

UNIVERSIDADE FEDERAL DE PELOTAS

Programa de Pós-Graduação em Biotecnologia



Dissertação

**Genix: Desenvolvimento de uma nova *pipeline*
automatizada para anotação de genomas
microbianos**

Frederico Schmitt Kremer

Pelotas, 2016

Frederico Schmitt Kremer

Genix: Desenvolvimento de uma nova *pipeline* automatizada para anotação de genomas microbianos

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

Orientador: Dr. Luciano da Silva Pinto

Comissão de Acompanhamento: Alan John Alexander McBride

Luciano Carlos da Maia

Pelotas, 2016

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação na Publicação

K92g Kremer, Frederico Schmitt

Genix: desenvolvimento de uma nova pipeline automatizada para anotação de genomas microbianos / Frederico Schmitt Kremer ; Luciano da Silva Pinto, orientador. — Pelotas, 2016.

59 f. : il.

Dissertação (Mestrado) — Programa de Pós-Graduação em Biotecnologia, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2016.

1. Bioinformática. 2. Ngs. 3. Servidor web. 4. Genômica microbiana. 5. Lamp. I. Pinto, Luciano da Silva, orient. II. Título.

CDD : 575.1

Elaborada por Maria Beatriz Vaghetti Vieira CRB: 10/1032

BANCA EXAMINADORA

Prof. Dr. Luciano da Silva Pinto (UFPel, CDTec)

Prof. Dr. Alan John Alexander McBride (UFPel, CDTec)

Prof. Dr. Luciano Carlos da Maia (UFPel, FAEM)

Agradecimentos

Acima de tudo, à minha mãe, Sandra, meu pai, Roberto (que Deus o tenha), e meu irmão, Oscar, por todo apoio, carinho e incentivo dado.

À minha namorada, Martina, pelo carinho, apoio, compreensão, *sabores do rei*, cervejas gourmet e *yoshakes*.

Aos meus amigos “dos tempos de CEFET”, Luiz Gustavo, Felipe, Cássia, Bruno (Bolaxa), Thais, Gabriel e Rai, pelas discussões sobre assuntos aleatórios, mesas de RPG e nerdices em geral.

Ao meu orientador, Professor Dr. Luciano Pinto, pelo apoio, pelos conhecimentos passados desde o meu período de iniciação científica, e por apostar no presente projeto.

À Universidade Federal de Pelotas e ao Programa de Pós-Graduação em Biotecnologia, pela estrutura que permitiu o desenvolvimento do presente trabalho.

Aos meus colegas de laboratório, Marcus, Julia, Rafael, Monize, André e Gabrielle, pelas discussões produtivas, cafés, amendoins, churrascos, piadas e ótimo ambiente de trabalho.

A todos os professores do Programa de Pós-Graduação em Biotecnologia e Bacharelado em Biotecnologia, pelos conhecimentos passados durante a minha formação.

Ao CNPq, pelo apoio financeiro durante o mestrado.

“If the implementation is hard to explain, it’s a bad idea”

The Zen of Python

Resumo

KREMER, Frederico Schmitt. **Genix: Desenvolvimento de uma nova *pipeline* automatizada para anotação de genomas microbianos**. 2015. Dissertação (Mestrado) – Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas, Pelotas.

O advento do sequenciamento de DNA de nova geração (NGS) reduziu significativamente o custo dos projetos de sequenciamento de genomas. Quanto mais fácil é de obter novos dados genômicos, mais acuradas deve ser a etapa de anotação, de forma a se reduzir a perda de informações relevantes e efetuar o acúmulo de erros que possam afetar a acurácia das análises posteriores. No caso dos genomas bacterianos, um grande número de programas para anotação já foi desenvolvido, entretanto, muitos destes softwares não incorporaram etapas para otimizar os seus resultados, como filtragem de proteínas falso-positivas/*spurious* e a anotação mais completa de RNA não-codificantes. O presente trabalho descreve o desenvolvimento do Genix, uma nova *pipeline* automatizada que combina a funcionalidade de diferentes *softwares*, incluindo Prodigal, tRNAscan-SE, RNAmmer, Aragorn, INFERNAL, NCBI-BLAST+, CD-HIT, Rfam e Uniprot, com a intenção de aumentar a afetividade dos resultados de anotação. Para avaliar a acurácia da presente ferramenta, foram usados como modelo de estudo os genomas de referência de *Escherichia coli* K-12, *Leptospira interrogans* cepa Fiocruz L1-130, *Listeria monocytogenes* EGD-e e *Mycobacterium tuberculosis* H37Rv. Os resultados obtidos pelo Genix foram comparados às anotações originais e as obtidas pelas ferramentas de anotação RAST e BASys, considerando genes novos, faltantes e exclusivos, informações de anotação funcional e predições de ORFs *spurious*. De forma a se quantificar o grau de acurácia, uma nova métrica, denominada *discrepância de anotação* foi também proposta. Na análise comparativa o Genix apresentou para todos os genomas o menor valor de discrepância, variando entre 0,96 e 5,71%, sendo o maior valor observado no genoma de *L. interrogans*, para o qual RAST e BASys apresentaram valores superiores a 14,0%. Além disso, foram identificadas proteínas *spurious* nas anotações geradas pelos demais programas, e, em menor número, nas anotações de referência, indicando que a utilização do Antifam permite um melhor controle do número de genes falso positivos. A partir dos testes realizados, foi possível demonstrar que o Genix é capaz de gerar anotação com boa acurácia (baixa discrepância), menor perda de genes relevantes (funcionais) e menor número de genes falso positivos.

Palavras-Chave: Bioinformática, NGS, Servidor Web, Genômica Microbiana, LAMP

Abstract

KREMER, Frederico Schmitt. **Genix: Development of a new automated pipeline for microbial genome annotation**. 2015. Dissertação (Mestrado) – Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas, Pelotas.

The advent of next-generation sequencing (NGS) significantly reduced the cost of genome sequencing projects. The easier it is to generate genomic data, the more accurate the annotation steps must be to avoid both the loss of information and the accumulation of erroneous features that may affect the accuracy of further analysis. In the case of bacteria genomes, a range of web annotation software has been developed; however, many applications have not incorporated the steps required to improve the output (eg: false-positive/spurious ORF filtering and a more complete non-coding RNA annotation). The present work describes the implementation of Genix, a new bacteria genome annotation pipeline that combines the functionality of the programs Prodigal, tRNAscan-SE, RNAmmer, Aragorn, INFERNAL, NCBI-BLAST+, CD-HIT, Rfam and UniProt, with the intention of increasing the effectiveness of the annotation results. To evaluate the accuracy of Genix, we used as models of study the reference genomes of *Escherichia coli* K-12, *Leptospira interrogans* strain Fiocruz L1-130, *Listeria monocytogenes* EGD-e and *Mycobacterium tuberculosis* H37Rv. the results obtained by Genix were compared to the original annotation and to those from the annotation pipelines RAST and BASys considering new, missing and exclusive genes, functional annotation information and the prediction of spurious ORFs. To quantify the annotation accuracy, a new metric, called “annotation discrepancy” was developed. In a comparative analysis, Genix showed the smallest discrepancy for the four genomes, ranging for 0.96 to 5.71%, the highest discrepancy was observed in the *L. interrogans* genome, for which RAST and BASys resulted in discrepancies greater than 14.0%. Additionally, several spurious proteins were identified in the annotations generated by RAST and BASys, and, in smaller number, in the reference annotations, indicating that the use of the Antifam database allows a better control of the number of false-positive genes. Based on the evaluations, it was possible to show that Genix is able to generate annotations with good accuracy (low discrepancy), low omission of relevant (functional) genes and a small number of false-positive genes.

Keywords: Bioinformatics, NGS, Webserver, Microbial Genomics, LAMP

Lista de Figuras

Figura 1. Exemplo de anotação do realizada pelo RAST para um genoma de <i>Leptospira interrogans</i>	25
Figura 2. Estrutura da <i>pipeline</i> de anotação do Genix.....	32
Figura 3. Página de registro do Genix.	38
Figura 4. Exemplo de uma página de projeto de anotação em andamento, incluindo um resumo dos parâmetros passados pelo usuário.	39
Figura 5. Exemplo de uma página de projeto de anotação finalizado no Genix, incluindo um resumo dos parâmetros passados pelo usuário, os <i>links</i> para o usuário efetuar o download dos resultados de anotação em diferentes formatos (tbl, gb, gff) e o <i>link</i> para a visualização da anotação através do Jbrowse.....	40
Figura 6. Visualização de uma anotação de genoma pela ferramenta JBrowse integrada ao Genix.	41
Figura 7. Registro protocolado dia 11 de novembro de 2015. Código do protocolo: BR5120150013114	58

Lista de Tabelas

Tabela 1. Opções (argumentos) aceitas pelo <i>script genix_annotation.py</i>	34
Tabela 2. Lista de genomas utilizados para a avaliação da ferramenta Genix	36
Tabela 3. Comparação da anotação original (referência) disponível no Genbank com os resultados obtidos pelas ferramentas Genix, RAST e BASys para o genoma de <i>M. tuberculosis</i> cepa H37Rv.	42
Tabela 4. Comparação da anotação original (referência) disponível no Genbank com os resultados obtidos pelas ferramentas Genix, RAST e BASys para o genoma de <i>E. coli</i> cepa K12.....	43
Tabela 5. Comparação da anotação original (referência) disponível no Genbank com os resultados obtidos pelas ferramentas Genix, RAST e BASys para o genoma de <i>L. interrogans</i> cepa Fiocruz L1-130 (Cromossomo I).	43
Tabela 6. Comparação da anotação original (referência) disponível no Genbank com os resultados obtidos pelas ferramentas Genix, RAST e BASys para o genoma de <i>L. monocytogenese</i> cepa EGD-e.	43

Lista de Abreviaturas

ACT – *Artemis Comparison Tool*
API – *Application Programming Interface*
CDS – *Coding DNA Sequence*
CGI – *Common Gateway Interface*
CM – *Covariance Model*
COG – *Cluster of Orthologous Groups*
DBG – *De Bruijn Graph*
DNA - *Deoxyribonucleic Acid*
GBK – *Genbank* (formato de arquivo)
GFF – *Gene Finding Format*
GO – *Gene Ontology*
HMM – *Hidden Markov Model*
LAMP – *Linux, Apache, MySQL and Python*¹
miRNA – *microRNAs*
MLST – *Multilocus Sequence Typing*
ncRNA – *non-coding RNA*
NGS – *Next Generation Sequencing*
OLC – *Overlap-Layout-Consensus*
ORF – *Open Reading Frame*
RBS – *Ribosome Binding Site*
RNA - *Ribonucleic Acid*
rRNA – *Ribosomal RNA*
SG – *String Graph*
TBL – *Feature Table*

¹ Esta expressão pode ser usada para arquiteturas que usem outros componentes, desde que a letra inicial seja preservada. Desta forma, a “M” pode representar também o banco de dados “MongoDB” e a letra “P” as linguagens Perl e PHP, por exemplo. Exemplos de variações desta abreviação incluem WAMP (Windows, Apache, MySQL and PHP) e LNMP (Linux, NGINX, MongoDB and Python).

tmRNA – *transfer-messenger RNA*

tRNA – *transfer RNA*

Sumário

1 INTRODUÇÃO GERAL	15
2 REVISÃO BIBLIOGRÁFICA	17
2.1 O Sequenciamento de DNA	17
2.2 Sequenciamento e Montagem de Genomas	19
2.3 Anotação de Genes e Genomas	20
2.3.1 Identificação de regiões codificantes (CDS, <i>Coding DNA sequences</i>)	21
2.3.2 Identificação de RNAs não codificantes	21
2.3.2.1 RNAmmer	22
2.3.2.2 tRNAscan-SE e Aragorn	22
2.3.2.3 INFERNAL e Rfam	23
2.3.3 Anotação estrutural	23
2.3.4 Anotação funcional	23
2.3.5 Anotação Automática	24
2.3.5.1 RAST	25
2.3.5.2 BASys	26
2.3.5.3 xBASE	26
2.3.5.4 Prokka	26
2.3.5.5 Eugene-PP	27
2.3.5.6 BG7	27
2.6 Visualização de anotações	27
2.7 Submissão para bancos de dados públicos	28
3 Sequenciamento de Genomas Microbianos	28
3 HIPÓTESE E OBJETIVOS	30
3.1 Hipótese	30
3.2 Objetivo Geral	30

3.3 Objetivos Específicos.....	30
4 metodologia.....	31
4.1 Desenvolvimento.....	31
4.1.1 Visão geral do <i>genix_server.py</i>	31
4.1.2 Construção do banco de dados (<i>get_database.sh</i>)	32
4.1.3 O script <i>genix_annotation.py</i>	33
4.1.4 <i>Front-end</i> e apresentação dos resultados de anotação.....	35
4.2 Comparação com outras ferramentas.....	36
5 RESULTADOS	38
5.1 Implementação.....	38
5.2 Validação	42
6 DISCUSSÃO	44
6 CONCLUSÃO.....	47
7 REFERÊNCIAS.....	48
8 ANEXOS	58
Anexo A – Registro do Genix protocolado no Instituto Nacional de Propriedade Industrial	58
Anexo B - Relação de genomas anotados pelo Genix já disponíveis no Genbank	

1 INTRODUÇÃO GERAL

O advento das plataformas de sequenciamento de DNA de nova geração, também denominadas *Next Generation Sequencing* (NGS), resultou em uma drástica redução do custo necessário para a obtenção de sequências genômicas e transcriptômicas. Por sua vez, esta redução acarretou em uma grande redução no tempo necessário para o sequenciamento de novos genomas, levando a um crescimento expressivo no volume de informação disponível em banco de dados públicos, como o Genbank e o Uniprot.

Após o sequenciamento, milhões de sequências de DNA são identificadas, que correspondem a fragmentos aleatórios que são processados pelos sequenciadores simultaneamente. O tamanho de fragmento lido varia de acordo com a plataforma, mas normalmente fica entre 30 e 700 pares de base. Com base nas sequências destes fragmentos, programas de montagem reconstroem, ao menos parcialmente, a(s) sequência(s) do(s) cromossomo(s) do organismo de interesse.

Com base na sequência do genoma é realizada a identificação dos genes e demais regiões funcionais (*features*). Neste processo, denominado anotação, diferentes programas são executados para se realizar a caracterização de regiões codificantes, genes responsáveis para RNAs estruturais, identificação de regiões repetitivas, dentre outras análises. Ao final, os dados de cada programa são integrados e as diferentes estruturas funcionais identificadas são representadas ao longo da sequência do genoma.

O processo de anotação é trabalhoso e pode envolver um grande número de análises, sendo por conta disso, em muitos casos executados através de ferramentas automatizadas denominadas *pipelines*, e pode ser de uso local, quando executadas na máquina do usuário, ou de uso *web*, quando executadas por um servidor remoto. As *pipelines* de uso local costumam apresentar uma maior versatilidade e permitir uma melhor configuração dos diferentes parâmetros do processo, mas também costumam exigir maiores conhecimentos de informática por parte do usuário, sendo executáveis, em sua maioria, unicamente em sistemas operacionais UNIX. Já as ferramentas de uso *web* costumam ser de uso mais

35 facilitado, mas tem resultados menos completos e uma menor possibilidade de
36 customização.

37

38 O sequenciamento de genomas de microbianos, sobretudo de organismos de
39 interesse industrial ou clínico, provê dados valiosos que pode ser utilizados como
40 bases para o desenvolvimento de uma grande variedade de estratégias
41 biotecnológicas. No caso de microorganismos patogênicos por exemplo, a
42 disponibilidade de sequências genômicas possibilita a utilização de técnicas de
43 vacinologia reversa para a identificação de novos alvos vacinais, prospecção de
44 genes que possam ser utilizados para tipagem molecular, identificação de genes de
45 resistência à antibióticos, identificação de novos alvos terapêuticos, dentre outras
46 abordagens.

47

48 A facilidade para a obtenção de sequencias de novos genomas torna necessária a
49 disponibilidade de ferramentas de anotação de fácil utilização e que sejam capazes
50 de gerar resultados completos e com boa acurácia. No presente trabalho é
51 apresentado o desenvolvimento do Genix, uma nova ferramenta de anotação de
52 genomas microbianos. Com base nos testes realizados através da comparação com
53 outras ferramentas, foi demonstrado que o Genix é capaz de gerar anotações mais
54 próxima das anotações de referência quando comparado às ferramentas RAST e
55 BASys, reduzindo o número de proteínas falso-positivas e falso-negativas, bem
56 como uma descrição mais detalhada dos RNA não codificantes (ncRNAs) presentes
57 no genoma.

58

59

2 REVISÃO BIBLIOGRÁFICA

2.1 O Sequenciamento de DNA

O sequenciamento de DNA consiste na identificação da sequência de nucleotídeos presentes em um ou mais fragmentos de interesse. As primeiras técnicas de sequenciamento de nucleotídeos foram desenvolvidas pouco mais de uma década após a descoberta da estrutura de dupla hélice do DNA por Watson & Crick (Watson & Crick, 1953), e consistiam na utilização de RNAses base específicas para a caracterização da sequência de RNA transportadores purificados (Holley *et al.*, 1965).

Em 1977, duas técnicas de sequenciamento de DNA foram descritas, de forma independente por Maxam & Gilbert (Maxam & Gilbert, 1977) e Sanger *et al* (Sanger *et al.*, 1977). Apesar de utilizarem uma um conjunto de reações químicas distintas, ambas resultavam na formação de um conjunto de fragmentos que poderiam ser separados através de eletroforese em gel. A determinação da sequência consistia na análise dos padrões de banda formados nas diferentes reações (Hutchison, 2007). Destas duas técnicas, a desenvolvida por Sanger *et al* acabou por se tornar o padrão para sequenciamento de DNA, sendo posteriormente otimizada com o uso de fluoróforos (permitindo a redução no número de reação de quatro para uma), e substituição da eletroforese em gel para eletroforese capilar, o que levou ao desenvolvimento dos primeiros sequenciadores de DNA (Karger & Guttman, 2009).

Os sequenciadores de DNA baseados no método de Sanger foram extensivamente utilizados para o sequenciamento dos primeiros genomas modelo, como *Haemophilus influenzae* (Fleischmann *et al.*, 1995), *Mycoplasma pneumoniae* (Himmelreich *et al.*, 1996), *Escherichia coli* (Blattner *et al.*, 1997), *Caenorhabditis elegans* (C. *elegans* Sequencing Consortium, 1998), *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000), *Homo sapiens* (Lander *et al.*, 2001; Venter *et al.*, 2001), *Oryza sativa* (Goff *et al.*, 2002), *Mus musculus* (Waterston *et al.*, 2002), dentre outros.

Em relação ao tamanho de um genoma, que pode variar de alguns milhões de bases no caso de bactérias, para alguns bilhões no caso de eucariotos, o tamanho e

número de fragmentos sequenciados, denominado *throughput*, nestes equipamentos, assim como o custo referente a cada base sequenciada, levou a necessidade da formação de consórcios entre diferentes institutos de pesquisa (Pareek *et al.*, 2011).

A crescente demanda por um maior *throughput* com menor custo, levou ao surgimento de novas metodologias de sequenciamento de DNA. No ano de 2004, a empresa Roche (www.roche.com/) lançou o primeiro sequenciador da linha 454, baseada na técnica de piro-sequenciamento (Ronaghi, 1998), uma abordagem paralela de alto-rendimento (*high throughput*) capaz de sequenciar milhares de fragmentos em uma única reação (Shendure & Ji, 2008). Posteriormente, outras empresas também desenvolveram suas próprias tecnologias proprietárias, resultando no lançamento das plataformas Illumina Solexa (www.illumina.com/), ABI SOLiD (www.appliedbiosystems.com/), Ion Torrent PGM (www.thermofisher.com/) e Helicos Heliscope (www.helicosbio.com/) (Liu *et al.*, 2012). Esta segunda fase do sequenciamento de DNA, caracterizada pelo alto rendimento, baixo custo (em relação às técnicas anteriores) e grande variedade de plataformas disponíveis, foi então inicialmente denominada sequenciamento de nova geração (do inglês *Next Generation Sequencing*, ou NGS), e posteriormente de sequenciamento de segunda geração (*Second Generation Sequencing*), devido ao surgimento das plataformas PacBio SMRT (www.pacb.com) e Oxford Nanopore (www.nanoporetech.com/), que foram por sua vez, classificadas como a terceira geração de sequenciadores. Entretanto, o termo NGS ainda é amplamente utilizada para se referir aos sequenciadores de segunda e terceira geração.

A disponibilidade de plataformas de sequenciamento capazes de gerar um grande volume de dados levou a um aumento expressivo no número de sequencias em bancos de dados públicos (Shendure & Ji, 2008), como o Genbank (Benson *et al.*, 2005) e o Uniprot (Apweiler *et al.*, 2004).

2.2 Sequenciamento e Montagem de Genomas

O sequenciamento de um genoma resulta na identificação da sequência de um grande número de fragmentos de DNA, uma vez que nenhuma técnica é capaz de sequenciar um genoma, mesmo que bacteriano, por completo. Desta forma, mesmo nas plataformas de sequenciamento de nova geração, é necessário que o genoma seja fragmentado, o que pode ser realizado através de métodos físicos ou enzimáticos. O tamanho recomendado para os fragmentos depende da plataforma de sequenciamento que será utilizada (Poptsova *et al.*, 2014).

Após o sequenciamento, as sequências identificadas para os fragmentos, denominadas “leituras” (*reads*), são sobrepostos de maneira que se reconstruir, ao menos parcialmente, a sequência original do genoma de interesse (Edwards & Holt, 2013). Este processo é denominado montagem, e requer que exista sobreposição entre os fragmentos para que possam ser construídas *contigs*. Por definição, uma *contig* é uma sequência ininterrupta (“contigua”) resultado da junção de dois ou mais fragmentos, podendo conter também bases não identificadas, comumente representadas através da letra “N” (Staden, 1979).

A montagem *de novo* de genomas sequenciados por tecnologias de segunda e terceira geração é computacionalmente mais complexa se comparada a montagem de genomas sequenciados pelo método de Sanger, sobretudo devido ao maior volume de dados a ser tratado e o tamanho menor de leitura (Baker, 2012). Para tornar viável a montagem de genomas sequenciados por estas novas tecnologias, novos algoritmos, como grafos de *overlap-layout consensus* (OLC), grafos de Bruijn (DBG) (Compeau *et al.*, 2011; Pevzner *et al.*, 2001) e grafos de *strings* (SG) (Myers, 2005) foram desenvolvidos, sendo implementados em ferramentas como Newbler (www.roche.com), Edena (Hernandez *et al.*, 2008) e MIRA (www.chevreux.org) (OLC), Velvet (Zerbino & Birney, 2008), Ray (Boisvert *et al.*, 2010), SOAPdenovo (Luo *et al.*, 2012), ABySS (Simpson *et al.*, 2009) e SPAdes (Bankevich *et al.*, 2012), e SGA (Simpson & Durbin, 2012) (SG).

Por conta das limitações impostas pelo grande número de *reads* de tamanho curto (Alkan *et al.*, 2010), os resultados de uma montagem *de novo* raramente representa uma montagem finalizada. A dificuldade de se resolver regiões repetitivas sem o uso

de etapas de sequenciamento complementar para o fechamento de *gaps*, resultou no crescimento no número de genomas depositados na forma de “rascunhos” (*draft*), que em muitos casos são suficientes para grande parte das análises de interesse, mas impossibilitam certos estudos de genômica comparativa e estrutural (Mardis *et al.*, 2002; Ricker *et al.*, 2012). Para reduzir a fragmentação das montagens *de novo*, diferentes metodologias foram desenvolvidas, como a integração de diferentes montagens (Lin & Liao, 2013; Nijkamp *et al.*, 2010; Sommer *et al.*, 2007; Yao *et al.*, 2012; Zimin *et al.*, 2008), ordenamento de *contigs/scaffolds* com base em um genoma de referência (Assefa *et al.*, 2009; Dias *et al.*, 2012; Galardini *et al.*, 2011; Lu *et al.*, 2014; Rissman *et al.*, 2009), fechamento *in silico* de *gaps* (Boetzer & Pirovano, 2012; Luo *et al.*, 2012; Piro *et al.*, 2014; Tsai *et al.*, 2010) e correção de erros de montagem (Otto *et al.*, 2010; Ronen *et al.*, 2012).

2.3 Anotação de Genes e Genomas

A anotação de um genoma consiste na identificação e caracterização das regiões funcionais, denominadas *features*, que estão presentes em sua sequência, o que pode incluir genes, promotores, terminadores, regiões de DNA repetitivo, operons, dentre outras (Edwards & Holt, 2013). A identificação das *features* pode ser realizada através do uso de dados experimentais, como alinhamento de sequências de transcritos (ex: RNA-Seq, *Expressed Sequence Tags*) ou proteínas, ou com base em ferramentas de predição *ab initio* (Yandell & Ence, 2012). Cada uma destas possíveis fontes de informação, seja experimental ou *ab initio*, é denominada evidência, e também é possível se combinar diferentes evidências para a geração de uma anotação consenso (Haas *et al.*, 2008). No caso de genomas eucarióticos, por exemplo, devido à complexidade da estrutura gênica, e a existência de fenômenos como o *splicing* alternativo, torna necessária a combinação de diferentes dados experimentais para a geração de uma anotação confiável. Por outro lado, a estrutura relativamente simples dos genes de procariotos permite que estes sejam identificados com boa acurácia, unicamente através de ferramentas *ab initio*.

2.3.1 Identificação de regiões codificantes (CDS, *Coding DNA sequences*)

As ferramentas *ab initio* para predição de genes codificantes em procariotos, como GLIMMER (Delcher *et al.*, 1999), Genemark.hmm (Borodovsky *et al.*, 2003), FGenesB (<http://www.softberry.com/>) e Prodigal (Hyatt *et al.*, 2010), utilizam modelos de estruturas de genes procarióticos para a realização das suas predições. No caso do GLIMMER e do Genemark.hmm, as estruturas são descritas na forma de *modelos ocultos de Markov* (HMM, *Hidden Markov Model*) (Durbin *et al.*, 1998), sendo a identificação de novos genes dependente do uso de um modelo já construído ou gerado por auto-treinamento do algoritmo. Já o Prodigal utiliza um algoritmo de computação dinâmica baseado na ocorrência de motivos de RBS, conteúdo C+G e tamanho da *Open Reading Frame* (ORF) para identificar as regiões com maior probabilidade de serem codificantes (Hyatt *et al.*, 2010).

Após a identificação, as ORFs podem ser comparadas com bancos de dados de genes, proteínas e domínios para a identificação de seus respectivos produtos. Ferramentas como BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009), BLAT (Kent, 2002), USEARCH (Edgar, 2010) e HMMER (Eddy, 2011) podem ser usadas para se alinhar as regiões de interesse contra bancos de dados com Genbank (Benson *et al.*, 2005), Uniprot (Apweiler *et al.*, 2004) e Pfam (Finn *et al.*, 2014) para se verificar se a região identificada possui similaridade com alguma proteína previamente caracterizada e identificar sua possível função. Além disso, é também possível se utilizar estas ferramentas para se remover possíveis erros no processo de predição de genes. Neste caso, ORFs falso positivas, denominadas, *spurious ORFs*, podem ser identificadas através da comparação com bancos de dados específicos, como o Antifam (Eberhardt *et al.*, 2012).

2.3.2 Identificação de RNAs não codificantes

Da mesma forma que genes que codificam para proteínas, genes de diferentes classes de RNAs não-codificantes (ncRNAs) (Ex: rRNAs, tRNAs, tmRNAs, miRNAs)

podem ser identificadas através ferramentas específicas de predição (Stadler, 2014). Abordagens mais simplificadas de identificação, como através de ferramentas de alinhamento local como o BLAST, permitem uma identificação rápida, mas desconsidera muitos aspectos estruturais, o que pode levar à um grande número de falso positivos (Lowe & Eddy, 1997). Desta forma, modelos probabilísticos, como HMMs, e preferencialmente baseados em estrutura secundária, como modelos de covariância (CM, *Covariance Models*), são preferidos para a identificação destes RNAs.

2.3.2.1 RNAmmer

A ferramenta RNAmmer utiliza como base o HMMER (Eddy, 2011), baseado em modelos ocultos de Markov, para a identificação de unidades de RNA ribossômico (rRNAs) de bactérias, archeas e eucariotos (Lagesen *et al.*, 2007). Neste caso, não são considerados aspectos estruturais devido à alta conservação deste grupo de RNAs em diferentes espécies (Tringe & Hugenholtz, 2008).

2.3.2.2 tRNAscan-SE e Aragorn

Os programas preditores de RNAs transportadores (tRNAs) tRNAscan-SE e Aragorn utilizam como base diferentes algoritmos. O tRNAscan-SE usam um modelo probabilístico denominado Covariance Model (CM), similar ao modelo oculto de Markov (HMM), mas que também considera a variação simultâneas entre diferentes posições para a estimativa de estruturas secundárias (Lowe & Eddy, 1997). Já o Aragorn usa como base uma simulação de estrutura secundária das regiões de *loop* do tRNAs para determinar se identificar na sequência sub-regiões que sejam compatíveis estruturalmente com esta classe de ncRNAs. Além de tRNAs, o Aragorn também realiza a busca de RNAs transportadores-mensageiros (tmRNAs), que consistem em RNA transportadores que possuem uma região de leitura aberta responsável pela síntese de um pequeno peptídeo de sinalização (Laslett, 2004).

2.3.2.3 INFERNAL e Rfam

O INFERNAL (Nawrocki *et al.*, 2009) é um pacote de ferramentas de alinhamento baseada em modelos de covariância. Um de seus programas, o cmsearch, permite que um conjunto de sequência seja comparado à um banco de dados de CMs para a identificação de regiões similares aos modelos, de forma análoga ao BLAST (Altschul *et al.*, 1990) e ao HMMER (Eddy, 2011). Para a anotação de ncRNAs em um genoma, é possível se utilizar esta ferramenta em conjunto com o Rfam, um banco de dados para estruturas de RNAs (Griffiths-Jones *et al.*, 2003). Entretanto, a busca através de CMs do INFERNAL é consideravelmente lenta se comparada ao BLAST, o que levou ao desenvolvimento de abordagens híbridas para reduzir o tempo necessário para a anotação, como a implementada na ferramenta rfam_scan.pl (<ftp://ftp.sanger.ac.uk/pub/databases/Rfam/>).

2.3.3 Anotação estrutural

A identificação de regiões funcionais, também denominados *motifs* ou domínios, em um determinado genes ou proteína é denominada anotação estrutural. No caso de proteínas, ferramentas como SignalP (peptídeos sinais) (Petersen *et al.*, 2011), TMHMM (hélices transmembrânicas) (Krogh *et al.*, 2001), Interproscan (domínios) (Zdobnov & Apweiler, 2001), e bancos de dados de regiões conservadas como Pfam (Finn *et al.*, 2014), PRODOM (Servant *et al.*, 2002) e SMART (Letunic *et al.*, 2012), podem ser usados para uma melhor caracterização estrutural.

2.3.4 Anotação funcional

É possível se estender as informações referentes à um determinado gene através da identificação das funções e processos biológicos associados a ele. Bancos de dados como *Clusters of Orthologous Groups* (COG) (Tatusov *et al.*, 2000) e o *Gene Ontology* (GO) (Tatusov *et al.*, 2000), organizam em estruturas hierárquicas estas funções, e usam um conjunto limitado e curado de termos para a identificação de

cada função e processo biológico. Além disso, algumas ferramentas como o BLAST2GO, permitem que a função de uma determinada proteína seja predita através da comparação dos seus resultados de BLAST contra um banco de dados de referência, com os termos GO (Conesa *et al.*, 2005).

Além da determinação da função de uma proteína, é possível também se reconstruir rotas metabólicas através das funções preditas para cada proteína do genoma. Ferramentas como KAAS (Moriya *et al.*, 2007), MinPaths (Ye & Doak, 2009) e PathPred (Moriya *et al.*, 2010) utilizam bancos de dados como o KEGG Pathways (Kanehisa & Goto, 2000) e o SEED (Overbeek *et al.*, 2014) como base para a identificação de proteínas ortologas relacionadas à rotas já elucidadas.

2.3.5 Anotação Automática

Como a anotação de um genoma pode envolver um grande conjunto de ferramentas e a integração de dados de diferentes fontes, sua realização de forma “manual” (não automatizada) tornaria o processo exaustivo e sujeito a falhas. Sendo assim, diversas ferramentas foram desenvolvidas para automatizar, ao menos parcialmente, a execução e integração dos resultados de cada ferramenta.

Para anotação de genomas bacterianos as ferramentas podem ser classificadas em *webserver*, como o RAST (Aziz *et al.*, 2008), xBASE (Chaudhuri & Pallen, 2006), BASys (Van Domselaar *et al.*, 2005) e o NCBI Prokaryotic Genome Annotation Pipeline (http://www.ncbi.nlm.nih.gov/genome/annotation_prok/), e as ferramentas de uso local, como o Prokka (Seemann, 2014), Eugene-PP (Sallet *et al.*, 2014), Maker (Cantarel *et al.*, 2008) e o BG7 (Pareja-Tobes *et al.*, 2012). Os *webserver* de anotação possuem como principal vantagem a fácil utilização, mesmo por usuários com pouca experiência com anotação de genomas e bioinformática, mas costumam oferecer pouca flexibilidade quanto à customização de seus parâmetros. Já as ferramentas de uso local, exigem um maior conhecimento técnico por serem executadas, em sua maioria, através de linhas de comandos e serem restritas a sistemas Linux / UNIX-Like / POSIX.

2.3.5.1 RAST

A ferramenta de anotação RAST (Aziz *et al.*, 2008) utiliza como base para a sua anotação o banco de dados de proteínas do SEED (Overbeek *et al.*, 2014) e possui uma ferramenta própria para identificação de regiões codificantes, apesar de também poder usar o GLIMMER (Delcher *et al.*, 1999) para *gene finding*. Além de genes codificantes, o RAST também identifica RNAs ribossomais e transportadores através de alinhamento de sequências, apesar de não realizar a busca de outras famílias de ncRNAs. Um exemplo de anotação do genoma de um isolado de *Leptospira interrogans*, contendo informações de rotas metabólicas previstas a partir do SEED está representado na Figura 1. Exemplo de anotação do realizada pelo RAST para um genoma de *Leptospira int*

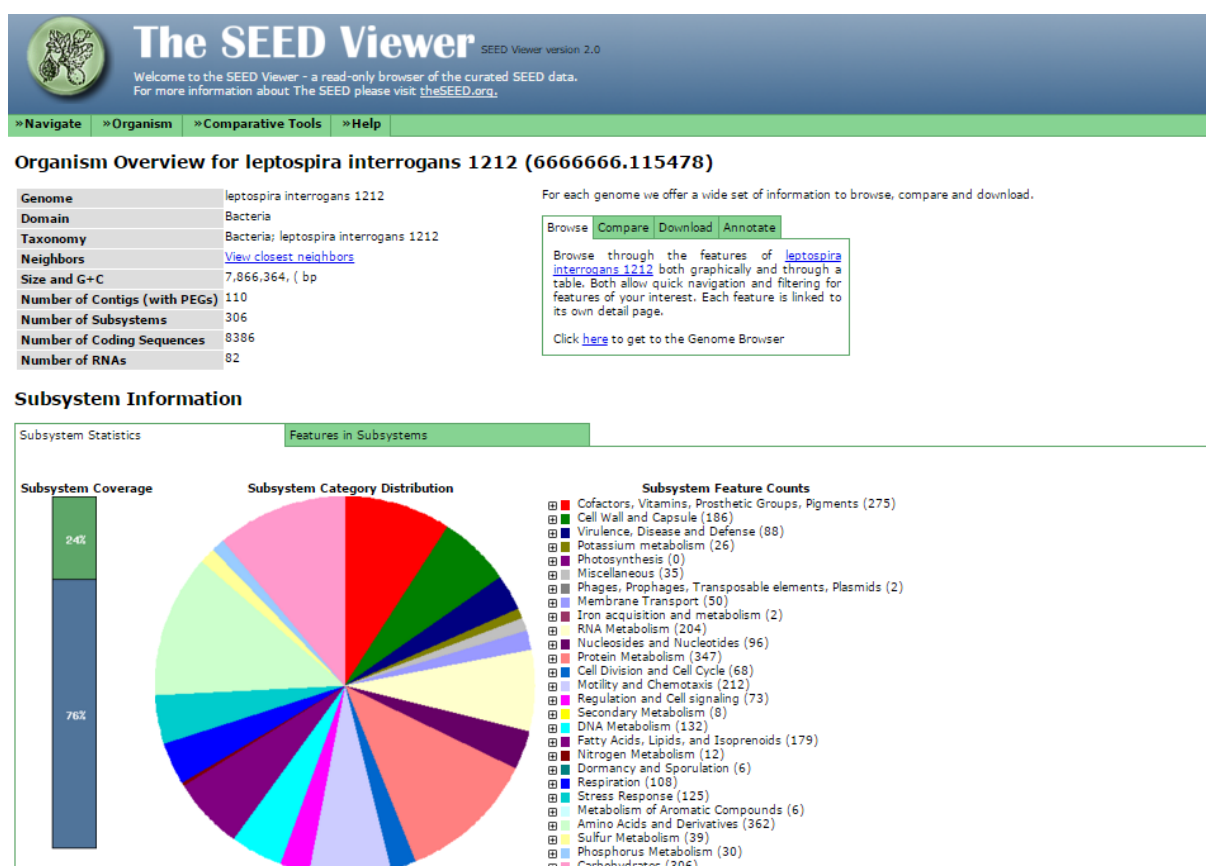


Figura 1. Exemplo de anotação do realizada pelo RAST para um genoma de *Leptospira interrogans*.

2.5.5.2 BASys

O BASys (Van Domselaar *et al.*, 2005) utiliza como base o Uniprot-Swissprot (Apweiler *et al.*, 2004), a porção de proteínas curadas do Uniprot, como base para anotação, também fazendo uso do GLIMMER (Delcher *et al.*, 1999) para a identificação das regiões codificantes. Após identificar as proteínas, um conjunto de *scripts* proprietários é executado para a identificação de domínios, termos GO, famílias de proteínas, peptídeo sinal, dentre outros *qualifiers*. A ferramenta também possui um *genome browser* integrado denominado CGView que permite a visualização de genomas circulares e geração de gráficos em diferentes formatos (Stothard & Wishart, 2005).

2.5.5.3 xBASE

A ferramenta xBASE (Chaudhuri & Pallen, 2006) também utiliza como base regiões codificantes identificadas pelo GLIMMER (Delcher *et al.*, 1999) e realiza a anotação com base em um genoma de referência já anotado. Para identificação de RNAs transportadores é utilizada a ferramenta tRNAscan-SE (Lowe & Eddy, 1997).

2.5.5.4 Prokka

A ferramenta de anotação Prokka (Seemann, 2014), de uso local, utiliza como base a predição de genes do Prodigal (Hyatt *et al.*, 2010), um conjunto de bancos de dados de proteínas derivado do Uniprot (Apweiler *et al.*, 2004) e modelos de Markov derivados do Pfam (Finn *et al.*, 2014) e TIGRFAM (Haft, 2003). Para identificação de ncRNAs, Aragorn (Laslett, 2004), INFERNAL (Nawrocki *et al.*, 2009) e RNAmmer (Lagesen *et al.*, 2007) podem ser utilizados.

2.5.5.5 Eugene-PP

A ferramenta Eugene-PP (Sallet *et al.*, 2014), de forma similar ao Prokka (Seemann, 2014), utiliza o Prodigal (Hyatt *et al.*, 2010) para a identificação de regiões codificantes, mas combina os resultados deste com dados de alinhamentos de Proteínas, transcritos e modelos probabilísticos baseados em genes de genomas de referência. Para a anotação de ncRNAs são usadas as ferramentas tRNAscan-SE (Lowe and Eddy, 1997), rfam_scan.pl (<ftp://ftp.sanger.ac.uk/pub/databases/Rfam/>) e RNAmmer (Lagesen *et al.*, 2007).

2.5.5.6 BG7

A *pipeline* de anotação BG7 (Pareja-Tobes *et al.*, 2012) utiliza a ferramenta tBLASTn do pacote NCBI-BLAST+ (Altschul *et al.*, 1990; Camacho *et al.*, 2009), tanto para o processo de *gene finding*, quando para a identificação das proteínas. Partindo-se de um conjunto de proteínas de referência, o programa realiza o alinhamento e processo os resultados para identificar possíveis quebras na fase de leitura (*frameshifts*) naturais (mutações) e artificiais (resultado de erros durante o sequenciamento ou na identificação das bases).

2.6 Visualização de anotações

A anotação de um genoma pode ser visualizada com auxílio de ferramentas denominadas *genome browsers*, como CGView (Stothard & Wishart, 2005), Artemis (Rutherford *et al.*, 2000), JBrowse (Skinner *et al.*, 2009), GenomeView (Abeel *et al.*, 2012), que representam graficamente a organização dos cromossomos (no caso de genomas finalizados) e *contigs/scaffolds* (no caso de genomas rascunho), e indicados as informações referentes à cada *feature*, possibilitando uma análise mais acurada. Outras ferramentas, como Circos (Krzywinski *et al.*, 2009), DNA Plotter (Carver *et al.*, 2009), BRING (Alikhan *et al.*, 2011) também possibilitam uma melhor customização na forma com a qual a anotação é representada, permitindo a geração de gráficos de alta qualidade para publicações científicas. Além disso, ferramentas

com o Artemis Comparison Tool (ACT) (Carver *et al.*, 2005), integram as funções de *genome browsers* com análise de sintenia, sendo útil para análises de genômica comparativa.

2.7 Submissão para bancos de dados públicos

Em muitos casos, como requisitos para publicação, artigos que relatam o sequenciamento de novos genes ou genomas deve incluir os respectivos códigos de acesso em bancos de dados públicos, como Genbank e EMBL. A submissão de sequências para estes bancos de dados segue um conjunto de recomendações e padrões internacionais, o que inclui o nome que será usado para as *features* e seus respectivos qualificadores (*qualifiers*), como são tratadas as regiões de *gaps*, quais informações foram usadas para a ligação das *scaffolds*, dentre outras (Pirovano *et al.*, 2015).

No caso de submissões de genomas para o Genbank, o processo envolve o cadastro da amostra no banco de dados BioSample, do projeto de sequenciamento, no banco de dados BioProject (Barrett *et al.*, 2012), a geração de um arquivo *genbank submission template* (.sbt) com dados dos autores do projeto (<https://submit.ncbi.nlm.nih.gov/genbank/template/submission/>), a formatação do arquivo contendo a anotação e as sequências através do programa TBL2ASN (<http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/>), e a submissão do resultado deste, um arquivo em formato .sqn, para o servidor do Genbank através do seu portal de submissão (<http://www.ncbi.nlm.nih.gov/>).

3 Sequenciamento de Genomas Microbianos

O sequenciamento de genomas de microorganismos, sobretudo aqueles de interesse clínico ou industrial, é de grande relevância para a pesquisa biológica, tanto básica quanto aplicada (Barbosa *et al.*, 2014). No caso de bactérias patogênicas, a análise do genoma possibilita o desenvolvimento de abordagens para identificação de novos alvos vacinais (Sette & Rappuoli, 2010), genes de resistência

416 à antibióticos (Köser *et al.*, 2014), genes relacionados ao processo de patogênese
417 (Chen *et al.*, 2012), marcadores para tipagem molecular e como o *Multilocus*
418 *Sequence Typing* (MLST) (Maiden *et al.*, 1998), dentre outros padrões moleculares.
419 Além disso, o sequenciamento do genoma de diferentes isolados e cepas de uma
420 mesma espécie ou gênero também fornece informações relevantes sobre a
421 variabilidade genética, adaptação e genes compartilhados (*core genome*) (Medini *et*
422 *al.*, 2005). O processo de anotação do genoma é uma etapa mandatória para que
423 muitas destas análises possam ser realizadas, devendo fornecer um resultado
424 acurado e completo dos genes e demais *features* presentes (Beckloff *et al.*, 2012).

425
426 Devido à popularização das plataformas de NGS, a necessidade por ferramentas de
427 anotação de fácil utilização, mesmo para pessoas com pouco experiência em
428 informática, levou ao surgimento de plataformas *web* que realizam este processo de
429 forma automatizada. Entretanto, os resultados gerados por estas ferramentas em
430 muitos casos se mostram limitados quando comparados aos obtidos por ferramentas
431 de uso local, que por sua vez, costumam exigir um maior conhecimento técnico para
432 sua utilização.

3 HIPÓTESE E OBJETIVOS

3.1 Hipótese

A integração de preditores de ncRNA e o uso do preditor de genes Prodigal e do banco de dados de proteínas *spurious* Antifam pode contribuir para a obtenção de uma anotação mais completa e acurada de genomas microbianos.

3.2 Objetivo Geral

Desenvolver uma nova ferramenta de anotação de genomas microbianos, visando uma anotação mais completa e acurada.

3.3 Objetivos Específicos

- Implementar a *pipeline* de anotação em Python integrando programas para predição de genes, busca em bancos de dados de sequências, predição de ncRNAs e clusterização de bancos de dados.
- Implementar o servidor *web* usando Apache HTTP Server, MySQL e um *back-end* desenvolvido em Python.
- Desenvolver uma interface para interação do usuário que inclua também um *genome browser* para visualização dos resultados da anotação.
- Aferir a acurácia da *pipeline* através da comparação dos seus resultados com os obtidos pelas ferramentas RAST e BASys para um conjunto de genomas de referência.

4 METODOLOGIA

4.1 Desenvolvimento

4.1.1 Visão geral do *genix_server.py*

O *back-end* do *webserver* foi construindo com *scripts* escritos em linguagem Python (<https://www.python.org/>) em que recebem os dados enviados pelo usuário a partir do módulo CGI, que realiza o processamento dos formulários POST. As informações passadas são então inseridas em um banco de dados MySQL (<https://www.mysql.com/>), que é acessado pelos *scripts* através do módulo MySQLdb. O *webserver* foi configurado utilizando o *Apache HTTP Server* (www.apache.org/) e roda em um servidor Linux Ubuntu 13.10 (<http://www.ubuntu.com/>).

Para a submissão de um genoma para anotação é necessário que o usuário realize previamente um cadastro, que tem como objetivo um maior controle sobre o acesso aos resultados de cada projeto. Após cada submissão um código de projeto denominado “*job_id*” é gerado e retornado ao usuário.

O gerenciamento dos projetos de anotação é realizado pelo *script* “*genix_server.py*”, que se conecta ao banco de dados e gerencia o processo de anotação, a construção do banco de dados de proteínas (através do *script* *get_database.sh*) e a passagem dos parâmetros passados para o *genix_annotation.py*. Apenas um genoma é anotado por vez, sendo o status do processo atualizado a cada etapa no banco de dados de forma que o usuário possa acompanhar o progresso do processo. A estrutura da pipeline está representada na Figura 2.

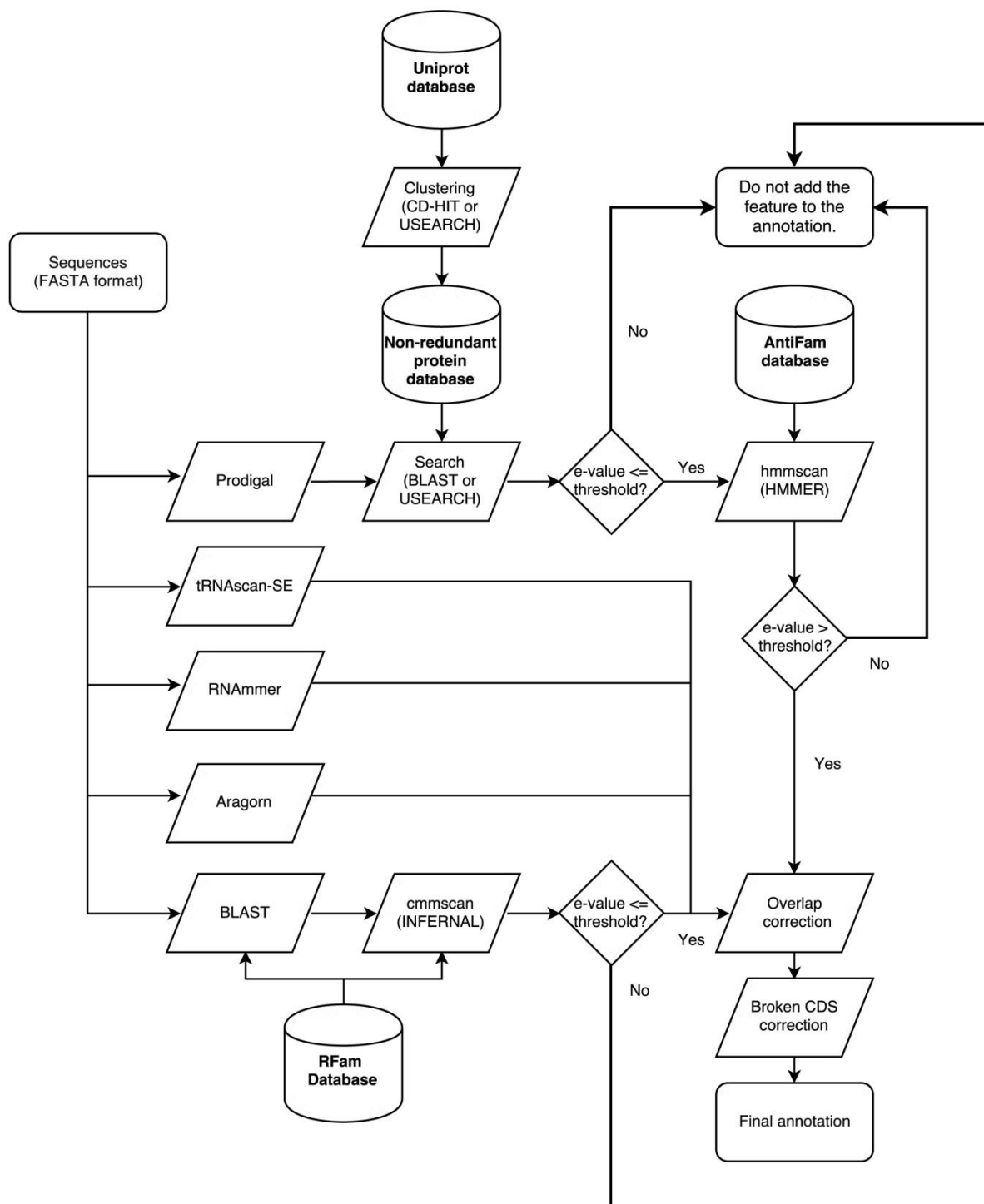


Figura 2. Estrutura da *pipeline* de anotação do Genix.

4.1.2 Construção do banco de dados (*get_database.sh*)

A construção do banco de dados é realizada pelo script *get_database.sh*, escrito em linguagem Bash, que recebe quatro argumentos: *tax_id*, prefixo banco de dados de saída, grau de identidade para clusterização do *dataset* e o programa que será usado para a clusterização, podendo ser escolhido o CD-HIT (Li & Godzik, 2006) ou o USEARCH (Edgar, 2010). O *tax_id* é um código de identificação padronizado utilizado por diferentes bancos de dados públicos, como NCBI, EMBL e Uniprot, abrangendo diferentes níveis taxonômicos (<http://www.ncbi.nlm.nih.gov/taxonomy>). O script utiliza este código para recuperar um *dataset* “bruto” (*raw dataset*) de sequências de proteínas a partir do Uniprot (Apweiler *et al.*, 2004). Após a geração deste primeiro *dataset*, a ferramenta de clusterização é executada usando o grau de identidade escolhida de forma a se gerar um arquivo FASTA com menor redundância. Por fim, os programas *makeblastdb* (Altschul *et al.*, 1990; Camacho *et al.*, 2009) e *usearch* (com a opção “*-makeudb_ublast*”) (Edgar, 2010) são executadas para a formatação dos bancos de dados de proteína para a etapa de anotação que será realizada posteriormente.

4.1.3 O script *genix_annotation.py*

O processo de anotação é realizado pelo script *genix_annotation.py*, que foi desenvolvido para poder receber uma ampla variedade de parâmetros (Tabela 1).

O script inicialmente cria um banco de dados *SQLite3* temporário mantido diretamente na memória RAM (*in memory*) (<https://www.sqlite.org/>) para armazenar os dados intermediários de anotação. Após isso o arquivo de entrada, em formato FASTA, que contém as sequências de genoma de interesse, é indexado e as regiões de *gap* são identificadas através de expressões regulares. Cada sequência recebe um código de identificação único (ID), sendo este usado como campo de identificação em um novo arquivo FASTA que é gerado para as análises posteriores. Este arquivo contém as sequências do arquivo original processadas de forma a se substituir todas as bases ambíguas pela letra ‘N’.

A identificação *open reading frames* (ORFs) é realizada através do *software* Prodigal (Hyatt *et al.*, 2010), sendo cada região identificada pelo programa comparada com o banco de dados de proteína gerado pelo *get_database.sh*, podendo ser usado para isso o BLASTx (Altschul *et al.*, 1990) do pacote NCBI-BLAST+ (Camacho *et al.*, 2009) ou o modo UBLAST do USEARCH (Edgar, 2010). As regiões que apresentam similaridade com o banco de dados e que apresentam um *e-value* menor que o *threshold* definido pelo usuário são mantidas e comparadas com o banco de dados de Hidden Markov Models (HMMs) do Antifam (Eberhardt *et al.*, 2012) com uso da ferramenta HMMER (Eddy, 2011). O Antifam consiste em um banco de dados de *spurious ORFs*, sequências de DNA que são comumente identificadas como possíveis regiões codificantes por ferramentas de predição de gene, mas que não verdade não codificam para nenhuma proteína e podem estar associadas a outras funções biológicas, como genes de RNAs estruturais ou sequências de repetição em *tandem*. Caso a sequência não apresentem nenhum hit no Antifam com *e-value* menor que o *threshold*, a sequência é indexada no banco de dados como um *Coding DNA Sequence* (CDS).

Tabela 1. Opções (argumentos) aceitas pelo *script genix_annotation.py*

Opção	Valor	Mandatário
-i	Caminho do arquivo FASTA de entrada	Sim
-o	Caminho do diretório de saída	Não
-t	Número de <i>threads</i>	Não
-dbp	Arquivo do banco de dados formatado	Sim
-antie	<i>Threshold</i> do <i>e-value</i> para o HMMER com o Antifam	Não
-blaste	<i>Threshold</i> do <i>e-value</i> para o BLAST de proteínas	Não
-blastrnae	<i>Threshold</i> do <i>e-value</i> para o BLAST de ncRNAs	Não
-infernale	<i>Threshold</i> do <i>e-value</i> para o INFERNAL	Não
-sub	<i>Template</i> de submissão para o Genbank	Não
-mg	Tamanho mínimo de região de "N"s para ser um gap	Não
-gc	Tabela de código genético do Prodigal	Não
-sl	Prefixo das <i>scaffolds</i>	Não
-lt	Prefixo das <i>locus_tags</i>	Não
-inst	Nome do centro de sequenciamento	Não
-sp	Programa de busca (BLAST ou UBLAST)	Não
-source	<i>String</i> com a <i>source information</i> (modelo NCBI)	Não
-gap	Informação usada unir as <i>scaffolds</i>	Não

A predição dos RNAs transportadores (tRNAs), ribossômicos (rRNAs) e transportadores-mensageiros (tmRNAs) é realizada com os programas tRNAscan-

SE (Lowe & Eddy, 1997), RNAmmer (Lagesen *et al.*, 2007) e Aragorn (Laslett, 2004), respectivamente. Outras classes de ncRNAs, assim como riboswitches, RNAs relacionados à RNases, sequências de CRISPRs, dentre outras *features* associadas à RNAs são preditas a partir do banco de dados do Rfam (Griffiths-Jones *et al.*, 2003). Para isso, uma busca por BLASTn (Altschul *et al.*, 1990; Camacho *et al.*, 2009) é realizada de forma a se identificar as famílias de RNAs que possam estar presentes no genoma, sendo posteriormente realizado uma busca complementar destas com uso do pacote INFERNAL (Nawrocki *et al.*, 2009). O uso do BLAST seguido do INFERNAL se deve ao fato de que o primeiro apresenta uma alta velocidade para o alinhamento, mas desconsidera aspectos estruturais e evolutivos, o que aumenta a chance de falso-positivos, enquanto o segundo apresenta uma baixa velocidade, apesar de ter uma acurácia maior para identificação de estruturas secundárias de RNAs. Uma abordagem similar é usada *script rfam_scan.pl* (<http://rfam.sanger.ac.uk/>).

Após a anotação das regiões codificantes e ncRNAs é realizada a integração dos dados. Nesta etapa, genes codificantes que apresentam ncRNAs sobrepostos são removidos, e as *features* são ordenadas nas sequências do genoma. O resultado da anotação é salvo em formato *Genbank* (GBK), *Gene Finding Format* (GFF) e *Feature Table* (TBL). Caso o usuário envie um arquivo *template* de submissão para o Genbank (SBT), um arquivo em formato *Sequin* (SQN) é gerado a com a ferramenta *tbl2asn* do NCBI (www.ncbi.nlm.nih.gov/genbank/tbl2asn2/).

4.1.4 Front-end e apresentação dos resultados de anotação

O *front-end* foi desenvolvido usando HTML, CSS3 e Javascript. Durante a submissão de um novo genoma, expressões regulares são utilizadas para verificar a consistência de cada informação digitada nos campos do formulário. Os resultados da anotação são disponibilizados para *download* em uma página de projeto que pode ser acessada se informando o *job_id* do projeto de interesse, e o login e senha do usuário que o submeteu. É possível também visualizar a anotação no próprio website através do *genome browser* Jbrowse (Skinner *et al.* 2009).

4.2 Comparação com outras ferramentas

Para se avaliar a acurácia da anotação gerada pelo Genix foram realizadas anotação para quatro genomas de referência (*RefSeq*): *Escherichia coli* cepa K12, *Leptospira interrogans* cepa Fiocruz L1-130 (cromossomo I), *Listeria monocytogenese* cepa EGD-e e *Mycobacterium tuberculosis* cepa H37Rv. A relação de códigos de acesso e contagem de genes na anotação original está apresentada na Tabela 2.

Tabela 2. Lista de genomas utilizados para a avaliação da ferramenta Genix

Organismo	Código do Genbank	Nº de CDSs
<i>E. coli</i> cepa K12	NC_000913.3	4319
<i>L. interrogans</i> cepa Fiocruz L1-130 (Cromossomo I)	AE016823.1	3394
<i>L. monocytogenese</i> cepa EGD-e	AL591824.1	2855
<i>M. tuberculosis</i> cepa H37Rv	AL123456.3	4031

Os resultados obtidos foram comparados à anotação original e às anotações geradas pelas ferramentas RAST (Aziz *et al.*, 2008) e BASys (Van Domselaar *et al.*, 2005). As métricas utilizada na comparação: contagem de genes novo (em relação à anotação original), contagem de genes faltantes (em relação à anotação original) e contagem de genes exclusivos (em relação à todas as anotações). Os genes que apresentaram divergência foram anotados funcionalmente pelo programa BLAST2GO (Conesa *et al.*, 2005) usando o banco de dados do *Uniprot-Swissprot* como base. As proteínas codificadas pelas anotações também foram analisadas pelo HMMER (Eddy, 2011) usando o banco de dados do Antifam (Eberhardt *et al.*, 2012). Por fim, uma métrica denominada “*discrepância de anotação*” foi proposta, consistindo na porcentagem de genes divergentes entre as anotações geradas e a anotação original disponível no Genbank. Considerando duas anotações para uma mesma sequência, uma sendo a “original” (A) e outra “nova” (N), a discrepância (D) para entre estas anotações (D_{NA}) é igual à soma do número de genes novos com os genes faltantes, dividido pela soma do número total de genes nestas anotações ($TG_A + TG_N$), como representado na fórmula:

$$D_{NA} = \frac{Gf_{AN} + Gn_{AN}}{TG_A + TG_N}$$

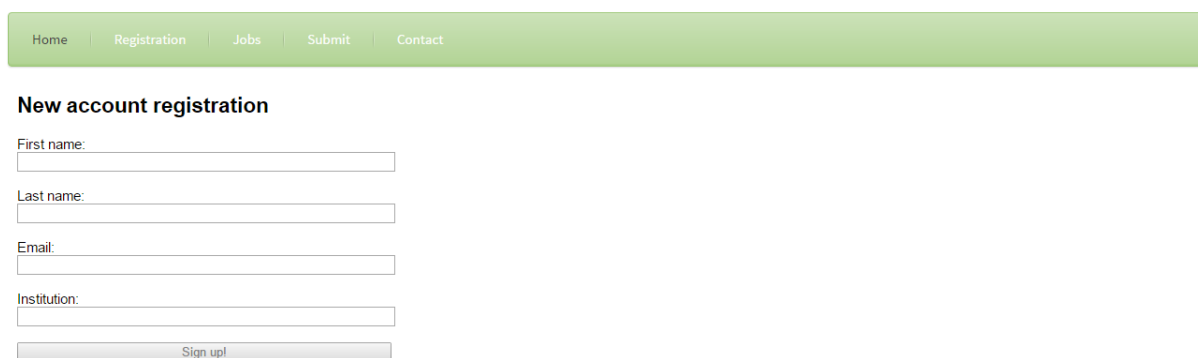
5 RESULTADOS

5.1 Implementação

O Genix foi implementando na forma de um servidor *web* (*webserver*), ao menos parcialmente, usando um modelo LAMP (Linux, Apache, MySQL e Python), e sendo as linguagens Perl e Bash e o sistema de banco de dados SQLite usados para tarefas pontuais. O site do programa foi hospedado no servidor do Laboratório de Bioinformática e Proteômica do Centro de Desenvolvimento Tecnológico, estando disponível através do endereço <http://labbioinfo.ufpel.edu.br/genix/>. Para a submissão de projetos de anotação (*jobs*), é necessária a realização de um cadastro de usuário que deve ser efetuado através do endereço <http://labbioinfo.ufpel.edu.br/genix/registration.html> (Figura 3)

Após o registro é gerado uma senha para que o usuário possa ter um maior controle sobre o acesso aos resultados de seus projetos. Não é necessário um e-mail institucional para a realização do cadastro.

[GENIX] Automated Bacterial Genome Annotation Pipeline



Home | Registration | Jobs | Submit | Contact

New account registration

First name:

Last name:

Email:

Institution:

Figura 3. Página de registro do Genix.

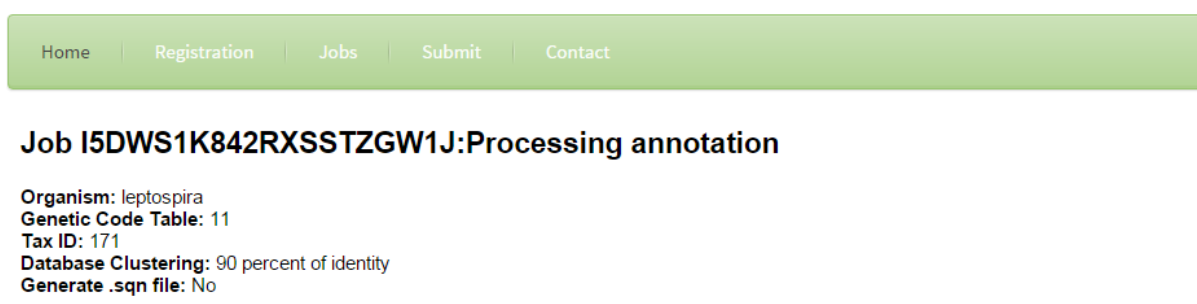
Ao submeter um genoma para a anotação, o usuário pode configurar diferentes parâmetros, como o *threshold* das ferramentas BLASTn, USEARCH, INFERNAL e HMMER, códigos de identificação para *scaffolds*, *locus_tag* e centro de sequenciamento, e dados de identificação do organismo (*source information*).

632 Também é possível se submeter arquivos no formato “.sbt” que são usados como
 633 *template* pelo programa tbl2asn do NCBI para a geração de arquivos de submissão
 634 para o Genbank.

635 Após a submissão, um código de acesso é gerado para o projeto (*job id*), sendo este
 636 usado pelo usuário para a recuperação dos resultados. O Genix realiza uma
 637 anotação por vez, sendo todos os projetos primeiramente indexados com o status
 638 “*queued*”. Os status “*processing database*” e “*processing annotation*” são usados
 639 para projetos em andamento, sendo referentes aos *scripts* “*get_database.sh*” e
 640 “*genix_annotation.py*”, respectivamente. Um exemplo de página de um projeto em
 641 andamento está apresentado na Figura 4, onde é possível ver o job id, o status do
 642 projeto e alguns dados informados pelo usuário durante a submissão do genoma.

643

[GENIX] Automated Bacterial Genome Annotation Pipeline



The screenshot shows the GENIX web interface. At the top is a green navigation bar with links: Home, Registration, Jobs, Submit, and Contact. Below this is a header for a specific job: "Job I5DWS1K842RXSSTZGW1J:Processing annotation". Underneath the header, several parameters are listed: "Organism: leptospira", "Genetic Code Table: 11", "Tax ID: 171", "Database Clustering: 90 percent of identity", and "Generate .sqn file: No".

644

645 **Figura 4.** Exemplo de uma página de projeto de anotação em andamento, incluindo um resumo dos
 646 parâmetros passados pelo usuário.

647

648 O status e os resultados de cada projeto podem ser verificados através do endereço
 649 <http://labbioinfo.ufpel.edu.br/genix/jobs.html>. Após o termino da anotação, os
 650 resultados podem ser recuperados nos formatos *Genbank format* (.gbk), *Gene*
 651 *Finding Format* (.gff) e *Feature Table* (.tbl). Caso tenha sido solicitada a geração de
 652 arquivos de submissão para o Genbank, os resultados do programa tbl2asn podem
 653 ser obtidos em formato compactado .zip. Um exemplo de anotação finalizada, com
 654 os links para download dos arquivos de resultado em diferentes formatos está
 655 apresentado na Figura 5.

[GENIX] Automated Bacterial Genome Annotation Pipeline

[Home](#)
[Registration](#)
[Jobs](#)
[Submit](#)
[Contact](#)

Job I5DWS1K842RXSSTZGW1J:Finished annotation

Organism: leptospira
Genetic Code Table: 11
Tax ID: 171
Database Clustering: 90 percent of identity
Generate .sqn file: No

Annotation Files:

[Sequences \(.fsa file\)](#)
[Features \(.tbl file\)](#)
[Annotation \(.gb\)](#)
[Annotation \(.gff\)](#)

[view annotation](#) (Powered by JBrowse)

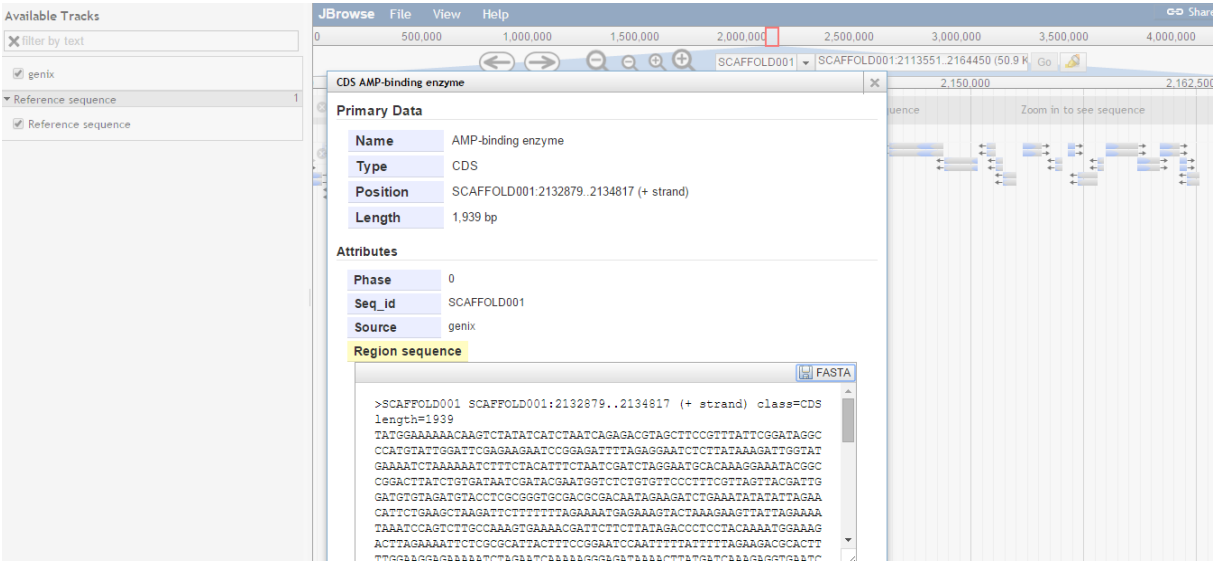
656

657 **Figura 5.** Exemplo de uma página de projeto de anotação finalizado no Genix, incluindo um resumo
 658 dos parâmetros passados pelo usuário, os *links* para o usuário efetuar o download dos resultados de
 659 anotação em diferentes formatos (tbl, gb, gff) e o *link* para a visualização da anotação através do
 660 Jbrowse.

661

662 Por fim, a visualização da anotação pode ser realizada na própria página se
 663 resultados através da ferramenta *JBrowse*, já integrada no Genix, através da opção
 664 “*view annotation*”. Na Figura 6 está apresentado um exemplo de visualização de
 665 genoma, com os dados de uma CDS sendo mostrados.

666



667

668 **Figura 6.** Visualização de uma anotação de genoma pela ferramenta JBrowse integrada ao Genix.

669

670

5.2 Validação

A validação da ferramenta foi realizada usando-se como base os genomas de *E. coli* cepa K12, *L. interrogans* cepa Fiocruz L1-130 (Cromossomo I), *L. monocytogenese* cepa EGD-e, *M. tuberculosis* cepa H37Rv, obtidos a partir do Genbank. As anotações geradas pelo Genix foram comparadas às obtidas pelas ferramentas RAST e BASys e às anotações de referência originais. Para cada genoma de referência, cada anotação foi avaliada quanto às suas CDSs totais, novas (em relação à referência), faltantes (em relação à referência) e exclusivas (em relação as demais). De cada uma destas categorias, também foi realizada a anotação funcional com a ferramenta BLAST2GO. Além disso, a partir do número de CDS total na anotação e na referência, e do número de CDSs faltantes e novas, foi calculado a *discrepância de anotação*. Por fim, também foi realizada a identificação de *spurious ORFs* a partir do banco de dados do Antifam. Os resultados obtidos para os genomas de *M. tuberculosis* cepa H37Rv, *E. coli* cepa K12, *L. interrogans* cepa Fiocruz L1-130 (Cromossomo I), *L. monocytogenese* cepa EGD-e estão apresentados nas tabelas Tabela 3, Tabela 4,

Tabela 5 e

Tabela 6.

Tabela 3. Comparação da anotação original (referência) disponível no Genbank com os resultados obtidos pelas ferramentas Genix, RAST e BASys para o genoma de *M. tuberculosis* cepa H37Rv.

Programa	CDS							Discrepância	Antifam
	T	N	F	E	NF	FF	EF		
Referência	4031	0	0	0	0	0	0	0,00%	9
Genix	3930	187	144	53	58	82	4	4,08%	0
RAST	4169	420	141	170	69	92	7	6,72%	10
BASys	4491	758	149	518	66	91	12	10,46%	14

T = Número total de Coding DNA Sequences (CDS); N = Número de CDSs novas (em relação à referência); F = Número de CDSs faltantes (em relação à referência); E = Número de CDSs exclusivas (em relação a todas anotações, incluindo a referência); NF = Número de CDSs novas com anotação funcional pelo BLAST2GO; FF = Número de CDSs faltantes com anotação funcional pelo BLAST2GO; EF = Número de CDSs exclusivas com anotação funcional pelo BLAST2GO.

Tabela 4. Comparação da anotação original (referência) disponível no Genbank com os resultados obtidos pelas ferramentas Genix, RAST e BASys para o genoma de *E. coli* cepa K12.

Programa	CDS							Discrepância	Antifam
	T	N	F	E	NF	FF	EF		
Referência	4319	0	0	0	0	0	55	0,00%	1
Genix	4116	85	144	38	87	125	7	2,66%	0
RAST	4241	266	172	214	48	137	9	5,00%	6
BASys	3906	29	221	2	88	159	2	2,95%	2

T = Número total de Coding DNA Sequences (CDS); N = Número de CDSs novas (em relação à referência); F = Número de CDSs faltantes (em relação à referência); E = Número de CDSs exclusivas (em relação a todas anotações, incluindo a referência); NF = Número de CDSs novas com anotação funcional pelo BLAST2GO; FF = Número de CDSs faltantes com anotação funcional pelo BLAST2GO; EF = Número de CDSs exclusivas com anotação funcional pelo BLAST2GO.

Tabela 5. Comparação da anotação original (referência) disponível no Genbank com os resultados obtidos pelas ferramentas Genix, RAST e BASys para o genoma de *L. interrogans* cepa Fiocruz L1-130 (Cromossomo I).

Programa	CDS							Discrepância	Antifam
	T	N	F	E	NF	FF	EF		
Referência	3394	0	0	0	0	0	3	0,00%	111
Genix	3137	171	214	34	29	5	2	5,71%	0
RAST	4253	1037	89	572	30	10	2	14,55%	459
BASys	3992	946	174	492	30	8	1	14,81%	272

T = Número total de Coding DNA Sequences (CDS); N = Número de CDSs novas (em relação à referência); F = Número de CDSs faltantes (em relação à referência); E = Número de CDSs exclusivas (em relação a todas anotações, incluindo a referência); NF = Número de CDSs novas com anotação funcional pelo BLAST2GO; FF = Número de CDSs faltantes com anotação funcional pelo BLAST2GO; EF = Número de CDSs exclusivas com anotação funcional pelo BLAST2GO.

Tabela 6. Comparação da anotação original (referência) disponível no Genbank com os resultados obtidos pelas ferramentas Genix, RAST e BASys para o genoma de *L. monocytogenes* cepa EGD-e.

Programa	CDS							Discrepância	Antifam
	T	N	F	E	NF	FF	EF		
Referência	2855	0	0	0	0	0	4	0,00%	6
Genix	2838	31	24	16	3	13	0	0,96%	0
RAST	2884	63	17	48	3	5	0	1,39%	5
BASs	1417	2	720	0	3	70	0	14,46%	5

T = Número total de Coding DNA Sequences (CDS); N = Número de CDSs novas (em relação à referência); F = Número de CDSs faltantes (em relação à referência); E = Número de CDSs exclusivas (em relação a todas anotações, incluindo a referência); N.F = Número de CDSs novas com anotação funcional pelo BLAST2GO; FF = Número de CDSs faltantes com anotação funcional pelo BLAST2GO; EF = Número de CDSs exclusivas com anotação funcional pelo BLAST2GO.

6 DISCUSSÃO

A ferramenta Genix foi desenvolvida de forma a ser de fácil utilização e, ao mesmo tempo, apresentar resultados mais completos e acurados em relação as atuais pipelines de anotação online de genomas procarióticos. Quanto à sua implementação e interface de uso, o Genix usa uma modelo *webserver*, de forma similar aos programas RAST (Aziz *et al.*, 2008), BASys (Van Domselaar *et al.*, 2005) e xBASE (Chaudhuri & Pallen, 2006). Em sua página é possível a realização do registro de usuário, submissão de genoma, acompanhamento do status do processo de anotação, recuperação dos resultados e visualização da anotação geração através da ferramenta JBrowse. Por utilizar uma *pipeline* própria para a geração do banco de dados de proteínas para o processo de anotação, combinada à uma etapa de remoção de redundâncias (que pode ser realizada pelo CD-HIT ou pelo USEARCH), é espero um ganho de performance substancial em relação ao use de bancos de dados genéricos, como o Uniprot-Swissprot, Uniprot-trEMBL e Genbank.

A interação de ferramentas de para predição de ncRNAs é outro diferencial, uma vez que ferramentas como RAST (Aziz *et al.*, 2008) e BASys (Van Domselaar *et al.*, 2005) oferecem poucos recursos para identificação destas famílias de RNA estruturais, limitando seus resultados aos RNAs transportadores e RNA ribossômicos. Ferramentas de anotação de uso local, como Prokka (Seemann, 2014) e Eugene-PP (Sallet *et al.*, 2014) também oferecem suporte à uma anotação mais acurada destes RNAs, mas a necessidade de configuração, a necessidade de serem executados em sistemas operacionais Linux / UNIX-Like e a interface por linha de comando pode ser um entrave para usuários com menor experiência em bioinformática.

Na análise comparativa dos resultados obtidos para os quatro genomas de referência, o Genix apresentou, em todos os casos, o menor valor de discrepância, indicando uma anotação mais fidedigna em relação às outras ferramentas de anotação. Além disso, para os genomas de *M. tuberculosis*, *E. coli* e *L. interrogans*, a ferramenta obteve o menor número de faltantes funcionais, indicando uma menor perda de genes que possuem de fato uma função biológica definida ou ao menos predita. Mesmo em genomas de referência, anotados e curados manualmente, um

grande número de genes identificados é classificado como codificante para proteínas hipotéticas. No genoma de *L. interrogans* sorovar Copenhageni cepa Fiocruz L1-130, por exemplo, mais de 40% da anotação original é composta por proteínas hipotéticas, o que tornou necessário o uso de ferramentas de anotação funcional para verificar o quão relevantes são os genes faltantes e novos nas diferentes anotações. Para este genoma, BASys e RAST apresentaram uma discrepância maior que 14%, enquanto o Genix apresentou 5,71 %. A ferramenta BASys também apresentou uma alta discrepância para o genoma de *M. tuberculosis*, com um valor de 10,46 %.

De fato, o grande número de proteínas hipotéticas no genoma de *L. interrogans* também justifica, ao menos em parte, o grande número de proteínas *spurious* identificadas na própria anotação de referência. Na análise realizada pelo Antifam para os diferentes genomas, todas as anotações (incluindo as de referência), apresentaram *hits* no banco de dados. Destas, a ferramenta que apresentou o maior número de *hits* foi o RAST, quando anotando o genoma *L. interrogans*, onde foram encontrados 459 *spurious* ORFs em um total de 4253, o que representa ~10,8% da anotação sendo constituída por proteínas erroneamente identificadas. Para este mesmo genoma, a própria anotação de referência apresentou 111 *hits*.

Na estrutura de sua *pipeline*, um dos principais diferenciais do Genix em relação às demais ferramentas é o uso do Prodigal (Hyatt *et al.*, 2010) ao invés GLIMMER (Delcher *et al.*, 1999), que é usado como padrão pelo BASys (Van Domselaar *et al.*, 2005) e como opcional pelo RAST (Aziz *et al.*, 2008), como programa de *gene finding*. Como demonstrado por Hyatt *et al.*, o Prodigal não só apresenta uma maior acurácia na identificação de sítios de ligação ao ribossomo (RBS – *Ribosome Binding Site*), como também na identificação correta do códon de iniciação, em relação às ferramentas Genemark (Borodovsky *et al.*, 2003) e GLIMMER (Delcher *et al.*, 1999). Além disso, a integração dos dados realizada ao final da anotação, onde são removidas CDSs que estão sobrepostas a genes de ncRNAs e outras *features* identificadas pelo INFERNAL, e a própria integração do Antifam como ferramenta de filtragem, contribuem para uma anotação mais acurada.

791 Como perspectivas futuras para o aprimoramento da presente ferramenta, inclui-se a
792 integração de novos programas para caracterização das proteínas preditas, o que
793 inclui o uso da ferramenta Interproscan, para anotação estrutural de proteínas,
794 através de sua *Application Programming Interface* (API) *online*, e dos bancos de
795 dados do Pfam e SMART, que podem ser acessados localmente através de arquivos
796 de HMMs. A integração de programas para anotação funcional, como o BLAST2GO
797 (Conesa *et al.*, 2005) e BLANNOTATOR (Kankainen *et al.*, 2012), capazes de
798 atribuir códigos do *Gene Ontology* (GO) (Ashburner *et al.*, 2000; Harris *et al.*, 2004),
799 para diferentes níveis funcionais, como processos biológicos e compartimentos
800 celulares, e de ferramentas para reconstrução de rotas metabólicas, como KAAS
801 (Moriya *et al.*, 2007), Minpaths (Ye & Doak, 2009) e PathPred (Moriya *et al.*, 2010),
802 também poderão auxiliar no aprimoramento dos resultados obtidos pela anotação.
803 Por fim, futuramente, também é pretendido o desenvolvimento de uma versão de
804 uso local da ferramenta, por linha de comando, e uma API para submissão de
805 genomas em *batch*.

6 CONCLUSÃO

- A implementação da *pipeline* de anotação do Genix integra as ferramentas USEARCH, BLAST, Prodigal, tRNAscan-SE, RNAmmer, ARAGORN, INFERNAL e HMMER, e os bancos de dados Uniprot, Antifam e Rfam.
- O *back-end* do Genix é executado em um servidor Apache HTTP, utiliza um banco de dados MySQL, e é gerenciado por *scripts* escritos em linguagem Python.
- A interação do usuário com a ferramenta é realizada através de um *front-end* desenvolvido em HTML/Javascript/CSS. A interface de uso integra o *genome browser* Jbrowse, possibilitando a visualização da anotação gerada no próprio navegador.
- O Genix é capaz de gerar uma anotação mais próxima das anotações de referência, com uma menor perda de dados relevantes.

823 **7 REFERÊNCIAS**

- 824 Abeel, T., Van Parys, T., Saeys, Y., Galagan, J. & Van de Peer, Y. (2012).
825 GenomeView: a next-generation genome browser. *Nucleic Acids Res.* 40, e12.
- 826 Alikhan, N.-F., Petty, N.K., Ben Zakour, N.L. & Beatson, S.A., (2011). BLAST Ring
827 Image Generator (BRIG): simple prokaryote genome comparisons. *BMC*
828 *Genomics* 12, 402.
- 829 Alkan, C., Sajjadian, S. & Eichler, E.E. (2010). Limitations of next-generation genome
830 sequence assembly. *Nat. Methods* 8, 61–65.
- 831 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local
832 alignment search tool. *J. Mol. Biol.* 215, 403–10.
- 833 Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S.,
834 Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A.,
835 O'Donovan, C., Redaschi, N. & Yeh, L.S.L. (2004). UniProt: the Universal
836 Protein knowledgebase. *Nucleic Acids Res.* 32, D115–9.
- 837 Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the
838 flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815.
- 839 Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis,
840 A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver,
841 L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin,
842 G.M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology.
843 The Gene Ontology Consortium. *Nat. Genet.* 25, 25–9.
- 844 Assefa, S., Keane, T.M., Otto, T.D., Newbold, C. & Berriman, M. (2009). ABACAS:
845 algorithm-based automatic contiguation of assembled sequences. *Bioinformatics*
846 25, 1968–1969.
- 847 Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma,
848 K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R.,
849 Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T.,
850 Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V.,
851 Wilke, A. & Zagnitko, O. (2008). The RAST Server: rapid annotations using
852 subsystems technology. *BMC Genomics* 9, 75.
- 853 Baker, M., 2012. *De novo* genome assembly: what every biologist should know. *Nat.*
854 *Methods* 9, 333–337.
- 855 Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin,
856 V.M., Nikolenko, S.I., Pham, S., Pribelski, A.D., Pyshkin, A. V, Sirotkin, A. V,
857 Vyahhi, N., Tesler, G., Alekseyev, M.A. & Pevzner, P.A. (2012). SPAdes: a new
858 genome assembly algorithm and its applications to single-cell sequencing. *J.*
859 *Comput. Biol.* 19, 455–77.
- 860 Barbosa, E.G., Aburjaile, F.F., Ramos, R.T., Carneiro, A.R., Le Loir, Y., Baumbach,
861 J., Miyoshi, A., Silva, A. & Azevedo, V. (2014). Value of a newly sequenced
862 bacterial genome. *World J. Biol. Chem.* 5, 161–8.
- 863 Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I.,
864 Kimelman, M., Pruitt, K.D., Resenchuk, S., Tatusova, T., Yaschenko, E. & Ostell,
865 J. (2012). BioProject and BioSample databases at NCBI: facilitating capture and
866 organization of metadata. *Nucleic Acids Res.* 40, D57–63.
- 867 Beckloff, N., Starkenburg, S., Freitas, T. & Chain, P. (2012). Bacterial genome

- 868 annotation. *Methods Mol. Biol.* 881, 471–503.
- 869 Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. (2005).
870 GenBank. *Nucleic Acids Res.* 33, D34–8.
- 871 Blattner, F.R., Plunkett G., Bloch C.A, Perna, N.T., Burland, V. Riley, M., Mollado-
872 Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W.,
873 Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. & Shao, Y. (1997). The
874 Complete Genome Sequence of *Escherichia coli* K-12. *Science* (80-.). 277,
875 1453–1462.
- 876 Boetzer, M. & Pirovano, W. (2012). Toward almost closed genomes with GapFiller.
877 *Genome Biol.* 13, R56.
- 878 Boisvert, S., Laviolette, F. & Corbeil, J. (2010). Ray: Simultaneous Assembly of
879 Reads from a Mix of High-Throughput Sequencing Technologies. *J Comput Biol*,
880 17(11), 1519-1533.
- 881 Borodovsky, M., Mills, R., Besemer, J. & Lomsadze, A. (2003). Prokaryotic gene
882 prediction using GeneMark and GeneMark.hmm. *Curr. Protoc. Bioinformatics*,
883 Wiley.
- 884 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. &
885 Madden, T.L. (2009). BLAST+: architecture and applications. *BMC*
886 *Bioinformatics* 10, 421.
- 887 Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C.,
888 Sánchez Alvarado, A. & Yandell, M. (2008). MAKER: an easy-to-use annotation
889 pipeline designed for emerging model organism genomes. *Genome Res.* 18,
890 188–96.
- 891 Carver, T., Thomson, N., Bleasby, A., Berriman, M. & Parkhill, J. (2009). DNAPlotter:
892 circular and linear interactive genome visualization. *Bioinformatics* 25, 119–20.
- 893 Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.-A., Barrell, B.G.,
894 Parkhill, J., 2005. ACT: the Artemis Comparison Tool. *Bioinformatics* 21, 3422–
895 3.
- 896 *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C.*
897 *elegans*: a platform for investigating biology. *Science* 282, 2012–8.
- 898 Chaudhuri, R.R. & Pallen, M.J. (2006). xBASE, a collection of online databases for
899 bacterial comparative genomics. *Nucleic Acids Res.* 34, D335–7.
- 900 Chen, Y.-T., Peng, H.-L., Shia, W.-C., Hsu, F.-R., Ken, C.-F., Tsao, Y.-M., Chen, C.-
901 H., Liu, C.-E., Hsieh, M.-F., Chen, H.-C., Tang, C.-Y., Ku, T.-H. (2012). Whole-
902 genome sequencing and identification of *Morganella morganii* KT pathogenicity-
903 related genes. *BMC Genomics* 13 Suppl 7, S4.
- 904 Compeau, P.E.C., Pevzner, P.A. & Tesler, G. (2011). How to apply de Bruijn graphs
905 to genome assembly. *Nat. Biotechnol.* 29, 987–91.
- 906 Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. & Robles, M. (2005).
907 Blast2GO: a universal tool for annotation, visualization and analysis in functional
908 genomics research. *Bioinformatics* 21, 3674–6.
- 909 Delcher, A.L., Harmon, D., Kasif, S., White, O. & Salzberg, S.L. (1999). Improved
910 microbial gene identification with GLIMMER. *Nucleic Acids Res.* 27, 4636–41.
- 911 Dias, Z., Dias, U. & Setubal, J.C., (2012). SIS: a program to generate draft genome

- 912 sequence scaffolds for prokaryotes. *BMC Bioinformatics* 13, 96.
- 913 Durbin, R., Eddy, S.R., Krogh, A. & Mitchison, G., (1998). Biological Sequence
914 Analysis. Cambridge University Press.
- 915 Eberhardt, R.Y., Haft, D.H., Punta, M., Martin, M., O'Donovan, C. & Bateman, A.
916 (2012). AntiFam: a tool to help identify spurious ORFs in protein annotation.
917 *Database (Oxford)*. 2012, bas003.
- 918 Eddy, S.R. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.* 7,
919 e1002195.
- 920 Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST.
921 *Bioinformatics* 26, 2460–1.
- 922 Edwards, D.J. & Holt, K.E., 2013. Beginner's guide to comparative bacterial genome
923 analysis using next-generation sequence data. *Microb. Inform. Exp.* 3, 2.
- 924 Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R.,
925 Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J. &
926 Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42,
927 D222–30.
- 928 Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F.,
929 Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A. & Merrick, J.M. (1995).
930 Whole-genome random sequencing and assembly of *Haemophilus influenzae*
931 Rd. *Science* 269, 496–512.
- 932 Galardini, M., Biondi, E.G., Bazzicalupo, M. & Mengoni, A. (2011). CONTIGuator: a
933 bacterial genomes finishing tool for structural insights on draft genomes. *Source*
934 *Code Biol. Med.* 6, 11.
- 935 Goff, S.A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., Glazebrook, J.,
936 Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C.,
937 Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J.,
938 Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W., Chen, L., Cooper,
939 B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh,
940 A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D.,
941 Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M.,
942 Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J.,
943 Macalma, T., Oliphant, A. & Briggs, S. (2002). A draft sequence of the rice
944 genome (*Oryza sativa* L. ssp. japonica). *Science* 296, 92–100.
- 945 Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S.R. (2003).
946 Rfam: an RNA family database. *Nucleic Acids Res.* 31, 439–41.
- 947 Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell,
948 C.R. & Wortman, J.R. (2008). Automated eukaryotic gene structure annotation
949 using EvidenceModeler and the Program to Assemble Spliced Alignments.
950 *Genome Biol.* 9, R7.
- 951 Haft, D.H. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res.*
952 31, 371–373.
- 953 Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K.,
954 Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult,
955 C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M.,
956 Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S.,

- Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T. & White, R. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, 258D–261.
- Hernandez, D., François, P., Farinelli, L., Osterås, M. & Schrenzel, J. (2008). *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res.* 18, 802–9.
- Himmelreich, R., Hilbert, H., Plagens, H., Pirkel, E., Li, B.-C., Herrmann & R. (1996). Complete Sequence Analysis of the Genome of the Bacterium *Mycoplasma Pneumoniae*. *Nucleic Acids Res.* 24, 4420–4449.
- Holley, R.W., Apgar, J., Everett, G.A., Madison, J.T., Marquisee, M., Merrill, S.H., Penswick, J.R. & Zamir, A. (1965). Structure of a ribonucleic acid. *Science* 147, 1462–5.
- Hutchison, C.A. (2007). DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res.* 35, 6227–6237.
- Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W. & Hauser, L.J., (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11, 119.
- Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.
- Kankainen, M., Ojala, T. & Holm, L. (2012). BLANNOTATOR: enhanced homology-based function prediction of bacterial proteins. *BMC Bioinformatics* 13, 33.
- Karger, B.L. & Guttman, A. (2009). DNA sequencing by CE. *Electrophoresis* 30, S196–S202.
- Kent, W.J., 2002. BLAT--the BLAST-like alignment tool. *Genome Res* 12, 656–64.
- Köser, C.U., Ellington, M.J., Peacock, S.J., 2014. Whole-genome sequencing to control antimicrobial resistance. *Trends Genet.* 30, 401–7.
- Kremer, F.S., Eslabão, M.R., Provisor, M., Woloski, R.D.S., Ramires, O. V, Moreno, L.Z., Moreno, A.M., Hamond, C., Lilenbaum, W. & Dellagostin, O.A. (2015). Draft Genome Sequences of *Leptospira santarosai* Strains U160, U164, and U233, Isolated from Asymptomatic Cattle. *Genome Announc.* 3, e00910–15.
- Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–80.
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. & Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–45.
- Lagesen, K., Hallin, P., Rødland, E.A., Staerfeldt, H.-H., Rognes, T. & Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35, 3100–8.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford,

- 1002 A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K.,
 1003 Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C.,
 1004 Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N.,
 1005 Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough,
 1006 R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman,
 1007 R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D.,
 1008 Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C.,
 1009 McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A.,
 1010 Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K.,
 1011 Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla,
 1012 A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D.,
 1013 Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson,
 1014 D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P.,
 1015 Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A.,
 1016 Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M.,
 1017 Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell,
 1018 J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock,
 1019 G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T.,
 1020 Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R.,
 1021 Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C.,
 1022 Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K.,
 1023 Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S.,
 1024 Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L.,
 1025 Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J.,
 1026 Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M. V.,
 1027 Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou,
 1028 M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H.,
 1029 Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H.,
 1030 Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman,
 1031 A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L.,
 1032 Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R.,
 1033 Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki,
 1034 Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones,
 1035 T.A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E. V.,
 1036 Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.
 1037 V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G.,
 1038 Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J.,
 1039 Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang,
 1040 S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A.,
 1041 Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J.,
 1042 Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J. & Szustakowski, J. (2001). Initial
 1043 sequencing and analysis of the human genome. *Nature* 409, 860–921.
- 1044 Laslett, D. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in
 1045 nucleotide sequences. *Nucleic Acids Res.* 32, 11–16.
- 1046 Letunic, I., Doerks, T. & Bork, P., 2012. SMART 7: recent updates to the protein
 1047 domain annotation resource. *Nucleic Acids Res.* 40, D302–5.
- 1048 Li, W., Godzi & A. (2006). CD-HIT: a fast program for clustering and comparing large
 1049 sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–9.
- 1050 Lin, S.-H. & Liao, Y.C. (2013). CISA: contig integrator for sequence assembly of

- 1051 bacterial genomes. *PLoS One* 8, e60843.
- 1052 Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M. (2012).
1053 Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.*
1054 2012, 251364.
- 1055 Lowe, T.M. & Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of
1056 transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–64.
- 1057 Lu, C., Chen, K.-T., Huang, S.-Y. & Chiu, H.-T. (2014). CAR: contig assembly of
1058 prokaryotic draft genomes using rearrangements. *BMC Bioinformatics* 15, 381.
- 1059 Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu,
1060 Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han,
1061 C., Cheung, D.W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y.,
1062 Yang, H., Wang, J., Lam, T.-W. & Wang, J. (2012). SOAPdenovo2: an
1063 empirically improved memory-efficient short-read *de novo* assembler.
1064 *Gigascience* 1, 18.
- 1065 Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q.,
1066 Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M. & Spratt, B.G.
1067 (1998). Multilocus sequence typing: a portable approach to the identification of
1068 clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci.*
1069 U. S. A. 95, 3140–5.
- 1070 Mardis, E., McPherson, J., Martienssen, R., Wilson, R.K. & McCombie, W.R. (2002).
1071 What is finished, and why does it matter. *Genome Res.* 12, 669–71.
- 1072 Maxam, A.M. & Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl.*
1073 *Acad. Sci. U. S. A.* 74, 560–4.
- 1074 Medini, D., Donati, C., Tettelin, H., Massignani, V. & Rappuoli, R. (2005). The
1075 microbial pan-genome. *Curr. Opin. Genet. Dev.* 15, 589–94.
- 1076 Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. & Kanehisa, M. (2007). KAAS: an
1077 automatic genome annotation and pathway reconstruction server. *Nucleic Acids*
1078 *Res.* 35, W182–5.
- 1079 Moriya, Y., Shigemizu, D., Hattori, M., Tokimatsu, T., Kotera, M., Goto, S. &
1080 Kanehisa, M. (2010). PathPred: an enzyme-catalyzed metabolic pathway
1081 prediction server. *Nucleic Acids Res.* 38, W138–43.
- 1082 Myers, E.W., 2005. The fragment assembly string graph. *Bioinformatics* 21 Suppl 2,
1083 ii79–85.
- 1084 Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R. (2009). Infernal 1.0: inference of RNA
1085 alignments. *Bioinformatics* 25, 1335–7.
- 1086 Nijkamp, J., Winterbach, W., van den Broek, M., Daran, J.-M., Reinders, M. & de
1087 Ridder, D. (2010). Integrating genome assemblies with MAIA. *Bioinformatics* 26,
1088 i433–9.
- 1089 Otto, T.D., Sanders, M., Berriman, M. & Newbold, C. (2010). Iterative Correction of
1090 Reference Nucleotides (iCORN) using second generation sequencing
1091 technology. *Bioinformatics* 26, 1704–7.
- 1092 Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., Edwards,
1093 R.A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A.R., Xia, F. &
1094 Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes
1095 using Subsystems Technology (RAST). *Nucleic Acids Res.* 42, D206–14.

- 1096 Pareek, C.S., Smoczynski, R. & Tretyn, A. (2011). Sequencing technologies and
1097 genome sequencing. *J. Appl. Genet.* 52, 413–35.
- 1098 Pareja-Tobes, P., Manrique, M., Pareja-Tobes, E., Pareja, E., Tobes, R. (2012). BG7:
1099 a new approach for bacterial genome annotation designed for next generation
1100 sequencing data. *PLoS One* 7, e49239.
- 1101 Petersen, T.N., Brunak, S., von Heijne, G. & Nielsen, H. (2011). SignalP 4.0:
1102 discriminating signal peptides from transmembrane regions. *Nat. Methods* 8,
1103 785–6.
- 1104 Pevzner, P.A., Tang, H. & Waterman, M.S. (2001). An Eulerian path approach to
1105 DNA fragment assembly. *Proc. Natl. Acad. Sci. U. S. A.* 98, 9748–53.
- 1106 Piro, V.C., Faoro, H., Weiss, V.A., Steffens, M.B.R., Pedrosa, F.O. & Souza, E.M.,
1107 Raittz, R.T. (2014). FGAP: an automated gap closing tool. *BMC Res. Notes* 7,
1108 371.
- 1109 Pirovano, W., Boetzer, M., Derks, M.F.L. & Smit, S. (2015). NCBI-compliant genome
1110 submissions: tips and tricks to save time and money. *Brief. Bioinform.* bbv104.
- 1111 Poptsova, M.S., Il'icheva, I.A., Nechipurenko, D.Y., Panchenko, L.A., Khodikov, M.
1112 V., Oparina, N.Y., Polozov, R. V., Nechipurenko, Y.D. & Grokhovsky, S.L.
1113 (2014). Non-random DNA fragmentation in next-generation sequencing. *Sci.*
1114 *Rep.* 4, 4532.
- 1115 Ricker, N., Qian, H. & Fulthorpe, R.R. (2012). The limitations of draft assemblies for
1116 understanding prokaryotic adaptation and evolution. *Genomics* 100, 167–175.
- 1117 Rissman, A.I., Mau, B., Biehl, B.S., Darling, A.E., Glasner, J.D., Perna, N.T. (2009).
1118 Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* 25,
1119 2071–3.
- 1120 Ronaghi, M., 1998. DNA SEQUENCING: A Sequencing Method Based on Real-Time
1121 Pyrophosphate. *Science* 80, 363–365.
- 1122 Ronen, R., Boucher, C., Chitsaz, H. & Pevzner, P. (2012). SEQuel: improving the
1123 accuracy of genome assemblies. *Bioinformatics* 28, i188–96.
- 1124 Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A. &
1125 Barrell, B. (2000). Artemis: sequence visualization and annotation.
1126 *Bioinformatics* 16, 944–945.
- 1127 Sallet, E., Gouzy, J., Schiex, T. (2014). EuGene-PP: a next-generation automated
1128 annotation pipeline for prokaryotic genomes. *Bioinformatics* 30, 2659–61.
- 1129 Sanger, F., Nicklen, S. & Coulson, A.R. (1977). DNA sequencing with chain-
1130 terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–7.
- 1131 Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*
1132 30, 2068–9.
- 1133 Servant, F., Bru, C., Carrère, S., Courcelle, E., Gouzy, J., Peyruc, D. & Kahn, D.
1134 (2002). ProDom: automated clustering of homologous domains. *Brief. Bioinform.*
1135 3, 246–51.
- 1136 Sette, A. & Rappuoli, R. (2010). Reverse vaccinology: developing vaccines in the era
1137 of genomics. *Immunity* 33, 530–41.
- 1138 Shendure, J., Ji & H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* 26,
1139 1135–45.

- 1140 Simpson, J.T. & Durbin, R. (2012). Efficient *de novo* assembly of large genomes
1141 using compressed data structures. *Genome Res.* 22, 549–56.
- 1142 Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M. & Birol, I.
1143 (2009). ABySS: a parallel assembler for short read sequence data. *Genome*
1144 *Res.* 19, 1117–23.
- 1145 Skinner, M.E., Uzilov, A. V, Stein, L.D., Mungall, C.J. & Holmes, I.H. (2009).
1146 JBrowse: a next-generation genome browser. *Genome Res.* 19, 1630–8.
- 1147 Sommer, D.D., Delcher, A.L., Salzberg, S.L. & Pop, M. (2007). Minimus: a fast,
1148 lightweight genome assembler. *BMC Bioinformatics* 8, 64
- 1149 Staden, R. (1979). A strategy of DNA sequencing employing computer programs.
1150 *Nucleic Acids Res.* 6, 2601–10.
- 1151 Stadler, P.F. (2014). Class-specific prediction of ncRNAs. *Methods Mol. Biol.* 1097,
1152 199–213.
- 1153 Stothard, P., Wishart, D.S. (2005). Circular genome visualization and exploration
1154 using CGView. *Bioinformatics* 21, 537–9.
- 1155 Tatusov, R.L., Galperin, M.Y., Natale, D.A., Koonin, E.V. (2000). The COG database:
1156 a tool for genome-scale analysis of protein functions and evolution. *Nucleic*
1157 *Acids Res.* 28, 33–6.
- 1158 Tringe, S.G. & Hugenholtz, P. (2008). A renaissance for the pioneering 16S rRNA
1159 gene. *Curr. Opin. Microbiol.* 11, 442–6.
- 1160 Tsai, I.J., Otto, T.D. & Berriman, M. (2010). Improving draft assemblies by iterative
1161 mapping and assembly of short reads to eliminate gaps. *Genome Biol.* 11, R41.
- 1162 Van Domselaar, G.H., Stothard, P., Shrivastava, S., Cruz, J.A., Guo, A., Dong, X.,
1163 Lu, P., Szafron, D., Greiner, R. & Wishart, D.S. (2005). BASys: a web server for
1164 automated bacterial genome annotation. *Nucleic Acids Res.* 33, W455–9.
- 1165 Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith,
1166 H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P.,
1167 Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H.,
1168 Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos,
1169 G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N.,
1170 Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos,
1171 R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A.,
1172 Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K.,
1173 Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M.,
1174 Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V.,
1175 Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong,
1176 F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A.,
1177 Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V, Milshina,
1178 N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch,
1179 D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X.,
1180 Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M.,
1181 Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao,
1182 S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T.,
1183 Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I.,
1184 Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher,
1185 S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S.,

- 1186 Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun,
1187 S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F.,
1188 Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T.,
1189 McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E.,
1190 Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D.,
1191 Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E.,
1192 Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S.,
1193 Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril,
1194 J.F., Guigó, R., Campbell, M.J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H.,
1195 Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N.,
1196 Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S.,
1197 Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J.,
1198 Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M.,
1199 Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser,
1200 K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J.,
1201 Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J.,
1202 Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros,
1203 W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M.,
1204 Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M.,
1205 Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M.,
1206 Wu, D., Wu, M., Xia, A., Zandieh, A. & Zhu, X. (2001). The sequence of the
1207 human genome. *Science* 291, 1304–51.
- 1208 Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P.,
1209 Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E.,
1210 Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B.,
1211 Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown,
1212 S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S.,
1213 Chiaromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins,
1214 F.S., Cook, L.L., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V.,
1215 Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K.D., Deri, J., Dermitzakis,
1216 E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn,
1217 D.M., Eddy, S.R., Elnitski, L., Emes, R.D., Eswara, P., Eyraas, E., Felsenfeld, A.,
1218 Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S.,
1219 Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N.,
1220 Goodstadt, L., Grafham, D., Graves, T.A., Green, E.D., Gregory, S., Guigó, R.,
1221 Guyer, M., Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs,
1222 A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I.,
1223 Jaffe, D.B., Johnson, L.S., Jones, M., Jones, T.A., Joy, A., Kamal, M., Karlsson,
1224 E.K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W.J.,
1225 Kirby, A., Kolbe, D.L., Korf, I., Kucherlapati, R.S., Kulbokas, E.J., Kulp, D.,
1226 Landers, T., Leger, J.P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd,
1227 C., Lucas, S., Ma, B., Maglott, D.R., Mardis, E.R., Matthews, L., Mauceli, E.,
1228 Mayer, J.H., McCarthy, M., McCombie, W.R., McLaren, S., McLay, K.,
1229 McPherson, J.D., Meldrim, J., Meredith, B., Mesirov, J.P., Miller, W., Miner, T.L.,
1230 Mongin, E., Montgomery, K.T., Morgan, M., Mott, R., Mullikin, J.C., Muzny, D.M.,
1231 Nash, W.E., Nelson, J.O., Nhan, M.N., Nicol, R., Ning, Z., Nusbaum, C.,
1232 O'Connor, M.J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G.,
1233 Pepin, K.H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C.S., Poliakov, A.,
1234 Ponce, T.C., Ponting, C.P., Potter, S., Quail, M., Reymond, A., Roe, B.A.,
1235 Roskin, K.M., Rubin, E.M., Rust, A.G., Santos, R., Sapojnikov, V., Schultz, B.,

- 1236 Schultz, J., Schwartz, M.S., Schwartz, S., Scott, C., Seaman, S., Searle, S.,
 1237 Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J.B., Slater, G., Smit,
 1238 A., Smith, D.R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C.,
 1239 Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J.,
 1240 Ucla, C., Ureta-Vidal, A., Vinson, J.P., Von Niederhausern, A.C., Wade, C.M.,
 1241 Wall, M., Weber, R.J., Weiss, R.B., Wendl, M.C., West, A.P., Wetterstrand, K.,
 1242 Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R.K.,
 1243 Winter, E., Worley, K.C., Wyman, D., Yang, S., Yang, S.-P., Zdobnov, E.M.,
 1244 Zody, M.C. & Lander, E.S. (2002). Initial sequencing and comparative analysis
 1245 of the mouse genome. *Nature* 420, 520–62.
- 1246 Watson, J.D. & Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for
 1247 deoxyribose nucleic acid. *Nature* 171, 737–8.
- 1248 Yandell, M. & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation.
 1249 *Nat. Rev. Genet.* 13, 329–42.
- 1250 Yao, G., Ye, L., Gao, H., Minx, P., Warren, W.C. & Weinstock, G.M. (2012). Graph
 1251 accordance of next-generation sequence assemblies. *Bioinformatics* 28, 13–6.
- 1252 Ye, Y. & Doak, T.G. (2009). A parsimony approach to biological pathway
 1253 reconstruction/inference for genomes and metagenomes. *PLoS Comput. Biol.* 5,
 1254 e1000465.
- 1255 Zdobnov, E.M. & Apweiler, R. (2001). InterProScan: an integration platform for the
 1256 signature-recognition methods in InterPro. *Bioinformatics* 17, 847–8.
- 1257 Zerbino, D.R. & Birney, E. (2008). Velvet: algorithms for de novo short read assembly
 1258 using de Bruijn graphs. *Genome Res.* 18, 821–9.
- 1259 Zimin, A. V, Smith, D.R., Sutton, G. & Yorke, J.A. (2008). Assembly reconciliation.
 1260 *Bioinformatics* 24, 42–5.
- 1261

1262 8 ANEXOS

1263

1264 Anexo A – Registro do Genix protocolado no Instituto Nacional de Propriedade

1265 Industrial

INPI

PEDIDO DE REGISTRO DE PROGRAMA DE COMPUTADOR

protocolo

INPI INSTITUTO NACIONAL DE PROPRIEDADE INDUSTRIAL
PROTÓTIPO GERAL
11/11/2015 016150001437
11:53 DERS
BR 51 2015 001311 4

IDENTIFICAÇÃO DO PEDIDO (Para uso do INPI)

Número do Pedido _____ Protocolo, Data e Hora _____

DADOS DO AUTOR DO PROGRAMA

Nº de Autores 2 Se mais de um, preencha a "Continuação", com todos os dados solicitados neste Quadro. Date e assine.

CPF* 023.766.470-45

Nome FREDERICO SCHMITT KREMER

Nome Abreviado, pseudônimo ou sinal convencional (se houver) _____

Data de Nascimento 29/07/1992 Nacionalidade Brasileiro

Endereço MAJOR CÍCERO DE GOES MONTEIRO

511

Cidade PELOTAS UF RS País Brasil

CEP 96.015-190 Telefone 33055814 FAX _____

E-mail fred.s.kremer@gmail.com

DADOS DO TITULAR DOS DIREITOS PATRIMONIAIS

Nº de Titulares 1 Se mais de um, preencha a "Continuação", com todos os dados solicitados neste Quadro. Date e assine.

CPF/CNPJ* 92242080000100

Nome/Razão Social UNIVERSIDADE FEDERAL DE PELOTAS

Nome abreviado, pseudônimo ou sinal convencional (se houver) UFPEL

Data de Nascimento _____ Nacionalidade/Origem _____

Endereço RUA GOMES CARNEIRO, 01 - ED. DELFIM MENDES SILVEIRA - 4º ANDAR - SALA 411

CAMPUS PORTO - CENTRO

Cidade PELOTAS UF RS País BRASIL

CEP 96.010-610 Telefone 5339211225 FAX _____

E-mail cit@ufpel.edu.br

☐ SIM, este Titular é Pessoa Jurídica. Caso afirmativo, assinale a melhor classificação:

☐ Órgão Público ☐ Sociedade com Intuito não Econômico ☐ Microempresa ☐ Software House

☒ Instituição Pública de Ensino ou Pesquisa ☐ Instituição Privada de Ensino ou Pesquisa ☐ Outras

ENDEREÇO PARA CORRESPONDÊNCIA E CONTATO (Preencha apenas o necessário)

Toda correspondência será enviada para: ☐ O Procurador ou ☒ O Titular acima ou

☐ Escaninho nº _____ ☐ Representação INPI em: _____ ☐ O Endereço abaixo:

Nome _____

Endereço _____

Cidade _____ UF _____ País _____

CEP _____ Telefone _____ FAX _____

E-mail _____

Modelo 1 (folha 1/2) E

1266

1267 Figura 7. Registro protocolado dia 11 de novembro de 2015. Código do protocolo:

1268 BR5120150013114

1269 **Anexo B - Relação de genomas anotados pelo Genix já disponíveis no**
 1270 **Genbank**
 1271

Organismo	Código de Acesso	Referência
<i>Leptospira kirschneri</i> str. 61H	JSVJ00000000	Manuscrito sendo revisado
<i>Leptospira santarosai</i> str. U160	LAYP00000000	(Kremer et al., 2015)
<i>Leptospira santarosai</i> str. U164	LAZM00000000	(Kremer et al., 2015)
<i>Leptospira santarosai</i> str. U233	LAZN00000000	(Kremer et al., 2015)
<i>Leptospira interrogans</i> str. Aceguá	LCZF00000000	Manuscrito sendo revisado
<i>Leptospira interrogans</i> str. Preá	LJBO00000000	Manuscrito sendo revisado
<i>Leptospira interrogans</i> str. RCA	LJBP00000000	Manuscrito sendo revisado
<i>Leptospira interrogans</i> str. Capivara	LJBQ00000000	Manuscrito sendo revisado

1272

1273