKAVTSEV IV, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. **Nucleic Acids Research.** 29(1):22-28, 2001.

URETA, A., ALVAREZ, B., ALVAREZ, A. RAMÓN, M.A. VERA, MARTINEZ-DRETS, G. Identification *of Acetobacter diazotrophicus, Herbaspirillum seropedicae* and *Herbaspirillum rubrisubalbicans* using biochemical and genetic criteria. **Plant and Soil** 172: 271-277, 1995.

URQUIAGA; CRUZ, K. H. S.; BODDEY, R. M. Contribution of Nitrogen Fixation to Sugar Cane: Nitrogen-15 and Nitrogen-Balance Estimates. **Soil Sci. Soc. Am. J. 56:105-114**, 1992.

VESSEY, J. K. Plant growth promoting rhizobacteria as biofertilizers. **Plant and Soil.** 255: 571–586, 2003.