UNIVERSIDADE FEDERAL DO PARANÁ

VANELY DE SOUZA

MONTAGEM DO DRAFT DO GENOMA DA BACTÉRIA *Herbaspirillum huttiense subsp. putei*

CURITIBA
2012

VANELY DE SOUZA

MONTAGEM DO DRAFT DO GENOMA DA BACTÉRIA *Herbaspirillum huttiense subsp. putei*

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

Orientador:

Roberto Tadeu Raittz, Dr.

Co-orientadora:

Leda Satie Chubatsu, Dra.

CURITIBA

2012

Dedico este trabalho a Deus que estava sempre comigo;

A meus pais exemplos de força e que foram à base da minha educação;

Aos meus avós por estarem sempre por perto;

A minhas irmãs por todo apoio e por toda minha admiração;

Ao Marcelo por todo carinho, atenção e incentivo;

Ao Mateus e ao Filipe por todos os momentos de distração.

AGRADECIMENTOS

A realização deste trabalho em muito se deve à colaboração e apoio de muitas pessoas, às quais transmito os mais sinceros agradecimentos.

Ao meu orientador, professor Dr. Roberto Tadeu Raittz pelos quatro anos em que faz parte da minha vida, por acreditar em mim e me mostrar o caminho da ciência, por todos os ensinamentos transmitidos e por toda paciência durante esse caminho.

A minha orientadora, professora Dra. Leda Satie Chubatsu por todo auxílio, explicações e sugestões no desenvolvimento do trabalho. Uma pessoa que passei a admirar ao longo desse projeto, pela atenção em cada dúvida que tive e por estar sempre pronta a passar seus conhecimentos.

Ao programa de Pós Graduação em Bioinformática pela oportunidade e estrutura cedida.

Ao Núcleo de Fixação de Nitrogênio – Departamento de Bioquímica e Biologia Molecular da UFPR e aos professores pelo envolvimento com o projeto e sugestões.

Ao INCT - Instituto Nacional de Ciência e Tecnologia de Fixação Biológica de Nitrogênio, na pessoa do professor Fábio de Oliveira Pedrosa.

Às professoras Jeroniza e Berenice pela dedicação em manter o programa e por sempre estarem disponíveis.

À CAPES e ao CNPQ pelo financiamento.

À Suzana, por toda sua atenção durante esses anos, estando sempre disposta a resolver nossos problemas, por sua amizade e paciência em ouvir todas as nossas histórias.

À Lea pela atenção em atender nossas dúvidas.

À Michelly pelo ínicio dessa caminhada.

Ao Eduardo e ao Rodrigo pelos ensinamentos.

Ao Vinicius por toda paciência com as inúmeras perguntas.

Ao Helisson por sempre ter prontamente uma resposta para qualquer pergunta.

Aos professores Lucas e Adriano pela companhia e ajuda no laboratório.

Ao professor Dieval por toda ajuda sempre desenvolvendo uma nova solução.

À Juliana, Angélica, Letícia e Aline pelos anos de graduação e amizade durante o mestrado.

Ao Sérgio por todos os momentos de distração e amizade durante esses dois anos.

Ao Lucas Ferreira pela companhia durante esse tempo.

Ao Ricardo por todo auxílio e disposição em ajudar.

Ao Alysson e Gustavo por todas as brincadeiras no laboratório, saídas e risadas. Foi muito bom ter vocês por perto.

À Paula e ao Leviston por dividirem comigo o desafio de montar esse grande quebra-cabeças.

À minha família por todos os momentos e por estarem sempre estar ao meu lado me apoiando. Amo muito todos vocês.

Ao meu namorado Marcelo, por estar sempre ao meu lado e por superar junto comigo os momentos difíceis. Por todo carinho e atenção quando mais precisei, apoiando e ajudando. Por todo amor e dedicação, que foram à força para chegar até o final. Te amo.

A Tamires, amiga de todas as horas e de muitos anos, obrigada por todo apoio e por todas as conversas.

À Deus, por tudo.

"Não é o que você faz, mas quanto amor você dedica no que faz que realmente importa."

Madre Tereza de Calcutá

RESUMO

O gênero Herbaspirillum foi descrito em 1986 e hoje é composto por 11 espécies, dentre elas a bactéria Herbaspirillum huttiense subsp. putei, isolada no Japão a partir de amostras de água de poço. Neste projeto obtemos uma seqüência parcial do genoma desta bactéria, com base em leituras curtas de dois seqüenciamentos SOLID, que juntos somam cerca de 205 milhões de leituras seqüenciadas. Como resultado da montagem de novo do primeiro conjunto de dados, obtivemos um total de 447 scaffolds e 2316 contigs que foram ordenados utilizando o programa jContigsort e representados num gráfico de dot plot. Os genomas de Herbaspirillum seropedicae e Herbaspirillum rubrisubalbicans foram usados como referência para os alinhamentos e ordenações. O fechamento de gaps foi inicialmente manual, baseado nessa atividade foram desenvolvidas funções em MATLAB para automatizar e facilitar esse processo. Os dados utilizados para o fechamento dos gaps foram obtidos através da montagem do segundo conjunto de dados gerados pelo sequenciamento SOLID e das montagens de suas duas tags separadamente. Depois das análises realizadas chegamos a um conjunto final com 37 scaffolds, que podem ser reduzidos a 33 se forem considerados os indícios de ligação presentes nas seqüências. O genoma da bactéria ficou com tamanho de aproximado a 5,7 Mb com conteúdo G+C de 62,3%. A partir do draft do genoma gerado, foi feita uma pré-anotação automática encontrando 5547 orfs utilizando o programa RAST e 6084 utilizando o HGF.

Palavras-chave : *Herbaspirillum*, *Herbaspirillum huttiense subsp. putei*, montagem genômica, SOLID, gaps.

ABSTRACT

The genus Herbaspirillum was described in 1966 and nowadays it is composed by 11 species, among them the bacteria Herbaspirillum huttiense subsp. putei, isolated in Japan from well water. This project has obtained a partial genome sequence of this bacteria, based on short reads from two SOLID runs, that resulted on 205 millions of sequenced reads. As a result of the de_novo assembling, we obtained an amount of 447 scaffolds and 2316 contigs, sorted using the JContigsort application and represented in a dot plot graph. The genomes of Herbaspirillum seropedicae and Herbaspirillum rubrisubalbicans were used as reference to alignments and sortings. The gap closing was initially manual and then, based on that activity some MATLAB functions were developed to automate and turn the process easier. The data used to close the gaps were obtained from the assembling of a second dataset, generated by SOLID sequencing and the assembling of its tags separately. After the analyzes, we reach a final dataset that contains 37 scaffolds, that might be reduced to 33 scaffolds considering the linkage between sequences. The final size of the bacteria's genome is nearly 5.7 Mb with G+C content of 62,3%. Based on the genome draft, we made an automatic pre-annotation where we can find 5547 orfs using software RAST and 6084 orfs using HGF.

Key-words: *Herbaspirillum*, *Herbaspirillum huttiense subsp. putei*, genomic assembler, SOLID.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 *Herbaspirillum*

O gênero *Herbaspirillum* foi descrito em 1986 por Baldani e colaboradores para classificar a espécie *Herbaspirillum seropedicae* (FIGURA 1). Algumas espécies descritas no gênero são fixadoras de nitrogênio e podem ser encontradas associadas a plantas de interesse comercial (SILVA, 2008).

O gênero é composto por bactérias Gram negativas pertencentes à classe das β-Proteobactérias. O gênero *Herbaspirillum* é composto atualmente de quatorze espécies: *H. seropedicae* (BALDANI et al., 1986); *H. rubrisulbalbicans* (BALDANI et al., 1996); *H. frisingense* (KIRCHHOF et al., 2001); *H. lusitanum* (VALVERDE et al., 2003); *H. chlorophenolicum* (IM et al., 2004); *H. huttiense subsp. huttiense*, (DING & YOKOTA, 2004); *H. hiltneri* (ROTHBALLER et al., 2006); *H. rhizosphaerae*(JUNG et al., 2007), *H. huttiense subsp. putei*, *H. autotrophicum* e *H. aquaticum* (DOBRITSA et al., 2010); *Herbaspirillum canariense*, *Herbaspirillum aurantiacum* e *Herbaspirillum soli* (CARRO et al. 2011).

Dentre essas, as espécies melhor estudadas são *H. seropedicae* (FIGURA 1) e *H. rubrisulbalbicans*, sendo que a primeira teve seu genoma seqüenciado (PEDROSA et. al, 2011), e está disponível para acesso no NCBI pelo número CP002039.

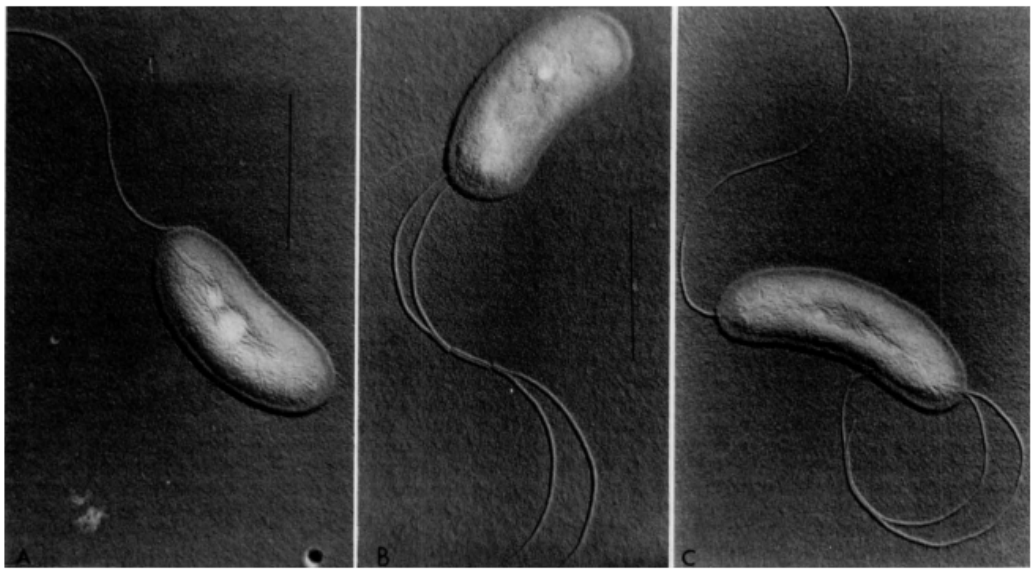


FIGURA 1 - CÉLULAS DE *Herbaspirillum seropedicae*, COM UM, DOIS OU TRÊS FLAGELOS EM UM OU AMBOS OS LADOS.
FONTE: BALDANI *et al*. (1986)

1.2 Herbaspirillum putei

A espécie *Herbaspirillum huttiense subsp. putei* (FIGURA 2) foi isolada a partir de amostras de água de poço em Osaka no Japão. Foi primeiramente classificada como *H. putei* (DING & YOKOTA, 2004), e posteriormente reclassificada como *H. huttiense subsp. putei* (DOBRITSA et al, 2010). É uma bactéria com conteúdo de G+C de 62,9% (DING & YOKOTA, 2004). A análise de seqüência do gene 16S rRNA indicou 99% de similaridade com *H. huttiense*, 98,1% com *H. rubrisulbalbicans* e 97,9% com *H. seropedicae*. Quanto a classificação taxonômica , *Herbaspirillum huttiense subsp. putei* pertence ao domínio Bactéria, filo Proteobacteria, classe Betaproteobacteria, ordem Burkholderiales, família Oxalobacteraceae, gênero *Herbaspirillum* e espécie *Herbaspirillum huttiense*.

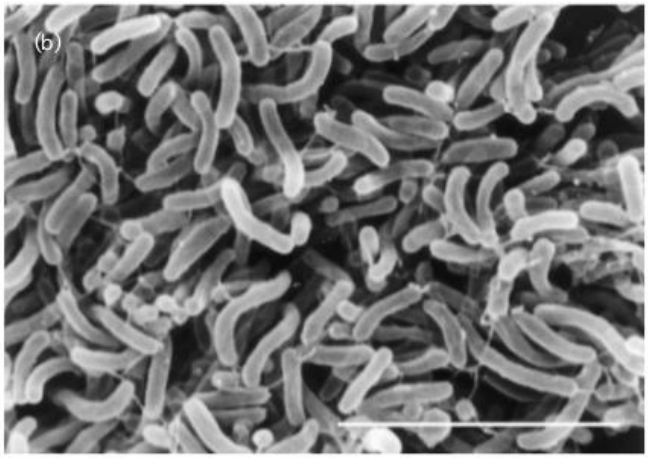


FIGURA 2 – MICROSCOPIA ELETRÔNICA DE CÉLULAS DE *Herbaspirillum huttiense subsp. putei*
FONTE: DING e YOKOTA *et al.* (2004)

Com base na árvore filogenética no rRNA 16S (DING E YOKOTA, 2004), podemos observar a FIGURA 3, que apresenta a relação de *Herbaspirillum huttiense subsp. putei* com *H. huttiense*, *H. rubrisulbalbicans* e *H. seropedicae*, organismos que estão mais próximos dessa espécie.

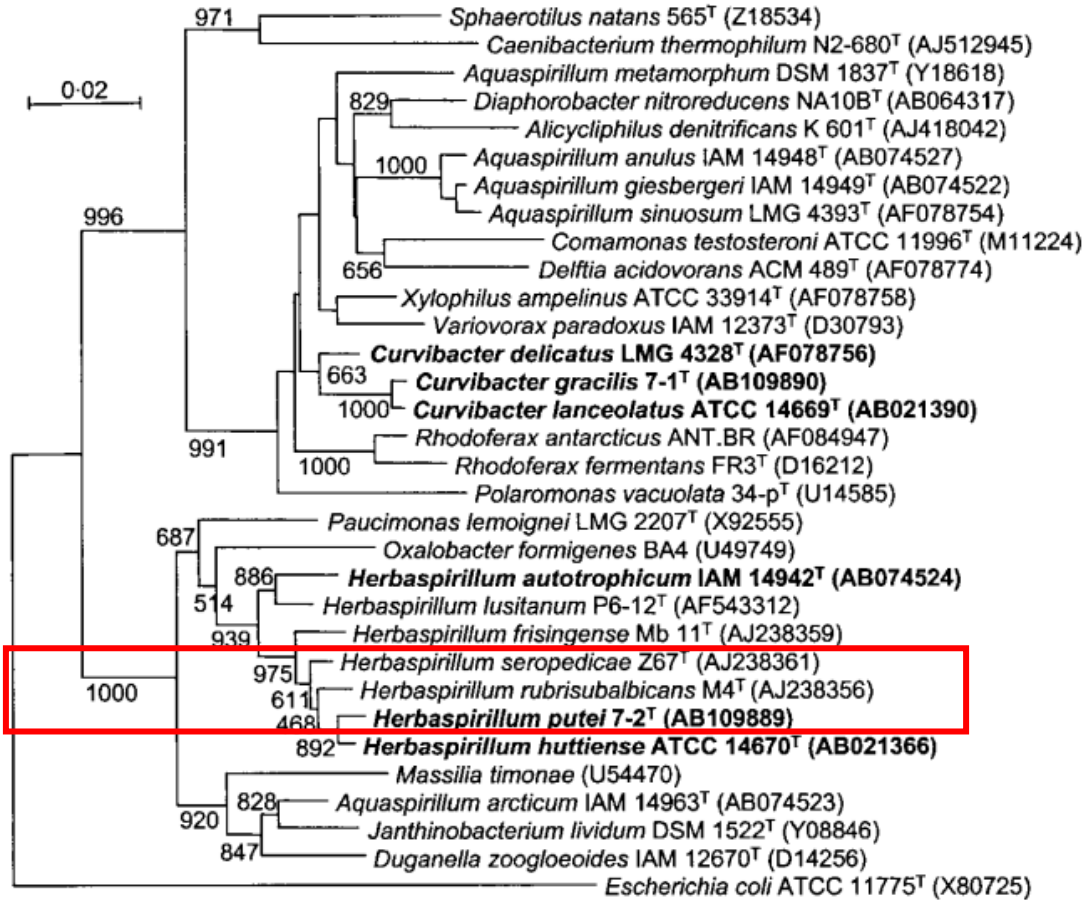


FIGURA 3 – ÁRVORE FILOGENÉTICA COM BASE EM SEQÜÊNCIAS DE GENES 16S rRNA.
FONTE: DING e YOKOTA (2004)

1.3 Sequenciamento de DNA

Na década de 1970, foram descritas as primeiras técnicas de seqüenciamento a de clivagem química (MAXAM et. al , 1977) e a de Sanger (SANGER, 1977) , que pode ser considerado um dos grandes feitos científicos (SOUZA e BRUSAMARELLO, 2009). Durante os anos novas técnicas foram surgindo como o pirosequenciamento da Roche/454, Illumina/Solexa e SOLID.

O primeiro genoma a ser seqüenciado foi o da bactéria *Haemophilus influenzae* em 1995 (FLEISCHMANN et al. 1995). Atualmente ajudado pelo enorme progresso na tecnologia, o seqüenciamento de genomas microbianos está avançando em um ritmo sempre crescente. (GALPERIN & KOONIN 2010). Com o seqüenciamento genômico, conseguimos obter de maneira ágil e com um menor custo informações importantes sobre um organismo.

Os procedimentos de seqüenciamento de DNA genômico envolvem a fragmentação e sequenciamento de milhares de pequenos segmentos de DNA que posteriormente deverão ser agrupado *in silico*, o que acaba gerando uma quantidade muito grande de dados. O principal desafio em seqüenciar genomas se encontra em realizar sua montagem.

Nos últimos anos as novas plataformas de seqüenciamento estão se tornando amplamente disponíveis, assim reduzindo o custo do seqüenciamento. O que antes era feito apenas em grandes centros agora pode ser feito por pesquisadores individualmente. Com a evolução de todas essas tecnologias o desafio a curto prazo é desenvolver protocolos robustos para a efetiva análise dos dados e também dar uma atenção especial a parte experimental. Essa nova geração de seqüenciamento vai acelerar as pesquisas biológicas, biomédicas e na parte da bioinformática, podendo assim analisar detalhadamente genomas e transcriptomas. (SHENDURE e JI, 2008).

Com o aumento de seqüências completas de genomas de bactérias, cresce também o número de genomas que vem sendo depositados nos bancos de dados públicos. Assim surgem muitos estudos para se determinar funcionalidades dessas estruturas genômicas. (HORIMOTO et al. 2001). Alguns dos *softwares* desenvolvidos para realizar a montagem e análise dos dados estão mal adaptados para trabalhar com as características específicas de dados de sequenciamento dessa nova geração. (NAGARAJAN et. al. 2010)

As informações obtidas a partir dos seqüenciamentos são os dados brutos para que a bioinformática realize seu trabalho, aliando a Ciência da Computação com a Biologia para assim chegar a um resultado final, um genoma completo com suas informações disponibilizadas.

1.3.1 Método de sequenciamento SOLID

O que diferencia o método de seqüenciamento SOLID (MCKERNAN et. al. 2006), dos demais métodos é que sua reação é catalisada por uma DNA ligase e não por uma polimerase. Nesse método os fragmentos de DNA são gerados e ligados aos adaptadores P1 e P2 que se ligam especificamente a uma microesfera.

O seqüenciamento ocorre por hibridização de sondas fluorescentes com alvo em cinco etapas diferentes (CARVALHO e SILVA, 2010).

Na primeira etapa, o primer (n) é utilizado, liberando as primeiras bases da sequência alvo para hibridização com a sonda. Uma das sondas do pool encontrará similaridade ao alvo ligando-se a ele. O sinal de fluorescência é lido, e as três últimas bases da sonda, incluindo o fluoróforo, são removidas. Inicia-se o segundo ciclo de hibridização e assim sucessivamente, até que o alvo seja todo coberto. A seqüência em fita dupla é desnaturada, e uma nova etapa de seqüenciamento é iniciada com o primer (n-1). Os ciclos de hibridização são repetidos, fornecendo informação de outras bases da seqüência alvo. Novas etapas de seqüenciamento com os primers (n-2), (n-3), e (n-4) são realizadas para que toda a seqüência alvo seja determinada. Todas as combinações possíveis de dinucleotídeos são marcadas nas sondas com apenas quatro fluoróforos. Assim, duas leituras são necessárias de cada base para que a sequência do dinucleotídeo da sonda seja resolvida. Esse processo inicia-se com a identificação da primeira base do alvo na segunda etapa de sequenciamento (primer n-1), que libera para hibridização com a sonda uma base já conhecida, a última base do adaptador (CARVALHO e SILVA, 2010).

O sitema SOLID apresenta alta precisão e velocidade, é uma plataforma de nova geração que permite conduzir múltiplos experimentos em uma única corrida, apresentando uma grande cobertura nas amostras.

1.3.2 Formato colorspace

Com a chegada da nova geração de seqüenciadores, o método SOLID, apresenta “reads” produzidos com maior cobertura, mas seu tamanho ainda fica pequeno, cerca de 50 pb. Para não apresentar problemas de precisão e altas taxas de erros nas seqüências, o sistema SOLID se baseia em ligações e apresenta grande qualidade nos reads, tendo como resultado de seu seqüenciamento o formato colorspace.

O sistema de colorspace apresenta 16 combinações de nucleotídeos que apresentam quatro fluoróforos. Para que se possa decodificar uma seqüência utilizamos uma matriz como a apresentada na FIGURA 4.

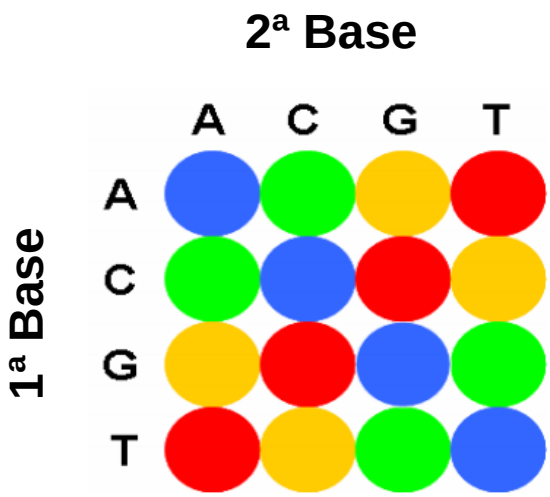
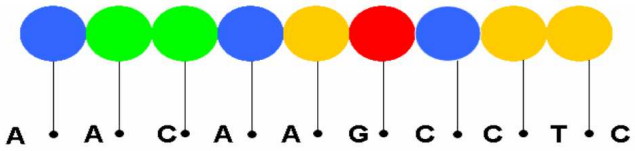


FIGURA 4 – MATRIZ PARA DECODIFICAR SEQÜÊNCIAS EM COLORSPACE.

PARA DECODIFICAR UMA SEQÜÊNCIA PODEMOS UTILIZAR A MATRIZ ACIMA SE CONHECERMOS PELO MENOS UMA DAS BASES. PODEMOS NOTAR QUE TANTO A -> T OU T -> A, QUANTO A COMPLEMENTAR DE UMA SEQÜÊNCIA, POR EXEMPLO A->G OU T -> C, APRESENTAM A MESMA COR.

FONTE: YUTAO et. al. (2012)

Com essa codificação o SOLID oferece uma melhor solução para se realizar a montagem completa de genomas. Para vizualizar a seqüência necessitamos utilizar a matriz apresentada anteriormente.



Se a primeira base apresentada é o A, utilizando a matriz localizamos a base e seguimos até a cor seguinte, apresentada no sinal de fluorescência que no exemplo é a cor azul assim sabemos que a próxima base é o A e assim sucessivamente até formar a seqüência final.

1.3.3 Formato base-space

O formato base-space é utilizado para que as seqüências de reads possam ser lidos pelo montador Velvet, as cores equivalem a: 0=A,1=C,2=G,3=T. Nesse formato as seqüências não tem nenhum significado biológico.

1.4 Montagem de Genoma

O avanço das tecnologias de seqüenciamento de nova geração ocasionou o desenvolvimento de novos métodos de montagem das seqüências do genoma, especialmente a montagem *de novo*. Devido ao pouco conhecimento da aplicação de todas essas ferramentas a escolha de um programa para montar genomas se torna uma tarefa difícil (ZHANG et. al. 2011).

Todos os diferentes procedimentos de seqüenciamento de DNA geram seqüências curtas quando comparadas ao genoma completo. Assim, a etapa conhecida como “montagem de genoma” envolve agrupar as diversas seqüências obtidas nas etapas de seqüenciamento numa determinada ordem que representa o genoma estudado. Desta forma, as seqüências obtidas na etapa de sequencimento de DNA, que envolvem experimentos em laboratório são chamadas de “reads” ou leituras (FIGURA 6). O número de nucleotídeos ou bases presentes nesses “reads”, podem variar em número de acordo com o método de seqüenciamento utilizado, por exemplo, cerca de 50pb para seqüenciamento do tipo SOLID e 400 pb para o sequenciamento 454. A FIGURA 5 apresenta um processo resumido de montagem de genoma.

Considerando que a quantidade de nucleotídeos deve ser superior várias vezes o tamanho previsto para o genoma, a dificuldade está em agrupar seqüências de 50 pb, por exemplo, na ordem adequada fornecendo um genoma da ordem de 5 Mb. Assim, claramente o procedimento de montagem do genoma deve ser automatizado com a utilização de programas de bioinformática adequados.

FIGURA 5 - REPRESENTAÇÃO RESUMIDA DA ORDEM DE MONTAGEM DO GENOMA ATÉ SUA FINALIZAÇÃO COM O PROCESSO DE ANOTAÇÃO AUTOMÁTICA DE SEQUÊNCIAS.

FONTE: A autora (2012)

Os programas que fazem a montagem do genoma variam de acordo com o tipo de dado utilizado, tamanho de “reads” aceitos, método utilizado para realizar a montagem, entre outros. Esses programas geram a partir dos *reads* sequências consenso que chamamos de *contigs* (FIGURA 6). Cada *contig* produzido pode ser ordenado e ter sua orientação encontrada ao longo de um cromossomo com informações adicionais, assim geramos a partir de vários *contigs* os *scaffolds* (FIGURA 4).

FIGURA 6 – VIZUALIZAÇÃO DOS CONCEITOS DE READS, CONTIGS E SCAFFOLDS EM SEQÜÊNCIAS DE GENOMAS.

FONTE: A autora (2012)

Para montar o genoma alguns dos programas “montadores” utilizam sobreposição de seqüências dos “reads”, assim, com todas as informações disponíveis ele tenta resolver este “quebra-cabeça” de seqüências. O objetivo final de todo projeto de genoma é identificar a ordem de cada nucleotídeo no conjunto original do genoma, assim, com toda essa complexidade de dados, é muito raro que um programa consiga reconstruir completamente o genoma em estudo.

Regiões com ausência de sobreposição são usualmente encontradas, essas regiões são chamadas de “GAP” e representam lacunas que são apresentadas ao

[1]final da montagem. A determinação da seqüência presente nesses gaps podem ser obtidas com experimentos de laboratório adicionais e com uma curadoria manual extensiva, assim podemos validar e corrigir a montagem final garantindo uma qualidade ao conjunto de genomas.[1]

Uma das abordagens para a montagem de genomas é a *de novo*, ou seja, partindo apenas dos “reads” sem que se tenha nenhum conhecimento da referência, busca-se completar a seqüência dos genomas. Utilizando a plataforma de seqüenciamento SOLID, o comprimento dos “reads” é pequeno e são produzidos em grande quantidade, assim apresentam uma grande cobertura em relação ao genoma. (ZERBINO e BIRNEY, 2008)

O desafio é fazer com que os algoritmos de montagem sejam otimizados para conseguir processar essa vasta quantidade de informações dos seqüenciamentos de nova geração que tem suas particularidades para processamento, pelo volume de dados e tipos de dados específicos a cada sequenciador. (CHAISSON e PEVZNER, 2008)

1.5 Anotação

O DNA armazena muitas informações biológicas dos organismos. A Bioinformática pode agregar novas informações, interpretações e análises para a anotação de um genoma.

Para que se tenha uma boa anotação, os bancos de dados biológicos têm um papel fundamental, pois neles se encontram depositadas inúmeras seqüências que podem servir como base para futuras anotações, através deles pode-se fazer comparações e buscar as similaridades das seqüências. Essa comparação de proteínas entre espécies é muito rica em se tratando de anotação funcional (LIBERMAN, 2004). A anotação envolve também formulação de hipóteses, testes, refinamentos e publicações (LIBERMAN, 2004; STEIN, 2001)

A anotação de um genoma pode ser dividida em três classes (FIGURA 7): anotação a nível de nucleotídeos, ou seja, buscamos a localização das seqüências e

[1] Disponível em: http://www.cbcb.umd.edu/research/assembly_primer.shtml. Acesso: 11/01/12

genes; anotação em nível de proteínas, onde queremos saber quais as funções dos genes presentes; e o nível de processo que busca saber quais as vias metabólicas esses genes estão presentes. (STEIN, 2001)

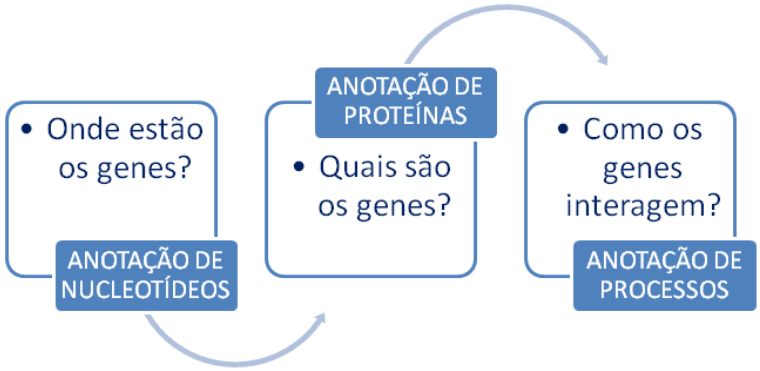


FIGURA 7 – A FIGURA APRESENTA O PROCESSO GERAL DE ANOTAÇÃO DE GENOMAS.
FONTE: Adaptação (STEIN 2001; TIEPPO, 2011)

Ao estudarmos um genoma a identificação de seus elementos é muito importante, assim podemos fazer a predição dos genes presentes e atribuir a eles suas funções biológicas e suas categorias funcionais. Para termos essas informações, dispomos de dois principais processos: anotações automáticas, onde as análises são feitas por meio de programas; e as anotações manuais, que passam pela curadoria envolvendo pesquisadores e toda sua experiência, para se alcançar os resultados.

É através da anotação que podemos disponibilizar para a comunidade científica, as informações para consultas de interesse para seus trabalhos. Assim, essa anotação passa a ser informativa sobre os processos de proteínas que já foram estudadas.

1.6 Objetivos

1.6.1 Objetivo Geral

Montar o draft da bactéria *Herbaspirillum huttiense subsp. putei* e realizar sua pré-anotação automática.

1.6.2 Objetivos específicos

- Utilizar para a montagem do genoma da bactéria *Herbaspirillum huttiense subsp. putei*, dados de seqüenciamento SOLID.
- Utilizar genomas de referência para ordenação dos resultados de contigs e scaffolds.
- Fechar as falhas de montagem utilizando dados de montagens auxiliares.
- Realizar a pré- anotação do draft da bactéria.

2 MATERIAIS E MÉTODOS

2.1 Conjunto de dados

O sequenciamento da bactéria *Herbaspirillum huttiense subsp. putei* foi realizado pelo Núcleo de Fixação de Nitrogênio do Departamento de Bioquímica e Biologia Molecular da UFPR, e cedido para este trabalho. Foram realizados dois seqüenciamentos, utilizando o método SOLID com bibliotecas de mate-paired, totalizando 205 119 350 reads seqüenciados em formato colorspace. Os “reads” são pareados e com tamanho de 50 pb. A TABELA 1 apresenta o detalhamento dos dois conjuntos de dados resultantes.

Com a finalização do genoma de *H. seropedicae* (PEDROSA et. al., 2011), várias bactérias, pertencentes a esse gênero vem sendo seqüenciadas neste mesmo departamento, pelo interesse no estudo desse gênero. Como são as primeiras a serem seqüenciadas no país, trazem com elas o desafio em utilizar esse novo método de seqüenciamento do tipo SOLID. Com a finalização dos genomas que vem sendo seqüenciados dentro do grupo, busca-se uma comparação mais abrangente dentro do gênero *Herbaspirillum*.

TABELA 1 – CARACTERÍSTICAS DOS CONJUNTOS DE DADOS DO SEQUENCIAMENTO SOLID

Características	**Corrida 05 _40pb**	**Corrida 1_50pb**
Tipo de leitura	SOLID	SOLID
Tipo de dado	Curto	Curto
Tamanho médio de leitura	50 pb	50 pb
Distância entre as extremidades pareadas	~ 1500	~ 1500
Total de leituras	102 768 904	102 350 446
Total de bases	5 138 445 200	5 117 522 300

FONTE: A autora (2012)

2.2 Alinhamento com o genoma de referência

Com o grande número de genomas depositados em bancos de dados públicos, podemos utilizar um genoma similar ao estudado para: auxiliar na montagem do genoma; realizar o alinhamento de seqüências e ordenar conjunto de dados baseando-se nessa referência.

Observando a proximidade das espécies de *Herbaspirillum seropedicae* e *Herbaspirillum rubrisubalbicans* com o genoma estudado, optamos pela escolha desses dois organismos para comparações em nosso estudo. O primeiro genoma está depositado no NCBI com número de acesso CP002039 e o segundo fez parte da pesquisa de mestrado do aluno Rodrigo Luis Alves Cardoso (CARDOSO, 2011), obtendo assim a seqüência do segundo genoma.

Utilizando o alinhamento realizado fizemos a ordenação dos dados utilizando o programa JContigSort (GUIZELINI et. al., 2011), que baseado no genoma de referência faz a distribuição e ordenação dos conjuntos de dados ao longo do genoma.

Podemos observar no fluxograma (FIGURA 8) abaixo os passos destes processos.

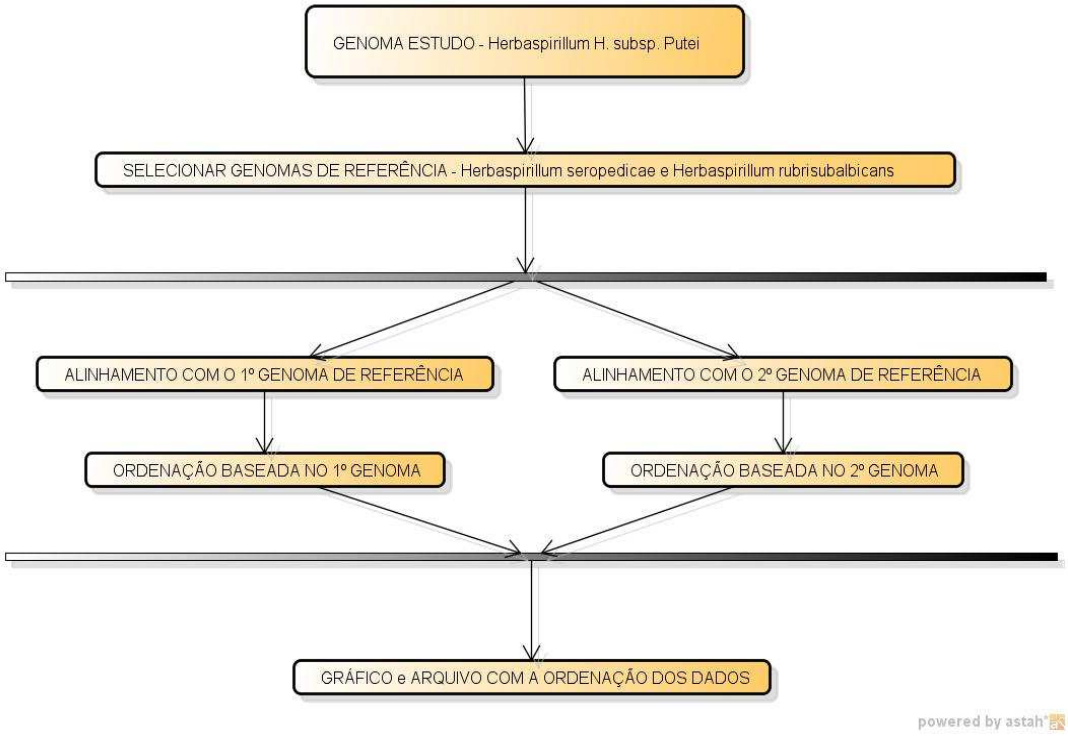


FIGURA 8 – FLUXOGRAMA DO PROCESSO DE ALINHAMENTO E ORDENAÇÃO DOS DADOS.
FONTE: A autora (2012)

2.3 Montagem automática do organismo

Para realizar a montagem automática do genoma, aconteceu um pré-processamento dos dados do conjunto de seqüenciamento corrida05_40pb, foi realizado um *trimming* nas bases finais dos "*reads*" desse conjunto, que em seguida foi submetido ao processo de montagem com a finalidade de chegar ao draft do genoma da bactéria *Hesbaspirillum huttiense subsp. putei*, para assim resultar em contigs e scaffolds finais da montagem.

2.3.1 Trimming de bases

O conjunto inicial de dados (corrida05_40pb) escolhido como principal foi analisado utilizando o programa Quality Assessment (RAMOS et. al, 2011) (FIGURA 9), que gera um gráfico da qualidade dos reads por posição de cada base. Com a utilização do programa o que se procurou foi eliminar as bases para que a montagem apresentasse uma melhor qualidade e com isso um menor número de contigs e scaffolds.

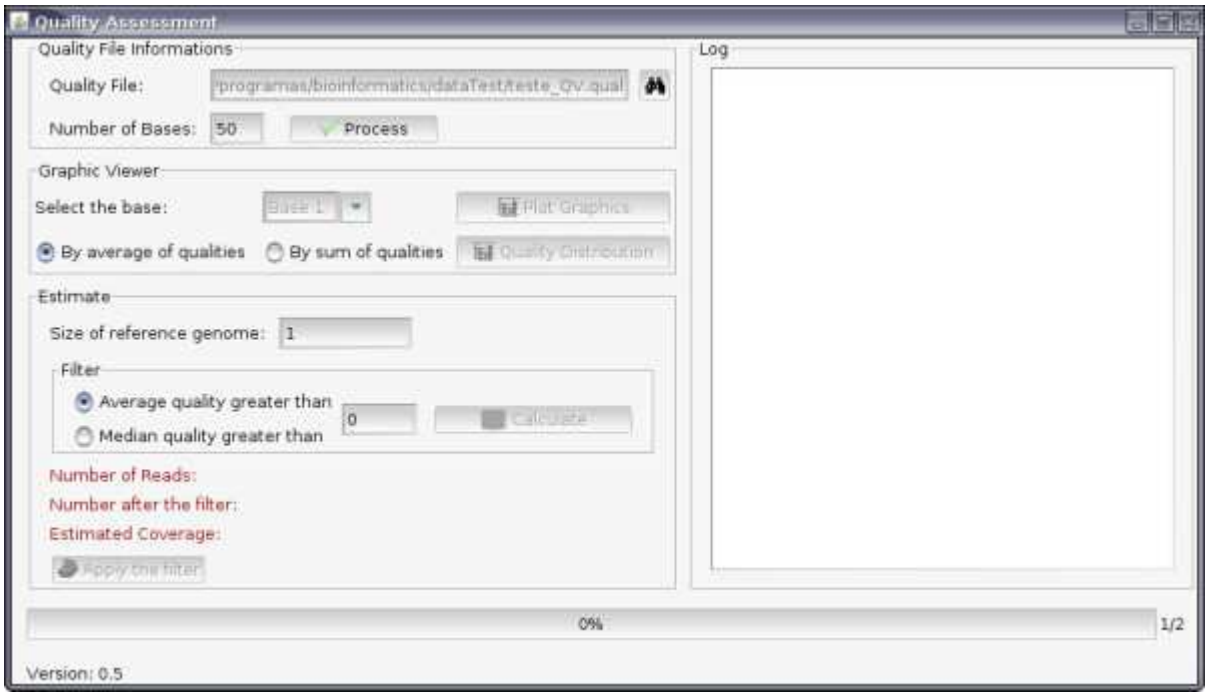


FIGURA 9 – TELA INICIAL DO PROGRAMA QUALITY ASSESSMENT.
FONTE: Quality assessment (RAMOS et. al. 2011)

O segundo conjunto de dados (corrida1_50pb) foi submetido a montagem sem nenhuma alteração, ou seja, ele complementou o conjunto que recebeu o *trimming* durante o processo de finalização da montagem.

2.3.2 Equipamento utilizado

Após finalizar a análise dos conjuntos de dados se fez necessário o início das montagens dos dados, foram desenvolvidas diferentes estratégias para utilizar os equipamentos disponíveis em cada fase do desenvolvimento do projeto.

A primeira estratégia foi a divisão dos dados em conjuntos menores para atender ao processamento disponível em servidores com 8 GB e 16 GB de memória RAM, pois o processamento não pode ser concluído com o total dos dados. Assim, esta estratégia foi descartada por não apresentar uma maneira eficiente em trabalhar com os dados do genoma. Para utilizar os dados sem dividi-los foi necessário a utilização de um cluster com 124 GB de RAM.

As montagens finais do genoma foram realizadas utilizando o cluster adquirido pelo Programa de Bioinformática da Universidade Federal do Paraná e as análises no computador disponibilizado ao projeto no próprio laboratório do programa. Com isso podendo obter os melhores resultados dentre as montagens. O cluster segue as especificações:

- Sistema operacional: GNU/Linux – Debian 6.0;
- Processador: 64 núcleos – Intel®Xeon® CPU E7 8837@267.Ghz.
- Memória RAM: 512 GB;
- HD: 200 GB – Storage 7.2 TB;
- Arquitetura: NUMA link 5.

2.3.3 Pipeline De novo

O pipeline *De novo* é um conjunto de softwares utilizados para montagens de genomas onde não se conhece a referência, ele faz a montagem dos “reads” e gera seqüências maiores como contigs e scaffolds, é geralmente utilizado para montagens de pequenos genomas, com até aproximadamente 30 Mb, trabalha com os “reads” gerados no sistema SOLID da Life technologies®. Ele faz uso da grande

quantidade de dados gerados pela plataforma SOLID e reconstrói o genoma baseado nos pequenos “reads”. O sistema de cores ajuda na correção de erros presentes nos “reads” para uma melhor montagem; o pipeline procura por regiões com cobertura não uniforme, e faz com que gere seqüências concenso altamente precisas.

Suporta o formato de mate-paired como o adotado no sequenciamento da bactéria estudada, *Herbaspirillum huttiense subsp. putei*. Com a correção de erros realizada pelo SAET, um dos programas desse *pipeline*, aumenta a média do comprimento dos contigs e do número de gaps fechados pelo ASiD, outro programa que compõe o *pipeline*.

Trabalha com a estimativa de parâmetros que melhor se encaixam ao seu conjunto de dados. Utiliza um programa montador de genomas acoplado, chamado de Velvet, que entra com sua versão 0.7.5.5, e é quem realmente fará a montagem do genoma. Otimiza o processo de montagem uma vez que busca de forma automática juntar programas que podem ser disparados por um único comando assim facilitando seu uso.

A FIGURA 10 apresenta um diagrama que mostra o fluxo básico do pipeline utilizado.

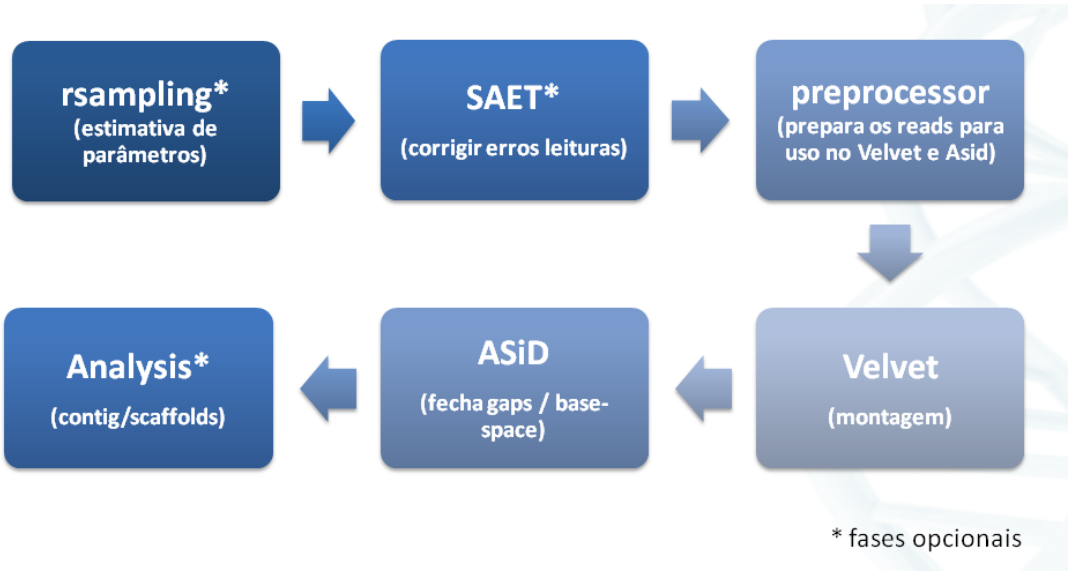


FIGURA 10 – DIAGRAMA QUE REPRESENTA O FLUXO DO PIPELINE.

O PIPELINE APRESENTADO SE DIVIDE EM 6 FASES DISTINTAS, CADA UMA DELAS SE REFERE A UM PROCESSO NOVO QUE OCORRE COM AS SEQÜÊNCIAS DE DNA COMO REPRESENTADO NA FIGURA.

FONTE: APPLIED BIOSYSTEMS (2012)

Segue uma breve descrição das fases do processo que envolve o pipeline:

- rsampling : onde ocorre a estimativa de alguns parâmetros como por exemplo: a estimativa precisa da cobertura por meio de sub-conjuntos que cria e avalia;
- SAET: utiliza a alta cobertura e os valores de qualidade precisos apresentados pelo SOLID e corrige as leituras em colorspace, em média tem uma boa redução da taxa de erro assim tornando o processo de montagem denovo mais preciso.
- preprocessor: prepara as leituras para serem utilizadas pelo montador Velvet, primeiro ele corta o primer das leituras, depois passa os números que representam o color space para pseudobases (0 = A, 1 = C, 2 = G e 3 = T), faz a orientação correta do fragmento.
- Velvet: nessa fase é onde realmente acontece a montagem do genoma, esse montador utiliza a abordagem de Brujin em cada nó, que representa a sobreposição dos k-mers. Oferece uma solução eficiente para a montagem de milhões de leituras compondo grandes contigs.
- ASiD: busca fechar os gaps entre os contigs primeiramente montados, quando obtém essa sobreposição, ele substitui os dois contigs por um outro mesclando os dois, e é nesse passo que são geradas as seqüências em base space;
- Analysis: realiza as análises sobre os dados gerados.

2.3.4 Velvet

O Velvet é um programa utilizado para montagem de genomas a partir de “reads” curtos produzidos pela nova geração de seqüenciadores. É muito empregado quando se estuda um novo organismo que ainda não tem um genoma de referência disponibilizado. O que diferencia esse programa é o gráfico de Brujin utilizado para a montagem, depois desse passo busca resolver as repetições encontradas. Ao final do processo gera uma saída com a montagem realizada e algumas estatísticas que são utilizadas para compreender melhor os dados. (ZERBINO, 2010)

Para cada conjunto de dados vamos encontrar os parâmetros que melhor se ajustam, alguns deles devem ser modificados logo ao se compilar o programa, diponibiliza ainda uma rotina para otimização de parâmetros. O Velvet trabalha com o tipo de dado colorspace, para a análise e montagem desse tipo de dado é preciso algumas configurações específicas, que vem pré-estabelecidas quando acoplado ao pipeline *De novo*. (ZERBINO, 2010)

Quando tratamos de processamento de dados do tipo colorspace estamos trabalhando com um formato muito específico de dados. São arquivos com propriedades muito diferentes por isso é recomendável utilizar esse processamento dentro do pipeline apresentado. (ZERBINO, 2010)

O processo básico de montagem do Velvet acontece em dois passos que envolvem os dois executáveis disponíveis dentro do pacote: o velveth e o velvetg, como o processamento utilizado envolve “reads” em *colorspace*, os executáveis são o velveth_de e o velvetg_de. O primeiro lê os arquivos com as seqüências e constrói uma espécie de dicionário com todas as palavras possíveis dentro do número informado como parâmetro, e vai definir os alinhamentos. Após esse passo, o velvetg_de, lê os alinhados e constrói o gráfico “de Brujin” a partir dos alinhamentos, remove erros e continua simplificando o gráfico e resolvendo as repetições.

2.3.4.1 Parâmetros Velvet

Ao iniciar o processo de montagem da bactéria *Herbaspirillum huttiense subsp. putei*, o ajuste de parâmetros foi feito de acordo com as análises do conjunto de dados, para que as montagens pudessem ser concluídas um fator determinante foi a escolha do parâmetro de *kmer*, que trata da quebra de palavras, ou seja, o tamanho que os “*reads*” serão divididos.

Dentro do trabalho observamos que o número de *kmer* baixo ocasionou problemas na montagem, ocorrendo de algumas vezes nem ser concluída. Outra observação que podemos fazer é que com esse mesmo ajuste, ou seja, *kmer* baixo o número de contigs ficou elevado.

Estabelecemos na montagem corrida05_40pb o valor de 28 para o *kmer* e na montagem corrida1_50pb o valor de 29, assim tendo um melhor desempenho obtendo uma resposta mais rápida do montador e resultados expressivos para a

análise. A TABELA 2, apresenta os parâmetros utilizados nas principais montagens do genoma.

TABELA 2 – PARÂMETROS UTILIZADOS NAS MONTAGENS DO GENOMA DE *Herbaspirillum huttiense subsp. putei.*

Parâmetro	**Corrida05_40pb**	**Corrida1_50pb**
kmer	28	29
ins_length	1500	1500
ins_length_sd	500	1000
min_contig_lgth	80	80
exp_cov	350	175

2.4 Fechamento de *gaps*

Depois do processo de montagem de um genoma obtemos um conjunto de dados resultante que contem alguns gaps, ou seja, falhas de montagem e que devem ser corrigidas para se obter a seqüência completa de um genoma. Os gaps podem estar entre contigs, que chamamos de gaps internos aos scaffolds, ou serem as prováveis ligações entre scaffolds, que chamamos de gaps externos a serem fechados.

Utilizamos como base para o fechamento de gaps a montagem corrida05_40pb. O fechamento de gaps dividiu-se em duas partes, primeiro o fechamento manual e segundo, o fechamento automático, que baseado em scripts desenvolvidos buscou facilitar e agilizar esse processo.

2.4.1 Fase manual

A fase manual do processo de fechamento de gaps, teve como estratégia utilizar as seqüências próximas ao gap, que no conjunto de dados é representado pela letra “N”, e a partir dessa informação buscar em outras seqüências de dados e nos bancos de dados do NCBI o complemento para as falhas da montagem.

Utilizamos o BLAST (ALTSCHUL, 1990) para realizar os alinhamentos contra os outros conjuntos de dados disponíveis da bactéria *H. huttiense subsp. putei* e contra a base de dados de proteínas não redundantes do NCBI.

Durante a análise dos dados por meio de blast, houve a necessidade de separar algumas bases que se encontravam próximas aos gaps, assim seria uma alternativa de agilizar e reduzir o escopo da busca, sempre respeitando um número mínimo de bases. Para automatizar esse processo, de localizar o gap e selecionar o número desejado de bases, foi desenvolvido um programa em Java, que recebe as seqüências e que resulta em um arquivo apenas com as seqüências de interesse. Gera um arquivo de log, que identifica qual a posição de cada gap no arquivo original (FIGURA 11), e um gráfico onde mostra a quantidade de gaps existente por scaffold (FIGURA12).

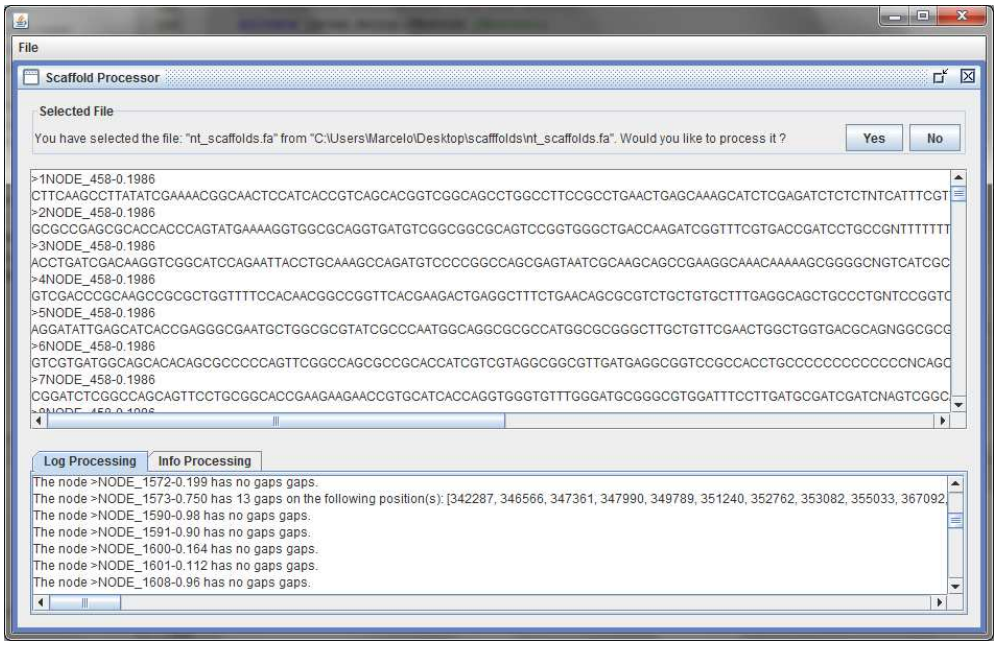


FIGURA 11 – PROGRAMA PARA SELECIONAR AS BASES DE INTERESSE PARA REALIZAR O BLAST.

FONTE: A autora (2011)

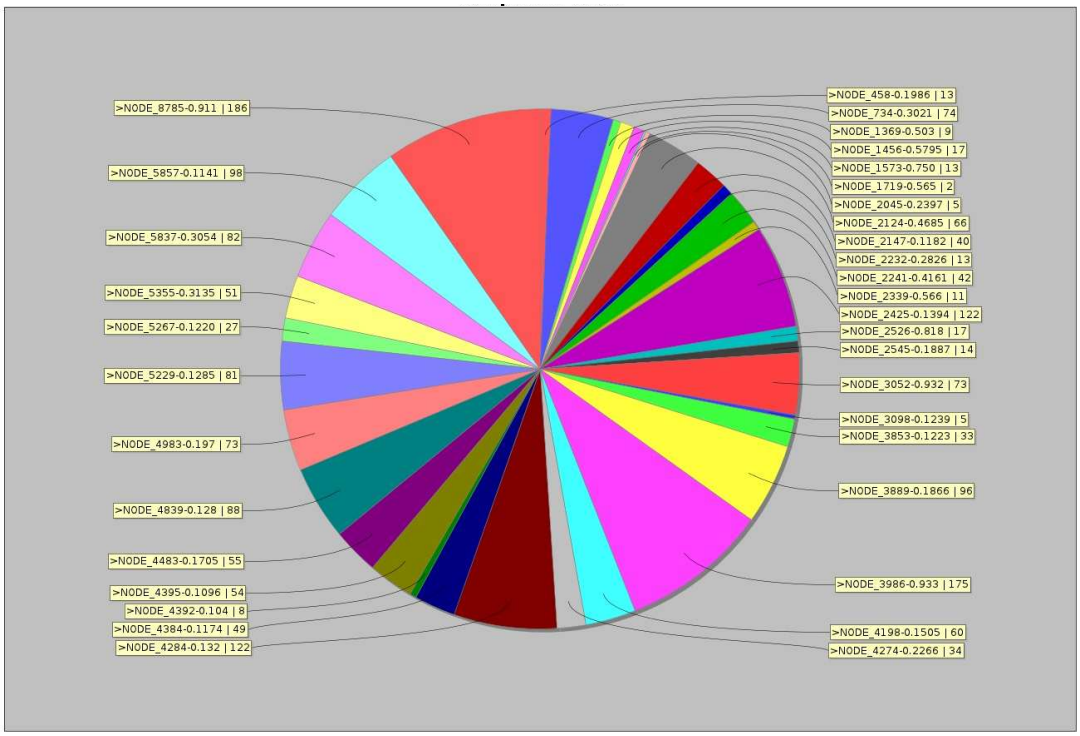


FIGURA 12 – GRÁFICO QUE APRESENTA A QUANTIDADE DE GAPS POR SCAFFOLD CONTIDO NO ARQUIVO.

FONTE: A autora (2011)

Utilizando esse conjunto com os dados separados e com as seqüências foco bem estabelecidas, utilizamos o BLAST (ALTSCHUL, 1990) para chegar a um resultado baseando-se no alinhamento final de retorno. Essa estratégia é valida também para verificação e validação de gaps fechados automaticamente.

2.4.2 Fase automática

Observando o processo realizado para o fechamento manual de gaps, vimos que o protocolo base de toda essa fase se repetia inúmeras vezes, assim buscamos automatizar essa fase deixando algo mais prático e intuitivo para o fechamento de gaps.

2.4.2.1 Proposta do script

O script foi desenvolvido, dentro do grupo de pesquisa, em MATLAB® e disponibilizado para testes com os dados disponíveis das montagens que vem sendo realizadas. Com o objetivo de fechar o maior número de gaps com os dados conhecidos utiliza uma seqüência de dados escolhida como *query* e outra seqüência como *subject*. O diagrama abaixo (FIGURA 13) representa o fluxo de processos realizados durante a execução do script.

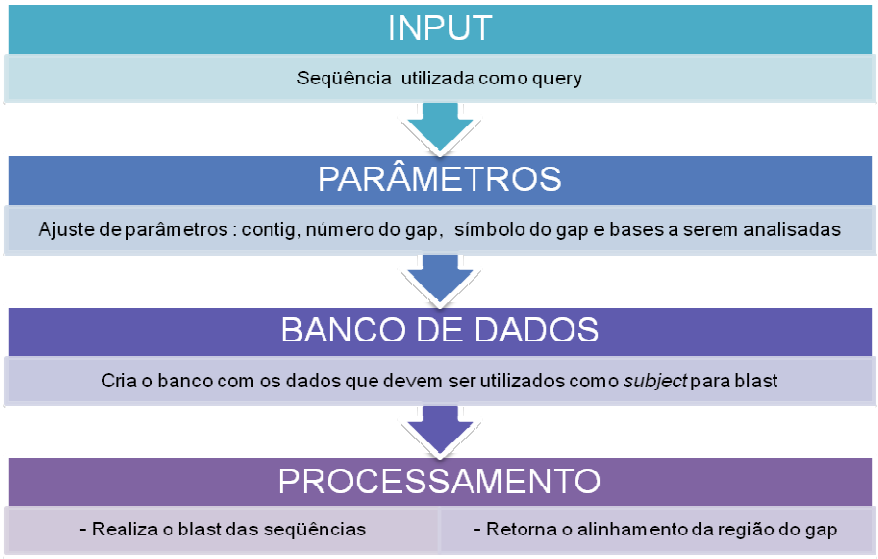


FIGURA 13 – DIAGRAMA DE PROCESSOS DO SCRIPT DE FECHAMENTO DE GAPS.
FONTE: A autora (2011)

Ao final desse processo temos como resultado o alinhamento do gap e a marcação de todas as possíveis regiões de fechamento de gaps.

2.4.2.2 Obtenção das seqüências

As seqüências utilizadas para o fechamento dos gaps foram as resultantes do processo de montagem, a montagem utilizada como *query* foi a corrida05_40pb e a segunda montagem da corrida1_50pb foi utilizada para gerar o banco de dados para

blast. Os gaps tem tamanho variável e são representados por N's, o velvet faz uma estimativa do tamanho do gap, costuma-se ter no mínimo 10pb representando cada gap por problema de compatibilidade com o banco do NCBI, isso ocorre mesmo se a distância for menor (ZERBINO, 2010).

Nos conjuntos de dados do projeto, ocorreu à representação do gap por apenas um N, o que dificultou na busca e no fechamento dos gaps. No entanto o script faz com que esse espaço de N's seja ajustado para tamanhos diferentes de gaps, ou seja, ele adéqua o alinhamento com o número que encontra de bases para fechar o gap, por exemplo, como mostra a FIGURA 14:

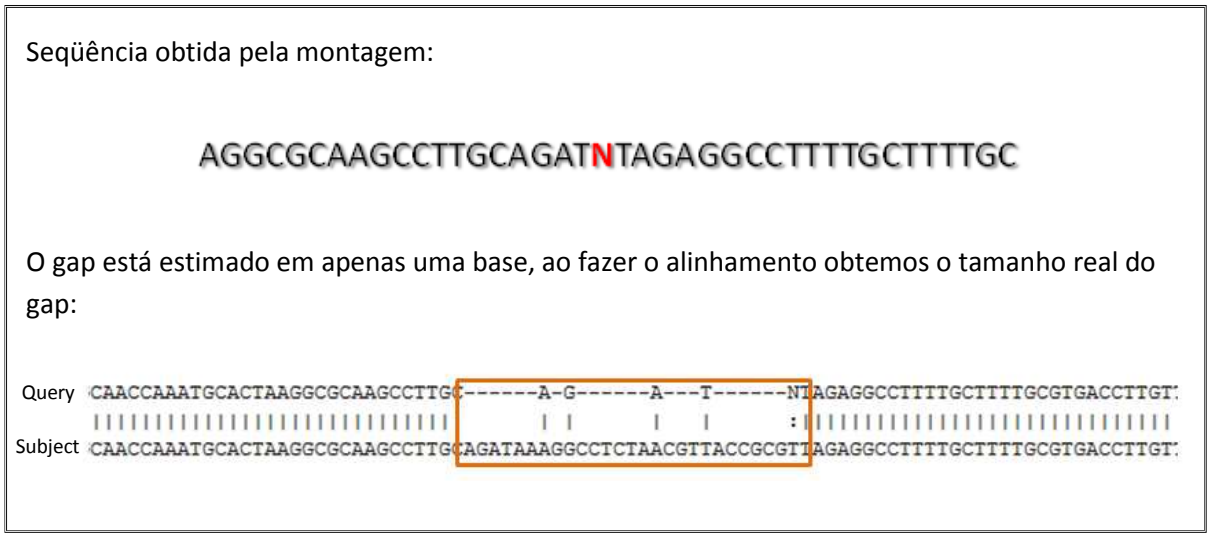


FIGURA 14 – EXEMPLO DO TAMANHO DO GAP ESTIMADO E O REAL ENCONTRADO PELO SCRIPT DE FECHAMENTO.

FONTE: A autora (2012)

Dessa forma utilizamos as seqüências disponíveis para fechar os gaps possíveis dentro do conjunto de dados, uma segunda abordagem pode utilizar bancos já conhecidos como o NR para realizar os blasts e encontrar as seqüências que fecham gaps. Uma amostra do conjunto de gaps fechados foi validada por meio de "*blasts*" *online* onde retornou em organismos próximos seqüências que o alinhamento também encontrou como positivas para gap fechado.

2.4.2.3 Bancos para blast

Para o fechamento de gaps nas seqüências foram utilizados o que chamamos de “bancos” para realizar o BLAST, ou seja, “bancos” são arquivos para busca criados pelo próprio software, a partir de seqüências que queremos como subject e que se utiliza para obter o resultado do alinhamento. Todos os gaps entre scaffolds não foram fechados levando em conta a falta de indícios de sobreposição e ligação entre eles. Para realizar essa busca, foram utilizadas as seqüências auxiliares das montagens realizadas com *H. huttiense subsp. putei*, registradas na TABELA 3, abaixo:

TABELA 3 – CONJUNTOS AUXILIARES DE DADOS PARA CRIAÇÃO DE BANCO DE DADOS PARA BLAST

Conjunto de dados	**N.º de contigs**	**N.º de scaffolds**
Corrida1_50pb	4 786	829
Ponta f3	8 295	----
Ponta r3	8 020	----

FONTE: A autora (2012)

2.5 Pré-anotação

Para atribuir características biológicas e obter as informações disponíveis nas seqüências, realizamos uma pré-anotação automática do conjunto final de 37 scaffolds da bactéria *Herbaspirillum huttiense subsp. putei*. Utilizamos duas plataformas para realizar a anotação do genoma: RAST – Rapid Annotation using Subsystem Technology (MEYER et. al, 2008) e o HGF - Hybrid Gene Finder (RAITTZ R.T. dados não publicados, 2011).

Dentro do desafio de analisar as inúmeras seqüências que vem sendo disponibilizadas, o RAST (MEYER et. al, 2008) procura proporcionar alto rendimento computacional para atribuições funcionais às seqüências. Controla o acesso de usuários para manter a privacidade dos dados submetidos, disponibilizando o resultado em vários formatos para download. Busca agilizar a anotação dos dados por meio da computação de allto desempenho, e acabar com esse que vem sendo um dos grandes gargalos da anotação.

O HGF (RAITTZ R.T. dados não publicados, 2011), busca encontrar orfs nos dados submetidos, seu diferencial é identificar orfs pequenas assim complementando os demais anotadores.

3 RESULTADOS E DISCUSSÃO

3.1 Trimming de bases

Com base nos resultados do programa Quality Assessment (RAMOS et. al, 2011) percebeu-se que as seqüências dos “reads” apresentavam uma queda de qualidade em suas bases finais, assim antes de submeter o conjunto de dados à montagem foi realizado um *trimming* nas 10 bases finais de cada “read”, passando de 50 pb para 40 pb. As FIGURAS 15 e 16, apresentam em gráficos a queda de qualidade dos “reads” percebidas nas duas tags do sequenciamento que chamamos de F3 e R3.

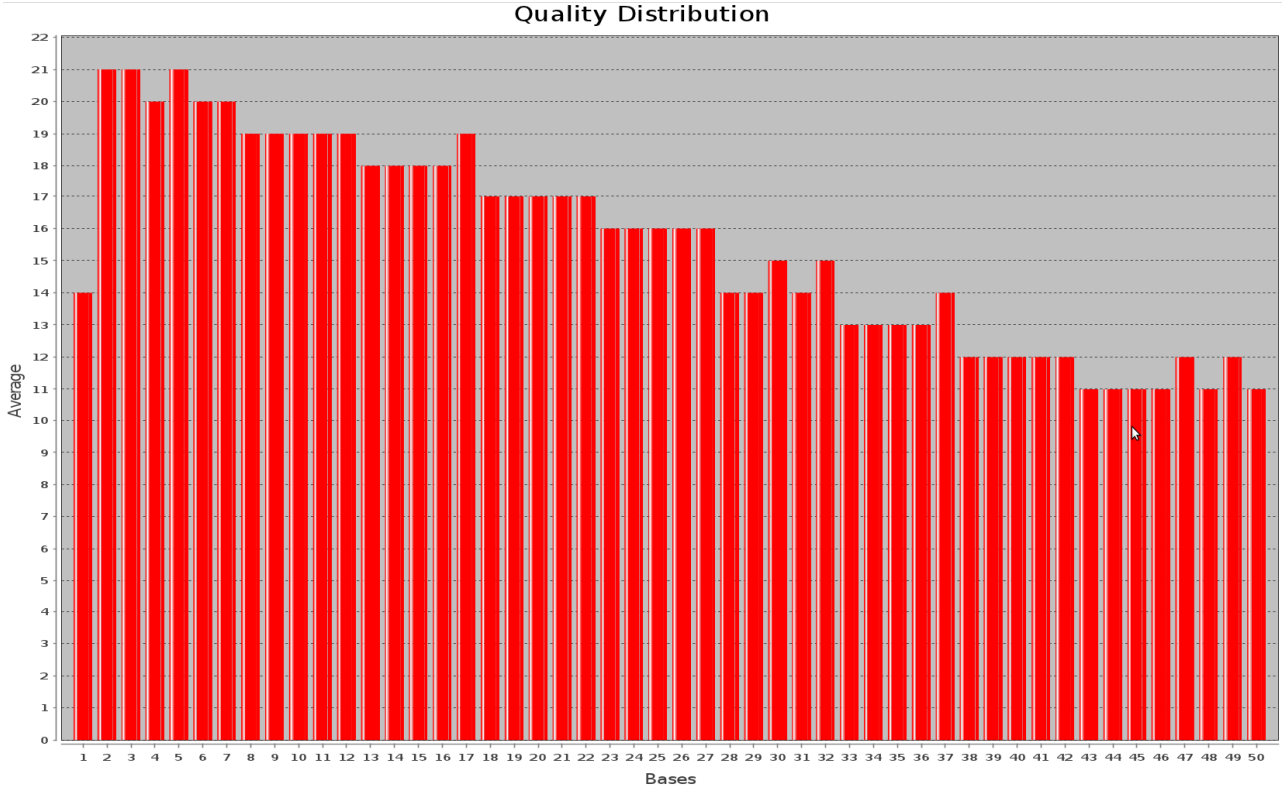


FIGURA 15 – GRÁFICO DE QUALIDADE DOS “*READS*” F3 POR BASE.

FONTE: Programa Quality assessment (RAMOS et. al. 2011)

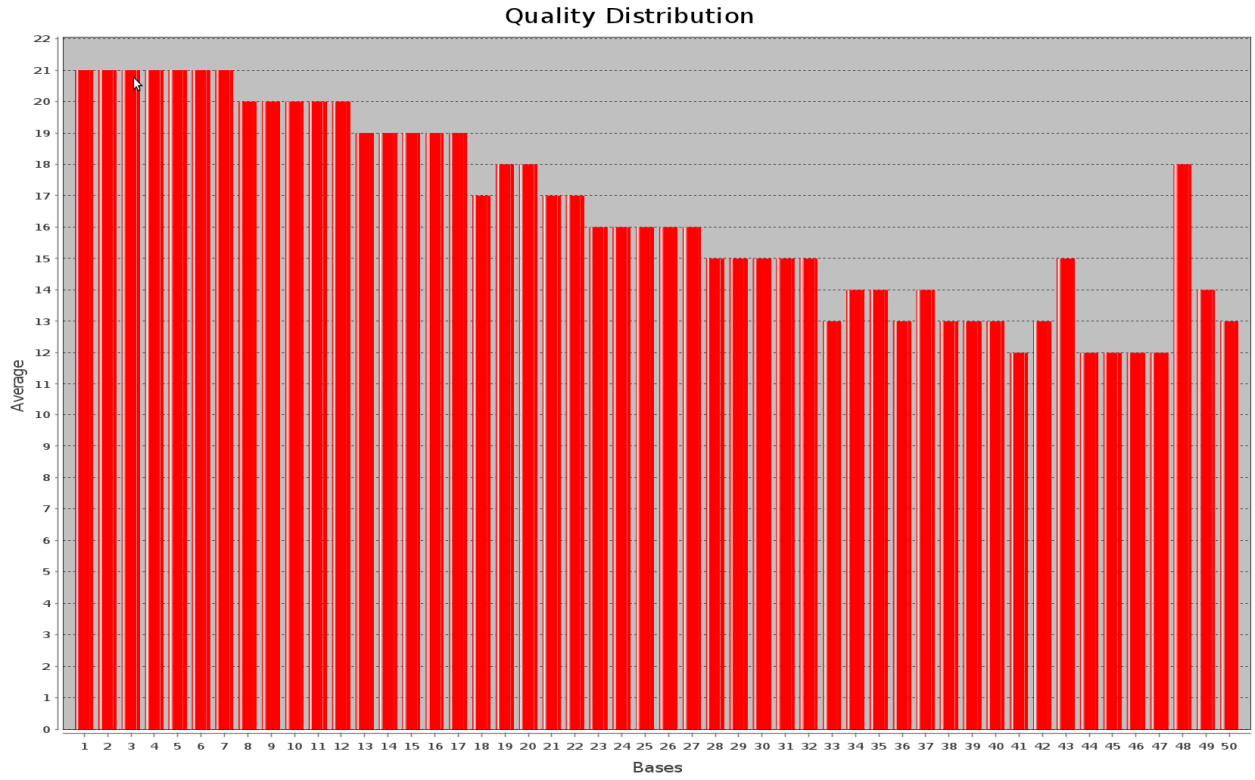


FIGURA 16 – GRÁFICO DE QUALIDADE DOS “*READS*” R3 POR BASE.

FONTE: Programa Quality assessment (RAMOS et. al. 2011)

Observando a qualidade apresentada nos gráficos buscou-se com a eliminação dessas 10 bases finais deixar a montagem com uma melhor qualidade e com isso após o término de todo o processo, gerar um menor número de contigs e scaffolds.

3.2 Conjuntos finais de scaffolds

Os conjuntos finais de scaffolds foram obtidos somando o resultado da montagem das duas corridas disponibilizadas, com o fechamento dos gaps. A TABELA 4 mostra o resultado dos scaffolds e a TABELA 5 apresenta o resultado dos contigs finais.

TABELA 4 – RESULTADO FINAL SCAFFOLDS

	Corrida05_40pb	**Corrida1_50pb**
Porcentagem de A	18	18
Porcentagem de C	31	31
Porcentagem de T	18	18
Porcentagem de G	31	31
Sum scaffolds	5780341	5864859
Num scaffolds	447	829
N50[1]	221930	235282
Max	566920	1115011

[1] N50 -> quer dizer que pelo menos 50% do genoma está representado em scaffolds de tamanho igual ou maior que o de N50.

FONTE: A autora (2012)

TABELA 5 – RESULTADO FINAL CONTIGS

	Corrida05_40pb	**Corrida1_50pb**
Porcentagem de A	18	18
Porcentagem de C	31	31
Porcentagem de T	18	18
Porcentagem de G	31	31
Sum contigs	5776603	58569945
Num contigs	2316	4786
N50[1]	5020	2116
Max	33410	11706

[1] N50 -> quer dizer que pelo menos 50% do genoma está representado em scaffolds de tamanho igual ou maior que o de N50.

FONTE: A autora (2012)

Durante o processo de montagem do genoma o programa montador utilizou para o primeiro conjunto de dados (corrida05_40pb) 63,4% dos reads disponíveis e para o segundo conjunto (corrida1_50pb) 50,8% dos reads, uma proporção boa levando em conta a presença de bases indefinidas nas seqüências.

Analisando o conjunto com o menor número de scaffolds, tomado como montagem principal, e complementado com os outros dados, pudemos observar que apenas 37 scaffolds, representavam aproximadamente 98% de todo o tamanho do genoma. Partindo dessa informação, geramos um arquivo contendo os 37 scaffolds significativos da montagem para prosseguir com as análises. Utilizar apenas os

contigs/scaffolds com mais de 1000 pb foi útil, assim funcionando como um filtro que facilitou o tratamento desses dados, o que foi retirado do conjunto principal foi utilizado para análises durante o processo.

Dentro desse conjunto de dados encontramos 1908 gaps, com a utilização dos scripts desenvolvidos e também do fechamento manual resultamos em 90 gaps fechados.

3.3 Ordenação do scaffolds

Os scaffolds da montagem SOLID foram ordenados ulizando como referência os genomas de *H. seropedicae* e *H. rubrisubalbicans*. Para realizar a ordenação utilizamos o software jContigSort (GUIZELINI et. al., 2011), que ordena os dados baseados na referência e gera um arquivo com a provável ordem dos contigs. Com essa ordenação e alinhamento utilizando o programa MUMer observamos que os genomas apresentam grandes regiões similares.

Com a ordenação dos scaffolds, resultados de alinhamentos, BLAST e as análises realizadas nos arquivos de coordenadas gerado pelo MUMer, pode-se indicar uma provável ligação entre 4 scaffolds o que reduz o conjunto de 37 para 33 scaffolds. As ligações estariam entre os scaffolds 2 e 37; 22 e 16; 31 e 23; 34 e 7 e precisam ser analisadas de maneira detalhada para que se comprove a ligação.

3.3.1 Herbaspirillum huttiense subsp. putei x Herbaspirillum seropedicae

O conjunto de scaffolds foi ordenado baseado no genoma de referência e realizado o alinhamento contra a *Herbaspirillum seropedicae* como observado na FIGURA 17 que apresenta o gráfico dotplot desse alinhamento.

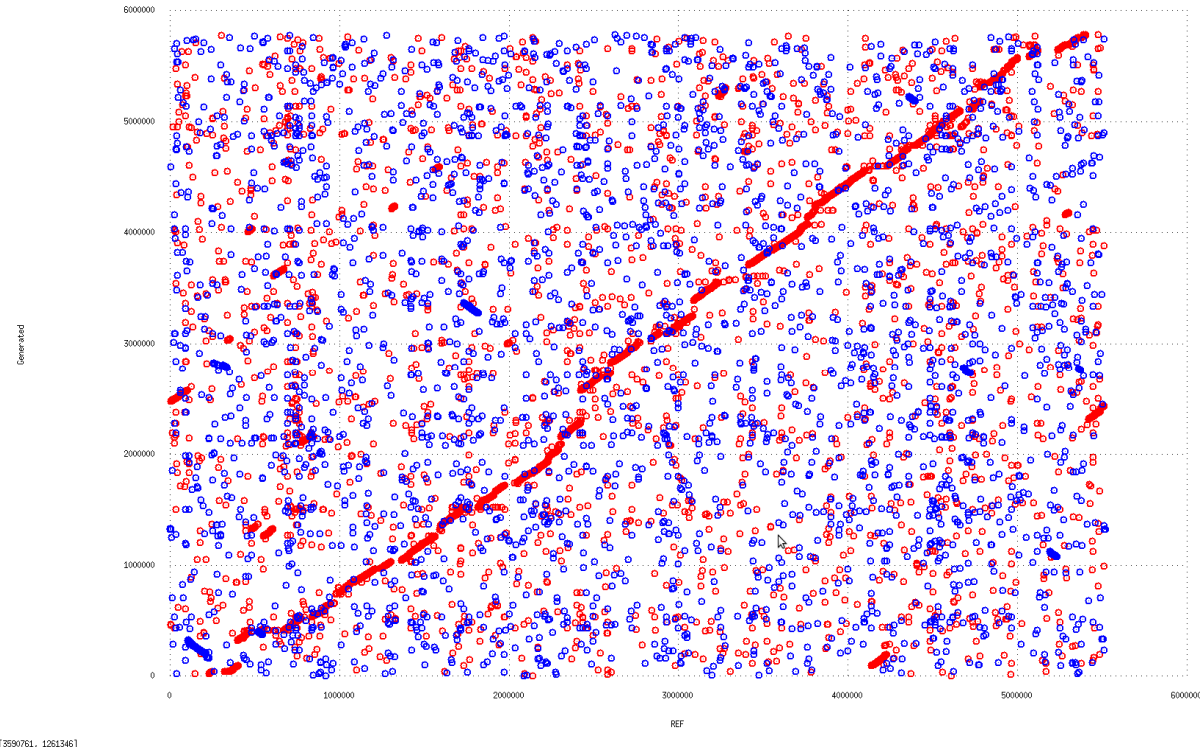


FIGURA 17 – GRÁFICO DOTPLOT DOS SCAFFOLDS DEPOIS DE ORDENADOS DE *H. huttiense subsp. putei* CONTRA O GENOMA COMPLETO DE *H. seropedicae,* MOSTRANDO OS ALINHAMENTOS COM A REFERÊNCIA.

FONTE: A autora (2011) com base em MUMmer

Com esse alinhamento podemos observar de maneira mais clara a alta identidade entre os dois genomas. A partir do draft do genoma de *Herbaspirillum huttiense subsp. putei* geramos o gráfico de GCSKEW acumulado apresentado na FIGURA 18 em comparação com o de *Herbaspirillum seropedicae*. A proximidade entre os genomas gera um padrão parecido entre eles no gráfico de GCSKEW acumulado apresentado.

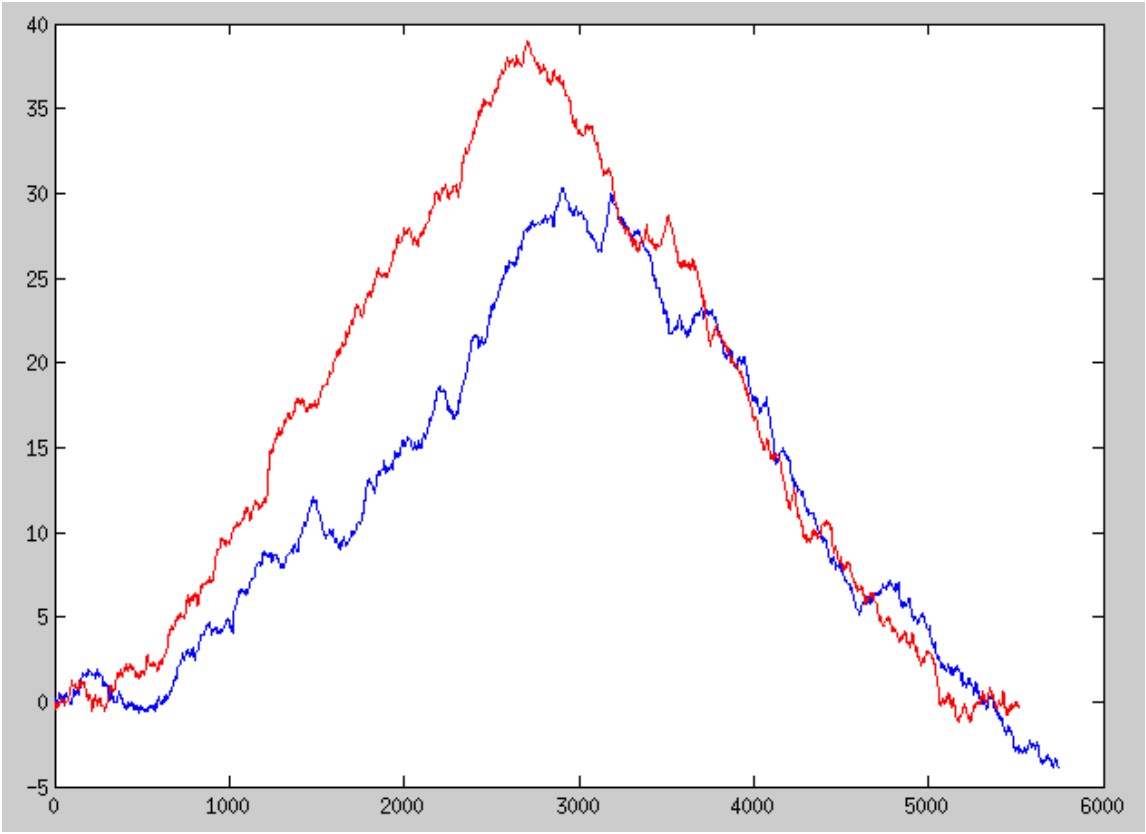


FIGURA 18 – GRÁFICO DO GCSKEW ACUMULADO DOS SCAFFOLDS ORDENADOS DE *H. huttiense subsp. putei.* A PRIMEIRA METADE ACUMULA-SE VALORES CRESCENTES E NA SEGUNDA VALORES DECRESCENTES.

FONTE: A autora baseado em MATLAB(2011)

3.3.2 Herbaspirillum huttiense subsp. putei x Herbaspirillum rubrisubalbicans

Na FIGURA 19, podemos observar o gráfico de dotplot do alinhamento do genoma com a segunda referência *Herbaspirillum rubrisubalbicans* utilizando o programa MUMmer.

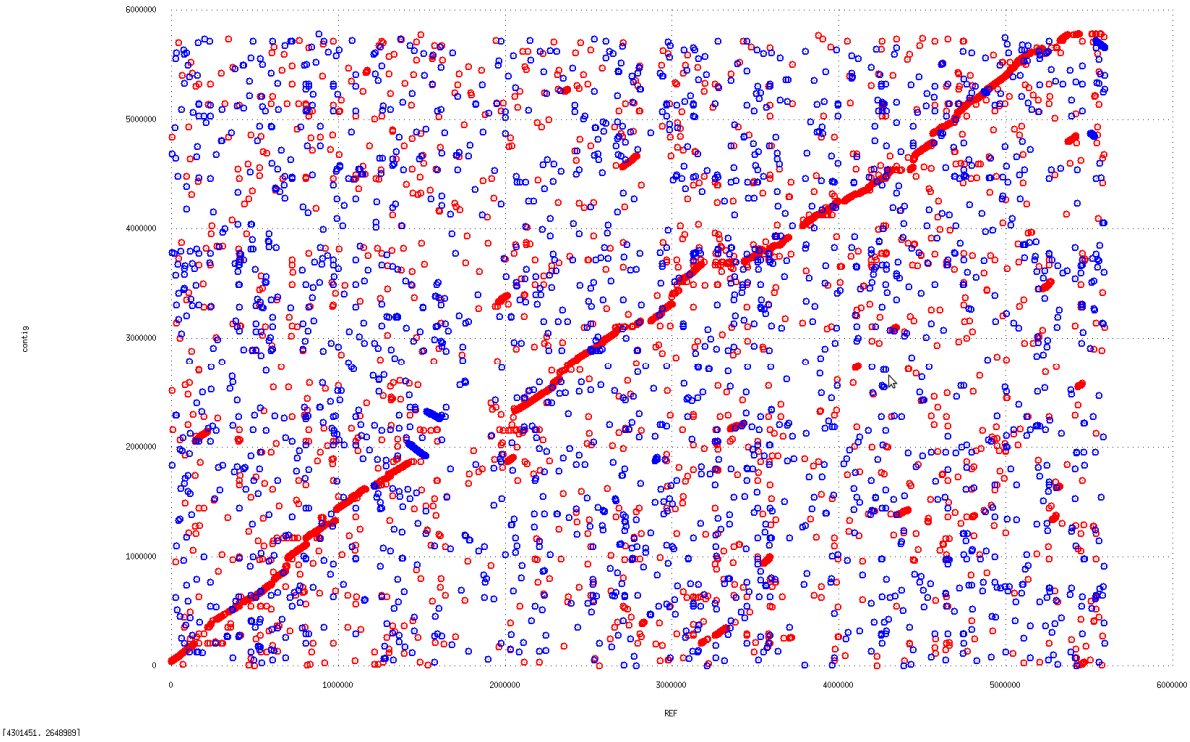


FIGURA 19 – GRÁFICO DOTPLOT DOS SCAFFOLDS DEPOIS DE ORDENADOS DE *H. huttiense subsp. putei* CONTRA O GENOMA COMPLETO DE *H. rubrisubalbicans*, MOSTRANDO OS ALINHAMENTOS COM A REFERÊNCIA.

FONTE: A autora (2011) com base em MUMmer

Observando também uma proximidade entre as duas espécies, a FIGURA 20, apresenta a comparação de *H. huttiense subsp. putei* e *H. rubrisubalbicans* com relação ao gráfico de GCSKEW acumulado.

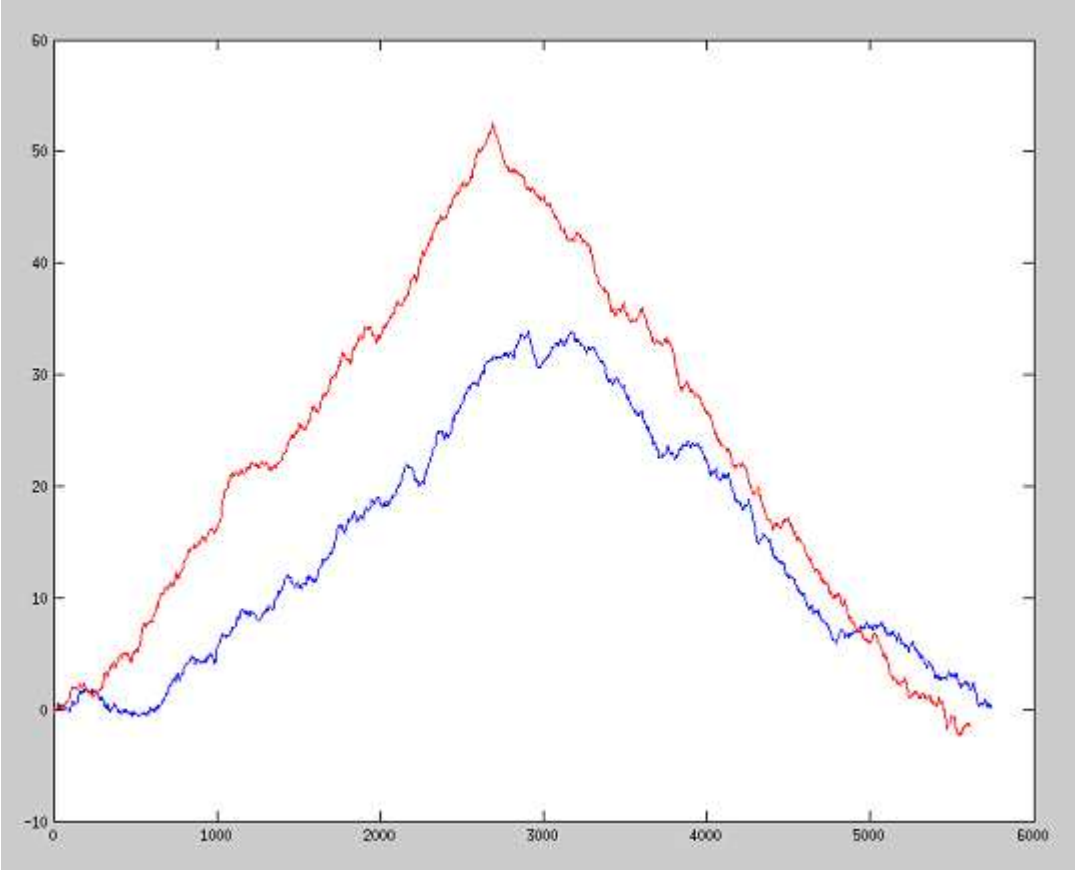


FIGURA 20 – GRÁFICO DO GCSKEW ACUMULADO DOS SCAFFOLDS ORDENADOS DE *H. huttiense subsp. putei.* O PADRÃO DO GRÁFICO BASEIA-SE NOS DADOS DISPONÍVEIS DE *H. rubrisubalbicans.*

FONTE: A autora baseada em MATLAB (2011)

3.4 Pré - Anotação

A pré-anotação pelo RAST dos 37 *scaffolds*, contendo 5 723 280 bases, gerou um total de 5 547 características anotadas. Como resultado obtivemos a anotação dos genes e alocação em subsistemas, tendo uma visão qualitativa dos genes anotados de acordo com sua função ou grupo (FIGURA 21). Já a realizada

pelo HGF resultou em 6 084 orfs anotadas. A diferença do número de orfs anotadas se deve a diferença de técnicas aplicadas pelos dois preditores.

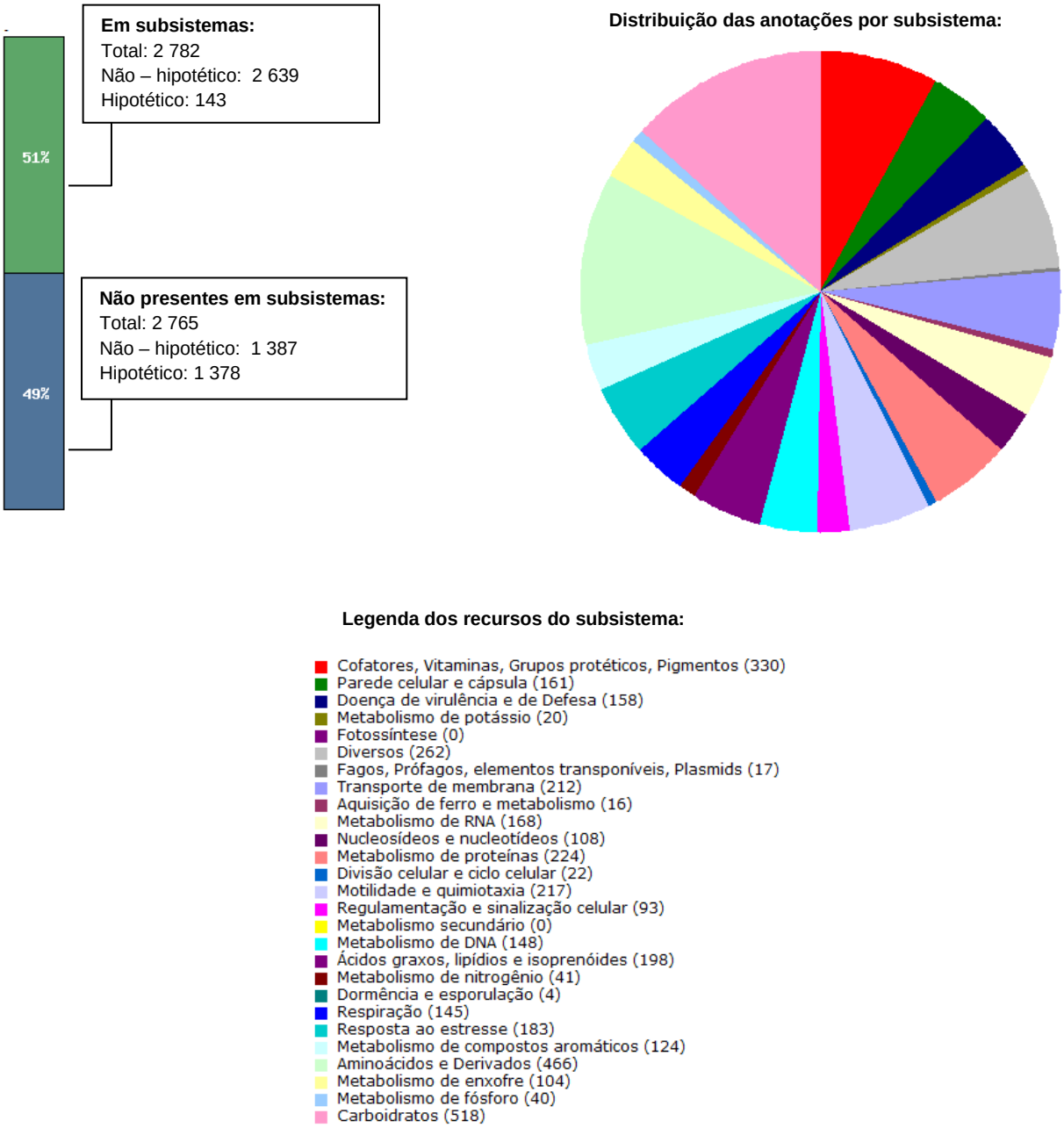


FIGURA 21 – DISTRIBUIÇÃO DOS GENES ANOTADOS PELA PLATAFORMA RAST DE ACORDO COM SUA CATEGORIA EM CADA SUBSISTEMA

FONTE: Adaptação RAST (2012)

Todo o processo de pré-anotação descrito neste trabalho, foi realizado por meios automatizados, assim, após a obtenção da seqüência completa desta bactéria é necessário que este genoma seja submetido a uma nova anotação e curadoria manual, para que se tenha precisão nos dados antes de seu depósito nos bancos de dados públicos.

4 CONCLUSÃO

O draft do genoma da bactéria *Herbaspirillum huttiense subsp. putei* obtido pela análise e processamento dos dados de contigs e scaffolds resultantes da montagem genômica resultou em 37 *scaffolds* que juntos somam 5 723 317 com conteúdo GC de 62,3 %, esses *scaffolds* podem ser reduzidos a 33 pelos indícios de ligação presentes nas seqüências. Apresentou grande similaridade com os genomas utilizados como referência *H. seropedicae* e *H. rubrisubalbicans*. O padrão do gráfico de Gcskew se manteve na ordenação dos scaffolds baseados em cada uma das referências.

Para o fechamento dos gaps apresentados no conjunto de dados, foram desenvolvidas novas estratégias para facilitar a forma manual, assim apresentando uma nova ramificação para trabalhos futuros, onde a bioinformática facilite nas pesquisas possibilitando que o fechamento de gaps, se torne mais ágil visto o crescimento do seqüenciamento genômico. Durante a finalização dos dados, não foram utilizadas técnicas de bancada em laboratório para obtenção de novas seqüências o que poderia posteriormente ajudar na finalização do genoma.

Assim, disponibilizamos uma anotação parcial do genoma bacteriano *Herbaspirillum huttiense subsp. putei*, para que em trabalhos futuros possa ser finalizada por meio de novas análises de dados, de novas seqüências obtidas e assim possa ser anotada por completo, depositada e disponibilizada para novas pesquisas.

REFERÊNCIAS

ALTSCHUL, S. F., et al. **Basic Local Alignment Search Tool**. Journal of Molecular Biology 215, 403–410 – 1990

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

BALDANI, J.I., POT, B., KIRCHHOF, G., FALSEN, E., BALDANI, V. L. D. OLIVARES, F. L., HOSTE, B. KERSTERS, K., HARTMANN, A.G., DOBEREINER, J. **Emended description of Herbaspirillum; inclusion of Pseudomonas rubrisubalbicans, a mild plant pathogen, as Herbaspirillum rubrisubalbicanscomb nov; and classification of a group of clinical isolates (EF group 1) as Herbaspirillum species 3.** International Journal of Systematic Bacteriology. v.46: 802-81 - 1996

CARDOSO, L. A. **Montagem genômica da bactéria endofítica diazotrófica**. 118 f. Dissertação (Mestrado em Bioinformática) - Universidade Federal do Paraná, Curitiba, 2011.

CARRO, L., RIVAS, R., LEON-BARRIOS, M., GONZALES-TIRANTE, M., VELÁZQUEZ, E., VALVERDE, A,. **Herbaspirillum canariense sp. nov., Herbaspirillum aurantiacum sp. nov. and Herbaspirillum soli sp. nov., three new species isolated in Tenerife (Canary Islands)**. International Journal of Systematic and Evolutionary Microbiology, 2011 : ijs.0.031336-0 v1.

CARVALHO, M. C. C. G., SILVA, D. C. G. **Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas**. Cienc. Rural [online]. 2010, vol.40, n.3, pp. 735-744. ISSN 0103-8478.

CHAISSON, M., PEVZNER, P. **Short read fragment assembly of bacterial genomes.** Genome Reserch, 2008, 18, 324

DING, L. and YOKOTA, A. **Proposals of Curvibacter gracilis gen. nov., sp. nov. and Herbaspirillum putei sp. nov. for bacterial strains isolated from well water and reclassification of [Pseudomonas] huttiensis, [Pseudomonas] lanceolata, [Aquaspirillum] delicatum and [Aquaspirillum] autotrophicum as Herbaspirillum huttiense comb. nov., Curvibacter lanceolatus comb. nov., Curvibacter delicatus comb. nov. and Herbaspirillum autotrophicum comb.**

nov. International Journal of Systematic and Evolutionary Microbiology, 2004. v.54, p.2223-2230.

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

ECKERT, B., WEBER, O.B., KIRCHHOF, G., HALBRITTER, A., STOFFELS, M., HARTMANN, A. **Azospirillum doebereinerae sp. nov., a nitrogen-fixing bacterium associated with the C4-grass Miscanthus**. International Journal of Systematic and Evolutionary Microbiology. 2001. 51: 17-26

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.

GALPERIN, M.Y., KOONIN, E.V., **From complete genome sequence to 'complete' understanding?** Trends Biotechnol. 2010; 28:398–406.

GUIZELINI, D; 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 (AB3C) and 3rd International Conference of the IberoAmerican Society for Bioinformatics (SolBio), 2011, Florianópolis. Abstract book, 2011.

HORIMOTO, K.; FUKUCHI, S.; MORI, K. **Comprehensive comparison between locations of orthologus genes on archaeal and bacterial genomes.** Bioinformatics, v. 17, p. 791-802, 2001.

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

JUNG, S.Y., LEE, M.H., OH, T.K., YOON, J.H. **Herbaspirillum rhizosphaerae sp. nov., isolated from rhizosphere soil of Allium victorialis var. platyphyllum.** International Journal of Systematic and Evolutionary Microbiology. 2007. 57: 2284–2288

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**. International Journal of Systematic and Evolutionary Microbiology, v. 51, p. 157-168, 2001.

LIBERMAN, F. **Análise dos fatores determinantes para a qualidade da anotação genômica automática.** 136 f. Dissertação (Mestrado em Biotecnologia e Ciências Genômicas) - Universidade Católica de Brasília, Brasília, 2004.

MAXAM, A. M., & GILBERT, W. A new method for sequencing DNA. Vol. 74, No. 2, pp. 560-564, February 1977 Biochemistry, 74(2), 560-564.

MCKERNAN, K. et al. **Reagents, methods, and libraries for bead-based sequencing**. US patent application 20080003571, 2006.

MEYER, F.; PAARMANN, D.; D'SOUZA ,M.; OLSON, R.; GLASS, EM.; KUBAL, M.; PACZIAN, T.; RODRIGUEZ, A.; STEVENS, R.; WILKE, A.; WILKENING, J. and EDWARDS, RA. **The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes.** BMC Bioinformatics 2008, 9**:**386.

MONTAGEM DE GENOMA: http://www.cbcb.umd.edu/research/assembly_primer.shtml. Acesso em 11/01/2012.

NAGARAJAN, N.; POP, M. **Sequencing and genome assembly using next-generation technologies**. Methods Mol Biol, 2010. 673:1–17

PEDROSA, F.O., MONTEIRO, R.A., WASSEM, R., CRUZ, L. M., AYUB, R. A., et al. **Genome of Herbaspirillum seropedicae strain SmR1, a specialized diazotrophic endophyte of tropical grasses**. PLoS Genetics. 2011; 7:e1002064.

RAMOS, R. T.J.; CARNEIRO, R. A.; 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 2011, 4:130 http://www.biomedcentral.com/1756-0500/4/130.

ROTHBALLER, M., SCHMID, M., KLEIN, I., GATTINGER, A., GRUNDMANN, S., HARTMANN, A. **Herbaspirillum hiltneri sp. nov., isolated from surface-sterilized wheat roots**. International Journal of Systematic and Evolutionary Microbiology. 2006. 56: 1341–1348

SANGER, F., *et al*. **DNA sequencing with chain-terminating inhibitors.** *Proceedings of the National Academy of Sciences* 74, 5463–5467, 1977.

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

SILVA, V. C. H. **Expressão e purificação de proteínas do sistema de secreção do tipo III de Herbaspirillum seropedicae.** 111 f. Dissertação (Mestrado em Ciências Bioquímicas) - Universidade Federal do Paraná, Curitiba, 2008.

SOUZA, A. L. F.; BRUSAMARELLO, L. C.C. **Sequenciamento de DNA: decifrando o manual de instruções dos seres vivos.** Genética na escola. Sociedade Brasileira de Genética. 03.03, 45-52. 2009

STEIN, L. **Genome annotation: from sequence to biology**. Nature reviews. Genetics. 2001. 2(7), 493-503.

TIEPPO, E. **Montagem e análise preliminar do genoma de Bradyrhizobium elkanii 587 utilizando leituras de sequências de DNA curtas.** 77 f. Dissertação (Mestrado em Bioinformática) - Universidade Federal do Paraná, Curitiba, 2011.

VALVERDE, A., VELAZQUEZ, E., GUTIERREZ, C., CERVANTES, E., VENTOSA, A., IGUAL, J.M. **Herbaspirillum lusitanum sp. nov., a novel nitrogenfixing bacterium associated with root nodules of Phaseolus vulgaris**. International Journal of Systematic and Evolutionary Microbiology. 2003. 53: 1979–1983

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

ZERBINO, D.R. **Using the Velvet de novo assembler for short-read sequencing technologies.** Curr. Protoc. Bioinformatics 31. 2010. 11.5.1–11.5.12.

ZHANG, W., CHEN, J., YANG, Y., TANG, Y., SHANG, J., et al. **A Practical Comparison of De NovoGenome Assembly Software Tools for Next-Generation Sequencing Technologies**. PLoS ONE 6(3): e17915. – 2011

YUTAO, F.; HEATHER E. P.; STEPHEN, F.M.; JINGWEI N. N. ; MICHAEL, D. R.; JOEL, A. M. ; KEVIN, J. M.; ALAN, P. B. **SOLID Sequencing and 2-Base Encoding**. Applied Biosystems. 2007.