UNIVERSIDADE FEDERAL DO PARANÁ

PAULA MAYUMI SAIZAKI

MONTAGEM DO *DRAFT* GENÔMICO DA BACTÉRIA *Herbaspirillum hiltneri* N3

CURITIBA

2012

PAULA MAYUMI SAIZAKI

MONTAGEM DO *DRAFT* GENÔMICO DA BACTÉRIA *Herbaspirillum hiltneri* N3

Dissertação apresentada ao Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de Mestre em Bioinformática.

Orientadora:
Dra. Maria Berenice Reynaud Steffens
Co-orientadora:
Dra. Jeroniza Nunes Marchaukoski

CURITIBA

2012

“If you don’t make mistakes,

You’re not working on hard enough problems

And that’s a big mistake.”

Frank Wilczek

AGRADECIMENTOS

Às minhas orientadoras Profa. Maria Berenice Reynaud Steffens e Profa. Jeroniza Nunes Marchaukoski, pela atenção, paciência, sugestões e ensinamentos que foram a base deste trabalho.

Ao Prof. Roberto Tadeu Raittz, pelo auxílio, puxões de orelha e principalmente pelo otimismo que sempre mostrou em meu trabalho quando nem eu mesma acreditava mais.

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

Ao Prof. Dieval Guizelini, pelo interesse e enriquecimento nos conhecimentos de montagem.

Ao Prof. Lucas Ferrari de Oliveira, por ser sempre prestativo dentro do laboratório e pelos ensinamentos de informática, particularmente em Linux.

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

Aos colegas de laboratório, pela amizade e por tornarem o dia a dia muito mais agradável e divertido.

À Vanely e ao Leviston, companheiros de montagem, que sempre me ajudaram quando precisei.

A todos os membros da coordenação do curso de Pós-Graduação em Bioinformática, pela simpatia e atenção.

Aos órgãos financiadores: CAPES, CNPq e REUNI.

A todas as pessoas que participaram do sequenciamento genômico e auxiliaram a produção desse trabalho, mesmo que de maneira indireta.

RESUMO

Este estudo teve objetivo de realizar a montagem e análise parcial do genoma da bactéria *Herbaspirillum hiltneri* estirpe N3^{T}. Esta bactéria foi isolada da raiz do trigo (*Triticum aestivum* var. Naxos) e caracterizada por ROTHBALLER e colaboradores (2006), na Alemanha. O genoma foi sequenciado pelo Núcleo de Fixação de Nitrogênio, sediado no Departamento de Bioquímica e Biologia Molecular da Universidade Federal do Paraná, utilizando a plataforma SOLiD^{TM} *Sequencing System*. Leituras curtas em *color space* provenientes do sequenciamento *Whole-genome shotgun* (WGS) em *fragments* (8,4 milhões) e *mate-paired* (107 milhões) foram montadas parcialmente através do *pipeline de novo accessory tools* 2.0. As melhores montagens de cada conjunto de dados foram integradas em uma montagem híbrida pelo montador Phrap, resultando em 5970 *contigs* e 283 *scaffolds.* O genoma montado apresentou cerca 5,04 Mpb e N50 de 19.633 pb. Em função da alta fragmentação do genoma na montagem final, foi aplicada uma estratégia alternativa para fechamento das 4.778 falhas presentes. A anotação parcial resultou em 3.818 *orfs* com alinhamentos positivos com genes do banco de NR, sendo a maior incidência com a bactéria *H. seropedicae* SmR1. Não foi possível localizar o operon *nif*HDK, o que sugere que este organismo seja incapaz de fixar nitrogênio. Entretanto, os genes *glnA, ntrC, ntrB, ntrY* and *ntrX*, envolvidos no metabolismo do nitrogênio, foram localizados no genoma. A validação do conteúdo biológico da montagem ocorreu pela identificação dos genes *housekeeping* codificadores de proteínas envolvidas na recombinação homóloga.

Palavras-chave: *Herbaspirillum hiltneri*, Sequenciamento de DNA, Genoma, Leituras curtas, Montagem de genomas.

ABSTRACT

This research had the purpose to make a partial assembly and analysis of the genome of the bacterium *Herbaspirillum hiltneri* N3^{T}. This bacterium was isolated from wheat root (*Triticum aestivum* var. Naxos) and characterized by ROTHBALLER and colleagues (2006), Germany. The genome was sequenced by the Nitrogen Fixation Center, Department of Biochemistry and Molecular Biology - Federal University of Paraná, using the SOLiD^{TM} Sequencing System platform. Short reads in *color space* format from the whole-genome shotgun sequencing (WGS) in *fragments* (8.4 million) and *mate-paired* (107million) were partially assembled using *de novo* pipeline accessory tools 2.0. The best assemblies from each data set were integrated within a hybrid assembly by the assembler Phrap, resulting in 5970 contigs and 283 *scaffolds*. The assembled genome showed size about 5,04 Mbp and N50 of 19.633 bp. Due to the genome's high fragmentation in the final assembly, it was applied an alternative strategy for closing 4778 gaps. The partial annotation resulted in 3818 *orfs* with positive alignments with the NR database genes and with the highest incidence with the bacterium *H. seropedicae* SmR1. It was not possible to find the *nif*HDK operon, suggesting that this organism is unable to fix nitrogen. However, *gnlA, ntrC, ntrB, ntrY* and *ntrX*, genes involved in nitrogen metabolism, were successfully found. Furthermore, the biological content validation was obtained by the localization of housekeeping genes related to homologous recombination.

Keywords: *Herbaspirillum hiltneri*, DNA sequencing, Short reads, Genome assembly.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

BLAST – *Basic Local Alignment Search Tool*

K-mer – (Semente) menor fração de uma sequência com tamanho pré-definido “k”

Mpb – Mega pares de bases

NCBI – *National Center for Biotechnology Information*

NGS – *Next-Generation Sequencing*

ORF – *(Open Reading Frame)* Fase de leitura aberta

pb – Pares de bases

PCR – *Polymerase Chain Reaction*

Primer – Oligonucleotídeo iniciador do processo de PCR

WGS – *Whole-genome shotgun*

SUMÁRIO

1 INTRODUÇÃO

1.1 Bioinformática

A Bioinformática, uma disciplina que combina conhecimentos de biologia e informática, surgiu em função da necessidade computacional de armazenamento e processamento do grande volume de informações proveniente de estudos biológicos de macromoléculas para, então, apreender sua função e organização estrutural. A bioinformática consiste no estudo via processos de informática de sistemas bióticos em múltiplos níveis (HOGEWEG et al., 2011). Possui três focos principais: o primeiro é a organização e manutenção de bancos dados de maneira a possibilitar o acesso por pesquisadores de todo o mundo, a exemplo do GenBank (BENSON et al., 2000) e o SWISS-PROT (BAIROCH et al., 2000). O segundo diz respeito ao desenvolvimento de novas ferramentas computacionais para solução de diferentes problemas. Atualmente há uma grande diversidade de softwares disponíveis com a finalidade de realizar estudos de dados biológicos, o que conduz ao terceiro foco: a utilização de ferramentas para análise e interpretação de resultados biológicos (LUSCOMBE et al., 2001). Como exemplos, têm-se os estudos de genoma, de expressão gênica e de modelagem de estrutura de proteínas, envolvendo desde a sequência de bases e aminoácidos até a conformação tridimensional.

1.2 Projetos Genoma

Um projeto genoma consiste no isolamento e obtenção da sequência completa de DNA de um dado organismo e da montagem, mapeamento e anotação de genes codificadores de RNAs, peptídeos e proteínas, visando o entendimento aprofundado do funcionamento do seu metabolismo e biologia. O primeiro genoma completamente sequenciado e montado foi o da bactéria *Haemophilus influenzae* (FLEISCHMANN et al., 1995). Desde então, milhares de genomas dos mais diversos organismos vêm sendo sequenciados, anotados e estudados.

Projetos genoma geralmente compreendem as seguintes etapas principais (Figura 1): a primeira é a laboratorial na qual um determinado organismo de interesse é isolado e cultivado (no caso de microrganismos). Então, é realizada a

extração e sequenciamento do DNA. A segunda etapa é computacional, composta pela montagem e anotação. Normalmente há ainda a última etapa, na qual o projeto retorna para a bancada para estudos, com vistas à validação dos genes previamente anotados, através de, por exemplo, transcriptoma e proteoma.

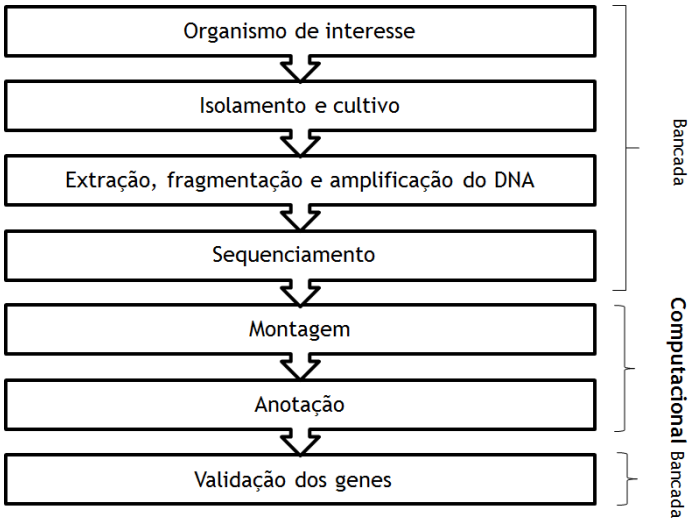


FIGURA 1: FLUXOGRAMA DE UM PROJETO GENOMA

Exemplo de fluxograma demonstrando as principais etapas do estudo de um genoma. Inicialmente são realizados procedimentos laboratoriais, seguido pela etapa de bioinformática e por úiltimo a validação dos resultados computacionais.

FONTE: A autora, 2012.

1.2.1 Sequenciamento automático do DNA

O primeiro método de sequenciamento de DNA foi desenvolvido por Sanger, Nickeln e Coulson (1977) e faz uso de 2’3’-dideoxinucleosídeos trifosfato (ddNTP), análogos dos 2´-deoxinuclosídeos trifosfatos normais, que agem como terminadores específicos da polimerização da cadeia de DNA, por inibição da atividade da enzima DNA polimerase. Na versão automatizada do método (ZIMMERMANN et al., 1988), cada ddNTP está ligado a uma molécula fluorescente que atribui ao fragmento interrompido uma cor particular. Os fragmentos de DNA são separados por eletroforese e a informação é captada diretamente pelo computador, que determina a sequência. Os sequenciadores baseados neste método produzem leituras longas, mas com tamanho variável (600 a 1.000 pares de bases). Esta metodologia é, atualmente, considerada como sequenciamento de primeira geração e foi utilizada para o sequenciamento de grande parte dos genomas depositados nos bancos públicos (METZKER, 2008).

1.2.1.1 Sequenciadores de nova geração

No início da década de 2.000 novas tecnologias de sequenciamento foram desenvolvidas e os chamados sequenciadores de nova geração (NGS - *Next-generation DNA Sequencing*) ganharam força no mercado. As plataformas mais populares são SOLiD™ *Sequencing System* (*Life* Technology Inc./Applied Biosystems, Califórnia, USA), 454GS-FLXTM (Roche, Basel, Switzerland) e Solexa® Genome Analyser (Illumina, San Diego, USA). A principal vantagem destas tecnologias é a produção de um grande volume de dados a baixo custo por base, pois em um só ciclo de sequenciamento estas plataformas são capazes de produzir bilhões de nucleotídeos, resultando na diminuição do custo de sequenciamento ao longo dos anos (METZKER, 2008). Em 2001, o custo aproximado do sequenciamento de um genoma humano era de 100 millhões de dólares; 10 anos depois este valor caiu para cerca de 11 mil dólares (Figura 2) (http://www.genomesonline.org). Em consequência, o número de estudos genômicos aumentou exponencialmente, como pode ser observado na figura 3 que apresenta o número de genomas depositados no *National Center for Biotechnology Information* (NCBI), no período de 1995 a 2011.

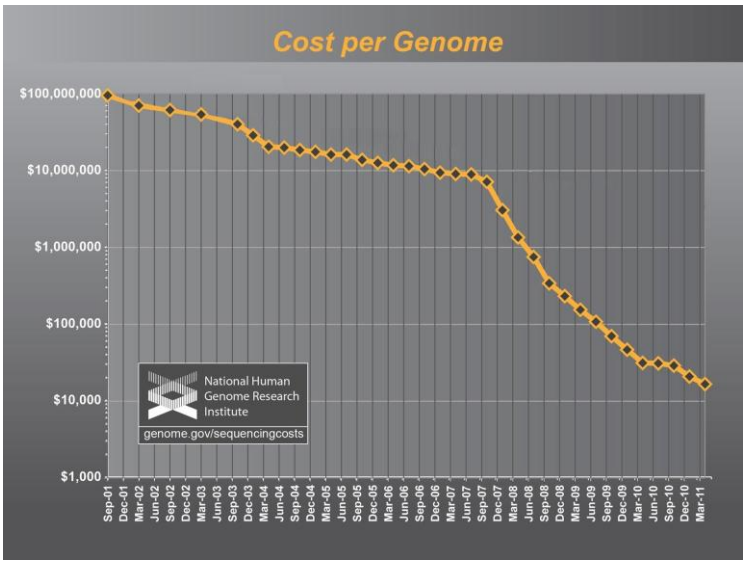


FIGURA 2: CUSTO POR SEQUENCIAMENTO DE GENOMA NA ÚLTIMA DÉCADA

Observa-se que a queda no custo do sequenciamento dos genomas (dólar) a partir de 2007 com a consolidação dos sequenciadores de nova geração.

FONTE: *National Human Genome Sequencing Research Institute* (http://www.genome.gov/)

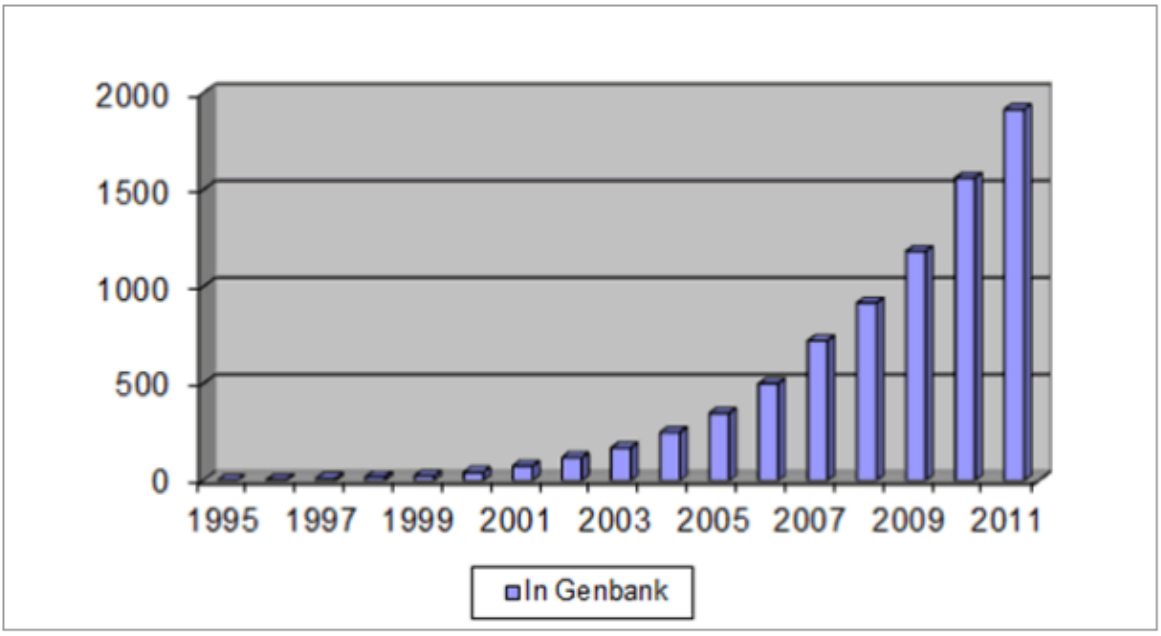


FIGURA 3 NÚMERO DE GENOMAS DEPOSITADOS NO NCBI AO LONGO DOS ANOS

O número de genomas sequenciados e montados aumentou exponencialmente no período de 1995 a 2011 como consequência da consolidação das tecnologias de nova geração.

Dados referentes a outubro de 2011.

FONTE: *Genoms Online Database* (http://www.genomesonline.org)

Entretanto, estes novos métodos apresentam novos desafios, como o tamanho das leituras (principalmente em tecnologias de leituras curtas), problemas de qualidade (MARTINEZ et al., 2010) e demanda computacional em função da quantidade de dados gerados. Diversos estudos evoluíram juntamente com os sequenciadores de nova geração, como projetos de ressequenciamento de genomas de diversos organismos, incluindo o humano, de sequenciamento de transcriptomas (RNA seq), de RNAs curtos, metagenômica e de proteína-DNA (CHiP-Seq) (SHENDURE et al., 2008).

1.2.1.2 Tecnologia SOLiD™ Sequencing System

Diferentemente da maioria das outras tecnologias, nas quais o sequenciamento ocorre via adição de bases pela ação da enzima DNA polimerase, na plataforma SOLiD™ *Sequencing System* (Sequencing by Oligonucleotide Ligation and Detection) o sequenciamento é baseado na ligação sequencial de nucleotídeos marcados e a sequência de DNA é gerada pela medida da ligação serial de um

oligonucleotideo pela enzima DNA ligase. O sequenciamento ocorre por hibridização de sondas fluorescentes com as sequências alvo. Este equipamento é particularmente conhecido pelo seu alto desempenho e produção massiva de dados (*ultra high throughput)* e por produzir sequências denominadas leituras curtas (*short reads).*

Para ser sequenciado na plataforma SOLiD™ *Sequencing System*, o DNA alvo deve ser fragmentado e uma biblioteca de DNA deve ser preparada através das seguintes metodologias: *fragments, mate-paired* ou *paired-end*. Os fragmentos de DNA são ligados nas extremidades com os adaptadores P1 e P2 e em seguida ligados à esferas (*beads*). No caso da técnica em *mate-paired*, os fragmentos de DNA estão distanciados por um adaptador interno de distância conhecida, o que resulta na formação de *scaffolds* (figura 4).

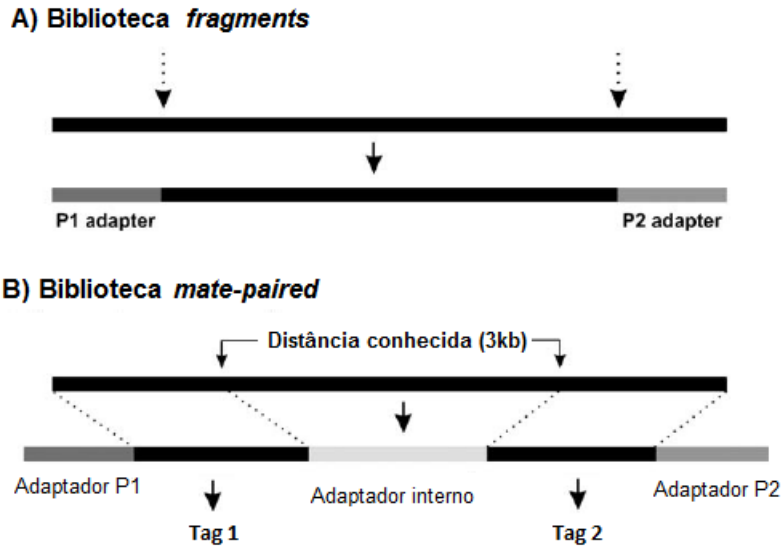


FIGURA 4 ESQUEMA PARA PREPARO DE BIBLIOTECAS DE DNA PARA O SEQUENCIAMENTO NA PLATAFORMA SOLiD™ *Sequencing System*. .

A) Biblioteca *fragments*: há a ligação de 2 adaptadores P1 e P2 nas pontas 5' e 3'. B) Biblioteca *mate-paired*: leituras (*tags*) de DNA são separadas por um adaptador interno de tamanho conhecido, sendo neste exemplo 3kb.

FONTE: Adaptado de ANSORGE, 2009.

O material genético é então amplificado em PCR em emulsão, no qual cada esfera é isolada em uma miscela proveniente da interação água e óleo contendo os reagentes necessários para o processo (figura 5A). Em seguida, os produtos são retirados e imobilizados, ou seja, ligados covalentemente, em uma placa de vidro (*slide*) onde ocorrerá o sequenciamento. *Primers* universais de tamanho *n* se anelam

com a cadeia de DNA a partir do adaptador P1 e o octâmero sonda (*probe)* marcado com fluorescência que contém a combinação das duas primeiras bases complementares é hibridizado e ligado pela enzima ligase, liberando uma das 4 possíveis cores do fluoróforo. As últimas bases são clivadas e novamente o octâmero correspondente é hibridizado. Esse ciclo de ligação se repete 10 vezes para a formação de leituras com 50 bases. Então a fita recém sintetizada é removida e o *primer* universal de tamanho *n* -1 é anelado, com o deslocamento de uma base. Tal etapa ocorre 5 vezes até que o *primer* de tamanho *n* - 4 entre em ação; assim, cada base é lida duas vezes, resultando na codificação di-base *color space* (figura 5B), onde para cada transição de base é conferido um valor de qualidade baseado nos sinais intensidade da luz emitida, denominado *base calling* (LEDERGERBER et al., 2010). Esta codificação permite maior acurácia ao sequenciamento e facilita consideravelmente a detecção de SNPs (*single nucleotide polymorphisms*) (ANSORGE, 2009; CARVALHO et al., 2010; METZKER, 2010, SHENDURE et al., 2008).

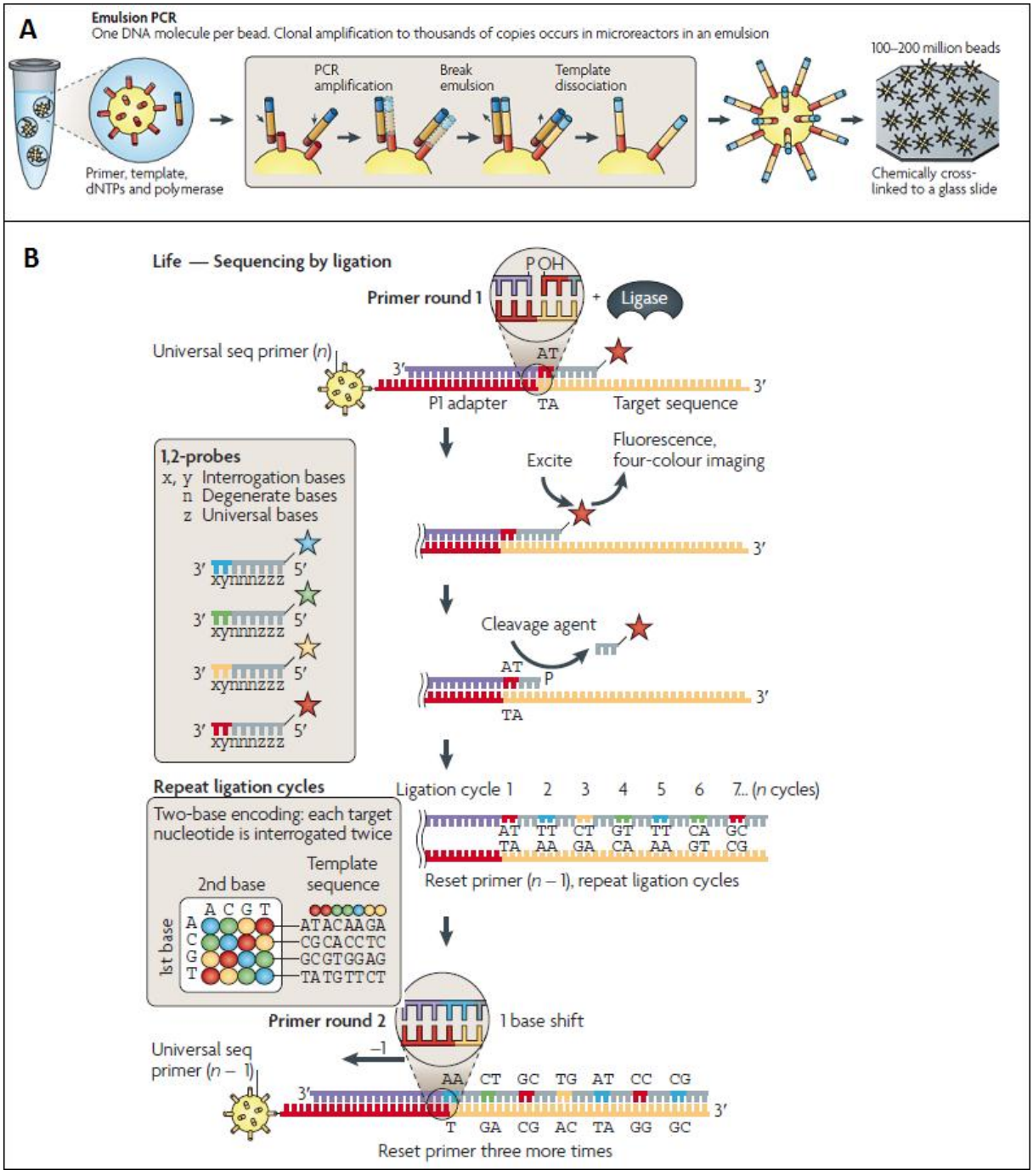


FIGURA 5 RESUMO ILUSTRATIVO DO SEQUENCIAMNETO NA PLATAFORMA SOLiD™ *Sequencing System*

A) Amplificação pela PCR em emulsão ocorre dentro de miscelas contendo o complexo esfera-DNA. O produto resultante é imobilizado em uma placa de vidro na qual é realizado o sequenciamento. B) *Primer* universal alinha e os octâmeros com as duas primeiras bases complementares hibridizam, são ligados e a fluorescência é medida. O ciclo de ligação se repete inúmeras vezes. Um novo *primer* com uma base a menos é alinhado e o ciclo se repete até chegar a 5 vezes, fazendo com que cada nucleotídeo seja lido duas vezes, o que origina a codificação di-base *color space.*

FONTE: METZKER, 2010.

1.2.2 Montagem e anotação de genomas

A estratégia mais utilizada para o sequenciamento genômico é o *whole-genome shotgun* (WGS), na qual as sequências são obtidas de maneira aleatória Depois da geração das leituras, estas podem ser mapeadas com um genoma de referência ou montadas através da estratégia *de novo*; a decisão de qual estratégia será utilizada depende do objetivo biológico pretendido, assim como custo, tempo e se há a existência prévia de dados de organismos próximos (MARTINEZ et al., 2010).

A montagem *de novo* (*ab initio)* visa juntar milhões de leituras de modo que todo o genoma sequenciado esteja representado e em sua ordem correta sem a necessidade de haver um genoma de referência prévio, ou seja, utilizando somente a informação das leituras. A maneira mais comum é via semente (*k-mer*), na qual as sequências são quebradas em porções menores em busca de sobreposição (figura 6) a fim de chegar à sequência consenso formando sequências contíguas (contigs). Em bibliotecas *mate-paired*, ocorre a formação de *scaffolds*, onde contigs estão montados na ordem correta, mas com falhas (*gaps*) entre eles. Tal técnica facilita consideravelmente o processo de montagem.

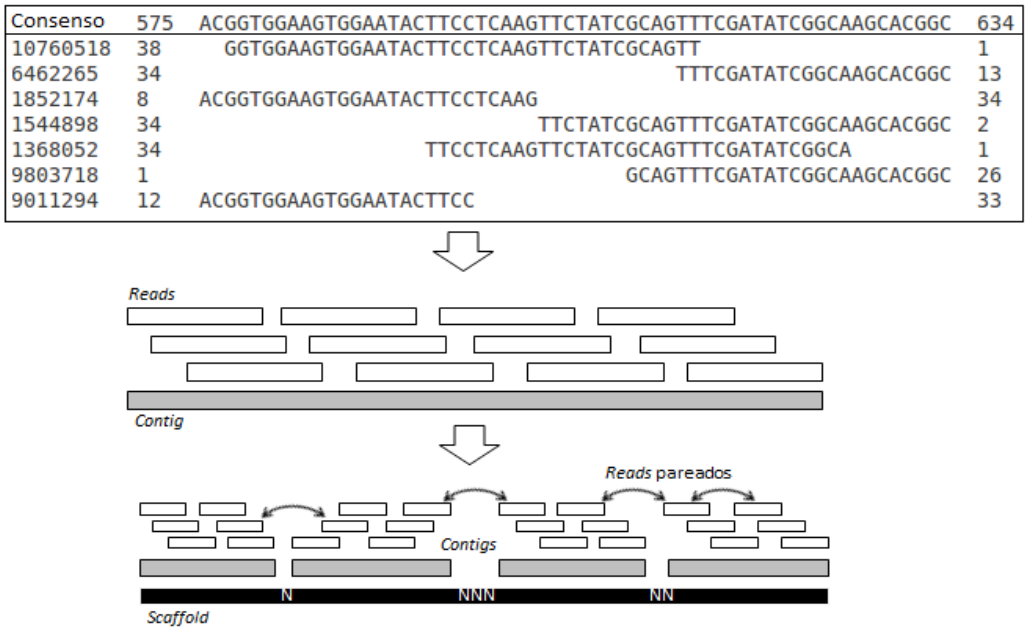


FIGURA 6: EXEMPLO DE MONTAGEM DE UMA SEQUÊNCIA CONSENSO DE DNA

Sobreposição das leituras e formação da sequência contig de acordo com a sequência consenso. Em bibliotecas *mate-paired* ocorre a formação de *scaffolds,* onde a falha está representada pela presença de “N”.

FONTE: Adaptado de CARDOSO, 2011.

Entre os montadores baseados em grafos de sobreposição direcionados ao método WGS, destacam-se: Arachne (BATZOGLOU et al., 2002), Celera (MYERS et al., 2000), Pcap (HUANG et al., 2003), EDENA (HERNANDEZ et al., 2008) e Phrap (GREEN, 1999). Entretanto, por esta estratégia de sobreposição ser bastante custosa computacionalmente, novos montadores vêm ganhando força juntamente com os sequenciadores de nova geração (TIEPPO, 2011), como o ABySS (SIMPSON et al., 2009) e o Velvet (ZERBINO et al., 2008). A escolha do montador mais adequado varia de acordo com o tipo de tecnologia de sequenciamento utilizado bem como o seu respectivo tipo de leitura, equipamento disponível para montagem e volume de dados.

A predição gênica ou anotação, que consiste na determinação dos genes codificadores de produtos funcionais assim como de seus constituintes regulatórios, pode ser considerada a primeira etapa para a interpretação biológica dos dados de montagem, levando futuramente ao estudo aprofundado do metabolismo. Para a real validação destas informações é necessária análise laboratorial posterior.

1.3 O Gênero *Herbaspirillum*

As bactérias pertencentes ao gênero *Herbaspirillum* são beta-proteobactérias aeróbicas gram negativas e com formato geralmente vibrióide. Possuem flagelos, que lhes confere motilidade e são comumente observadas em associação endofítica com gramíneas, banana e abacaxi (BALDANI et al., 1986; BALDANI et al., 1992, CRUZ et al., 2001; SCHMID et al., 2006).

Atualmente existem doze espécies descritas do gênero: *Herbaspirillum seropedicae* (BALDANI *et al.,* 1986); *Herbapirillum rubrisubalbicans* (BALDANI et al., 1996), *Herbaspirillum frisingense* (KIRCHHOF *et al.,* 2001), *Herbaspirillum lusitanum* (VALVERDE et al., 2003), *Herbaspirillum autotrophicum*, e *Herbaspirillum huttiensis* (DING & YOKOTA, 2004), *Herbaspirillum chlorophenolicum* (IM et al*.,* 2004), *Herbaspirillum hiltneri* (ROTHBALLER *et al.,* 2006), *Herbaspirillum rhizosphaerae* (JUNG et al., 2007), *Herbaspirillum aquaticum* (DOBRITSA et al*.,* 2010), *Herbaspirillum canariense* (CARRO et al., 2011), *Herbaspirillum aurantiacum* (CARRO et al., 2011) e *Herbaspirillum soli* (CARRO et al., 2011).

1.3.1 Herbaspirillum seropedicae

A primeira espécie caracterizada do gênero *Herbaspirillum* foi o *Herbaspirillum seropedicae* (BALDANI et al., 1986), uma bactéria endofítica cuja principal característica é a capacidade de fixar o nitrogênio atmosférico durante a associação com gramíneas (ELBELTAGY et al., 2001; RONCATO-MACCARI et al., 2003). Foi encontrada na colonização em arroz, cana de açúcar, sorgo, milho, entre outros vegetais de importância agronômica. Apresenta potencial de biofertilizante e, assim sendo, constitui uma importante alternativa aos fertilizantes e adubos químicos (CHOUDHURY et al., 2004).

H. seropedicae estirpe SmR1 é o único organismo deste gênero com o genoma completamente sequenciado, montado, anotado e publicado (PEDROSA et al., 2011). Os genomas de outras espécies se encontram em fase de montagem e anotação, entre elas o *H. rubrisubalbicans* (CARDOSO, 2011).

1.3.2 Herbaspirillum hiltneri

Bactérias da espécie *Hesbaspirillum hiltneri* foram isoladas da raiz do trigo (*Triticum aestivum* var. Naxos) e caracterizadas por ROTHBALLER e colaboradores (2006) na Alemanha. Apresentam tamanho celular em torno de 1,6 a 2,0 µm e diâmetro com cerca de 0,5 a 0,6 µm (Figura 7)

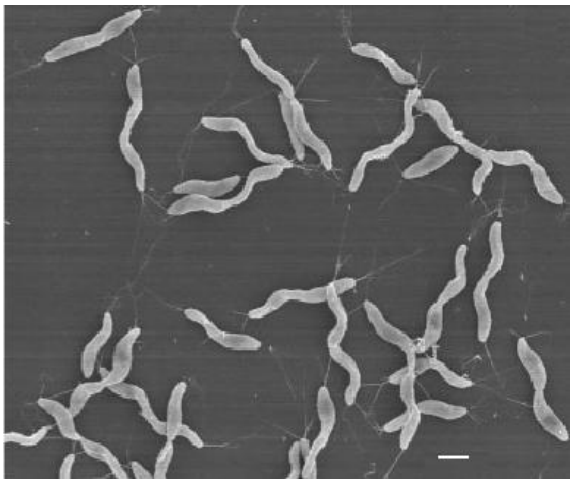


FIGURA 7: MORFOLOGIA DO *H. hiltneri* $N3^T$.

Fotografia via micrografia eletrônica. É possível observar a presença de flagelos polares. Barra: 1 µm.

FONTE: ROTHBALLER et al., 2006.

Foram determinadas três cepas: N3^T, N5 e N9, sendo N3 ^{T} a estirpe tipo. O organismo filogeneticamente mais próximo foi o *Herbaspirillum lusitanun*, apresentando 99,9% de identidade com o gene *16S rRNA.* A Figura 8 mostra a sua classificação na árvore filogenética de acordo com o gene *23S rRNA* (ROTHBALLER et al., 2006). Os autores afirmam que esta bactéria aparenta ser incapaz de fixar nitrogênio em condições de laboratório e que os genes *nif*D e *nif*H não puderam ser identificados por técnicas de PCR mesmo com a utilização de diversos *primers*. O conteúdo G+C no DNA dos isolados foi de 60,9% a 61,5% (ROTHBALLER et al., 2006).

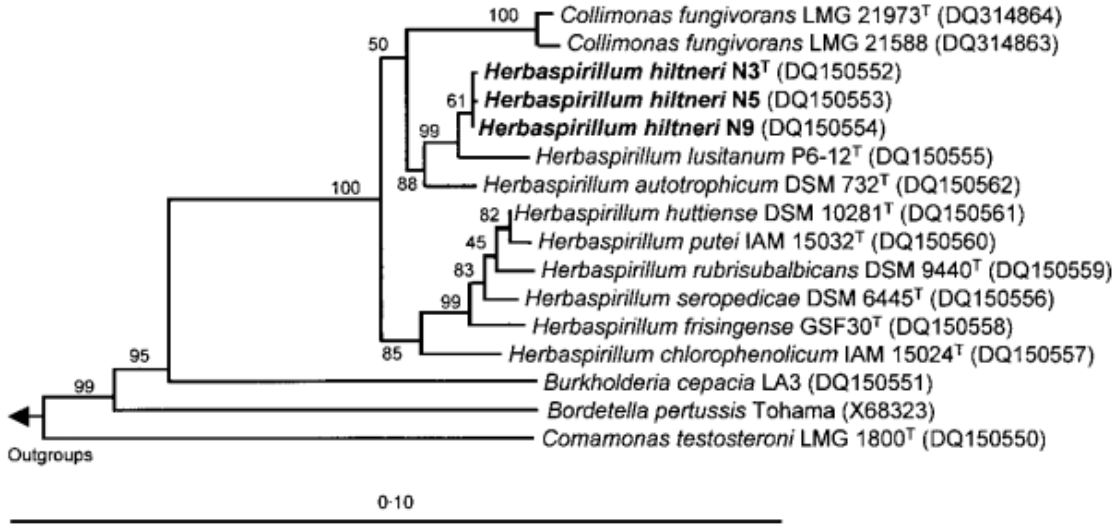


FIGURA 8 ÁRVORE FILOGENÉTICA BASEADA EM DADOS DE SEQUENCIAMENTO DO GENE *23S rRNA*

Ávore filogenética do gênero *Herbaspirillum* e espécies próximas. Estão sendo mostradas as três estirpes do *Herbaspirillum hiltneri*: N3^T, N5 e N9. Classificação baseada na análise do sequenciamento do 23S.

FONTE: ROTHBALLER et al., 2006.

1.4 Fixação Biológica do Nitrogênio

O nitrogênio é um elemento essencial à manutenção e desenvolvimento da vida. Apesar de 78% do nitrogênio disponível fazer parte da composição da atmosfera terrestre, somente os organismos diazotróficos são capazes de utilizá-lo na forma gasosa (N_2). Neste grupo estão as bactérias que realizam a fixação biológica, onde o dinitrogênio é convertido em uma forma assimilável, como sais de nitrato e amônio, que os demais seres vivos, como os vegetais, conseguem metabolizar.

1.4.1 Genes *nif*

O cluster dos genes *nif* e *fix* constitui a base das proteínas relacionadas à fixação do nitrogênio em bactérias (HENNECKE, 1990). O operon *nif*HDK é um importante componente genético em bactérias diazotróficas, pois codifica as proteínas estruturais da nitrogenase (AVTGES et al., 1983), enzima responsável pela redução do nitrogênio atmosférico em amônio (CHEN et al., 1996). A bactéria *Klebsiella pneumoniae* foi o primeiro organismo diazotrófico no qual foram caracterizados os genes *nif* (MERRICK, 1983), que são encontrados também em outras bactérias fixadoras de nitrogênio, entre elas o *Herbaspirillum seropedicae* e *Azospirillum brasiliense* (CHUBATSU et al., 2011; ZHANG et al., 1994). A organização estrutural do cluster *nif* de *H. seropedicae* estirpe SmR1 assim como seus principais componentes, está representada na figura 9 (CHUBATSU et al., 2011). Proteínas Ntr e Gln estão envolvidas na ativação dos genes *nif*, regulando sua expressão e, consequentemente, a fixação do nitrogênio.

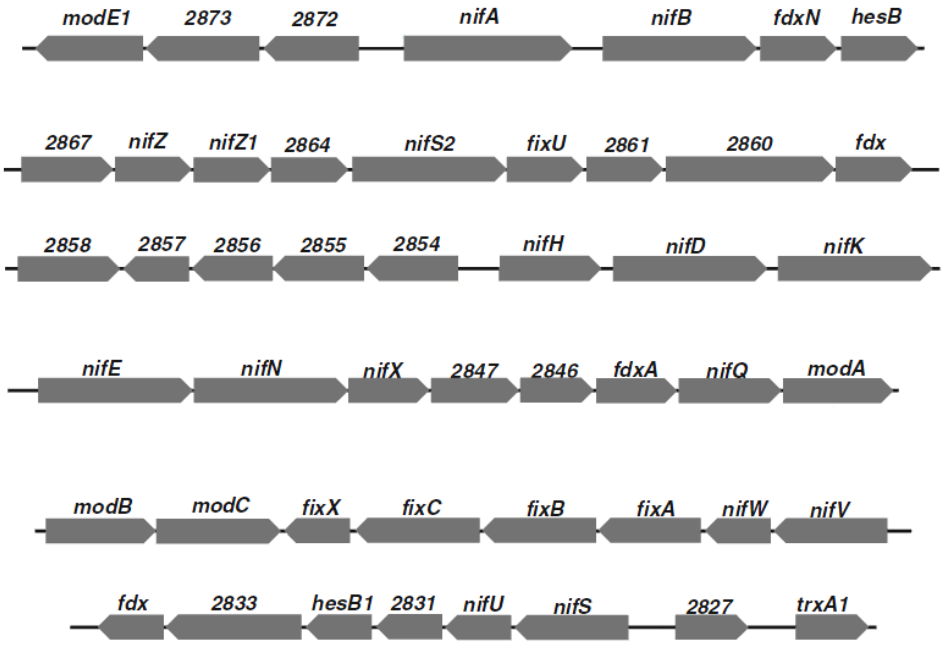


FIGURA 9 ORGANIZAÇÃO ESTRUTURAL DO CLUSTER *nif* E *fix* NO GENOMA DO *Herbaspirillum seropedicae* SmR1.

FONTE: CHUBATSU et al., 2011

1.5 Genes de Recombinação e Reparo do DNA

A preservação da informação genética é imprescindível para a perpetuação de uma espécie, fazendo-se necessário um complexo mecanismo de reparo do DNA para a correção de erros em seu conteúdo. Os erros mais comuns são o mau pareamento, excisão de bases e a quebra na fita simples ou dupla (SNUSTAD et al., 2001). Em procariotos, há pelo menos um mecanismo específico com a finalidade de evitar e/ou corrigir tais erros, no qual diversos genes são expressos, entre eles os genes *rec*, *mut*, *uvr* e *lex*A (WITKIN, 1976).

Os genes *rec* estão diretamente envolvidos no processo de recombinação genética homóloga que é a base dos processos de reparo do DNA em bactérias (NELSON et al., 2002).

Quando o DNA é fortemente danificado por agentes mutagênicos, um sistema de reparo denominado Sistema SOS é induzido e uma série de genes é ativada, resultando na recombinação homóloga e na produção de proteínas de replicação que garantem a preservação da informação genética. Tal resposta é uma tentativa de escapar dos efeitos letais de um DNA muito danificado. Diversos genes são expressos nesta resposta, entre eles os *uvr, rec, umu e lexA* (JANION, 2008; McKenzie et al., 2000).

O gene *rec*A codifica para a proteína RecA, uma enzima que apresenta diversas funções, como recombinação genética, reparação pós replicativa e indução da Resposta SOS. Consiste em um gene altamente conservado, e, portanto, um importante parâmetro para classificação filogenética e evolução em procariotos e eucariotos, inclusive em organismos complexos de todos os reinos (BRENDEL et al., 1996). Juntamente com o gene *16S rRNA*, a análise do gene *recA* pode auxiliar em estudos taxonômicos (KARLIN et al., 1995; ROCHA et al.,2005).

Os genes *recA e recX* de *Herbaspirillum seropedicae* SmR1 já foram caracterizados estrutural e funcionalmente (GALVÃO, 2005; STEFFENS et al., 1993). O mutante *rec*A apresentou elevada sensibilidade a agentes mutagênicos enquanto o mutante *rec*X apresentou sensibilidade parcial, confirmando o papel fundamental da proteína RecA na manutenção da viabilidade celular e a participação de RecX na Resposta SOS.

1.6 Objetivos

1.6.1 Objetivo geral

- Apresentar uma montagem parcial do genoma da bactéria *Herbaspirillum hiltneri* $N3^T$ a partir de leituras curtas em *color space* e provenientes de sequenciamento utilizando a plataforma SOLiD™ *Sequencing System*.

1.6.2 Objetivos específicos

- Realizar montagens baseadas em sequenciamentos em *mate-paired* e *fragments;*
- Integrar as melhores montagens em *mate-paired* e *fragments* em uma montagem híbrida;
- Ordenar os contigs e *scaffolfds* de acordo com o genoma de referência *Herbaspirillum seropedicae* SmR1;
- Realizar a anotação parcial do genoma;
- Buscar no *draft* do *H. hiltneri* obtido genes relacionados ao metabolismo do nitrogênio e à recombinação homóloga.

2 METODOLOGIA

2.1 Origem dos Dados – Microrganismo e Sequenciamento

A bactéria *Herbaspirillum hiltneri* estirpe tipo N3^{T} foi gentilmente cedida pelo Dr. Anton Hartman, do Centro Nacional de Pesquisa de Saúde e Meio Ambiente (Neuherberg, Alemanha).

O sequenciamento do genoma do *Herbaspirillum hiltneri* foi realizado no Núcleo de Fixação de Nitrogênio, situado no Departamento de Bioquímica e Biologia Molecular da Universidade Federal do Paraná, utilizando o equipamento The SOLiD^{TM} *Sequencing System*. Foram realizados dois sequenciamentos distintos: o primeiro conjunto de dados foi proveniente da técnica em *mate-paired*, onde as leituras são denominadas *tags*. Foram produzidas 53.725.333 leituras na primeira *tag* (F3) e 53.360.039 leituras na segunda *tag* (R3), totalizando 107.085.372 leituras. As duas *tags* estão distanciadas entre si pelo adaptador interno que tem cerca de 1.500 pb. O segundo conjunto de dados foi obtido do sequenciamento em *fragments* e foram obtidas 8.387.256 leituras. Todas as sequências apresentaram 50 bases.

A profundidade de cobertura foi calculada pela equação 1 e indica o número de vezes que genoma está representado no sequenciamento. O tamanho do genoma foi baseado no genoma do *H. seropedicae* que é de 5,5 milhões de nucleotídeos, uma vez que o tamanho do *H. hiltneri* é desconhecido.

$$Cobertura = \frac{Tamanho\ da\ leitura * Número\ de\ leituras}{Tamanho\ do\ genoma} \quad (1)$$

2.2 Análise dos Dados Brutos

A análise dos dados brutos produzidos no sequenciamento direcionou a estratégia de montagem. Nesta etapa foi possivel verificar o estado das leituras em um primeiro contato e tal observação foi realizada através da listagem dos arquivos FASTA e da sua respectiva qualidade. Posteriormente, os dados foram indexados a fim de se analisar todo o conteúdo em diversos pontos dos arquivos através da utilização do software MATLAB (The Language Of Technical Computing).

Outro software utilizado foi o FASTQC[1] que é capaz de realizar diversas análises estatísticas, como frequência de bases (absoluta e ao longo da leitura), de valores de qualidade, repetições por *k-mer,* etc. Como o arquivo de entrada foi originalmente desenvolvido para o sequenciador Illumina® e deve ser em fastq, foi necessário converter as leituras em *color space* para tal formato. Para tanto foi utilizado um *script* desenvolvido no Programa de Pós Graduação em Bioinformática da UFPR[2] (GUIZELINI, D., dados não publicados).

2.2.1 Análise de qualidade

A qualidade das leituras produzidas em um sequenciamento é de suma importância para a montagem do genoma, uma vez que quanto maior qualidade dos dados, maior a sua confiabilidade. Diferentemente de outros sequenciadores de alto desempenho, o SOLiD™ não dispõe de um programa interno que realize o processo de "pré filtração" de dados de má qualidade (SASSON et al., 2010), o que resulta em grande quantidade de leituras pouco informativas e descartáveis. Como consequência, requer maior atenção na análise da qualidade das sequências e, em posse desta análise pode-se, quando necessário, realizar o *trimming,* ou seja, a poda de algumas bases da ponta 3' bem como o descarte das leituras de baixa qualidade. A poda já é um procedimento estabelecido e bastante utilizado nos dados de sequenciamento de nova geração, em especial com leituras curtas (DiGUISTINI et al., 2009; QU et al., 2009).

Para a análise qualitativa foram utilizados os programas Matlab 2010 e Quality Assessment Software. Através do Quality Assessment foi realizada a análise de qualidade média por base ao longo de cada leitura (RAMOS et al., 2011).

---

[1] Disponível em http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

[2] http://www.bioinfo.ufpr.br/

2.2.2 Análise de mapeamento

Estudos de similaridade das leituras frente a um genoma de referência foram realizados utilizando o software Mosaik 1.1.0017 (STRÖMBERG, et al., 2009). O mapeamento das leituras foi realizado com os genomas de referência do *Herbaspirillum seropedicae* SmR1, *H. rubrisubabicans* e a *Collimonas fungivorans.* Esta última foi escolhida em função da sua proximidade filogenética (figura 8). As análises de comparação genômica realizadas neste estudo foram baseadas em *H. seropedicae* estirpe SmR1, um mutante natural resistente ao antibiótico estreptomicina, por ser a única representante do gênero *Herbaspirillum* já completamente sequenciada, anotada e publicada (PEDROSA et al., 2011).

Dotplot é o diagrama baseado em métodos estatísticos no qual o alinhamento é realizado entre duas sequências, seja aminoácido ou núcleotídios, a fim de se obervar a sua similaridade (GIBBS et al., 1970). Para tal estudo dos resultados de montagem, ou seja, dos conjuntos de contigs e/ou *scaffolds*, foi utilizado uma função para Matlab desenvolvida no Programa de Pós-Graduação em Bioinformática UFPR (RAITTZ, R. T., dados não publicados).

2.3 Montagem Parcial do Genoma de *H. hiltneri*

2.3.1 Montagem utilizando dados brutos em color space

Para a obtenção da montagem parcial foi utilizado o fluxograma de execução (*pipeline*) "*de novo* accessory tools 2.0", disponível no *site* da *Life Tech/Applied Biosystems*[3], desenvolvido especificamente para tecnologia SOLiD™. Consiste das seguintes etapas, respectivamente:

a) SAET - SOLiD™ Accuracy Enhancement Tool v.2.2

Etapa anterior à montagem; objetiva realizar a correção de possíveis erros nas leituras, considerando cada valor de qualidade correspondente e a sequência consenso.

[3] Disponível em http://solidsoftwaretools.com/gf/project/denovo/

b) Preprocessor v.1.0

Converte os dados de *color space* (.csfasta) para *double encoded* (.de), arquivo de entrada para o programa Velvet. Além disso, retira a primeira base de cada leitura e descarta as leituras sem seu respectivo par, no caso de *mate-paired*.

c) Velvet v.0.7.55

Software de montagem indicado pela *Life Technologies – Applied Biosystems* para a montagem dos dados do SOLiD™. O Velvet é apropriado para leituras curtas e é baseado no Grafo de Bruijn. Inicialmente o Velveth é executado para a construção da tabela *hash* baseado no *k-mer* e em seguida é a vez do Velvetg realizar a manipulação do grafo. Quanto maior a semente, maior será a especificidade da montagem, mas o número de leituras utilizadas será menor. Para a realização de uma boa montagem, é imprescindível determinar o *k-mer* ideal, que considera o equilíbrio entre estes parâmetros (ZERBINO et al., 2008).

d) ASiD - Assembly Assistant for SOLiD v.1.0

Fecha falhas entre contigs em *scaffolds*.

e) Postprocessor v. 1.0

Converte os contigs em *basespace* com significado biológico.

f) Analyse 1.0

Analisa as estatísticas dos contigs e *scaffolds* montados, incluindo porcentagem de cada base, n50, cobertura, contig máximo, mínimo e médio.

A montagem necessita de testes com diversos parâmetros para aperfeiçoar o resultado. Tais parâmetros variam de acordo com o genoma e o tipo de sequenciamento, fazendo com que se faça necessário o seu ajuste de acordo com os arquivos de saída do montador. O fluxograma com as etapas realizadas com as melhores montagens está representado na figura 10.

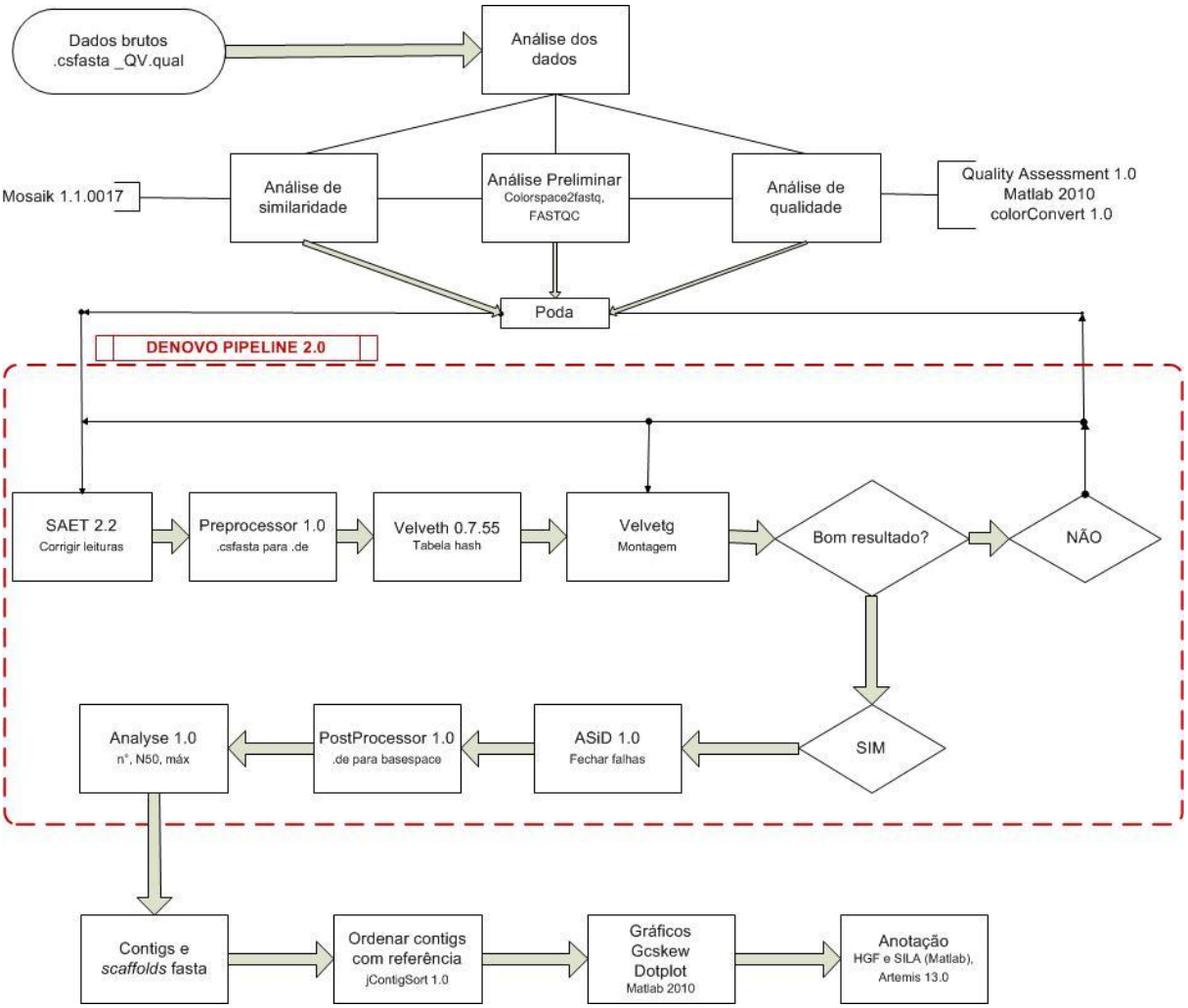


FIGURA 10 FLUXOGRAMA PARA CADA MONTAGEM REALIZADA A PARTIR DAS LEITURAS EM *COLOR SPACE.*

Inicialmente foi realizado o estudo das leituras brutas para então serem definidos os parâmetros de montagem para o *de novo pipeline*. Os arquivos de saída, conjuntos de contigs e *scaffolds*, foram ordenados, anotados e analisados.

2.3.2 Montagem híbrida

Para integrar os melhores resultados de contigs e *scaffolds* das montagens em *fragments* e *mate-paired*, respectivamente, foi realizada uma montagem híbrida utilizando o montador Phrap v. 1.080812 (GREEN et al., 1999) do pacote Phred/Phrap/Consed. Esta etapa teve a finalidade de unir os dados provenientes de diferentes fontes em um único conjunto, partindo do princípio que eles se complementam (CERDEIRA et al., 2011). Neste caso, foram dados provenientes de sequenciamentos realizados com técnicas distintas. A fim de se eliminar dados repetidos, foi utilizado o programa Simplifier 0.3, uma ferramenta que elimina

sequências redundantes de um conjunto de contigs gerados pela NGS (RAMMOS et al., 2012). A figura 11 mostra o fluxograma no qual ocorre a união dos dados híbridos e os procedimentos seguintes.

2.4 Avaliação da Montagem

Para estudar a confiabilidade e coerência de cada montagem, cada conjunto de contigs e *scaffolds* foi ordenado, avaliado pela visualização dos gráficos de GCskew e dotplot, e anotado.

2.4.1 Ordenação dos contigs

Para identificar a ordem correta dos contigs, estes foram comparados e alinhados com o genoma de referência *Herbaspirillum seropedicae* SmR1 utilizando o jContigSort 1.0 (GUIZELINI et al., 2011).

2.4.2 GC Skew acumulado

A análise do GCskew acumulado é um importante parâmetro gráfico para visualização e validação da montagem. Em organismos cujo DNA cromossomal se encontra na forma circular, pode se observar claramente duas regiões distintas: a primeira com excesso da base C sobre G e outra posterior que apresenta o comportamento inverso de excesso de G sobre C. O ponto de inflexão, ou seja, o pico do gráfico, que separa tais regiões é a origem de replicação (LOBRY, 1999; ARAKAWA et al., 2007). Esta tendenciosidade estatística sistemática pode ser facilmente visualizada plotando a soma da equação 2 em um gráfico em intervalos pré definidos, o que comprova a polarização da composição nucleotídica do genoma (ARAKAWA et al., 2007).

$$GCskew = \frac{(C - G)}{(C+G)} \qquad (2)$$

A plotagem do gráfico foi realizada através de uma função (RAITTZ, R.T., dados não publicados) no software Matlab, utilizando a soma da equação 2 em intervalos de 1.000 pb.

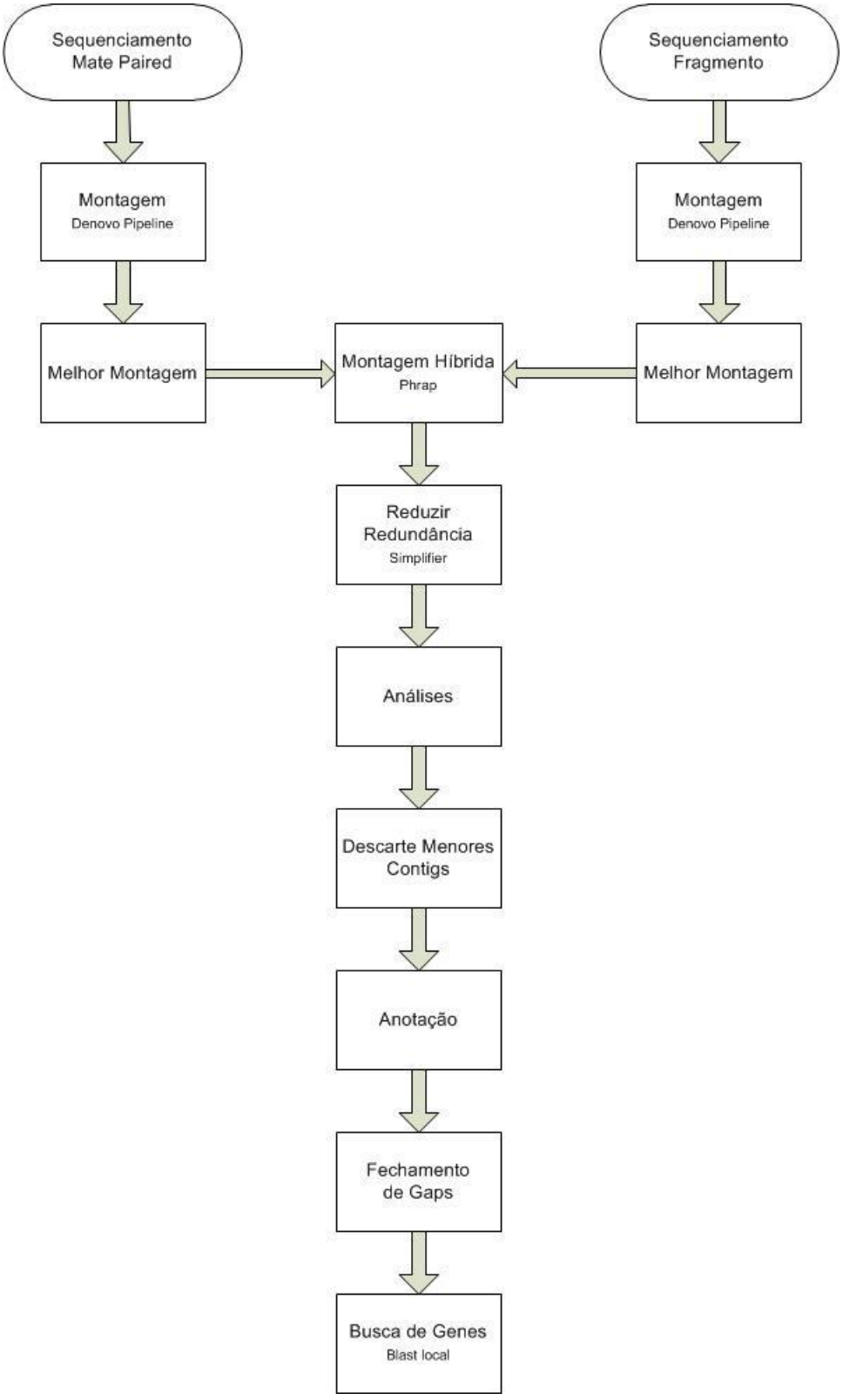


FIGURA 11 FLUXOGRAMA A PARTIR DOS CONJUNTOS DE CONTIGS E *SCAFFOLDS*

Os melhores resultados das montagens em *color space* foram integrados via montagem híbrida. Depois foi realizada uma análise parcial deste conjunto de dados, anotação, estudo do fechamento de falhas e busca de genes de interesse.

2.5 Anotação Parcial

A anotação foi realizada pelo programa HGF (*hydrid gene finder)* juntamente com o SILA (*sequence-indexed local aligner*), ambos desenvolvidos no Programa de Pós-Graduação em Bioinformática da UFPR (RAITTZ, R.T., dados não publicados). Estes conjuntos de funções são executados juntamente com o Matlab e realizam o alinhamento pelo Blast local de cada fase de leitura aberta ou *orf* (*open reading frame*) a procura de possíveis genes baseados no banco de NR disponível no site do NCBI. Assim, o melhor *score* de cada *orf* é identificado de acordo com a sua denominação no banco. Há uma escala de cores, na qual cinza corresponde a nenhum resultado compatível com a sequência, vermelho com a identificação positiva da *orf* e rosa como *score* significativo, mas não alto o suficiente para se fazer uma afirmação. Este último caso requer atenção diferenciada e deve ser estudado separadamente.

O arquivo de saída é no formato gbk e foi visualizado e analisado no programa Artemis 13.0, uma ferramenta que possibilita a visualização e anotação do DNA, no contexto de sequencia e em suas seis fases de leitura (RUTHEFORD et al., 2000; BERRIMAN et al., 2003).

2.6 Estudo do Fechamento de Falhas

O fechamento de falhas consiste na união de contigs no interior de *scaffolds* na região contendo bases indeterminadas. A estratégia mais comum é a realização do alinhamento do *scaffold* frente a um banco de dados constituído de conjuntos de contigs, leituras ou genes e eliminar as falhas onde há presença de sobreposição da sequência em ambos os lados (figura 12).

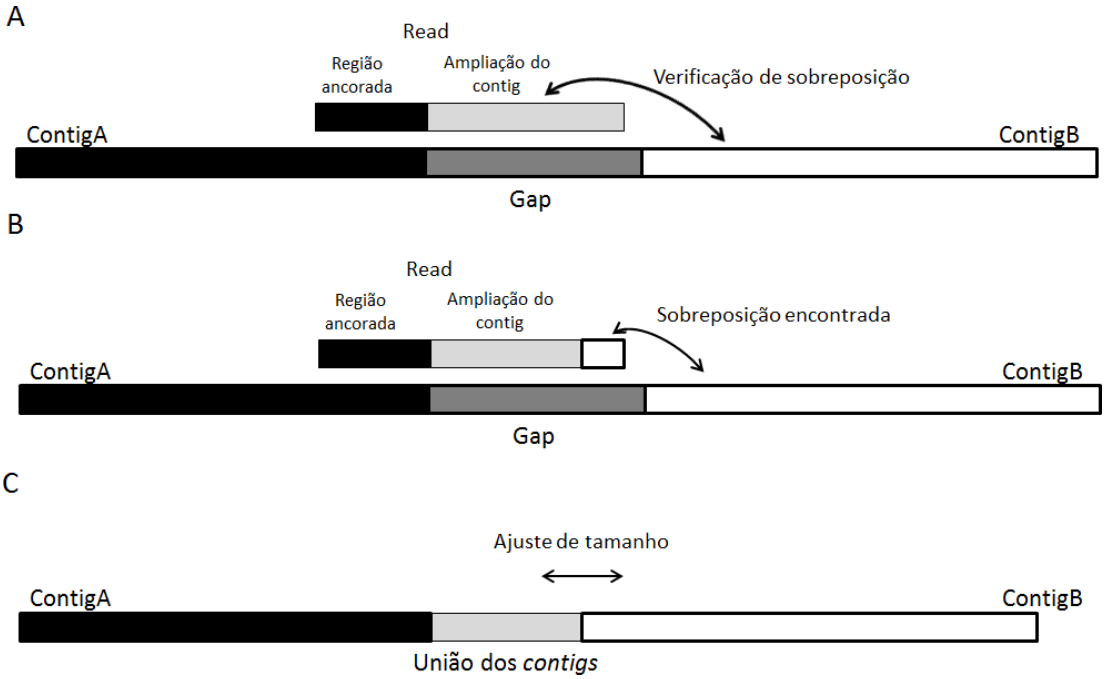


FIGURA 12 DEMONSTRAÇÃO DO FECHAMENTO DE UMA FALHA

Em A) há um *scaffold* composto por dois contigs separados por uma falha. Uma leitura ancora na extremidade do ContigA e a sua ampliação é comparada com o contigB. Em B) uma parte da leitura se sobrepõe com a extremidade do contigB. Em C) Ocorre a união dos contigs A e B com a inserção da leitura e o fechamento da falha. Houve um ajuste de tamanho da distância entre os contigs, pois comumente o tamanho da falha é estimado pela informação de pareamento, mas é não exato.

FONTE: CARDOSO, 2011.

No estudo para fechamento de falhas manual foi utilizado a ferramenta de busca de alinhamento básico local - BLAST - (ALTSCHUL et al., 1990) versão 2.2.24+[4], onde o *scaffold* foi utilizado como *Query* (pesquisa) e alinhado contra um banco de dados, a fim de se encontrar regiões de sobreposição nas falhas.

Outra estratégia utilizada foi via um conjunto de funções para Matlab desenvolvido no Programa de Pós-Graduação em Bioinformática (RAITTZ, R.T., dados não publicados), o que possibilitou encontrar a posição da falha automaticamente e visualizá-la para então eliminar os “N’s”. Nesta etapa, a utilização dos dados em *fragments* foi fundamental, uma vez que dados heterogêneos facilitam o processo de união de contigs.

[4] Disponível em http://www.ncbi.nlm.nih.gov

2.7 Busca de Genes

Para localizar genes específicos de interesse foi utilizado o blast local versão 2.2.24+. Para tanto, foi feito um banco de dados com a melhor montagem híbrida e as sequências nucleotídicas dos genes foram comparadas com esta biblioteca. Os resultados com alto índice de significância estatística (*bitscores* superiores a 200) foram considerados positivos, ou seja, que há um forte indício da presença dos genes no genoma. Já alinhamentos com *bit scores* inferiores a 200 foram descartados e considerados ausentes. Este valor foi baseado na versão online do Blast que considera que *scores* acima de 200 conferem alta de similaridade.

Esta etapa foi também uma forma de confirmar o conteúdo biológico da montagem e pode ser utilizada para auxiliar na união de contigs. Os genes procurados foram os ribossomais do *H. hiltneri*; o cluster *nif* e o cluster r*ec,* sendo estes últimos baseados em *H. seropedicae* SmR1. Todos os genes são oriundos do banco de genes do NCBI.

3 RESULTADOS E DISCUSSÃO

Os dados obtidos do sequenciador SOLiD™ pelas bibliotecas de *mate-paired* (conjunto das pontas F3 e R3) e *fragments,* foram analisados, montados utilizando diferentes estratégias e anotadas.

3.1 Análise das Leituras

O primeiro estudo realizado foi referente à quantidade de leituras presentes e qual a sua respectiva cobertura. A profundidade de cobertura total foi de 1.041 vezes (tabela 1).

TABELA 1 NÚMERO E PROFUNDIDADE DE COBERTURA DAS LEITURAS DOS SEQUENCIAMENTOS

Dados de Sequenciamento	Leituras (milhões)	Profundidade de Cobertura
Mate-paired	107,08	973,45x
Fragments	8,39	76,26x
Mate-paired + Fragments	114,59	1.041,73x

Na análise visual foi observada a presença de diversos pontos ao longo das leituras, ou seja, bases indeterminadas nas quais não houve sinal luminoso detectado no equipamento. Para cada ponto lhe é conferido um valor negativo (-1) como qualidade (figura 13), resultando no decaimento da qualidade média da leitura e descarte de toda a sequência, uma vez que o *color space* é baseado na codificação di-base.

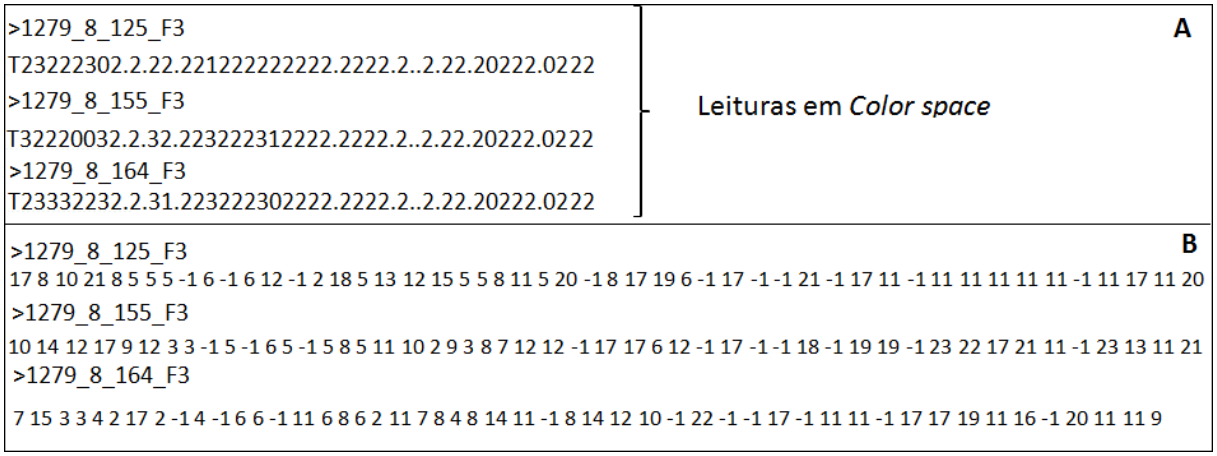


FIGURA 13 DETECÇÃO DA PRESENÇA DE PONTOS NAS LEITURAS

Nas leituras em *color space* foram encontradas bases indefinidas codificadas como pontos (A), que recebem valores negativos referentes à qualidade (B).

Outro desafio foi a presença de bases repetidas na ponta 3' da leitura. Estas caudas poliméricas aparecem sistematicamente na grande maioria das leituras e em número variado, chegando frequentemente a constituir até metade da sequência (figura 14). Não foi observada nenhuma tendenciosidade para um nucleotídeo específico.

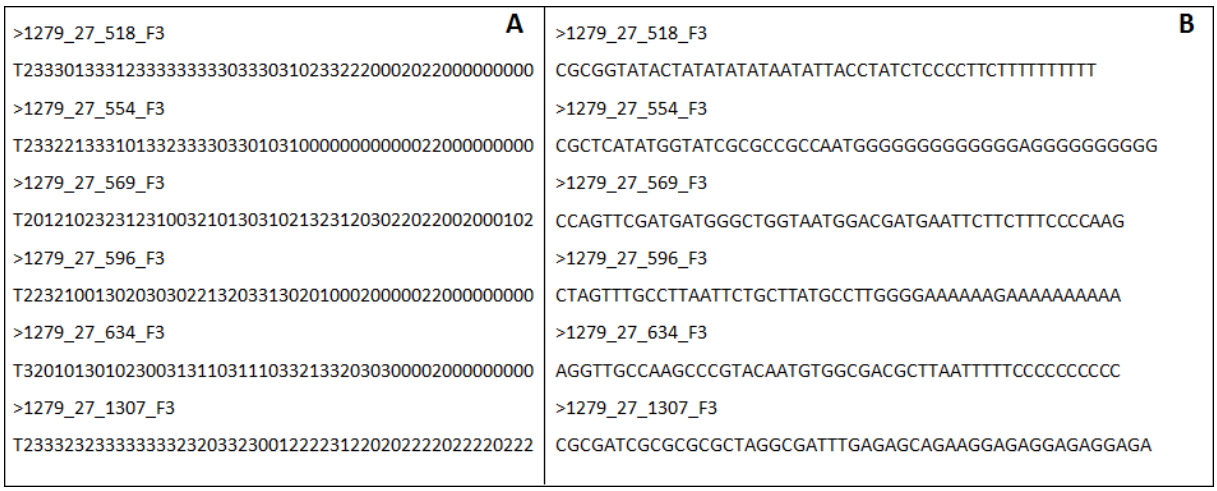


FIGURA 14 PRESENÇA DE BASES REPETIDAS NA PONTA 3' DAS LEITURAS

Foram observadas bases repetidas na ponta 3' das leituras. Em A) observa-se que as sequências codificadas em *color space* frequentemente recebem o valor zero. Em B) essas mesmas sequências estão em fasta, facilitando a visualização.

Através do programa FASTQC foi possível visualizar graficamente estas repetições, como representado no gráfico da figura 15. Esta análise é baseada na contagem da frequência de cada *k-mer* de 5 bases na biblioteca de leituras. Então este número de pentâmero observado é dividido pela frequência média esperada por

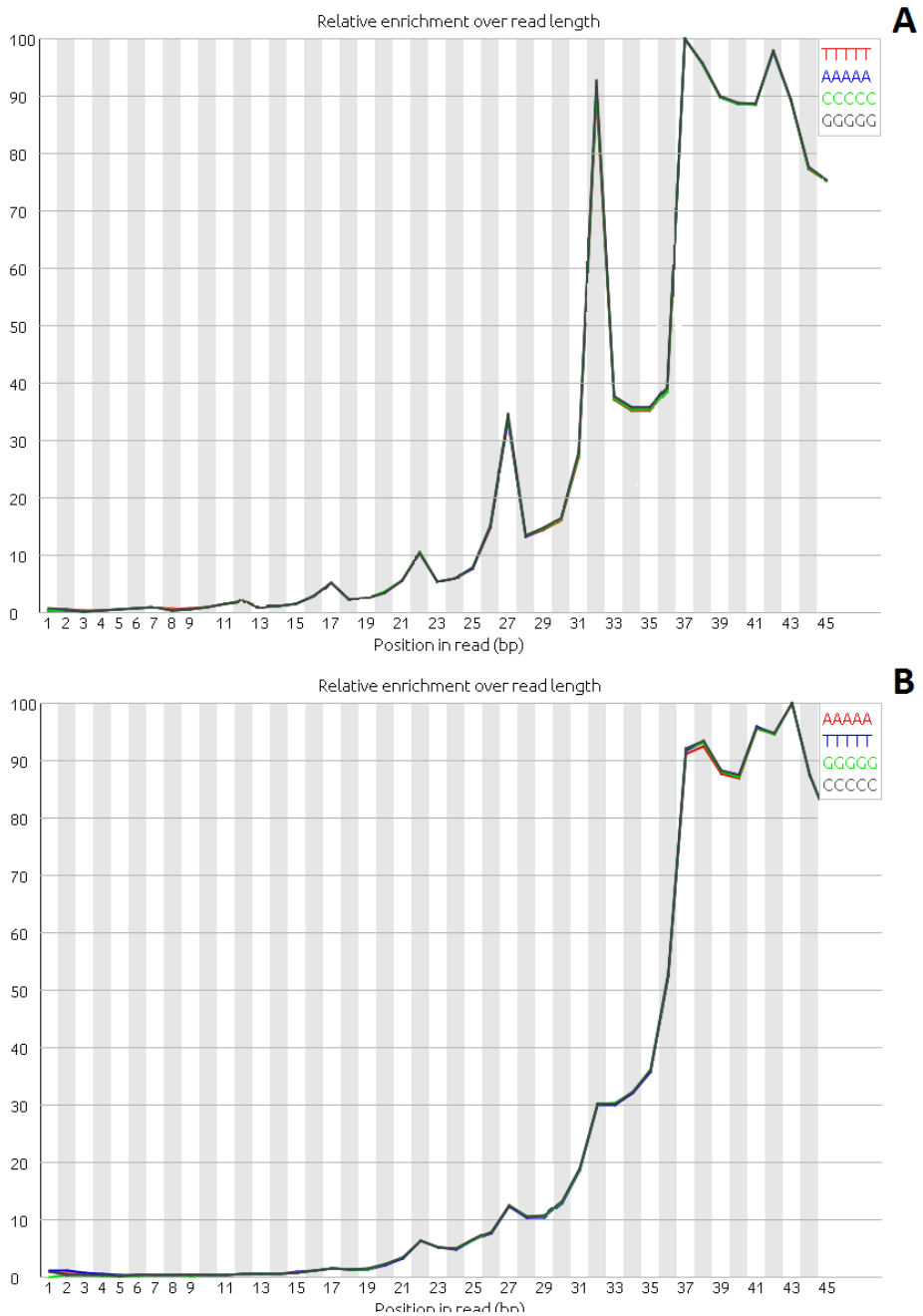


FIGURA 15 TENDENCIOSIDADE POR *K-MERS* NA PONTA 3' DA *TAG* F3 (A) E R3 (B).

O gráfico gerado pelo programa FASTQC representa a frequência dos quatro pentâmeros mais comuns (AAAAA,TTTTT,GGGGG e CCCCC) ao longo da posição da leitura (eixo X). Na ponta 3' é observado um pico (demonstrado pelo valor 100 no eixo Y como o maior valor observado para esse conjunto de dados), indicando que é a região com maior quantidade de bases repetidas.

base para esta sequência. As repetições de TTTTT, AAAAA, CCCCC e GGGGG foram as mais frequentes e, em sua maioria, na ponta 3'. Estas regiões não correspondem a dados reais e foram causadas por problemas ainda não identificados, que podem ter ocorrido durante o sequenciamento ou na montagem da biblioteca. A partir desta observação, se fez necessário um estudo aprofundado da qualidade. No sequenciamento do genoma em *fragments* comprovou-se a inconsistência das regiões repetidas, já que estas não foram observadas. Ambos os problemas discutidos ocorreram expressivamente e apenas no sequenciamento em *mate-paired*.

Outro estudo realizado no FASTQC foi da frequência das bases (A,C,T e G), e o resultado foi de 52% de G+C. A figura 16 mostra a frequência de cada base ao longo da leitura. Em todos os conjuntos foi observado que no final das leituras ocorre a queda do conteúdo GC, praticamente se igualando ao de TA, o que contabiliza cerca de 25% de frequência de cada base e ocorreu mais expressivamente nos dados em *mate-paired*.

Todos estes erros encontrados nas leituras durante a análise preliminar, tanto os pontos como o comportamento diferenciado no final da leitura, contribuíram para a idéia de poda de bases da ponta 3' com o objetívo de eliminá-los.

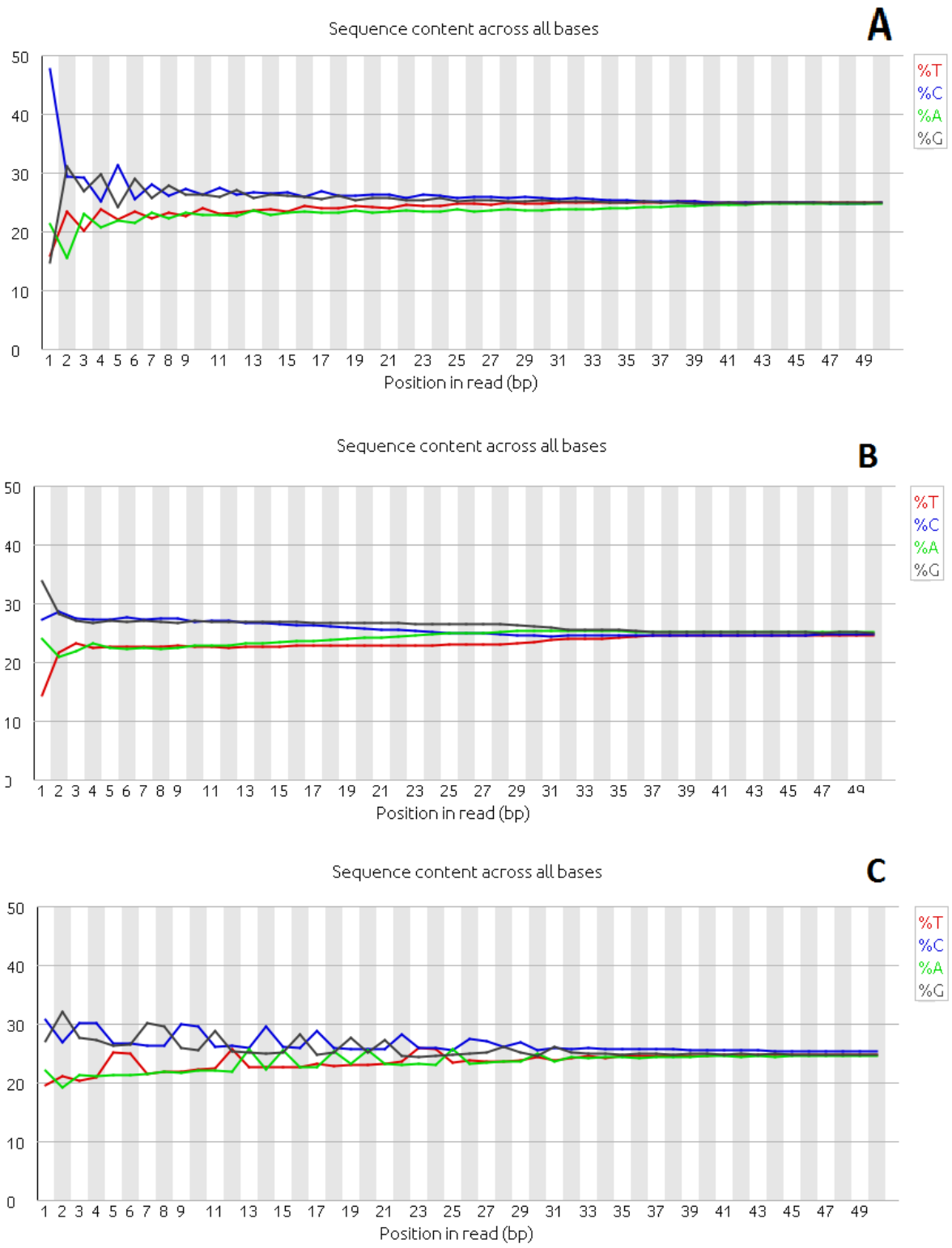


FIGURA 16 FREQUÊNCIA DE CADA BASE AO LONGO DA LEITURA

Estão retratadas as frequências médias (eixo das ordenadas) de A,T,C e G ao longo das leituras (eixo das abscissas) da A) *tag* F3, B) *tag* R3 e C) *fragments*. Em todos os conjuntos é observada a queda do conteúdo GC de acordo com a posição (pb) e igualamento com a porcentagem de AT. Gráficos gerados pelo software FASTQC.

3.1.1 Análise de qualidade e poda das sequências

Com o programa Quality assessment foi gerado um gráfico onde se demonstra a qualidade média por localização na leitura (figura 17). Em F3 é observada a baixa qualidade da leitura desde o início representada pelo valor 14. No final houve um pico a partir da 38ª base. Tal comportamento atípico também ocorreu com a *tag* R3 a partir da 33ª posição. Estes valores mais altos no final na leitura correspondem às bases repetidas nas pontas 3' discutidas anteriormente. Já os dados referentes ao sequenciamento em *fragments* (figura 17-C) apresentaram boa qualidade, iniciando de 26 e decaindo ao longo da leitura como esperado (CHAISSON et al., 2008; QU et al., 2009).

Os resultados das análises preliminares das leituras juntamente com os estudos qualitativos justificaram a retirada das bases na ponta 3', já que os erros ocorreram mais expressivamente no final das sequencias. Este procedimento de poda foi realizado de maneira variável e foram realizados testes com leituras de diferentes tamanhos: 25, 30, 35, 40 e 50, sendo este último correspondente à leitura inteira.

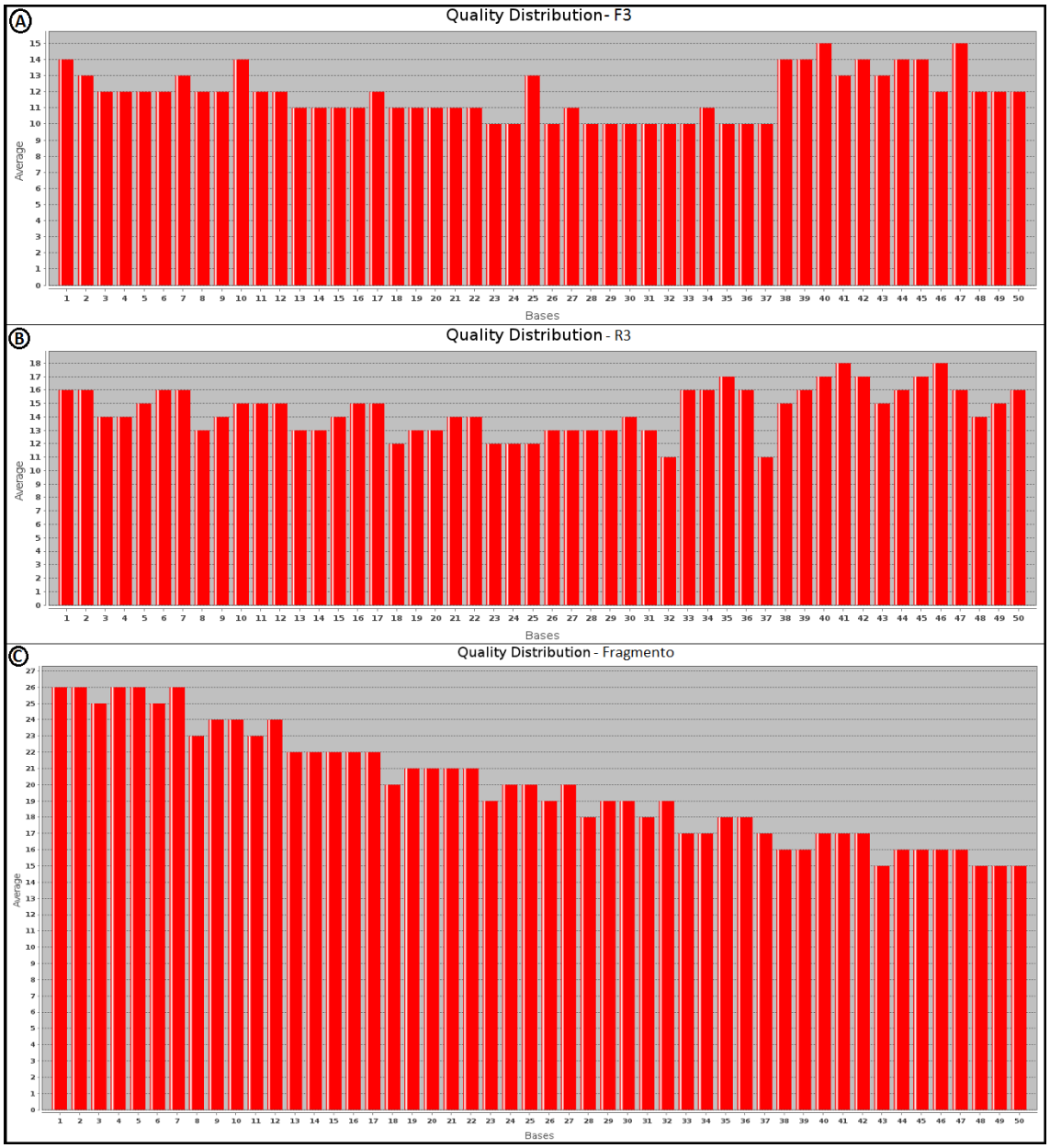


FIGURA 17 DISTRIBUIÇÃO DO VALOR DE QUALIDADE MÉDIA AO LONGO DA LEITURA

Os gráficos gerados no programa Quality Assessment demonstram o valor médio da qualidade no eixo y de acordo com a posição da base na leitura no eixo x: A) Ponta F3 do sequenciamento em *mate-paired*. B) Ponta R3 do sequenciamento *mate-paired* e C) Sequenciamento em *fragments*.

3.1.2 Análise de mapeamento

Utilizando o software Mosaik e leituras podadas para 35 bases, foi realizado o mapeamento com três genomas. A maior similaridade foi com o *H. seropedicae* SmR1, onde foi obtido 4,8% de identidade. O mesmo procedimento foi realizado com o *H. rubrisubalbicans* e foi obtido 4,3% de alinhamento. Já com a *Collimonas fungivorans* a similaridade foi de 3,3%. As leituras em *fragments* foram mapeadas

com o melhor conjunto de contigs em *mate-paired* e foi obtido 46% de alinhamento. Todos estes resultados de valores baixos são aceitáveis uma vez que o SOLiD™ produz grande quantidade de leituras passíveis de descarte, causando baixa porcentagem de mapeamento quando comparado com outras tecnologias. Estudos mostraram que mesmo entre organismos da mesma espécie comumente cerca de metade das leituras não alinha (SHEN et al., 2008; SUZUKI et al., 2011).

3.2 Montagem do Genoma

Foram realizadas montagens utilizando três conjuntos de dados: inicialmente o conjunto *mate-paired*, em sequida *fragment* e por último o melhor resultado de cada um deles constituiu a montagem híbrida (tabelas 2 a 4).

Todas as montagens resultaram em números bastante altos de contigs e, consequentemente, de tamanho reduzido. Este resultado era esperado em função do tipo de tecnologia utilizado, pois normalmente os sequenciamentos de nova geração baseados em leituras curtas tendem a apresentar maior desafio na etapa de montagem (MARTINEZ et al., 2010). Foram realizados diversos ensaios variando diferentes parâmetros a fim de se conseguir o melhor resultado. O tamanho médio do genoma observado nas montagens foi em torno de 5 milhões de nucleotídeos.

3.2.1 Montagem parcial do sequenciamento em mate-paired

Os principais parâmetros de montagem foram o tamanho da leitura após a poda e o *k-mer*. Na tabela 2 estão apresentados os 6 melhores resultados identificados de MP1 a MP6.

A montagem utilizando leituras com 40 bases não se mostrou eficiente, pois o Velvet utilizou apenas 7,3% das leituras que resultaram em um tamanho do genoma de 1,6 megabases, valor muito abaixo do esperado. O ensaio MP2 foi realizado para aumentar estes números. Assim, quando abaixou-se o *k-mer* de 31 para 29, o número de leituras utilizadas aumentou 2,2 vezes e o tamanho passou para 3,7 milhões de bases, mas ainda são valores considerados baixos. Isso ocorreu provavelmente em função das bases repetidas que permaneceram após a retirada de dez bases. Este resultado demandou uma poda mais expressiva.

TABELA 2 RESULTADOS DAS MONTAGENS EM *MATE-PAIRED*

PARÂMETROS	**MP1**	**MP2**	**MP3**	**MP4**	**MP5**	**MP6**
Conjunto de leituras (milhões)	107	107	107	107	107	22
Tamanho da leitura (pb)	40	40	30	30	25	40
k-mer	31	29	21	19	19	27
Leituras usadas (milhões)	7,82	16,96	44,22	50,93	46,44	6,0/13,5
Scaffolds	-	19499	8307	-	-	-
Contigs	6742	14792	13104	19315	17650	12425
N50 contigs (pb)	262	292	717	331	421	1847
N50 *scaffolds* (pb)	-	299	18927	-	-	-
Maior contig (pb)	2125	4869	8112	2930	5693	51134
Maior *scaffold* (pb)	-	22239	170031	-	-	-
Tamanho do genoma (Mpb)	1.59	3.71	5.19	5.08	5.09	3.14

Resultado das montagens realizadas com as leituras do sequenciamento em *mate-paired*. Estão mostrados o conjunto inicial de leituras para a entrada no *pipeline* da *de novo*, tamanho da leitura utilizado após a poda, *k-mer*, número de leituras utilizado pelo Velvet, número *de scaffolds* e contigs gerados, N50, maior contig e *scaffold* e tamanho do genoma (soma de todas as sequências montadas em milhões de pares de bases (Mpb)).

A montagem MP3 formou *scaffolds* com êxito e em menor número, apenas 8.307. Além disso, possui a maior sequência montada, com 170 mil bases e foi considerada a melhor montagem em *mate-paired.* A figura 18 contém os gráficos de alinhamento dotplot e GCskew com o genoma de referência do *H. seropedicae*, após a sua ordenação. Foi observada uma região com aproximadamente 600 mil bases na qual não houve alinhamento e que causou o acúmulo de sequências no início do eixo das abscissas. Uma explicação para isso é que a montagem apresenta sequências muito pequenas e/ou não informativas.

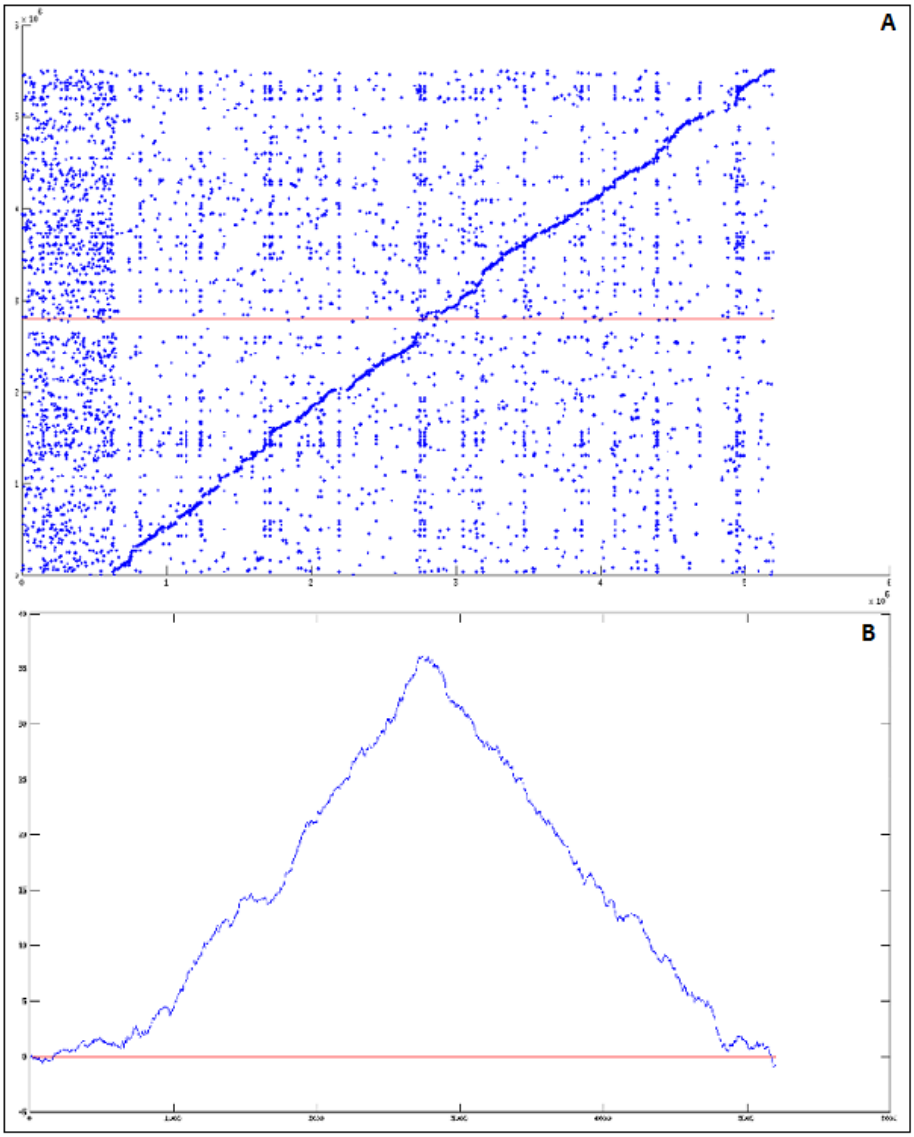


FIGURA 18 GRÁFICOS DE A) DOTPLOT E B) GCSKEW DO CONJUNTO DE *SCAFFOLDS* DA MONTAGEM MP3

Foi utilizado o genoma de referência de *H. seropedicae* SmR1. Em ambos os gráficos o eixo x representa a posição na montagem (Mpb) e o eixo y mostra a posição de A) genoma de referência e B) GCskew.

A montagem MP4 foi uma tentativa de aperfeiçoar o resultado anterior através da diminuição do *k-mer*. Consequentemente foi obtido maior aproveitamento de leituras, em torno de 47,6%. Entretanto, não houve formação de *scaffolds* e o número de contigs foi o mais alto, cerca de 20 mil.

No próximo teste (MP5) foi feita uma poda mais radical, descartando-se metade da leitura. O resultado foi um número extremamente alto de contigs, cerca de 17 mil.

O último ensaio (MP6) foi realizado com a utilização de um filtro de qualidade através do Quality Assessment (RAMMOS et al., 2010) no qual foram selecionadas as leituras com melhores valores de qualidade médios de cada *tag*, totalizando cerca de 22 milhões, ou seja, 11 milhões de F3 e 11 milhões de R3. Esta estratégia se mostrou pouco eficiente, pois somente as leituras que possuem o seu respectivo *pair mate* foram consideradas pelo *pipeline*, totalizando cerca de 13,5 milhões de leituras, enquanto o restante foi descartado. O Velvet então utilizou cerca de 6 milhões de leituras, que somaram cerca de 3 Mpb.

Não foi possível se obter resultado com êxito utilizando as leituras íntegras, (50 pares de bases), em função da limitação computacional, agravada pela baixa qualidade da ponta 3´. Os erros ocorreram na etapa de montagem durante o Velvet.

3.2.2 Montagem parcial do sequenciamento em fragments

A tabela 3 mostra as 10 melhores montagens do sequenciamento em *fragments*, denominadas MF1 a MF10. MF1 a MF7 foram montagens sem poda, ou seja, com 50 pb e MF8 a MF10 foram leituras podadas contendo 35 pb. Em geral as montagens apresentaram tamanho genômico um pouco menor do que as em *mate-paired*. O mesmo comportamento de número alto de contigs e estatísticas de N50 e contig máximo reduzidos se repetiu. O lado positivo foi o aumento de leituras utilizadas, chegando até a 60%. Foram então realizados testes onde o principal parâmetro modificado foi o *k-mer.* Vale ressaltar que não houve formação de *scaffolds* em função do tipo da metodologia de preparo da biblioteca de DNA.

TABELA 3 RESULTADOS DAS MONTAGENS EM *FRAGMENTS*

Parâmetros	**MF1**	**MF2**	**MF3**	**MF4**	**MF5**	**MF6**	**MF7**	**MF8**	**MF9**	**MF10**
N° leituras (mi)	8,387 mi									
Tamanho Leitura (pb)	50							35		
k-mer	19	21	23	25	27	29	31	19	21	25
Leituras usadas (mi)	4,83	5,11	4,91	4,7	4,26	3,37	3,18	4.23	4,8	5,06
Contigs	9165	11664	11369	11764	12335	12712	12800	15392	15739	17935
N50 (pb)	299	427	443	435	412	395	379	294	338	245
Maior contig	6123	3526	5709	5365	6059	7332	7843	6061	4311	1579
Tamanho do genoma (Mpb)	3.43	4.71	4.71	4.77	4.83	4.81	4.70	3.64	4.09	3.88

Resultado das montagens realizadas com as leituras do sequenciamento em *fragments*. Estão mostrados o conjunto inicial de leituras para a entrada no *pipeline* da *de novo*, tamanho da leitura, *k-mer*, número de leituras utilizado pelo Velvet, número de contigs gerados, N50, maior contig montado e tamanho do genoma (soma de todas os contigs).

As montagens realizadas com leituras podadas resultaram em um número de contigs muito superior às de leituras íntegras, variando de 15 a quase 18 mil, além de apresentar tamanho com no mínimo 1 Mpb a menos do esperado. Portanto, esta estratégia foi descartada.

O melhor resultado foi obtido na MF3, na qual se utilizou a leitura integral e *k-mer* 23. O seu alinhamento e GCskew acumulado pode ser observado na figura 19. Mais uma vez uma concentração de pontos no eixo horizontal foi observada, desta vez em torno de 500 mil bases.

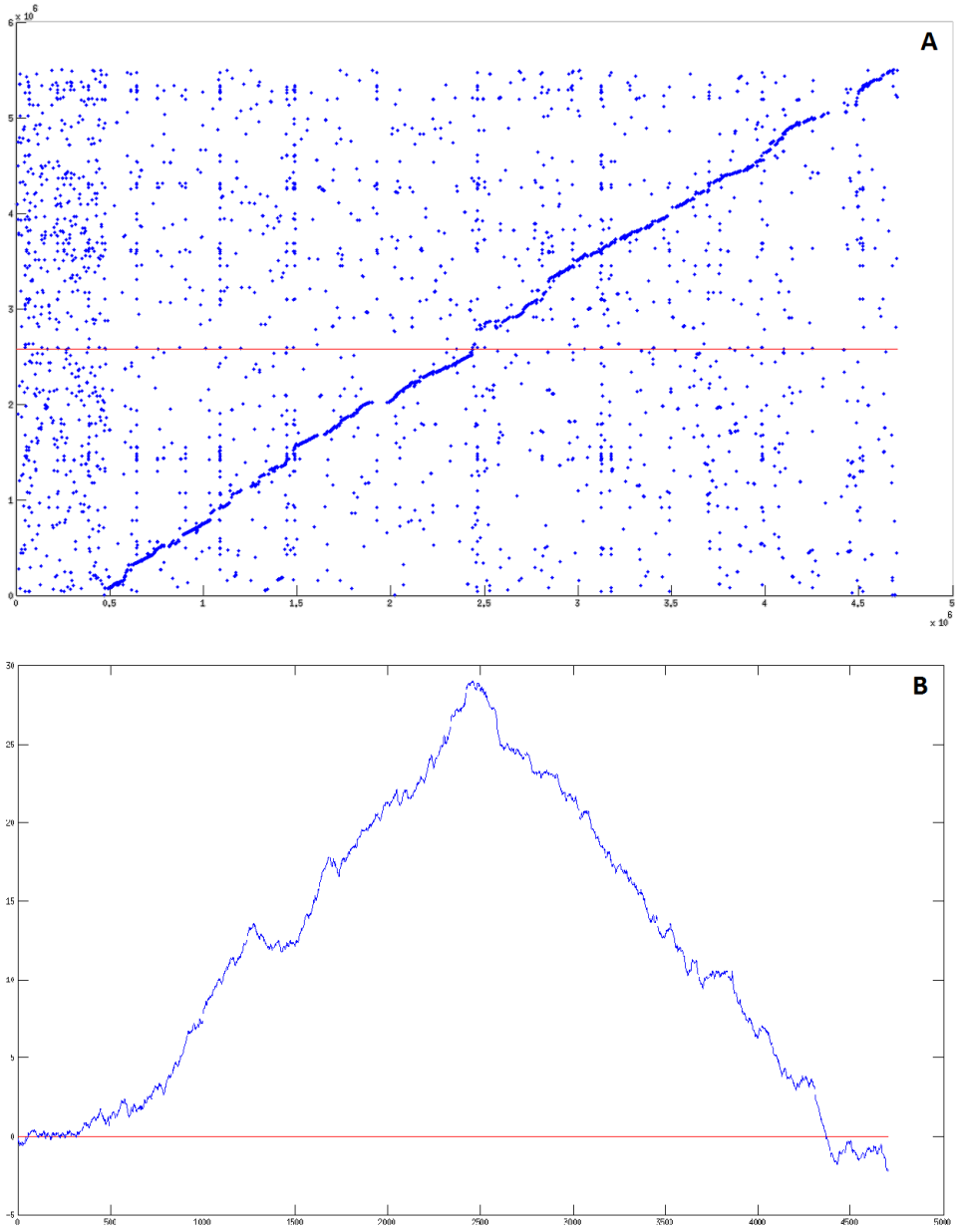


FIGURA 19 DOTPLOT E GCSKEW DA MONTAGEM MF3 EM *FRAGMENTS.*
Nestes gráficos foi utilizado o *H. seropedicae* SmR1 como referência.

Apesar das divergências das montagens, como a diferença do tamanho genômico, houve fortes indícios de concordância entre os dados, como pode ser visualizado no gráfico de dotplot da figura 20, onde os contigs de *fragments* foram ordenados e alinhados contra a melhor montagem em *mate-paired* MP3. No entanto, novamente o acúmulo de pontos foi observado, mas desta vez tanto no eixo das abscissas como nas ordenadas e, portanto, concluiu-se que ambas as montagens apresentam sequências muito pequenas e/ou não informativas.

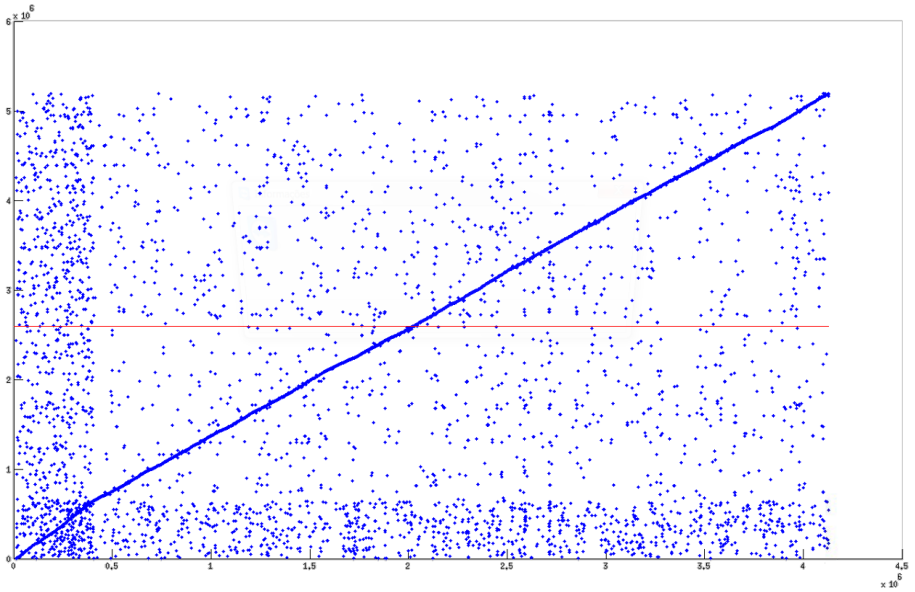


FIGURA 20 ALINHAMENTO DA MONTAGEM DE *FRAGMENTS* CONTRA *MATE-PAIRED*.

Os contigs resultantes do sequenciamento em *fragments* foram ordenados de acordo com a melhor montagem em *mate-paired* MP3. Uma forte concordância entre os dados pode ser observada.

3.2.3 Montagem híbrida

A fim de se unir os dados das montagens de *mate-paired* e de *fragments*, foi realizada a montagem híbrida utilizando os melhores resultados provenientes de cada um dos dois sequenciamentos. Foram realizados três ensaios de montagens híbridas (MH1 a MH3) utilizando diversos conjuntos de dados de ambos os sequenciamentos e os resultados estão apresentados na tabela 4. Concluiu-se que a montagem híbrida 1 foi a melhor dentre as três, pois resultou no menor número de contigs, 7.956, maior N50 e menor tamanho genômico (7,3 megabases). Entretanto, este valor é maior do que o esperado.

TABELA 4 RESULTADOS DA MONTAGEM HÍBRIDA

	MH 1	**MH 2**	**MH 3**
Mate-paired	MP3	MP5	MP3
Fragments	MF3	MF2	MF5
		MF6	
N° contigs inicial	19.676	53.313	20.642
N° contigs final	7.956	11.178	9.858
Tamanho (pb)	7.346.695	7.458.539	8.039.357
N50 (pb)	6.321	801	2.824

Utilizando o programa Simplifier 3.0 com o objetivo de reduzir a redundância, foi possível diminuir o número de contigs de 7.956 para 3.391 e o tamanho para 6,1 megabases (tabela 5).

TABELA 5 DADOS PARA MONTAGEM HÍBRIDA ANTES E APÓS A REDUÇÃO DA REDUNDÂNCIA

	Antes	**Após**
N° contigs	7.956	3.391
Tamanho do genoma (pb)	7.346.695	6.138.101
N50 (pb)	6.321	13.439

Entretanto, ainda houve um excesso de cerca de um milhão de bases no tamanho do genoma. Tais bases correspondem a dados redundantes que não foram identificados pelo Simplifier ou então sequências muito pequenas e não informativas. Portanto, optou-se por aplicar uma linha de corte de 700 pb, na qual todos os contigs menores do que esse valor foram descartados. Esta estratégia já foi observada em estudos anteriores, mostrando melhora nos resultados (TIEPPO, 2011). Após o corte, o tamanho do genoma caiu para cerca de 5 megabases, valor próximo do tamanho estimado (tabela 6). O acúmulo de bases no início do eixo das abscissas, observado em todos os gráficos de dotplot anteriores, foi eliminado com esta estratégia (figura 21). Pela plotagem do gráfico GCskew foi possível observar o aperfeiçoamento da montagem. Na figura 22 está apresentado o seu novo perfil mais padronizado e organizado em termos de ascensão e declínio. Como já determinado em outros estudos, a distância do pico até as extremidades é próxima a 50% do tamanho do genoma (GREGORIEV et al., 1998). A montagem final

apresentou 1.192 sequências, sendo 5.970 contigs e 283 *scaffolds.* O alto número de contigs é consequência da presença de um número elevado de de falhas (*gaps*).

TABELA 6 DADOS PARA MONTAGEM HÍBRIDA ANTES E APÓS A APLICAÇÃO DA LINHA DE CORTE DE 700 PB

	Antes	**Após**
N° sequências	3.391	1.192
Tamanho do genoma (Mpb)	6,14	5,04
Menor contig (pb)	371	701
N50 (pb)	13.439	19.633

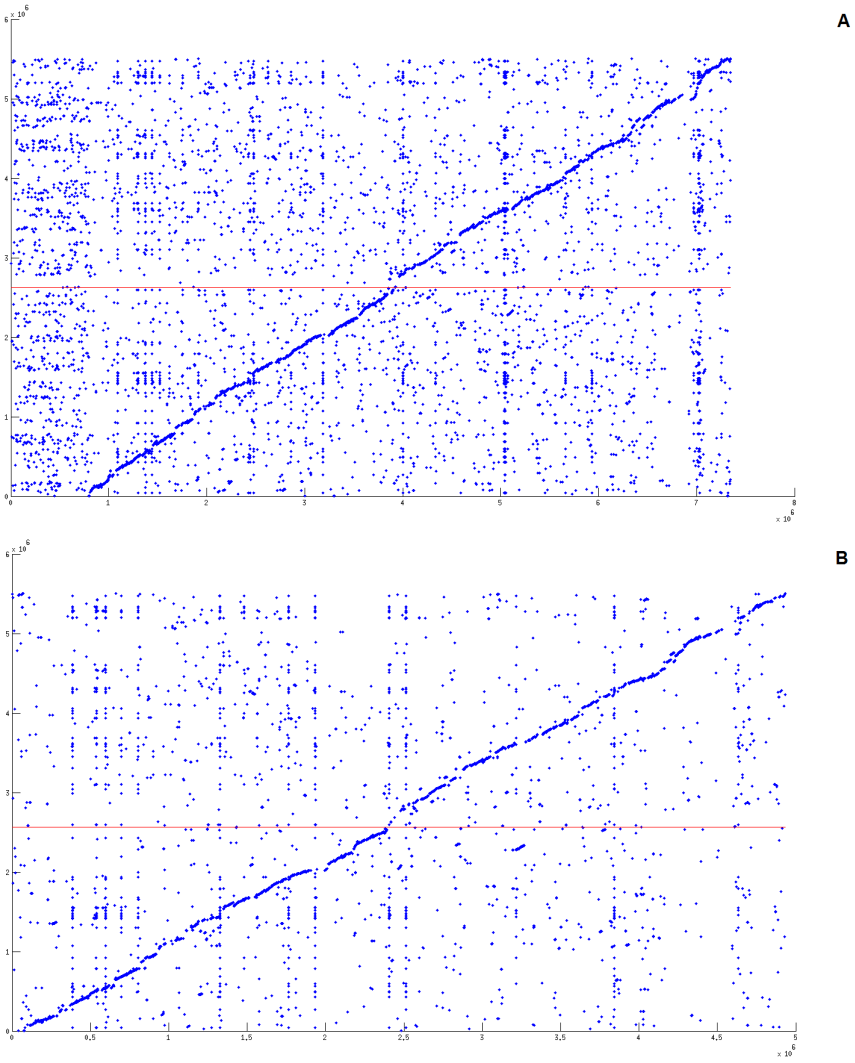


FIGURA 21 GRÁFICOS COMPARATIVOS DO DOTPLOT DOS DADOS ENTRE ANTES E APÓS AS ESTRATÉGIAS DE FINALIZAÇÃO DAS MONTAGENS.

Os gráficos demonstraram a otimização do resultado após a eliminação de redundância via Simplifier e descarte de sequências inferiores a 700 pb. A) Dotplot antes e B) após tais procedimentos.

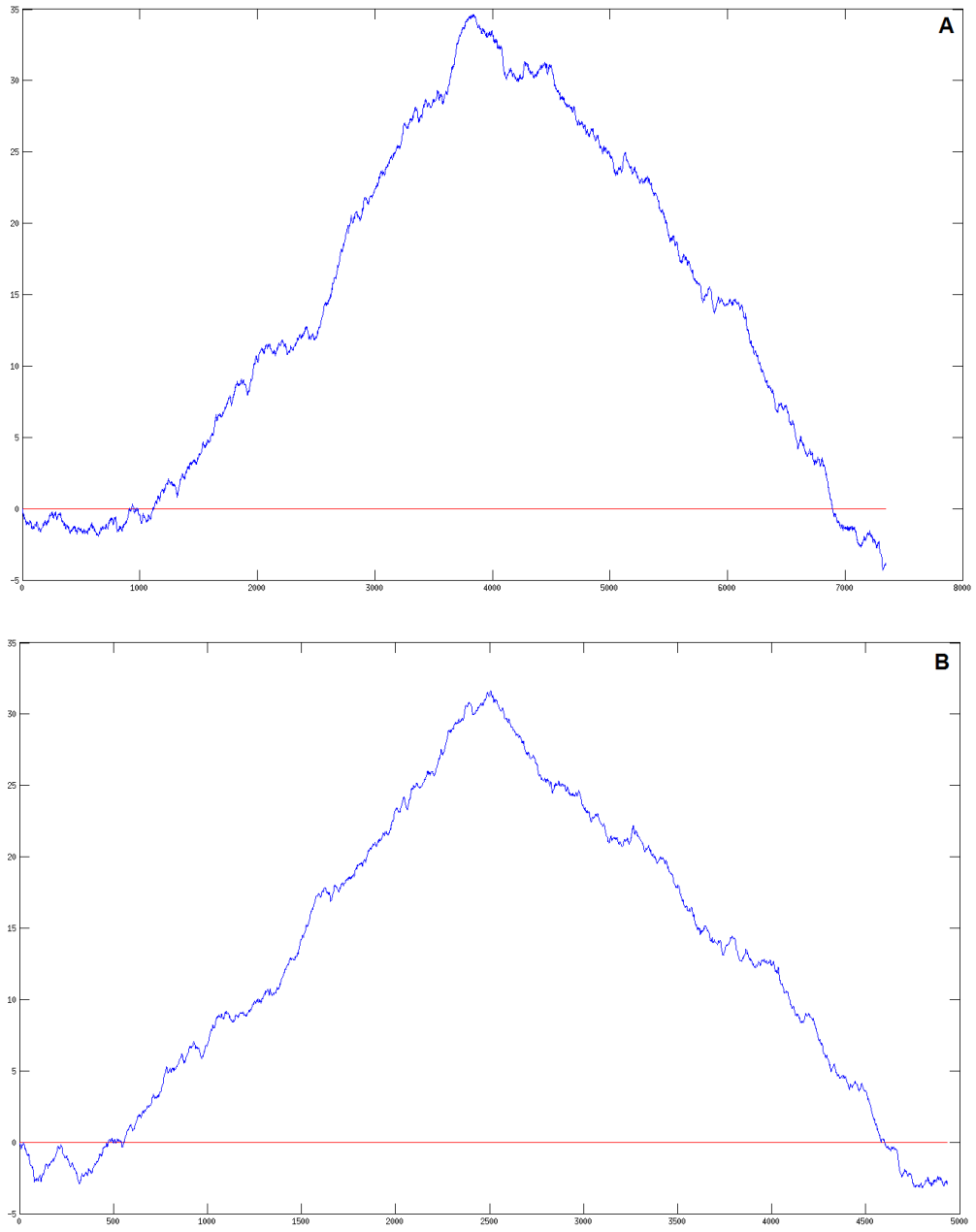


FIGURA 22 GRÁFICOS COMPARATIVOS DO GCSKEW ACUMULADO ENTRE ANTES E APÓS A ESTRATÉGIAS DE FINALIZAÇÃO DAS MONTAGENS.

Os gráficos demonstraram a otimização e padronização após a eliminação de redundância via Simplifier e descarte de sequências inferiores a 700 pb. A) GCskew antes e B) após tais procedimentos.

3.3 Fechamento de Falhas

Em posse da melhor montagem, foi realizado o cálculo do número de falhas, o qual resultou em 4.778. Este elevado valor é resultante do alto índice de fragmentação do genoma, fato que dificultou o processo de união de contigs dentro dos *scaffolds.* Então, optou-se por realizar um estudo de estratégias para a eliminação das falhas.

Na abordagem manual com a utilização do Blast local foi montado um banco de dados com os contigs da montagem em *fragments* e os *scaffolds* foram usados como *query*. Este procedimento se mostrou pouco eficiente, uma vez que foram raras as vezes em que havia cobertura na região de bases indefinidas. Na grande maioria das vezes, não houve contig que corrigisse a região de falha com alinhamento significativo de ambos os lados da mesma (figura 23).

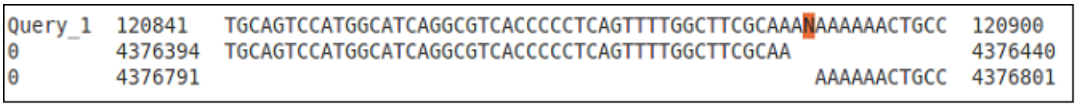


FIGURA 23 EXEMPLO DE ALINHAMENTO NEGATIVO DE CONTIGS NO *SCAFFOLD* PARA FECHAMENTO DE FALHAS

A abordagem tradicional para eliminar falhas não mostrou bons resultados, pois na grande maioria dos casos não houve cobertura na região de bases indefinidas, demonstrada pela letra N em laranja na *query.*

Diante destes resultados optou-se por testar uma estratégia alternativa que envolve um conjunto de *scripts* no Matlab, onde cada falha é encontrada automaticamente, estudada individualmente e alinhada com o banco de dados de contigs. A grande vantagem deste método é a possibilidade de plotagem dos dois melhores resultados em um gráfico, de maneira similar ao dotplot, o que facilita bastante a sua visualização (figura 24). Outro ponto positivo é a busca automática de falhas nas quais há sobreposição do contig em ambos os lados da mesma, facilitando o processo de fechamento, além de torná-lo mais rápido. Segundo esta análise, seria possível fechar 3.70 falhas que correspondem a cerca de 7,74% do total.

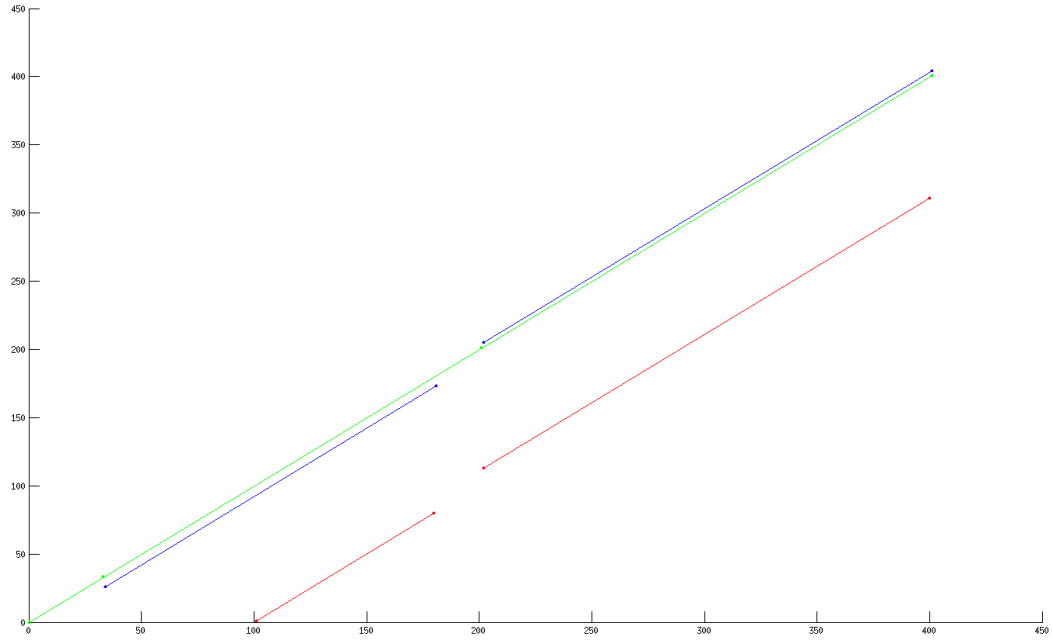


FIGURA 24 ALINHAMENTO POSITIVO PELA ESTRATÉGIA ALTERNATIVA VIA MATLAB

Em verde está representada a *query*, sendo o ponto a falha no interior do *scaffold.* Em azul e em vermelho estão demonstrados o primeiro e o segundo melhores alinhamentos de um contig do banco de dados, respectivamente. É observado que ambos os contigs podem ser utilizados para fechar a falha já que apresentaram alinhamentos nos dois lados da falha.

3.4 Anotação Parcial

Após a anotação automática no HGF e SILA, o genoma do *H. hiltneri* com seus possíveis genes foi analisado no Artemis (figura 25). Foram encontradas 6.904 *orfs*, sendo que 3.818 apresentaram resultado significativo no alinhamento do Blast. O maior número de genes alinhados foi obtido com o *H. seropedicae*, o único representante do gênero com o qual foi comparado. Outras bactérias apareceram com frequência, entre elas *Collimonas sp., Herminiimonas sp.* e a *Burkholderia sp*., sugerindo proximidade filogenética entre elas.

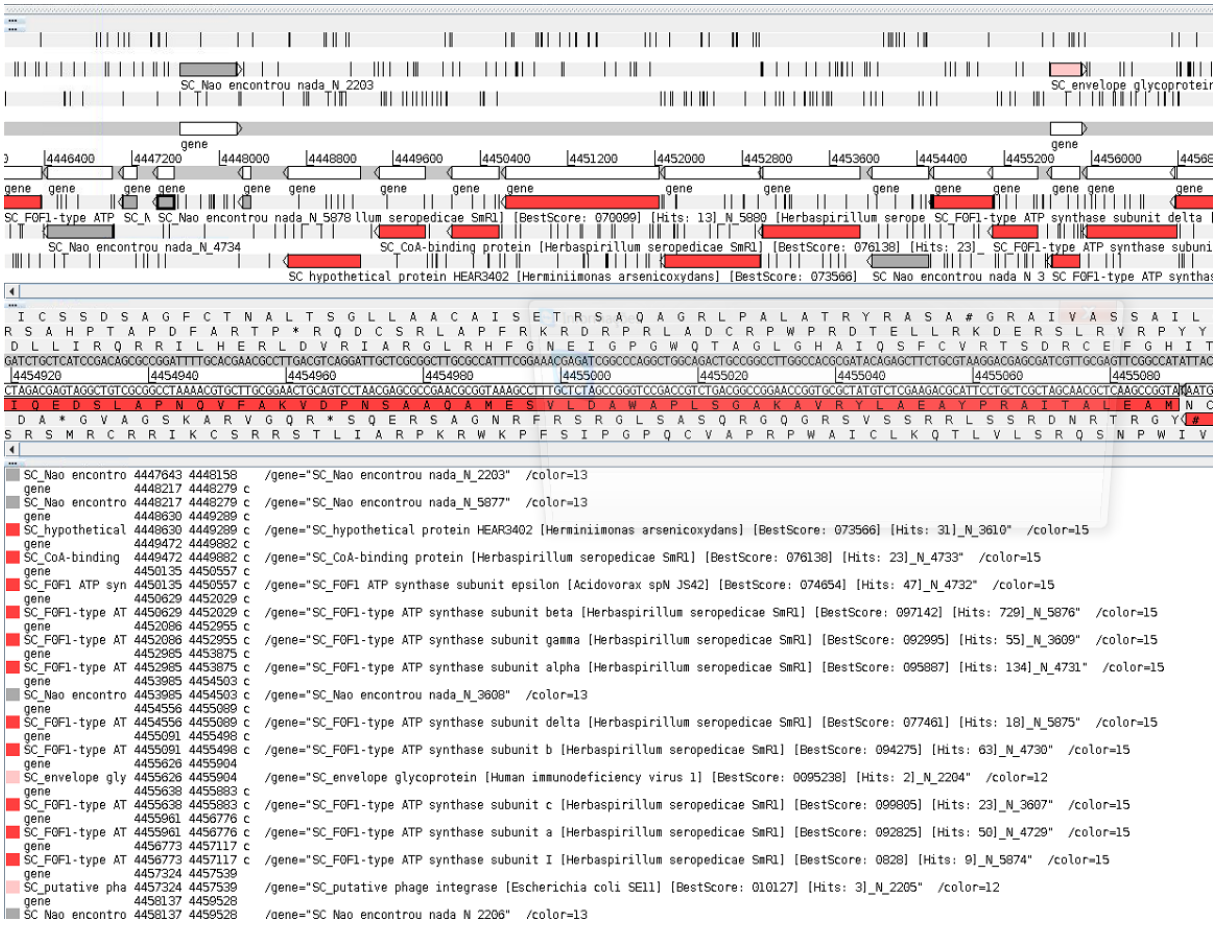


FIGURA 25 VISUALIZAÇÃO DA ANOTAÇÃO DO GENOMA DO *H. hiltneri* NO ARTEMIS.

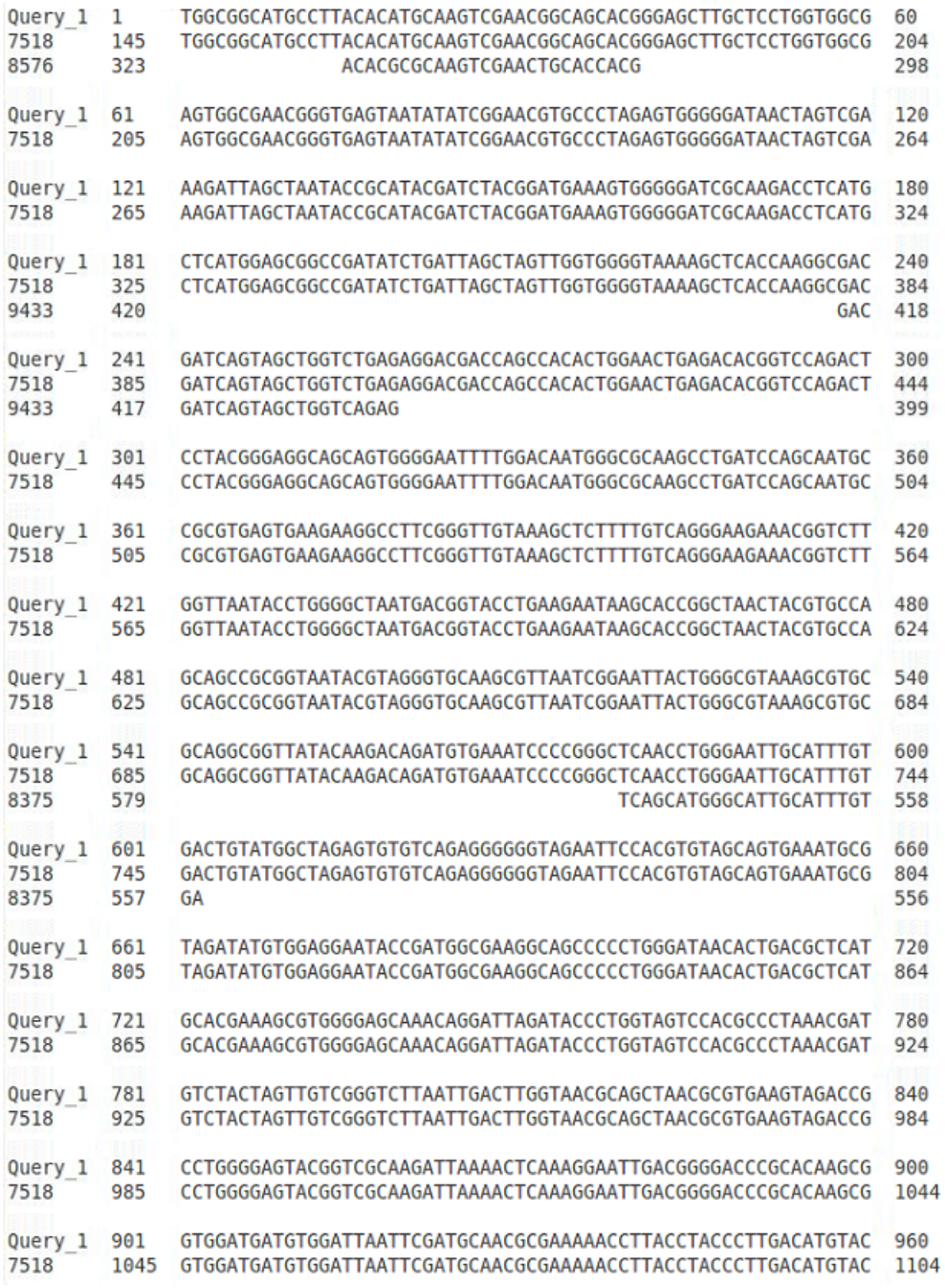
É possível visualizar os possíveis genes já alinhados e identificados de acordo com o *score* no Blast. Muitas *orfs* encontram sua similaridade com o *H. seropedicae* SmR1 em vermelho. Outras aparentam ser regiões não condificadoras, como está demonstrado em cinza na figura e com a frase “Não encontrou nada”.

3.5 Busca de Genes

A busca de genes de interesse foi realizada via Blast local. O critério para a confirmação da presença ou não de determinado gene foi de acordo com o *bitscore* obtido.

3.5.1 Genes ribossomais

Os genes *rRNA16S* (figura 26) e *23S* do *Herbaspirillum hiltneri* N3^{T} foram encontrados com êxito na montagem, ambos com 100% de identidade, confirmando a espécie e estirpe do genoma estudado. Uma análise comparativa do gene *16S rRNA* foi realizada com os organismos mais próximos filogeneticamente (tabela 7)



```
Query_1  1     TGGCGGCATGCCTTACACATGCAAGTCGAACGGCAGCACGGGAGCTTGCTCCTGGTGGCG  60
7518     145   TGGCGGCATGCCTTACACATGCAAGTCGAACGGCAGCACGGGAGCTTGCTCCTGGTGGCG  204
8576     323                 ACACGCGCAAGTCGAACTGCACCACG                    298

Query_1  61    AGTGGCGAACGGGTGAGTAATATATCGGAACGTGCCCTAGAGTGGGGGATAACTAGTCGA  120
7518     205   AGTGGCGAACGGGTGAGTAATATATCGGAACGTGCCCTAGAGTGGGGGATAACTAGTCGA  264

Query_1  121   AAGATTAGCTAATACCGCATACGATCTACGGATGAAAGTGGGGGATCGCAAGACCTCATG  180
7518     265   AAGATTAGCTAATACCGCATACGATCTACGGATGAAAGTGGGGGATCGCAAGACCTCATG  324

Query_1  181   CTCATGGAGCGGCCGATATCTGATTAGCTAGTTGGTGGGGTAAAAGCTCACCAAGGCGAC  240
7518     325   CTCATGGAGCGGCCGATATCTGATTAGCTAGTTGGTGGGGTAAAAGCTCACCAAGGCGAC  384
9433     420                                                        GAC  418

Query_1  241   GATCAGTAGCTGGTCTGAGAGGACGACCAGCCACACTGGAACTGAGACACGGTCCAGACT  300
7518     385   GATCAGTAGCTGGTCTGAGAGGACGACCAGCCACACTGGAACTGAGACACGGTCCAGACT  444
9433     417   GATCAGTAGCTGGTCAGAG                                          399

Query_1  301   CCTACGGGAGGCAGCAGTGGGGAATTTTGGACAATGGGCGCAAGCCTGATCCAGCAATGC  360
7518     445   CCTACGGGAGGCAGCAGTGGGGAATTTTGGACAATGGGCGCAAGCCTGATCCAGCAATGC  504

Query_1  361   CGCGTGAGTGAAGAAGGCCTTCGGGTTGTAAAGCTCTTTTGTCAGGGAAGAAACGGTCTT  420
7518     505   CGCGTGAGTGAAGAAGGCCTTCGGGTTGTAAAGCTCTTTTGTCAGGGAAGAAACGGTCTT  564

Query_1  421   GGTTAATACCTGGGGCTAATGACGGTACCTGAAGAATAAGCACCGGCTAACTACGTGCCA  480
7518     565   GGTTAATACCTGGGGCTAATGACGGTACCTGAAGAATAAGCACCGGCTAACTACGTGCCA  624

Query_1  481   GCAGCCGCGGTAATACGTAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGC  540
7518     625   GCAGCCGCGGTAATACGTAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGC  684

Query_1  541   GCAGGCGGTTATACAAGACAGATGTGAAATCCCCGGGCTCAACCTGGGAATTGCATTTGT  600
7518     685   GCAGGCGGTTATACAAGACAGATGTGAAATCCCCGGGCTCAACCTGGGAATTGCATTTGT  744
8375     579                                     TCAGCATGGGCATTGCATTTGT  558

Query_1  601   GACTGTATGGCTAGAGTGTGTCAGAGGGGGGTAGAATTCCACGTGTAGCAGTGAAATGCG  660
7518     745   GACTGTATGGCTAGAGTGTGTCAGAGGGGGGTAGAATTCCACGTGTAGCAGTGAAATGCG  804
8375     557   GA                                                           556

Query_1  661   TAGATATGTGGAGGAATACCGATGGCGAAGGCAGCCCCCTGGGATAACACTGACGCTCAT  720
7518     805   TAGATATGTGGAGGAATACCGATGGCGAAGGCAGCCCCCTGGGATAACACTGACGCTCAT  864

Query_1  721   GCACGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCCTAAACGAT  780
7518     865   GCACGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCCTAAACGAT  924

Query_1  781   GTCTACTAGTTGTCGGGTCTTAATTGACTTGGTAACGCAGCTAACGCGTGAAGTAGACCG  840
7518     925   GTCTACTAGTTGTCGGGTCTTAATTGACTTGGTAACGCAGCTAACGCGTGAAGTAGACCG  984

Query_1  841   CCTGGGGAGTACGGTCGCAAGATTAAAACTCAAAGGAATTGACGGGGACCCGCACAAGCG  900
7518     985   CCTGGGGAGTACGGTCGCAAGATTAAAACTCAAAGGAATTGACGGGGACCCGCACAAGCG  1044

Query_1  901   GTGGATGATGTGGATTAATTCGATGCAACGCGAAAAACCTTACCTACCCTTGACATGTAC  960
7518     1045  GTGGATGATGTGGATTAATTCGATGCAACGCGAAAAACCTTACCTACCCTTGACATGTAC  1104
```

FIGURA 26 ALINHAMENTO COM 100% DE IDENTIDADE COM O GENE *rRNA 16S* DO *H. hiltneri* N3$^{T.}$

Exemplo de alinhamento positivo utilizando Blast local. O gene *16S rRNA* foi encontrado no contig de número 7.518.

TABELA 7 ANÁLISE DE SIMILARIDADE DO GENE *rRNA 16S* DO *H. hiltneri* COM OUTROS MICRORGANISMOS

Bactéria	Identidade
H. hiltneri	100 %
H. lusitanum	99 %
H. rhizosphaerae	98 %
H. autotrophicum	98 %
H seropedicae	98 %
H. firingense	97 %
H. chlorophenolicum	97%
H. huttiense	97 %
H. rubrisubalbicans	97 %
Hermiimonas saxobsidens	97 %
H. puttei	97 %
Hermiimonas arenicoxydans	97 %
Collimonas fungivorans	96%
Burkholderia megapolitana	92%

3.5.2 Genes do metabolismo do nitrogênio

Os genes *nifH e nifD* não foram encontrados no genoma do *H. hiltneri* e este resultado foi concordante com o estudo de ROTHBALLER et al., 2010. Isso sugere que o microrganismo em questão não seja um diazotrofo, uma vez que estes genes, juntamente com o gene *nifK*, codificam as proteínas estruturais da enzima nitrogenase. Na figura 27 está apresentado o resultado da busca do gene *nifK* utilizando o blast local. O seu *bitscore* ficou em 44, um valor pouco expressivo e resultante de pequenos alinhamentos locais e aleatórios, o que levou a conclusão da ausência desta sequência no genoma.

Outros genes relacionados à fixação do nitrogênio foram procurados, como o *nif*A, *nifH, nif*B, *fix*C e *fix*X, mas não foram encontrados. No entanto, os genes *gln*A e *ntr*Y, *ntr*C, *ntr*B, também envolvidos no metabolismo do nitrogênio e na regulação de expressão gênica, foram identificados com êxito e com alto índice de alinhamento, ou seja, com *bitscores* maiores que 700 (tabela 8). O *ntrX* apresentou índice de alinhamento intermediário, mas alto o suficiente para comprovar a sua presença no genoma.

FIGURA 27 ALINHAMENTO PARA BUSCA DO GENE *nifK* NO GENOMA DO *H. hiltneri*

O tamanho das sequências alinhadas foi muito baixo, não passando de 32 bases. O *bit score* ficou em 44.6, valor considerado baixo e pouco representativo.

TABELA 8 CLASSIFICAÇÃO DOS GENES RELACIONADOS AO METABOLISMO DO NITROGÊNIO DE ACORDO COM O *BIT SCORE*

0 - 200	**200 – 700**	**> 700**
nifH	*ntrX*	*ntrC*
nifD		*glnA*
nifK		*ntrB*
nifA		*ntrY*
nifB		
fixC		
fixX		
narX		

3.5.3 Genes de recombinação e reparo do DNA

Os genes de recombinação e reparo do DNA foram escolhidos para serem buscados pelo fato de consistirem em genes *housekeeping* (de manutenção), pois, diferentemente do cluster *nif*, a célula não sobrevive sem este complexo sistema de reparo de DNA. Assim, como estes genes devem obrigatoriamente estar presentes dentro do genoma da bactéria, eles foram utilizados para validar a estratégia de busca de genes para confirmação do conteúdo biológico da montagem.

Todos os genes envolvidos no processo de recombinação e reparo do DNA que foram procurados foram encontrados com êxito. Entre eles, os genes *recB, recJ, recG, recQ, mutM, mutS, uvrA e uvrB* apresentaram alto índice de alinhamento (*bit score* acima de 700) (tabela 9). Já os genes *recA, recN, recO, recR lexA, mutL, mutT, mutY e recX* apresentaram resultados considerados intermediários no Blast local (tabela 9). Isso pode ser decorrente tanto do alto número de contigs como de varaiações naturais na sequência nucleotídica, uma vez que os genes buscados não são da mesma espécie.

TABELA 9 CLASSIFICAÇÃO DOS GENES DE RECOMBINAÇÃO DE ACORDO COM O *BIT SCORE*

0 - 200	**200 – 700**	**> 700**
-	*recA*	*recB*
	lexA	*recJ*
	recX	*recQ*
	recR	*recG*
	recN	*mutS*
	recO	*mutM*
	mutL	*uvrA*
	mutY	*uvrB*
	mutT	

Na figura 28 está mostrado o alinhamento positivo do gene *uvr*A, onde é observado que este está distribuído em dois contigs distintos com uma região sobreposta. Esse tipo de comportamento ocorreu com frequência e resulta na queda do *bit score,* pois o seu cálculo é baseado em cada alinhamento local de cada contig. Estes resultados positivos comprovam que este procedimento de busca de genes pode ser utilizado tanto para validação da montagem e confirmação do conteúdo biológico como para a união de contigs e/ou *scaffolds.* Nenhum gene deste grupo foi considerado ausente, ou seja, com *bit score* inferior a 200.

```
Query_1  1     ATGGAAGAAATTCGCATTCGCGGCGCGCGCACGCATAACCTCAAGAATATCAACCTAGAC  60
5950     911   ATGGAAGAAATTCGCATTCGCGGCGCACGCACGCACAATCTGAAGAACATCAGCCTCGAC  970

Query_1  61    CTCCCCCGCAACAAGCTCATCGTCATCACCGGGCTGTCCGGTTCGGGCAAGTCATCGCTG  120
5950     971   TTACCCCGCAACAAACTGATCGTGATTACCGGCCTGTCCGGCTCGGGCAAGTCGTCGCTG  1030

Query_1  121   GCCTTCGATACGCTCTATGCCGAGGGTCAGCGCCGCTACGTGGAGTCGCTGTCGGCCTAT  180
5950     1031  GCATTCGACACTCTGTATGCAGAAGGCCAGCGCCGCTATGTCGAGTCGCTGTCATCCTAT  1090

Query_1  181   GCGCGCCAGTTCCTGCAACTGATGGAAAAGCCCGATGTGGACATGATCGAAGGCCTGTCC  240
5950     1091  GCCCGCCAGTTCCTGCAACTGATGGAAAAACCCGACGTCGACATGATCGAAGGCCTGTCG  1150

Query_1  241   CCGGCGATTTCCATCGAGCAGAAGGCGACCTCGCACAACCCGCGCTCCACCGTGGGCACG  300
5950     1151  CCGGCGATCTCGATCGAGCAGAAGGCGACCTCGCACAATCCGCGTTCGACCGTCGGCACC  1210

Query_1  301   GTCACCGAGATCCACGACTACCTGCGCCTGCTGTACGCCCGCGTCGGCACGCCCTACTGC  360
5950     1211  GTCACGGAAATTCACGACTATCTGCGCCTGCTGTACGCCCGCGTCCGC              1258
2668     1                          TATCTGCGCCTGCTGTACGCCCGCGTCGGTACGCCTTATTGC  42

Query_1  361   CCCGACCACCCCGAACATCCGCTGGAGGCCCAATCGGTCTCGCAGATGGTCGATGCCGTG  420
2668     43    CCCGACCATCCGGAAAATCCGCTGGCGGCGCAATCGGTGTCGCAGATGGTCGACGCCGTG  102

Query_1  421   CTGGCGCTGCCGGAAGACACCAAGCTGATGATCATGGCGCCGGTGGTAGCCAACCGCAAG  480
2668     103   CTGGCCATGCCGGAAGACACCAAGCTGATGATCCTGGCGCCGGTTGTCGCCAACCGCAAG  162

Query_1  481   GGCGAGCATGCCGACCTGTTCGAAGAGATGCAGGCCCAGGGCTTCGTGCGCTTCCGCATC  540
2668     163   GGCGAACACGTCGACCTGTTCGAGCAGATGCAGGCGCACGGCTTCGTGCGTTTCCGCATC  222

Query_1  541   CAGAGCGGCACCGGCACGGCCAAGGTGTATGAAGTCGATGACCTGCCCAAGCTCAAGAAG  600
2668     223   CAGAGCGGCACCGGCACGGCAAAAGTCTACGAGATCGATGACCTGCCGAAGCTCAAGAAA  282
```

FIGURA 28: ALINHAMENTO DE DOIS CONTIGS DIFERENTES COM O GENE *uvrA*

A primeira base do gene está no final do contig 5950 enquanto a porção final do gene se encontra no início do contig 2668. É observada uma região de sobreposição que possibilida a união destes contigs. Após a união, o *score* relativo a esse contig aumenta.

4 CONCLUSÃO

A melhor montagem do genoma da bactéria *Herbaspirillum hiltneri* N3^T em *mate-paired* foi a MP3, na qual foi realizada a poda nas leituras em 30 pb e *k-mer* de 21 e resultou em 8.307 *scaffolds* e 13.104 contigs. Já as leituras em *fragments*, que não foram podadas, produziram 11.369 contigs em MF3.

Uma montagem final híbrida do genoma de *H. hiltneri* N3^T realizada a partir de leituras curtas, provenientes da plataforma SOLiD^{TM} *Sequencing System,* apresentou 283 *scaffolds* e 5.970 contigs, N50 de 19.633 pb e tamanho genômico de aproximadamente 5,04 Mpb. Apesar alta fragmentação do genoma, as sequências puderam ser ordenadas com base no genoma de referência da bactéria *Herbaspirillum seropedicae* SmR1, e através dos gráficos de dotplot e GCskew foi possível concluir que este resultado é coerente. A anotação parcial resultou em 3.818 *orfs* contendo alinhamentos positivos com genes do banco de NRs via Blast local.

A montagem final apresentou 4.778 falhas. A aplicação de uma estratégia alternativa para o fechamento de falhas via conjunto de funções no Matlab se mostrou bastante promissora, visto que esta poderá facilitar o processo de união de contigs através da automatização do processo.

A busca de genes de envolvidos no metabolismo de nitrogênio e na recombinação homóloga foi um ótimo parâmetro para análise da informação biológica da montagem. O estudo sugere que a bactéria *H. hiltneri* N3^T não seja um fixador de nitrogênio em função da ausência do operon *nif*HDK. Outros genes relacionados ao metabolismo do nitrogênio e à recombinação homóloga foram localizados com êxito, baseados em seus resultados de Blast local em *bit scores*. A confirmação da existência dos genes citados deverá ser feita por análise laboratorial.

REFERÊNCIAS

ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W., LIPMANL, D.J. Basic Alignment search tool. **J. Mol. Biol.** Vol 215, 3:403-10, 1990.

ANSORGE, W.J. Next-generation DNA sequencing techniques. **New Biotech.** Vol. 25, 4:195-203, 2009.

ARAKAWA, K., TOMITA, M. The GC Skew index: A measure of genomic compositional asymmetry and degree of replicational selection. **Evolutionary Bioinformatics**, 3:159-168, 2007.

AVTGES, P., SCOLNIK, P.A., HASELKORN, R. Genetic and physical map of the sctructural genes (*nifH,D,K)* coding for the nitrogenase complex of *Rhodopseudomonas capsulate.* **Journ. of Bacteriol.** Vol. 156, 1:251-256, 1983.

BAIROCH, A., APWEILER, R.The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. **Nucleic Acids Res.** 28(1):45-48, 2000.

BALDANI, J.I.; POT, B.; KIRCHHOF, G.; FALSEN, E.; BALDANI, V.L.D.; OLIVARES, F.L.; HOSTE, B.; KERSTERS, K.; HARTMANN, A.; GILLIS, M.; DÖBEREINER, J. Emended description of *Herbaspirillum*; inclusion of [*Pseudomonas*] *rubrisubalbicans*, a mild plant pathogen, as *Herbaspirrillum rubrisubalbicans* comb. nov.; and classification of a group of clinical isolates (EF Group 1) as species 3. **Int. J. Syst. Bacteriol**., v. 46, p. 802-810, 1996.

BALDANI, J.I.; BALDANI, V.L.D.; SELDIN, L.; DÖBEREINER, J. Characterization of *Herbaspirillum seropedicae* gen. Nov., sp. Nov., a Root- Associated Nitrogen-Fixing Bacterium, **Int. J. Syst Bacteriol.**, p.86-93, 1986.

BATZOGLOU, S.; JAFFE, D.B.; STANDEY, K.; BUTLER, J.; GNERRE, S.; MAUCELI, E.; BERGER, B.; MESINOV, J.P.; LANDER, E.S. ARACHNE: A Whole-Genome Shotgun Assembler. **Genome Research**, v. 12, p. 177-189, 2002.

BENSON, D.A., KARSCH-MIZRACHI I., LIPMAN, D.J., OSTELL, J., RAPP, B.A. WHEELER, D.L. GenBank. **Nucleic Acids Res.** Vol. 28, 1:15-8. 2000.

BERRIMAN, M., RUTHFORD, K. Viewing and annotating sequence data with Artemis. **Brief. Bioinform.**, vol 4:124-132, 2003.

BRENDEL, V., BROCCHIERI, L., SANDLER, S. CLARK, A.J., KARLIN, S. Evolutionary comparisons of RecA-like proteins across al major kingdoms of living organisms. **J. Mol. Evol**, 44:528-541, 1997.

CARDOSO, R.L.A. Montagem genômica da bactéria endofítica diazotrófica *Herbaspirillum rubrisubalbicans* M4. Dissertação (Bioinformática). **Universidade Federal do Paraná**, 2011.

CARRO, L., RIVAS, R., LÉON-BARRIOS, M, GONZALEZ-TIRANTE, M., VELÁZQUES, E., VALVERDE, A. *Herbaspirillum canariense* sp.nov., *Herbaspirillum aurantiacum* sp.nov. and *Herbaspirillum soli* sp.nov., three new species isolated in Tenerife (Canary Islands). **Int. J. Syst. Bacteriol.** July 25, 2011.

CARVALHO, M.C.C., SILVA, D.C.G.S. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas. **Ciênc. Rur.** Vol. 40, 3:735-744, 2010.

CERDEIRA, L.T., CARNEIRO, A.R., RAMOS, R.T.J., ALMEIDA, S.S., D'AFONSECA, V., SCHNEIDER, M.P.C., J., BAUMBACH, J., TAUCH, A., McCULLOCH, J.A., AZEVEDO, V.A.C., SILVA., A. Rapid hybrid *de novo* assembly of a microbial genome using only short reads: *Corynebacterium pseudotuberculosis* 119 as case study. **Journ. of Microbiol. Methods.** 86:218-223, 2011.

CHAISSON, M.J., BRINZA, D., PEVZNER, P.A. De novo fragment assembly with short *mate-paired* reads. Does read length matter? **Gen. Res.** 19:336-346, 2009.

CHEN, H.M., HUANG, T.C., CHIEN, C.Y. Nucleotide sequence of *nif HDK* operon in the aerobic nitrogen-fixing unicellular *Synechococcus* RF-1. **Bot. bull. Acad. Sin.** 37:99-105, 1996.

CHOUNDHURY, A.T.M.A., KENNEDY, I.R. Prospects and potentials for systems of biological nitrogen fixation in sustainable rice production. **Biol Fertil Solis.** 39:219-227, 2004.

CHUBATSU, L.S., MONTEIRO, R.A., SOUZA, M.E., OLIVEIRA, M.A.S., YATES, M.G., WASSEM, R., BONATTO, A.N., HUERGO, L.F., STEFFENS, M.B.R., RIGO, L.U., PEDROSA, F.O. Fixation control in *Herbaspirillum seropedicae*. **Plant Soil**, May, 2011.

CRUZ, L. M.; SOUZA, E. M.; WEBER, O. B.; BALDANI, I. J.; DOBEREINER, J.; PEDROSA, F. O. 16S ribosomal DNA characterization of nitrogen-fixing bacteria isolated from banana (*Musa* spp.) and pineapple (*Ananas comosus* (L) Merril). **App. and Environ. Microbiol.**, v. 67, n. 5, p. 2375-2379, 2001.

DiGUISTINI, S., LIAO, N.Y., PLATT, D., ROBERTSON, G., SEIDEL, M., CHAN, K, S., DOCKING, T.R., BIROL, I., HOLT, R.A., HIRST, M., MARDIS, E., MARRA, M.A., HAMELIN, C.R., BOHLMANN, J., CREUIL, C., JONES, S.J.M. *De novo* genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. **Gen. Biol.** 20:R94, 2009.

DING, L.; YOKOTA, A. Proposals of *Curvibacter gracilis* gen. nov., sp. nov. and *Herbaspirillum putei* sp. nov. for bacterial strains isolated from well water and reclassification of [*Pseudomonas*] *huttiensis*, [*Pseudomonas*] *lanceolata*, [*Aquaspirillum*] *delicatum* and [*Aquaspirillum*] *autotrophicumas Herbaspirillum huttiense* comb. nov., *Curvibacter lanceolatus* comb. nov., *Curvibacter delicatus* comb. nov. **Int. J. Syst. Bacteriol**, 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., and description of *Herbaspirillum aquaticum* sp. nov. **Int. J. Syst. Evol. Microbiol**., v. 60, p. 1418-1426, 2010.

ELBELTAGY, A., NICHIOKA, K., SATO, T., SUZUKI, H., YE, B., HAMADA, T., ISAWA, T., MITSUI, H., MINAMISAWA, K. Endophytic colonization and in planta nitrogen fixation by a *Herbaspirillum* sp. Isolated from wild rice species. **Appl. Environ. Microbiol.** Vol. 67, 11:5285-5293, 2001.

FLEISCHMANN, R.D.; ADAMS M.D.; WHITE, O.; CLAYTON, R.A.; KIRKNESS, E.F.; KERLAVAGE, A.R.; BULT, C.J.; TOMB, J.F.; DOUGHERTY, B.A.; MERRICK, J.M.; *et al*. Whole-genome random sequencing and assembly of *Haemophilus influenza* Rd. **Science**, v. 269, n. 5223 p. 496-512,1995.

GALVÃO, C. W. Caracterização bioquímica das proteínas RecA e RecX de *Herbaspirillum seropedicae*. Tese (Bioquímica), **Universidade Federal do Paraná**, 2005.

GIBBS, A.L, McINTYRE, G.A. The diagram, a method for comparing sequences and its use with amino acid and nucleotide sequences. **Eur. J. Biochem.** 16:1-11, 1970.

GREEN, P. Phrap. <http://phrap.org/>, 1999.

GREGORIEV, A. Analyzing genomes with cumulative skew diagrams. **Nucleic Acids Res.**, Vol 26: 2286-2290, 1998.

GUIZELINI, Dieval ; PEDROSA, F. O. ; TIBAES, J. H. ; MARCHAUKOSKI, J. ; STEFFENS, M. B. R. ; SOUZA, E. M. de ; SOUZA, V. ; RAITTZ, R. T. . jContigSort: a new computer application for contigs ordering. In**: 7th International Conference of the Brazilian Association for Bioinformatics and Compuational Biology**, Abstract book, 2011.

HENNECKE, H. Nitrogen fixation genes involved in the *Bradyrhizobium japonicum* – soybeans symbiosis. **FEBS letters.** Vol. 268, 2:422-426, 1990.

HERNANDEZ, D. FRANÇOIS, P., FARINELLI, L., OSTERAS, M., SCHRENZEL, J. De novo bacterial genome sequencing: MIllions of very shor reads assembled on a desktop computer. **Genome Res.** 18:802-809, 2008.

HOGEWEG, P. The roots of bioinformatics in theoretical biology. **PLoS Comp. Biol.** Vol. 7, 3:e1002021, 2011

IM, W.; BAE, H.; YOKOTA, A.; LEE, S.T. *Herbaspirillum chlorophenolicum* sp. nov., a 4-chlorophenol-degrading bacterium. **Int. J. Syst. Bacteriol**, v. 54, p. 851-855, 2004.

HUANG, X.; WANG, J.; ALURU, S.; YANG, S.P.; HILLIER, L. PCAP: a whole-genome assembly program. **Genome Res**., v. 13, p. 2164–2170, 2003.

JANION, C. Inducible SOS response system of DNA repair and mutagenesis in *Escherichia coli.* **Int. J. Biol. Sci.** Vol. 4, 6:338-344, 2008.

JUNG, S.; LEE, M.; OH, T.; YOON, J. *Herbaspirillum rhizophaerae* sp. nov., isolated from rhizosphere soil of *Allium victorialis* var. *platyphyllum*. **Int. J. Syst. Bacteriol**, v. 57, p. 2284-2288, 2007.

KARLIN, S., WEINSTOCK, G.M., BRENDE, V. Bacterial classifications derived from recA protein sequence comparisons. **Journ. Bacteriol.** Vol. 177, 23:6881-6893, 1995.

KIRCHHOF, G.; ECKERT, B.; STOFFELS, M.; BALDANI, J.I.; REIS, V.M.; HARTMANN, A. *Herbaspirillum frisingense* sp. nov., a new nitrogen-fixing bacterial species that occurs in C4-fibre plants. **Int. J. Syst. Bacteriol**, v. 51, p. 157-168, 2001.

LEDERGERBER, C., DSEEIMOZ, C. Base-calling for next-generation sequencing platforms. **Brief. in Bioinform.**, jan, 2011.

LOBRY, J.R.; Genomic landscapes. **Microbiology today**, Vol 26: 164-165, 1999.

LUSCOMBE, N.M.; GREENBAUM, D.; GERSTEIN, M. What is bioinformatics? An introduction and overview. **Yearbook of Medical Informatics**, p. 83-100, 2001.

MARTINEZ, D.A., NELSON, M. A. The next generation becomes now generation. **PLoS Genet.** 6(4):e1000906, 2010.

McKENZIE, G., HARRIS, R.S., LEE, P.PL., ROSENBERG, S.M. The SOS response regulates adaptive mutation. **PNAS.** Vol. 97, 12:6646-6651, 2000.

MERRICK, M. J. Nitrogen control of the *nif* regulon in *Klebsiella pneumoniae* involvement of the *ntrA* gene and analogies between *ntrC* and *nifA*. **EMBO J**., v. 2, p. 39-44, 1983.

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

NELSON, D.L., COZ, M.M. Lehninger Princípios de Bioquímica. **Ed. Sarvier**. 3 ed, p. 856-862, 2002.

MYERS, E.W.; SUTTON, G.G.; DELCHER, A.L.; DEW, I.M.; FASULO, D.P.; *et al*. A Whole-Genome Assembly of *Drosophila*. **Science**, v. 287, p. 2196-2204, 2000.

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, 2011.

QU, W., HASHIMOTO, S., MOTISHITA, S. Efficiente frequency-based de novo short-read clustering for error trimming in next-generation sequencing. **Gen. Res.** 19:1309-1315, 2009.

RAMOS, R.T.J.; CARNEIRO, A.R.; BAUMBACH, J.; AZEVEDO,V.; SCHNEIDER, M.P.C.; SILVA, A. Analysis of quality raw data of second generation sequencers with Quality Assessment Software. **BMC Research Notes**, 4:130, 2011.

RAMMOS, R., CARNEIRO, A.R., AZEVEDO, V., SCHNEIDER, M.P.C. SILVA, A. Simplifier: the web tool to eliminate redundant NGS contigs. **Adv. Integr. OMICS Appl Biotechnol.** ISSN/ISBN: 20479174, 2012.

ROCHA, E. P.C., CORNET, E., MICHEL, B. Comparative and evolutionary analysis of the bacterial homologous recombination systems. **PLoS Genetics.** Vol. 1, 2:0247-0259, 2005.

RONCATO-MACARI, L.D.B.; RAMOS, H.J.O.; PEDROSA, F.O.; ALQUINI, Y.; YATES, M.G.; RIGO, L.U.; STEFFENS, M.B.R.; SOUZA, E.M. Endophytic *Herbaspirillum seropedicae* expresses *nif* gene in gramineous plants. FEMS **Microbiol. Ecol.** 45: 39-47, 2003.

ROTHBALLER, M.; SCHMID, M.; KLEIN, I.; GATTINGER, A.; GRUNDMANN, S.; HARTMANN, A. *Herbaspirillum hiltneri* sp. Nov., isolated from surfaced-sterilized wheat roots, **Int. J. Syst. Bacteriol**, Vol 56, 1341-1348, 2006.

RUTHEFORD, K., PARKHILL, J. CROOK, J., HORSNELL, T., RICE, P., RAJANDREAM, M.A., BARREL, B. Artemis: sequence visualization and annotation. **Bioinformatics**. Vol. 16 no. 10, 944-945, 2000.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors (DNA polymerase/nucleotide sequences/bacteriophage 4X174). **Biochemistry**, v. 74, n. 12, p. 5463-5467, 1977.

SASSON, A.; MICHAEL, T.P. Filtering error from SOLiD Output. **Bioinformatics**, Vol. 26 (6):849-850, 2010.

SCHIMID, M.; BALDANI, J.I.; HARTMANN, A. The Genus *Herbaspirillum.* **Prokaryotes** 5:141–150, 2006.

SHEN, Y., SUMMER, S., LIU, Y., HOBERT, O., PETER, I. Comparing platforms for *C. elegans* mutant identification using high-throuput whole-genome sequencing. **PLoS ONE.** Vol. 3, 12:e4012, 2008.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. **Nature Biotechnology**, Vol. 26, p. 1135-1145, 2008.

SIMPSON, J. T., WONG, K., JACKMAN, S. D., SCHEIN, J. E., JONES, S. J., BIROL, I. **ABySS: A parallel assembler for short read sequence data**. *Genome Research, 19*, 1117-1123, 2009.

SNUSTAD, D. P., SIMMONS, M. J., Fundamentos de Genética. **Ed. Guanabara Koogan. S.A.** 2° ed, 311-340, 2001.

STEFFENS, M. B. R.; RIGO, L. U.; FUNAYAMA, S.; SOUZA, E. M.; MACHADO, H. B.; PEDROSA, F. O. Cloning of a *recA*-like gene from the diazotroph *Herbaspirillum seropedicae* strain Z78. **Can. J. Microbiol**., 39, 1096-1102, 1993.

SUZUKI, S., ONO, N., FURUSAWA, C., YING, B.W., YOMO, T. Comparison of sequence reads obtained from three next-generating sequencing platforms. **PLoS ONE.** Vol. 6, 5:e19534, 2011.

STRÖMBERG, M., LEE, W.P. MOSAIK read alignment and assembly program. **<http://bioinformatics.bc.edu/marthlab/Mosaik>**, 2009.

TIEPPO, E. Montagem e análise preliminar do genoma de *Bradyrhizobium elkanii* 587 utilizando leituras curtas. Dissertação (Bioinformática). **Universidade Federal do Paraná**, 2011.

TULLI, R., FISCHER.R. HASELKORN, R. The *ntr* genes of *Escherichia coli* activate the *hut* and *nif* operons of *Klebsiella pneumonia*. **Gene.** Vol. 19, 1:109-116, 1982.

VALVERDE, A.; VELÁZQUEZ, E.; GUTIÉRREZ, C.; CERVANTES, E.; VENTOSA, A.; IGUAL, J. *Herbaspirillum lusitanum* sp. nov., a novel nitrogen-fixing bacterium associated with root nodules of *Phaseolus vulgaris*. **Int. J. Syst. Bacteriol**, v. 53, p. 1979-1983, 2003.

WITKIN, E.M. Ultraviolet mutagenesis and inducible DNA repairs in *Escherichia coli*. **Bacteriol. Rev.** Vol. 40, 4:869-907, 1976.

ZERBINO, D.R.; BIRNEY, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. **Genome res.**, 18: 821-829, 2008.

ZHANG, Y., BURRIS, R.H., LUDDEN, P.W., ROBERTS,G.P. Regulation of nitrogen fixation in *Azospirillum brasilense*. **FEMS Microbio. Lett.** 152(2):195-204, 1997.

ZIMMERMANN, J. VOSS, H., STEGEMANN, J. ANSORGE, W. Automated Sanger dideoxy sequencing reaction protocol. **FEBS Lett.** 20; 233(2):432-65, 1988.