UNIVERSIDADE FEDERAL DE PELOTAS
Programa de Pós-Graduação em Biotecnologia
Centro de Desenvolvimento Tecnológico

Dissertação

Sequenciamento, montagem e anotação do genoma de um novo isolado de *Leptospira borgpetersenii*

Marcus Redü Eslabão

Pelotas, 2012

MARCUS REDÜ ESLABÃO

SEQUENCIAMENTO, MONTAGEM E ANOTAÇÃO DO GENOMA DE UM NOVO ISOLADO DE *Leptospira borgpetersenii*

Dissertação apresentada ao Programa de Pós-Graduação em Biotecnologia da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Mestre em Ciências (área do conhecimento: Bioinformática).

Orientador: Odir Antonio Dellagostin

Pelotas, 2012

Dados de catalogação na fonte:
Ubirajara Buddin Cruz – CRB 10/901
Biblioteca de Ciência & Tecnologia - UFPel

E76s Eslabão, Marcus Redü

Sequenciamento, montagem e anotação do genoma de um novo isolado de *Leptospira borgpetersenii* / Marcus Redü Eslabão. – 32f. : tab. – Dissertação (Mestrado). Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas. Centro de Desenvolvimento Tecnológico, 2012. – Orientador Odir Antônio Dellagostin.

1.Biotecnologia. 2.Sequenciamento de nova geração. 3.*Leptospira borgpetersenii.* 4.Genômica. 5.Bioinformática. I.Dellagostin, Odir Antônio. II.Título.

CDD: 614.56

Banca examinadora:

Dr. Odir Antônio Dellagostin, Universidade Federal de Pelotas (Presidente)
Dr. Alan John Alexander McBride, Universidade Federal de Pelotas
Dr. Luciano Carlos da Maia, Universidade Federal de Pelotas
Dr. Luciano da Silva Pinto, Universidade Federal de Pelotas

AGRADECIMENTOS

À Universidade Federal de Pelotas que através do Centro de Biotecnologia abriu a oportunidade de aprendizado e desenvolvimento.

À Rede Paraense de Genômica e Proteômica - UFPA e ao Laboratório de Biologia Celular e Molecular - UFMG, que foram parceiros neste projeto.

Ao meu orientador Dr. Odir Antonio Dellagostin, que acreditou no meu potencial, me guiou pelo campo acadêmico e proporcionou todas as condições necessárias para o desenvolvimento do meu trabalho.

Aos Dr. Artur Silva e Dr. Vasco Azevedo, por abrirem as portas e me acolher em seus laboratórios.

Aos Doutorandos Rommel Ramos, Adriana Carneiro e Anderson Santos, que me guiaram pelo campo de sequenciamento, montagem e anotação.

Aos meus estagiários Frederico Kremer, Mariana Pereira e Jessica Plaça, que me acompanharam e trabalharam em todos os processos deste trabalho.

Aos Dr. Everton Silva e Michel Fagundes que isolaram, identificaram e cultivaram a bactéria utilizada neste trabalho.

Aos demais integrantes do Centro de Biotecnologia que participaram direta ou indiretamente.

À minha família e a minha namorada que sempre me apoiaram de todas as formas possíveis.

RESUMO

ESLABÃO, Marcus Redü. **Sequenciamento, montagem e anotação do genoma de um novo isolado de *Leptospira borgpetersenii*.** 2012. 32f. Dissertação – Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas, Pelotas.

A leptospirose é uma zoonose negligenciada com distribuição global. A doença é causada por bactérias patogênicas do gênero *Leptospira*, as quais acometem humanos e vários animais domésticos e silvestres, acarretando graves problemas à saúde humana e prejuízos na pecuária. O presente trabalho teve como objetivo sequenciar o genoma da *Leptospira borgpetersenii* sorogroupo Ballum cepa 4E, isolada de camundongo doméstico (*Mus musculus*), um dos principais reservatórios deste gênero. A sequência completa do genoma foi determinada através do sistema SOLiD™, onde foram obtidas mais de 85 milhões de leituras com tamanho de 50 pb cada. Essas leituras foram utilizadas para obtenção de *scaffolds* dos dois cromossomos presente neste organismo, através de montagem *ab initio* com os softwares Velvet e Edena; e posterior orientação das contigs com o software G4All. Com a conclusão da montagem, o cromossomo maior apresentou o tamanho de 3.071.053 pb, 40,58% de conteúdo GC, 36 tRNA, 4 rRNA e 2.908 fases de leitura abertas (ORF). Para o cromossomo menor o total de bases foi de 305.940 pb, conteúdo GC de 40,25%, 277 ORFs, nenhum tRNA e rRNA foram preditos. Foi observada uma redução do cromossomo maior da cepa 4E em ralação ao cromossomo maior da cepa L550, onde 99 genes da cepa L550 não estão presentes na cepa 4E e cerca de 394 kb de região não codificante também foi perdida. A principal hipótese para a redução é o efeito da presença de um grande número de elementos móveis, processo observado no genoma de outras cepas da espécie *L. borgpetersenii.* O método Applied Biosystems SOLiD™ 4 permitiu a determinação da sequência do genoma de *L. borgpeter*senii cepa 4E, com ampla cobertura e acurácia. Os metodos de montagem *ab initio* utilizados proporcionaram aproveitar ao máximo as sequencias geradas.

Palavras-chave: Sequenciamento. *Leptospira*. Genoma. Next-Generation Sequencing.

ABSTRACT

ESLABÃO, Marcus Redü. **Sequencing, assembly and genome annotation of a new isolated of Leptospira borgpetersenii.** 2012. 32f. Dissertação – Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas, Pelotas.

Leptospirosis is a neglected zoonosis with global distribution. The disease is caused by pathogenic bacteria of the genus *Leptospira*, which affect humans and various domestic and wild animals, causing serious problems to human health and damage to livestock. The objective of this study was to determine the genome sequence of *Leptospira borgpetersenii* serogroup Ballum strain 4E, isolated from domestic mice (*Mus musculus*), one of the main reservoirs of this genus. The complete genome sequence was determined using SOLiD^{TM} system, which generated over 85 million 50 bp reads. These reads were used to obtain scaffolds of the two chromosomes present in this organism through the *ab initio* sequence assembly with Velvet and Edena softwares and orientation of contigs with G4All software. With completion of the assembly process, the large chromosome was 3,071,053 bp, GC content of 40.58%, 36 tRNA, 4 rRNA and 2,908 open reading frames (ORF). The small chromosome has 305,940 bp, GC content of 40.25%, 277 ORFs, no tRNA or rRNA. A reduction in the large chromosome of 4E strain was observed compared to the large chromosome of L550 strain, where 99 genes of L550 strain are not present in the 4E strain and about 394 kb of non-coding region was also lost. The main hypothesis for this reduction is the effect of the presence of a large number of mobile genetic elements. Genome reduction has been observed in other strains of *L. borgpetersenii*. The Applied Biosystems SOLiD™ 4 method allowed determination of the genome sequence of *L. borgpetersenii* strain 4E, with wide coverage and accuracy. The *ab initio* assembly methods used allowed for complete utilization of the sequences generated.

Key Words: Sequencing. Leptospira. Genome. Next-Generation Sequencing.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

DNA	*Deoxyribonucleic acid* (Ácido desoxirribonucleico)
GB	Giga pares de bases
IS	*Insertion sequence* (sequência de inserção)
NGS	*Next-generation sequencing* (sequenciamento de nova geração)
OLC	*Overlap-layout-consensus*
ORF	*Open Reading Frame* (janela aberta de leitura)
pb	Pares de base
PCR	*Polymerase Chain Reaction* (reação em cadeia da polimerase)
RNA	*Ribonucleic acid* (ácido ribonucleico)

Sumário

1 INTRODUÇÃO

1.1 Leptospirose: aspectos gerais

A leptospirose é uma zoonose de distribuição global causada por bactérias patogênicas do gênero *Leptospira,* as quais acometem vários animais domésticos e silvestres. O ciclo de transmissão desta doença envolve a interação entre reservatórios animais, um ambiente favorável e grupos humanos ou animais suscetíveis. Os fatores de risco associados à infecção dependem, portanto, de características da organização espacial, dos ecossistemas e das condições de vida e trabalho da população (MURHEKAR et al., 1998).

Os humanos são infectados pela penetração de leptospiras nas mucosas e na pele lesada ou íntegra, quando em contato com água contamina. A doença pode se apresentar nas formas subclínicas ou formas graves com alta letalidade. Ela, na maioria dos casos, se inicia abruptamente com febre, mal-estar geral e cefaleia. A forma anictérica aparece em 60% a 70% dos casos. A doença pode ser discreta, de inicio súbito com febre, cefaleia, dores musculares, anorexia, náuseas e vômitos. Dura de um a vários dias, sendo frequentemente rotulada como síndrome gripal ou virose. Uma infecção mais grave pode ocorrer. Na forma ictérica, a fase septicêmica evolui para uma doença ictérica grave, disfunção renal, fenômenos hemorrágicos, alterações cardíacas e pulmonares, associadas a taxas de letalidade que variam de 5% a 20% (LEVETT, 2001).

Na pecuária nacional e mundial a ocorrência da leptospirose acarreta prejuízos econômicos (VASCONCELOS, 1997). Esta doença pode se manifestar tanto na forma esporádica quanto a endêmica. Os mamíferos domésticos de produção, trabalho e companhia são susceptíveis e acometidos por leptospiras tanto nas áreas urbanas como rurais (VASCONCELOS, 1997). Nas espécies de interesse zootécnico (ovinos, bovinos, caprinos, equinos e suínos) a leptospirose está relacionada a distúrbios reprodutivos causando abortamentos, natimortalidade, esterilidade e queda de fertilidade. Dependendo do sorovar envolvido e de fatores relativos ao hospedeiro, como grau de imunidade e estado fisiológico, pode ocasionar grandes prejuízos econômicos nos rebanhos com mortalidade de animais jovens e queda no ganho de peso e na produção de leite (FAINE ET AL,

1999;GUIMARÃES, 1982) A vacinação é uma importante ação preventiva contra a infecção dos animais por leptospiras. Para esse fim, vacinas inativadas são as mais utilizadas (DELLAGOSTIN et al., 2011). Porém vacinas convencionais, constituídas de células inteiras inativadas (bacterinas) não proporcionam proteção efetiva contra os diferentes sorovares causadores da leptospirose (ADLER; DE LA PENA, 2010;LEVETT, 2001).

O gênero *Leptospira* possui dezenove espécies, sendo usualmente classificada com base na sorologia em sorogrupos e sorovares. Destas espécies treze são patogênicas: *L. alexanderi, L. alstonii, L. borgpetersenii, L. inadai, L. interrogans, L. fainei, L. kirschneri, L. licerasiae, L. noguchi, L. santarosai, L. terpstrae, L. weilii, L. wolffii,* podendo ser classificada em mais de 260 sorovares, e seis espécies são saprófitas: *L. biflexa, L. meyeri, L. yanagawae, L. kmetyi, L. vanthielii*, podendo ser classificadas em mais de 60 sorovares (ADLER; DE LA PENA, 2010;FAINE ET AL, 1999). Dentre as espécies citadas acima foram listadas no Brasil até o ano de 2007 as espécies patogênicas: *L. santarosai, L. interrogans, L. kirshneri* e *L. borgpetersenii* (SILVA et al., 2009), contudo, estudos realizados na cidade de Pelotas-RS resultaram em quatro novos isolados de *L. noguchii* (SILVA et al., 2007;SILVA et al., 2009), até então não reportado no Brasil e quatro novos isolados de *L. borgpetersenii* (DA SILVA et al., 2010).

Atualmente apenas seis genomas desta ampla diversidade do gênero *Leptospira* estão disponíveis, sendo eles: *L. biflexa* cepa Paris e *L. biflexa* cepa Ames (PICARDEAU et al., 2008); *L. interrogans* cepa L1-130 e *L. interrogans* cepa Lai (NASCIMENTO et al., 2004); *L. borgpetersenii* cepa L550 *e L. borgpetersenii* cepa JB197 (BULACH et al., 2006).

Para que haja o estudo e criação de novas tecnologias, como vacinas recombinantes e diagnósticos moleculares, se faz necessária uma fonte ampla e confiável de dados onde o pesquisador poderá conhecer pontualmente o organismo no qual está trabalhando. O sequenciamento e anotação do genoma são fundamentais para um entendimento mais aprofundado de um determinado ser, onde a falta destas informações implica na inviabilidade ou dificuldade no desenvolvimento das tecnologias citadas.

1.2 Sistema de sequenciamento

"*Next-generation sequencing*" ou plataformas de nova geração são métodos de sequenciamento capazes de gerar milhões de leituras de pequenas, com elevada acurácia e cobertura apresentando ainda redução de custo e tempo de sequenciamento, gerando em poucas horas a mesma quantidade de dados que seria gerada por centenas de sequenciadores do tipo capilar. Dentre esta novas plataformas podemos citar SOLiD™ (Applied Biosystems), 454 GS FLX (Roche) e Illumina (Genome Analyzer). Diversas problemáticas surgiram com estas novas tecnologias, dentre elas, realizar montagem de genomas com leituras curtas em regiões repetitivas e de baixa complexidade e também a grande quantidade de dados a serem processados (METZKER, 2010;SCHUSTER, 2008).

No sistema SOLiD™, em sua verão 4, é possível obter até 300 GB por rodada, e leituras com tamanho de 35 pb a 50 pb, possuindo uma acurácia máxima de 99,94%. Esta plataforma apresenta três metodologias para construção de bibliotecas genômicas: *fragments*, *mate-pair* e *paired-end*. A escolha do tipo de biblioteca é imprescindível, pois altera completamente a performance do equipamento em tamanho de leituras, tempo de sequenciamento e total de bases lidas (Tabela 1) (APPLIED BIOSYSTEMS® SOLID™ 4 SYSTEM, 2010). Seu sistema de sequenciamento emprega adaptadores ligados aos fragmentos de DNA a ser sequenciado, similar a outras plataformas de nova geração. Posteriormente esses fragmentos com adaptadores são colocados em uma emulsão, juntamente com esferas magnéticas (*beads*), e submetidos a uma homogeneização, onde espera-se que cada fragmento de DNA com os adaptadores se ligue a uma *bead* em uma gotícula de água. Logo após, uma PCR em emulsão é realizada e uma amplificação deste fragmento DNA ocorre, cobrindo assim a *bead* com diversas cópias deste fragmento. A PCR em emulsão é depositada sobre uma placa de vidro, podendo conter até oito amostras diferentes. Diferentemente de outras plataformas, o SOLiD™ utiliza a enzima DNA ligase para sequenciar o fragmento de DNA amplificado e sondas que possuem duas bases específicas (di-base), três degeneradas e um fluoróforo. O processo de sequenciamento pode ser teoricamente dividido em três partes, para melhor compreensão sendo elas: iniciação, amplificação e troca de primer.

Iniciação: um primer universal é ligado ao adaptador do fragmento de DNA mais próximo a *bead*, sendo que, este estenda-se até o final do tamanho do adaptador (Figura 1.1).

Amplificação: A DNA ligase, a partir de um primer universal, atua unindo as sondas onde as duas bases específicas parearam com o fragmento de DNA ligado a *bead* (sequência molde)(Figura 1.1), no momento da ligação pela enzima ligase o fluoróforo é liberado emitindo um sinal luminoso (Figura 1.2) que é capturado pelo sistema óptico do sequenciador. As três bases degeneradas permanecem ao lado das duas bases específicas (Figura 1.4), assim a ligase vai atuando até o final da sequencia molde (Figura 1.5).

Troca de primer: Ao final deste ciclo de ligações, o primer universal e a fita formada pela enzima ligase são removidos. Um novo primer universal contendo uma base a menos é ligada à sequencia molde (Figura 1.6) e todo processo de amplificação é repetido (Figura 1.7). A troca de primer ocorre cinco vezes, isso é necessário para ler todas as bases da sequencia molde (Figura 2).

Cada sinal fluorescente, liberado com o processo de ligação, representa a leitura de duas bases para cada cor (Figura 3). Ao final do processo, um arquivo com as cores lidas (csfasta) e um arquivo de qualidade Phred é gerado, contendo as informações de cada di-base.

Tabela 1. Performance do sequenciador SOLiD^{TM} 4 de acordo com o tipo de biblioteca genômica.

Tipo de biblioteca	Tamanho de leitura	Dias de sequenciamento	Total de bases geradas
“Mate-Paired”	2 x 35 pb	8 - 9	50 – 70 GB
	2 x 50 pb	12 – 16	80 – 100 GB
“Paired-End”	50 pb x 25 pb	11 – 13	55 – 70 GB
“Fragment”	1 x 35 pb	3.5 – 4.5	25 – 35 GB
	1 x 50 pb	6 – 8	40 – 50 GB

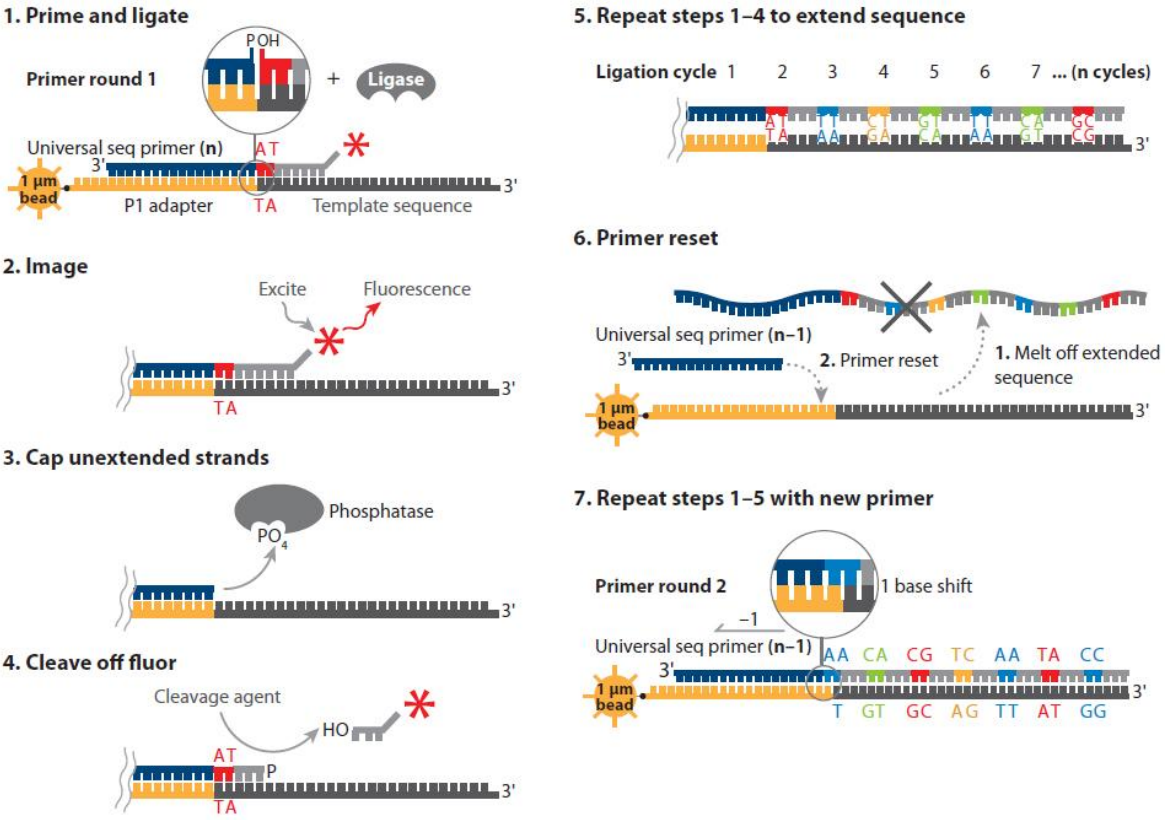

Figura 1. Processo de sequenciamento SOLiD^{TM} System (MARDIS, 2008)

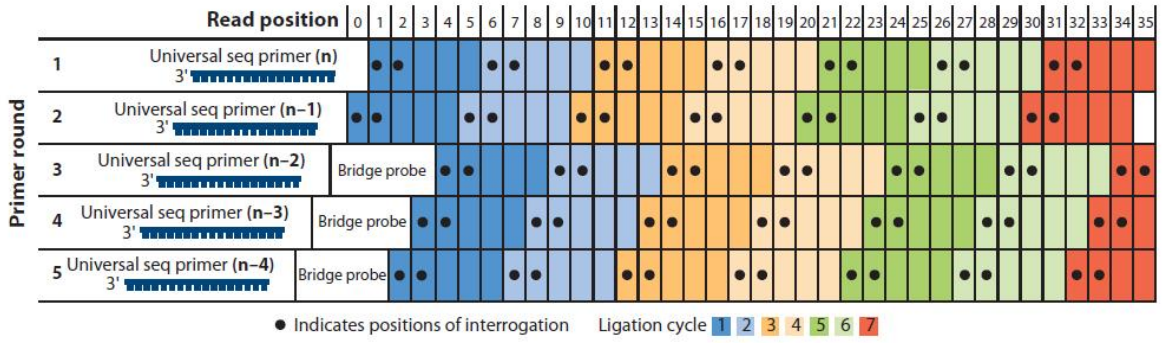

Figura 2. Representação das bases lidas de acordo com a troca de primer (MARDIS, 2008)

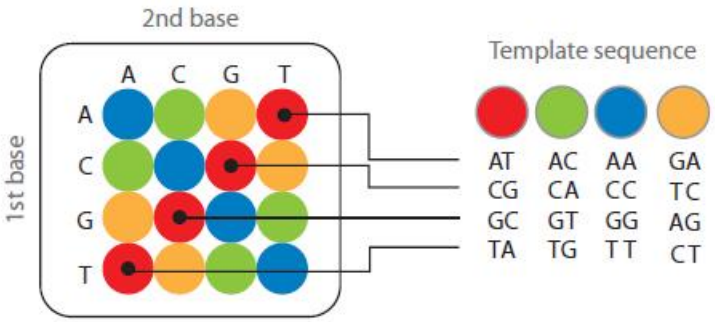

Figura 3. Código de cores que representa cada par de bases lido (MARDIS, 2008)

1.3 Montagem do genoma

O processo de montagem consiste em unir as leituras geradas pelo sequenciamento, levando em consideração a identidade entre elas. Para isso, as sequências são estendidas por meio de sobreposição da extremidade inicial de uma com a extremidade final da outra, até obter a reconstituição da sequência original (SCHUSTER, 2008). A união das leituras obtidas pelo sequenciamento é alinhada através de identidade formando uma sequência maior denominada “contig”. Para o processo de montagem, podemos seguir duas abordagens: *de novo* (*ab initio*) e montagem por referência (MILLER; KOREN; SUTTON, 2010).

1.3.1 Montagem *ab initio*

A montagem *ab initio* leva em consideração a identidade entre as leituras obtidas no sequenciamento. O processo baseia-se em um alinhamento onde o tamanho das sobreposições é definido como *k-mer*, gerando “contigs”. Os algoritmos montadores de genomas NGS baseiam-se geralmente em grafos, onde os vértices representam leituras e os arcos a sobreposição entre as leituras. Podemos dividir os montadores de genomas em: “Greedy graph”, caminho Euleriano e “Overlap-layout-consensus” (MILLER; KOREN; SUTTON, 2010).

1.3.1.1 Greedy graph

Nesta abordagem uma leitura deve-se alinhar com outra leitura com o melhor alinhamento possível. O processo é repetido até que todas as combinações sejam testadas, e a estratégia para formar o grafo leva em conta somente o tamanho das sobreposições entre as leituras (MILLER; KOREN; SUTTON, 2010).

1.3.1.2 Caminho Euleriano

Nos algoritmos baseados no caminho Euleriano ou grafo de Bruijn as leituras são fragmentas n-mers, onde cada fragmento ou n-mer representa um pedaço da leitura original. Com os fragmentos é montado um grafo de Bruijn, onde cada aresta corresponde a um fragmento da leitura original. O nó de origem corresponde ao prefixo menos uma base da região de sobreposição e seu nó de destino ao sufixo menos uma base da região de sobreposição. A reconstrução da fita original é formada a partir do caminho que percorre todas as pontes somente uma vez. Esse método requer servidores com grande capacidade de memória (Figura 4) (MILLER; KOREN; SUTTON, 2010;ZERBINO; BIRNEY, 2008).

1.3.1.3 *Overlap-layout-consensus (OLC)*

Este método baseia-se em grafos de sobreposição. O processo pode ser dividido em três etapas: identificação das sobreposições, geração do grafo e alinhamento das sequências. Na identificação de sobreposição, as leituras são alinhadas em “pair-wise” ou par a par, onde o k-mer é calculado para todas as leituras, o alinhamento é criado através das leituras que compartilham as melhores sobreposições. O OLC é utilizado amplamente para os dados de sequenciamento pelo método de Sanger, mas atualmente montadores para leituras de NGS também estão utilizando este método (Figura 4) (HERNANDEZ et al., 2008;MILLER; KOREN; SUTTON, 2010).

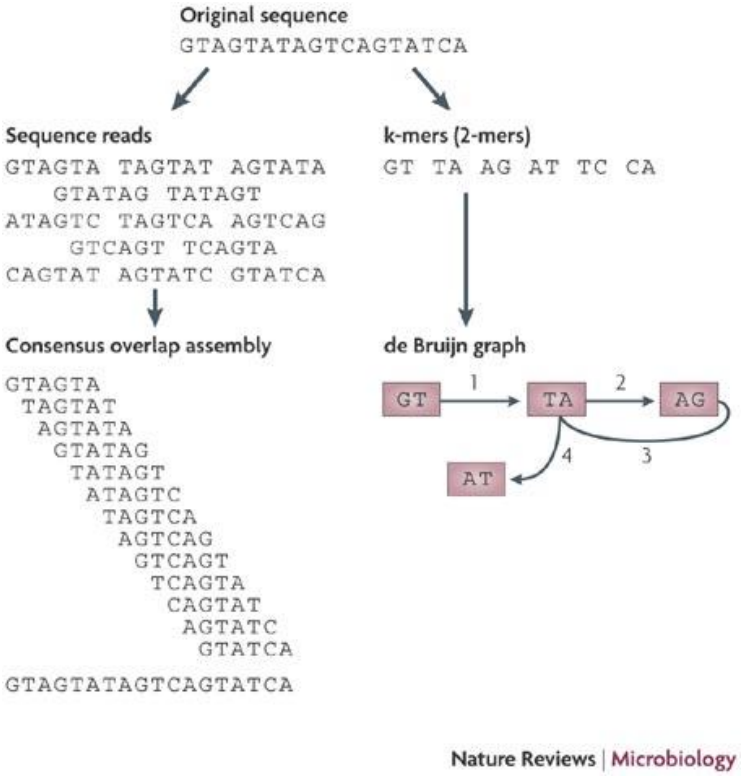

Figura 4. Representação dos modelos OLC e grafo de Bruijn (MACLEAN; JONES; STUDHOLME, 2009)

1.3.2 Montagem por referência

Basicamente as leituras são alinhadas contra um genoma de referência, que deve ser escolhido por proximidade filogenética em relação ao organismo sequenciado. O alinhamento leva em consideração a identidade entre leituras e referência, permitindo lacunas (“gaps”) e bases não idênticas entre leitura e referência (“mismatches”). Um problema observado nesta abordagem são regiões repetitivas no genoma de referência, onde, independente do número de repetições, todas serão mapeadas.

1.4 Anotação do genoma

O processo de anotação de um genoma consiste em atribuir o máximo de informação a um genoma. Inicialmente possíveis regiões codificadoras (ORF) são preditas, podendo levar em conta cálculos *ab initio,* similaridade com outro organismo filogeneticamente aparentado ou simplesmente procurar por um códon de

iniciação e outro de terminação com um numero “n” de bases entre eles. Esses métodos são usualmente fundidos, aumentado assim a confiabilidade da predição de ORFs. Ao terminar o processo de localização das ORFs, diversas predições podem ser feitas de forma individual a cada uma das regiões codificadoras, como por exemplo, predições de RNA transportadores e ribossomais. Da mesma forma individualizada é possível atribuir funções com base em bancos de dados específicos, levando em consideração a similaridade entre a sequência da ORF do novo organismo e do organismo já sequenciado (AZIZ et al., 2008).

2 OBJETIVOS

2.1 Objetivo geral

Sequenciar e anotar o genoma do organismo *Leptospira borgpetersenii* sorogrupo Ballum cepa 4E.

2.2 Objetivos específicos

- Obter o código genético através do sequenciamento pelo método Applied Biosystems SOLiD™ 4.
- Montar a sequência do genoma com o auxilio de softwares especializado.
- Anotar o genoma com auxilio de softwares especializados.
- Conferir manualmente as informações apontadas pelo software.
- Publicar as sequências em bancos de dados on-line.

3 MATERIAIS E MÉTODOS

3.1 Cultivo e extração DNA

A cepa utilizada neste trabalho (DA SILVA et al., 2010) foi cultivada em meio EMJH (Difco) enriquecido com suplemento (Difco) a 28 °C durante 7 dias e seu crescimento acompanhado através de contagem em câmara de Petroff-Hausser. A cultura obtida passou por análise sorológica, com base em painel com anti-soro de coelho e análise genética, com o sequenciamento do gene 16S (DA SILVA et al., 2010). Após a confirmação, foi feita a extração de DNA genômico utilizando o protocolo adaptado do kit Bacterial Genomic DNA purification (Invitrogen). Foi realizada uma quantificação da concentração do DNA genômico através de um espectrofotômetro, para isto, o DNA foi diluído em água com fator de diluição 1:500, logo após uma leitura no espectrofotômetro com comprimento de onda a 260nm foi realizada e aplicada a formula, (resultado da leitura) x 50 x fator de diluição, onde o resultado da quantificação é obtido em $\mu g/ml^{-1}$.

3.2 Sequenciamento

O sequenciamento da cepa foi realizando na Universidade Federal do Pará, utilizando o método Applied Biosystems SOLiD 5500, com base em bibliotecas do tipo Fragment library utilizando o kit SOLiD™ Fragment Library Construction e seguindo protocolo do fabricante.

3.3 Montagem

3.3.1 Qualidade das sequências

Todas as sequências passaram por um filtro de qualidade onde todas as leituras que obtiveram a média da qualidade inferior a phred 20 foram descartadas. Para isso o software Quality Assessment foi utilizado (RAMOS et al., 2011).

3.3.2 Correção dos erros

As leituras que não foram descartadas pelo filtro phred 20 foram submetidas à correção de possíveis erros de sequenciamento através do algoritmo Saet (http://solidsoftwaretools.com).

3.3.3 Montagem *ab initio*

Para montagem *ab initio* foram utilizadas as estratégias “Overlap-layout consensus” através do software Edena (HERNANDEZ et al., 2008) e “de Bruijn graph” através do software Velvet (ZERBINO; BIRNEY, 2008), cuja complementaridade dos resultados foi observada anteriormente por Hernandez e colaboradores e vem sendo utilizada com sucesso na montagem de genomas bacterianos (CERDEIRA et al., 2011). Os melhores resultados gerados por ambos os programas foram unidos em um único arquivo e submetido ao programa Simplifier (Rommel Ramos, dados não publicados) removendo possíveis redundâncias geradas pela concatenação dos dados da montagem *ab initio*.

3.3.4 Mapeamento dos contigs

Para reconstruir a sequência original as contigs geradas ao final do processo de montagem *ab initio* foram orientada através do software G4All (Rommel Ramos, dados não publicados) utilizando como referência o genoma da bactéria *L. borgpetersenii* serovar Hardjo-bovis cepa L550 (BULACH et al., 2006).

3.3.5 Correção das gaps

Com a orientação das contigs, algumas lacunas (“gaps”) foram observadas na fita que reconstitui a sequência original. Essas regiões poderiam ser geradas por erros de sequenciamento, baixa cobertura, erros na montagem ou simplesmente por não existir no novo organismo sequenciado. Uma lista destas regiões foi gerada e cada uma das “gaps” foi revisada individualmente de forma manual com o auxilio do software CLC Genomics Workbench (http://www.clcbio.com) utilizando a abordagem IMAGE (TSAI; OTTO; BERRIMAN, 2010).

3.4 Anotação funcional

A localização das ORFs foi realizada com o programa FgenesB (http://linux1.softberry.com/) e para anotação do genoma o Square DNA annotator (Marcus Redü Eslabão, dados não publicados) foi utilizado. Este *pipe line* utiliza os bancos de dados BLAST nr e Swiss-prot para caracterizar as regiões codificadoras. Para aferir a anotação do Square, o algoritmo RAST (AZIZ et al., 2008) foi empregado. As predições de RNAs transportadores e RNAs ribossomais foram realizadas pelos programas tRNAscan-SE (LOWE; EDDY, 1997) e RNAmmer (LAGESEN et al., 2007), respectivamente. Para a visualização e edição do genoma foi utilizado o programa Artemis (RUTHERFORD et al., 2000).

3.5 Comparação do tamanho do genoma

Para comparar o tamanho do genoma da cepa deste trabalho com a cepa de referência, foi feito um alinhamento da sequencia de DNA da cepa L550 contra cepa 4E, através do site Webact (ABBOTT et al., 2005). Os resultados foram carregados no software ACT(CARVER et al., 2005), onde os genes da cepa L550 foram sobrepostos a sua sequencia de DNA, e um filtro foi aplicado, onde somente os genes da cepa L550 cujo sequencia de nucleotídeos não estavam presentes na cepa 4E foram selecionados.

O total de bases contidas em regiões codificantes e o total de bases não contidas em regiões codificantes foram analisados com o auxilio do software Artemis (RUTHERFORD et al., 2000).

4 RESULTADOS

O processo de sequenciamento com o método Applied Biosystems SOLiD™ 4 System gerou um total de 85.302.595 leituras com 50 pb cada, totalizando mais de quatro bilhões de pares de base lidos. Para gerar esta quantia de dados, um tempo de aproximadamente quinze horas é necessário neste modelo de sequenciador, porém em uma única rodada são sequenciados até oito genomas procariotos o que aumentou o tempo total deste sequenciamento para 6 dias.

Os dois arquivos gerados no sequenciamento, um contendo as sequencias e outro referente à qualidade de cada uma das bases, foram submetidos ao filtro Phred 20 (mais de 99% de precisão). Após este processo restaram 52.973.349 leituras, ou seja, mais de 37% de leituras foram descartadas por baixa qualidade.

No processo de montagem *ab initio,* com o montador Velvet, e com os parâmetros *coverage cutoff* 11, k 11 e *expected coverage* 260, foram gerados 5.148 contigs com mediana (N50) de 1.033 pb. Enquanto que com o software Edena, o com os parâmetros overlap 33, coverage cutoff 11 e depth 18, foram gerados 9.891 contigs com N50 de 605 pb.

Um relatório de problemas foi gerado, incluindo *gaps* e *frameshifts*, para primeira versão onde constatou-se um total de 4823 “gaps” no cromossomo 1 e 98 “gaps” no cromossomo 2. Após a revisão manual o número de “gaps” para o cromossomo 1 caiu para 362 e para o cromossomo 2 caiu para 51.

Ao final do processo o cromossomo maior apresentou o tamanho de 3.071.053 pb, 40,58% de conteúdo GC, 36 tRNA, 4 rRNA e 2908 ORFs. Para o cromossomo menor o total de bases foi de 305.940 pb, conteúdo GC de 40,25%, nenhum tRNA, nenhum rRNA e 277 ORFs.

O alinhamento dos genomas das cepas L550 e 4E (Figura 5) resultou na constatação de que 99 genes presentes no cromossomo maior da cepa L550 não estão presentes no cromossomo maior da cepa 4E, sendo o total de bases contidas nestes genes de 91.266 pb. Estes 99 genes foram classificados através de seus produtos (Figura 6), onde a maior parte foi identificada como hipotética ou transposase das famílias ISLbp1 e *IS1477*, os demais genes que apresentavam produto identificado e/ou nome de gene foram citados (Tabela 2). A análise da quantidade de bases contidas em regiões codificantes e não codificantes das cepas L550 e 4E resultaram em uma redução de 394.785 pb na região não codificante da cepa 4E em relação a região não codificante da cepa L550.

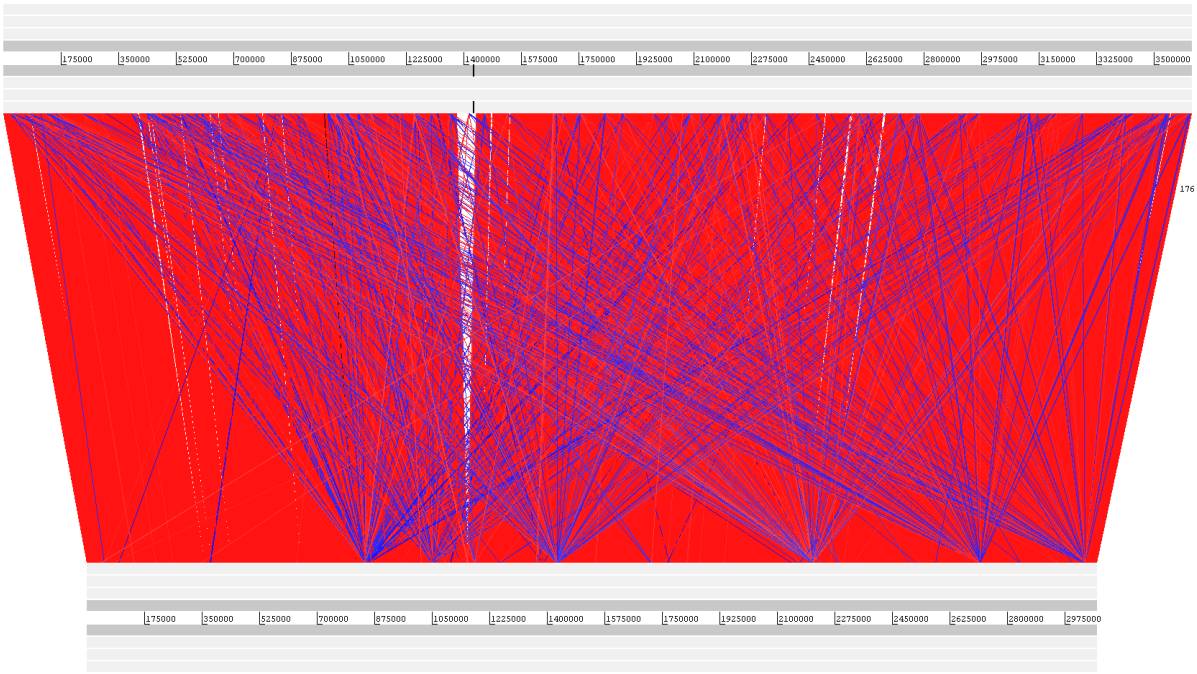

Figura 5. Representação da sintenia gerada pelo software Webact. Na parte superior está representado o genoma da cepa L550 e na parte inferior a cepa 4E. As barras em vermelho representam regiões idênticas, as linhas azuis regiões com sequências invertidas e as regiões brancas corresponde a ausência de sequências.

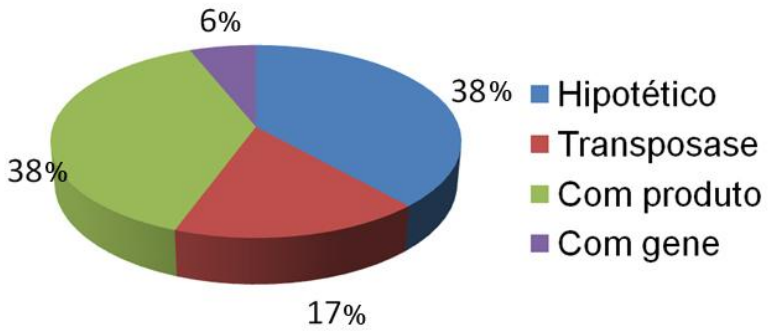

Figura 6. Categoria dos 99 genes contidos no cromossomo maior da *L. borgpetersenii* cepa L550 que não estão contidos no cromossomo maior da *L. borgpetersenii* cepa 4E. Classificados de acordo com o produto anotado no genoma.

Tabela 2 – Genes que contém produto identificado e estão presentes no cromossomo maior da *L. borgpetersenii* cepa L550 que não estão presentes no cromossomo maior da *L. borgpetersenii* cepa 4E.

Identificação	Descrição
LBL_0360	Lipoprotein
LBL_0586	Lipoprotein
LBL_0695	AraC family transcription regulator
LBL_1085	RNA-directed DNA polymerase polymerase
LBL_1086	Transcriptional regulator
LBL_1087	Transcriptional regulator
LBL_1166	Alcohol dehyodrogenase
LBL_1167	Glycosyltransferase
neuB-2	N-acetylneuraminic acid (sialicacid) synthetase
LBL_1169	Cytidylyltransferase
LBL_1171	Carbamoyl transferase
LBL_1172	Pyridoxal phosphate-dependent aminotransferase
LBL_1173	Carbamoyl transferase
LBL_1174	Dehydrogenase
LBL_1175	Pyridoxal-phosphate-dependent aminotransferase
LBL_1177	Acetyltransferase
LBL_1178	Zinc-binding dehydrogenase
LBL_1181	Methylase/methyltransferase
LBL_1182	Aminopeptidase
LBL_1183	Cytidylyltransferase
LBL_1185	Short chain dehydrogenase
LBL_1186	Aryl-alcohol dehydrogenase-related oxidoreductase
LBL_1187	N-acetylneuraminic acid (sialicacid) synthetase
LBL_1190	Pyridoxal-phosphate-dependent aminotransferase
LBL_1191	ABC transporter permease/ATP-binding protein
nagB	Glucosamine-6-phosphate deaminase
gmhA-2	Phosphoheptose isomerase
LBL_1197	2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase
kdsB-2	3-deoxy-manno-octulosonate cytidylyltransferase
LBL_1199	Oxidoreductase family protein
LBL_1200	Inositol monophosphatase family protein
LBL_1206	Methylase/methyltransferase
LBL_1207	Dehydrogenase
LBL_1210	PP-loop superfamily protein
hisH	Imidazole glycerol phosphate synthase subunit
hisF-2	Imidazoleglycerol phosphate synthase subunit
LBL_1214	Nucleoside-diphosphate-sugar epimerase
LBL_1215	Nucleoside-diphosphate-sugar pyrophosphorylase
LBL_1216	Glucose galactose epimerase
LBL_1217	Sugar oxidoreductase
LBL_1218	N-acetyl glucosamine/N-acetyl galactosamine epimerase
LBL_1219	UDP-N-acetylglucosamine 2-epimerase
LBL_1220	Glycosyltransferase
LBL_2192	PilT domain-containing protein

5 DISCUSSÃO

Dos seis genomas disponíveis para o gênero *Leptospira,* dois são da espécie *L. borgpetersenii*. O trabalho que anunciou o sequenciamento das cepas L550 e JB197 mostrou uma redução significativa no tamanho do genoma de *L. borgpetersenii* comparado à *L. interrogans (NASCIMENTO et al., 2004)*. Diversas evidências demonstram que a espécie *L. borgpetersenii* está passando por um processo de redução do seu genoma, mediado por elementos de sequência de inserção (IS) (BULACH et al., 2006). Estes elementos garantem grande plasticidade ao genoma procarioto, sendo capazes de diversos tipos de rearranjos como, por exemplo, deleções, inversões e fusões de replicon (MAHILLON; LEONARD; CHANDLER, 1999). Seguindo o padrão da espécie, o novo genoma da cepa 4E também apresentou uma grande redução em relação ao genoma de *L. interrogans,* e uma redução de 543.393 pb do cromossomo maior em relação à cepa de referência.

Entre as hipóteses que foram levantadas para explicar a redução do genoma, podemos citar a falta de cobertura no genoma, que foi rapidamente descartada, pois foram geradas mais de 56 milhões de leituras com 50 pb de extensão cada e qualidade superior a 99%, totalizando mais de 2,8 bilhões de pares de bases lidos. Estatisticamente esta cobertura representa aproximadamente 800 vezes o tamanho do genoma. Outra hipótese seria erros na montagem, para isto duas montagens *ab initio* foram utilizadas, utilizando abordagens diferentes e complementares entre si, o que reduz significativamente a chance de erros de montagem. E a última hipótese os elementos móveis do genoma, que demonstra ser a mais plausível para explicar este processo de redução, onde além destes elementos estarem presente nos dois genomas conhecidos de *L. borgpetersenii*, uma rápida consulta ao banco de dados ISfinder (SIGUIER et al., 2006), revelou a presença de diversos elementos IS, que precisam ser confirmados individualmente. Porém a análise da diferença de quantidade de bases na região não codificante do cromossomo maior novo genoma e do genoma de referência revelaram uma redução de 394.785 pb na região não codificante da 4E em relação a L550, que explica de onde ocorreu a maior parte da redução deste genoma.

O número de *gaps* restante mesmo após a curadoria manual da montagem também pode ser explicado pelo elevado número de elementos móveis no genoma, porém alguns destes possíveis *gaps* foram gerados com base na predição de genes, onde os genes que apresentavam *frameshitf* também foram incluídos no relatório de problemas. Levando em consideração os genomas já sequenciados da *L. borgpetersenii* o número próximo de 250 pseudogenes é observado, boa parte das gaps contidas no relatório de problemas deste trabalho pode representar apenas pseudogenes, mesmo assim, todos os problemas passarão por uma revisão pontual, para confirmação desta hipótese, onde será feito o alinhamento de todas as leituras do novo organismo contra cada um dos pseudogenes da *Leptospira* de referência.

Após a confirmação de todas as hipóteses para explicar a redução deste genoma, será necessária a compreensão do que foi perdido em relação a outras leptospiras, Para isso, todas as sequências serão submetidas ao software Blast2go (CONESA et al., 2005). Os dados gerados contribuirão para um entendimento da distribuição dos genes dentro de determinadas funções biológicas, podendo ser comparado com os dados já publicados referentes às demais leptospiras. Além da distribuição dos genes, é possível comparar filogeneticamente o genoma deste trabalho com as espécies já sequenciadas, observar rotas metabólicas e mecanismos de patogenicidade.

O genoma anotado e comparado é uma fonte ampla de dados que contribui para estudos que visam o desenvolvimento de novas tecnologias como vacinas recombinantes e métodos de diagnóstico molecular.

6 CONCLUSÕES

O método Applied Biosystems SOLiD™ 4 permitiu a determinação da sequência do genoma de *L. borgpete*rsenii cepa 4E, com cobertura total de 800 vezes e acurácia de 99,94% na leitura das bases. A estratégia de montagem utilizando duas abordagens, *Overlap-layout consensus* e *Bruijn graph*, permitiram tirar o máximo de proveito da cobertura e precisão gerada pelo sequenciador SOLiD™ 4, devido à complementariedade dos dados gerados por essas abordagens.

O genoma da cepa 4E apresentou uma redução comparada aos outros genomas já sequenciados de *Leptospira,* onde 99 genes e cerca de 394 kb de região não codificantes foram perdidas, onde a principal hipótese para explicar esta redução é o grande número de elementos móveis, e a observação de um processo de redução do genoma na espécie *Borgpetersenii.*

REFERÊNCIAS

ABBOTT, J. C.; AANENSEN, D. M.; RUTHERFORD, K.; BUTCHER, S.; SPRATT, B. G. WebACT--an online companion for the Artemis Comparison Tool. **Bioinformatics.***,* v.21, n.18, p.3665-3666, 2005.

ADLER, B.; DE LA PENA, M. A. *Leptospira* and leptospirosis. **Veterinary microbiology***,* v.140, n.3-4, p.287-296, 2010.

AZIZ, R. K.; BARTELS, D.; BEST, A. A.; DEJONGH, M.; DISZ, T.; EDWARDS, R. A.; FORMSMA, K.; GERDES, S.; GLASS, E. M.; KUBAL, M.; MEYER, F.; OLSEN, G. J.; OLSON, R.; OSTERMAN, A. L.; OVERBEEK, R. A.; MCNEIL, L. K.; PAARMANN, D.; PACZIAN, T.; PARRELLO, B.; PUSCH, G. D.; REICH, C.; STEVENS, R.; VASSIEVA, O.; VONSTEIN, V.; WILKE, A.; ZAGNITKO, O. The RAST Server: rapid annotations using subsystems technology. **BMC.Genomics***,* v.9, p.75, 2008.

BULACH, D. M.; ZUERNER, R. L.; WILSON, P.; SEEMANN, T.; MCGRATH, A.; CULLEN, P. A.; DAVIS, J.; JOHNSON, M.; KUCZEK, E.; ALT, D. P.; PETERSON-BURCH, B.; COPPEL, R. L.; ROOD, J. I.; DAVIES, J. K.; ADLER, B. Genome reduction in *Leptospira borgpetersenii* reflects limited transmission potential. **Proc.Natl.Acad.Sci.U.S.A***,* v.103, n.39, p.14560-14565, 2006.

CARVER, T. J.; RUTHERFORD, K. M.; BERRIMAN, M.; RAJANDREAM, M. A.; BARRELL, B. G.; PARKHILL, J. ACT: the Artemis Comparison Tool. **Bioinformatics.***,* v.21, n.16, p.3422-3423, 2005.

CERDEIRA, L. T.; CARNEIRO, A. R.; RAMOS, R. T.; DE ALMEIDA, S. S.; D'AFONSECA, V.; SCHNEIDER, M. P.; BAUMBACH, J.; TAUCH, A.; MCCULLOCH, J. A.; AZEVEDO, V. A.; SILVA, A. Rapid hybrid de novo assembly of a microbial genome using only short reads: *Corynebacterium pseudotuberculosis* I19 as a case study. **J.Microbiol.Methods***,* v.86, n.2, p.218-223, 2011.

CONESA, A.; GOTZ, S.; GARCIA-GOMEZ, J. M.; TEROL, J.; TALON, M.; ROBLES, M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics.***,* v.21, n.18, p.3674-3676, 2005.

DA SILVA, E. F.; FELIX, S. R.; CERQUEIRA, G. M.; FAGUNDES, M. Q.; NETO, A. C.; GRASSMANN, A. A.; AMARAL, M. G.; GALLINA, T.; DELLAGOSTIN, O. A. Preliminary characterization of *Mus musculus*-derived pathogenic strains of *Leptospira borgpetersenii* serogroup Ballum in a hamster model. **Am.J.Trop.Med.Hyg.***,* v.83, n.2, p.336-337, 2010.

DELLAGOSTIN, O. A.; GRASSMANN, A. A.; HARTWIG, D. D.; FELIX, S. R.; DA SILVA, E. F.; MCBRIDE, A. J. Recombinant vaccines against Leptospirosis. **Human Vaccines***,* v.7, n.11, p.1215-1224, 2011.

HERNANDEZ, D.; FRANCOIS, P.; FARINELLI, L.; OSTERAS, M.; SCHRENZEL, J. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. **Genome research***,* v.18, n.5, p.802-809, 2008.

LAGESEN, K.; HALLIN, P.; RODLAND, E. A.; STAERFELDT, H. H.; ROGNES, T.; USSERY, D. W. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. **Nucleic Acids Researsh***,* v.35, n.9, p.3100-3108, 2007.

LEVETT, P. N. Leptospirosis. **Clinical Microbiology Reviews***,* v.14, n.2, p.296-326, 2001.

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

MACLEAN, D.; JONES, J. D.; STUDHOLME, D. J. Application of 'next-generation' sequencing technologies to microbial genetics. **Nature Reviews Microbiology***,* v.7, n.4, p.287-296, 2009.

MAHILLON, J.; LEONARD, C.; CHANDLER, M. IS elements as constituents of bacterial genomes. **Researsh Microbiology***,* v.150, n.9-10, p.675-687, 1999.

MARDIS, E. R. Next-generation DNA sequencing methods. **Annual Review Of Genomics And Human Genetics***,* v.9, p.387-402, 2008.

METZKER, M. L. Sequencing technologies - the next generation. **Nature Reviews Microbiology***,* v.11, n.1, p.31-46, 2010.

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

MURHEKAR, M. V.; SUGUNAN, A. P.; VIJAYACHARI, P.; SHARMA, S.; SEHGAL, S. C. Risk factors in the transmission of leptospiral infection. **Indian Journal of Medical Research***,* v.107, p.218-223, 1998.

NASCIMENTO, A. L.; KO, A. I.; MARTINS, E. A.; MONTEIRO-VITORELLO, C. B.; HO, P. L.; HAAKE, D. A.; VERJOVSKI-ALMEIDA, S.; HARTSKEERL, R. A.; MARQUES, M. V.; OLIVEIRA, M. C.; MENCK, C. F.; LEITE, L. C.; CARRER, H.; COUTINHO, L. L.; DEGRAVE, W. M.; DELLAGOSTIN, O. A.; EL-DORRY, H.; FERRO, E. S.; FERRO, M. I.; FURLAN, L. R.; GAMBERINI, M.; GIGLIOTI, E. A.; GOES-NETO, A.; GOLDMAN, G. H.; GOLDMAN, M. H.; HARAKAVA, R.; JERONIMO, S. M.; JUNQUEIRA-DE-AZEVEDO, I. L.; KIMURA, E. T.; KURAMAE, E. E.; LEMOS, E. G.; LEMOS, M. V.; MARINO, C. L.; NUNES, L. R.; DE OLIVEIRA, R. C.; PEREIRA, G. G.; REIS, M. S.; SCHRIEFER, A.; SIQUEIRA, W. J.; SOMMER, P.; TSAI, S. M.; SIMPSON, A. J.; FERRO, J. A.; CAMARGO, L. E.; KITAJIMA, J. P.; SETUBAL, J. C.; VAN SLUYS, M. A. Comparative genomics of two *Leptospira interrogans* serovars reveals novel insights into physiology and pathogenesis. **Journal of Bacteriology***,* v.186, n.7, p.2164-2172, 2004.

PICARDEAU, M.; BULACH, D. M.; BOUCHIER, C.; ZUERNER, R. L.; ZIDANE, N.; WILSON, P. J.; CRENO, S.; KUCZEK, E. S.; BOMMEZZADRI, S.; DAVIS, J. C.; MCGRATH, A.; JOHNSON, M. J.; BOURSAUX-EUDE, C.; SEEMANN, T.; ROUY, Z.;

COPPEL, R. L.; ROOD, J. I.; LAJUS, A.; DAVIES, J. K.; MEDIGUE, C.; ADLER, B. Genome sequence of the saprophyte *Leptospira biflexa* provides insights into the evolution of *Leptospira* and the pathogenesis of leptospirosis. **PLoS.One.**, v.3, n.2, p.e1607, 2008.

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

RUTHERFORD, K.; PARKHILL, J.; CROOK, J.; HORSNELL, T.; RICE, P.; RAJANDREAM, M. A.; BARRELL, B. Artemis: sequence visualization and annotation. **Bioinformatics.**, v.16, n.10, p.944-945, 2000.

SCHUSTER, S. C. Next-generation sequencing transforms today's biology. **Nature Methods**, v.5, n.1, p.16-18, 2008.

SIGUIER, P.; PEROCHON, J.; LESTRADE, L.; MAHILLON, J.; CHANDLER, M. ISfinder: the reference centre for bacterial insertion sequences. **Nucleic Acids Researsh**, v.34, n.Database issue, p.D32-D36, 2006.

SILVA, E. F.; BROD, C. S.; CERQUEIRA, G. M.; BOURSCHEIDT, D.; SEYFFERT, N.; QUEIROZ, A.; SANTOS, C. S.; KO, A. I.; DELLAGOSTIN, O. A. Isolation of *Leptospira noguchii* from sheep. **Veterinary microbiology**, v.121, n.1-2, p.144-149, 2007.

SILVA, E. F.; CERQUEIRA, G. M.; SEYFFERT, N.; SEIXAS, F. K.; HARTWIG, D. D.; ATHANAZIO, D. A.; PINTO, L. S.; QUEIROZ, A.; KO, A. I.; BROD, C. S.; DELLAGOSTIN, O. A. *Leptospira noguchii* and human and animal leptospirosis, Southern Brazil. **Emerging Infectious Diseases**, v.15, n.4, p.621-623, 2009.

TSAI, I. J.; OTTO, T. D.; BERRIMAN, M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. **Genome Biology**, v.11, n.4, p.R41, 2010.

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