ReNoteWeb - Plataforma Web para Melhoramento de Montagem e Enriquecimento da Anotação de Genomas Procariotos

Gislenne da Silva Moia[1], Adonney Allan de Oliveira Veras[2].

[1]Faculdade de Engenharia de Computação, Universidade Federal do Pará campus Tucuruí (CAMTUC-UFPA), Pará, Brasil

[2]Faculdade de Computação, Universidade Federal do Pará campus Castanhal (FACOMP/CCAST), Pará, Brasil

*allanveras@ufpa.br

RESUMO

A redução do tempo e custos para realizar o sequenciamento de DNA, resultou no aumento significativo no depósito de informações biológicas em bancos de dados públicos, como NCBI (National Center for Biotechnology Information). Além disso, impulsionou a genômica, assim como inúmeras outras análises das ciências ômicas. A produção de grande volume de dados por execução culminou na necessidade do desenvolvimento de algoritmos capazes de manusear estas informações e auxiliar nas análises, a exemplo, tem-se a montagem e anotação de genomas procariotos. Ao longo dos anos, vários pipelines e ferramentas computacionais foram desenvolvidos com o objetivo de automatizar essa tarefa e consequentemente reduzir o tempo total para se conhecer o conteúdo gênico de um determinado organismo, principalmente de organismos não-modelo, e colaborar com a identificação de possíveis alvos com aplicabilidade biotecnológica. No caso das ferramentas de anotação automática, observa-se a alta acurácia dos resultados. Entretanto, isto não exime de se realizar o processo de curadoria manual, onde as informações inferidas no processo automático são checadas e enriquecidas pelos curadores. Esta tarefa demanda um tempo que é diretamente proporcional à quantidade de produtos gênicos do organismo alvo do estudo.

Com o intuito de auxiliar neste processo é apresentado a ferramenta web denominada ReNoteWeb, dotada de uma interface simples e intuitiva, para realização do processo de melhoria da montagem com a possibilidade de identificar produtos ausentes da sequência

genômica original ou resultado obtido em montagem prévia. Além disto, o ReNoteWeb é capaz de realizar o processo de anotação de todos os produtos, baseado em informações obtidas em bases de dados externas de alta acurácia. A engine responsável pela realização do tratamento dos dados foi desenvolvida em JAVA e a plataforma web utiliza os recursos do framework Yii, a anotação produzida por esta plataforma visa redução do tempo total no processo de curadoria manual.

Para validação da ferramenta, foram utilizados vinte organismos. A eficiência foi constatada por meio da comparação da anotação desses mesmos organismos disponíveis no banco de dados do NCBI e anotação realizada na plataforma RAST. A ferramenta está disponível em: http://biod.ufpa.br/renoteweb/.

PALAVRAS-CHAVES: Genoma, Anotação, Montagem, Melhoria da Montagem, ferramenta web.

INTRODUÇÃO

A evolução nas tecnologias de sequenciamento NGS (Next Generation Sequencing), promoveram contínuas reduções nos custos de sequenciamento e impulsionou as análises ômicas como um todo, disponibilizando uma maior quantidade no número de genomas disponíveis. Consequentemente, estimulou o desenvolvimento de novos algoritmos e ferramentas computacionais capazes de lidar com o volume maior de dados produzidos por estas plataformas e auxiliar na realização de inúmeras análises, a exemplo a montagem e anotação de genomas (Kremer et al., 2017).

A montagem de um genoma consiste no processo de reconstrução de genomas a partir de sequências de DNA oriundas do processo de sequenciamento (Koren et al., 2014; Miller et al., 2010). Diversos programas foram desenvolvidos para automatizar esse processo, como exemplo, SPAdes (Bankevich et al., 2012), Velvet (Zerbino, 2010), Megahit (Li et al., 2015), entre outros.

Contudo, o conhecimento sobre o conteúdo gênico presente nesse organismo é revelado após a sua anotação, esse processo pode ser feito de duas maneiras, automática ou manual. Na anotação automática (Richardson & Watson, 2013), às informações biológicas são atribuídas as ORFs (Open Reading Frame) por meio de várias estratégias, como, a

similaridade entre a sequência utilizada como *query* e a presente no banco de dados definida como *subject*, e dentre as ferramentas que executam esta tarefas podemos citar o PROKKA (Seemann, 2014), DFAST (Tanizawa et al., 2018), RATT (Otto et al., 2011), RAST (Aziz et al., 2008) e MAKER (Cantarel et al., 2008).

Na etapa de curadoria manual, estas informações são avaliadas e podem ser enriquecidas pelos curadores, a fim de se evitar a propagação de erros (Odell et al., 2017; Pfeiffer & Oesterhelt, 2015). Isto é realizado por meio de busca em bancos de dados curados, como UniProt (Bateman et al., 2017), Refseq (Haft et al., 2018), entre outros.

Observa-se, no entanto, que inúmeras aplicações para montagem e anotação foram desenvolvidas para um ambiente desktop e especificamente para serem executadas sobre o sistema operacional Linux. Geralmente a execução destas ferramentas é feita por meio de linhas de comandos extensas e complexas, sendo que a quantidade de parâmetros é diretamente proporcional à complexidade da tarefa a ser realizada. Além disso, os recursos computacionais e humanos atuantes na área de computação dos pequenos grupos de pesquisa são insuficientes para realizar tais tarefas, sendo impactadas diretamente na produção científica (Lantz et al., 2018).

Em decorrência disto, a plataforma web ReNoteWeb foi desenvolvida sendo dotada de uma interface gráfica intuitiva. A ferramenta possibilita serviços automatizados para realizar o processo de melhoria da montagem, com a possibilidade de identificação de produtos ausentes da sequência genômica original. E por meio do processo de anotação dos resultados obtidos, possibilita a agregação de conhecimento sobre o organismo alvo do estudo através do enriquecimento das informações gênicas obtidas em banco de dados de alta acurácia.

MATERIAIS E MÉTODOS

Linguagem de programação e banco de dados

A interface web do ReNoteWeb foi desenvolvida através da framework Yii (https://www.yiiframework.com/), o SGBD responsável pelo controle de projetos utilizado foi o MySQL versão 8.0.23 (https://www.mysql.com/).

Os módulos responsáveis pelo processo de melhoria de montagem e anotação são Pan2Hgene (Silva de Oliveira et al., 2021) e CODON Software (Merlin et al., 2021) respectivamente, ambos foram desenvolvidos utilizando a linguagem de programação JAVA (http://www.oracle.com/).

Pipeline

O pipeline de execução do ReNoteWeb (Figura 01) é composto por dois módulos principais: (i) melhoria de montagem e (ii) anotação.

Os arquivos de entrada para o módulo (i) são: sequência completa ou em rascunho do genoma no formato FASTA, e o arquivo contendo as leituras brutas no formato FASTQ (single ou paired). O arquivo de leituras brutas é mapeado contra o arquivo FASTA previamente montado, para identificação de leituras que não apresentem *match* denominadas de unmapped reads. Na etapa seguinte, estas leituras são utilizadas no processo de montagem *de novo* usando o montador SPAdes (Bankevich et al., 2012). O resultado desta etapa é anotado através do software Prokka (Seemann, 2014). As sequências de CDS presentes no arquivo de anotação são extraídas e alinhadas contra o arquivo FASTA de entrada. Os produtos que não obtiverem *match*, ou seja ausentes da sequência original, são adicionados ao final do arquivo da sequência do Genoma no formato FASTA. (Silva de Oliveira et al., 2021).

O arquivo de entrada necessário para este módulo (ii) é sequência do genoma completo ou em rascunho no formato FASTA. Neste módulo, uma Máquina de Estados Finitos é utilizada no processo de identificação das CDS. Posteriormente, as sequências destas possíveis CDS são mapeadas contra o banco de dados do Uniprot, de acordo com os valores de parâmetros fornecidos pelo usuário. Os dados de mapeamento que contém nome do produto, valor de identidade, similaridade, nome do organismo ao qual o produto está vinculado são baixados, após isto são executados uma série de filtros para resolver regiões de sobreposição de CDS, e determinar, baseado nos valores de identidade e acurácia, qual a informação de produto vai constar na anotação. Como resultado o usuário pode exportar um arquivo no formato EMBL ou um projeto no formato do CODON software para posterior curadoria manual.

Vale destacar que esse processo de anotação é automatizado e os resultados produzidos nesta etapa necessitam de curadoria manual posterior que pode ser realizada

diretamente em outros programas, como por exemplo o CODON Software (Merlin et al., 2021).

Os módulos deste pipeline podem ser executados de maneira independente, a escolha do usuário, tornando possível a execução inclusive de ambos, o que caracteriza o pipeline completo.

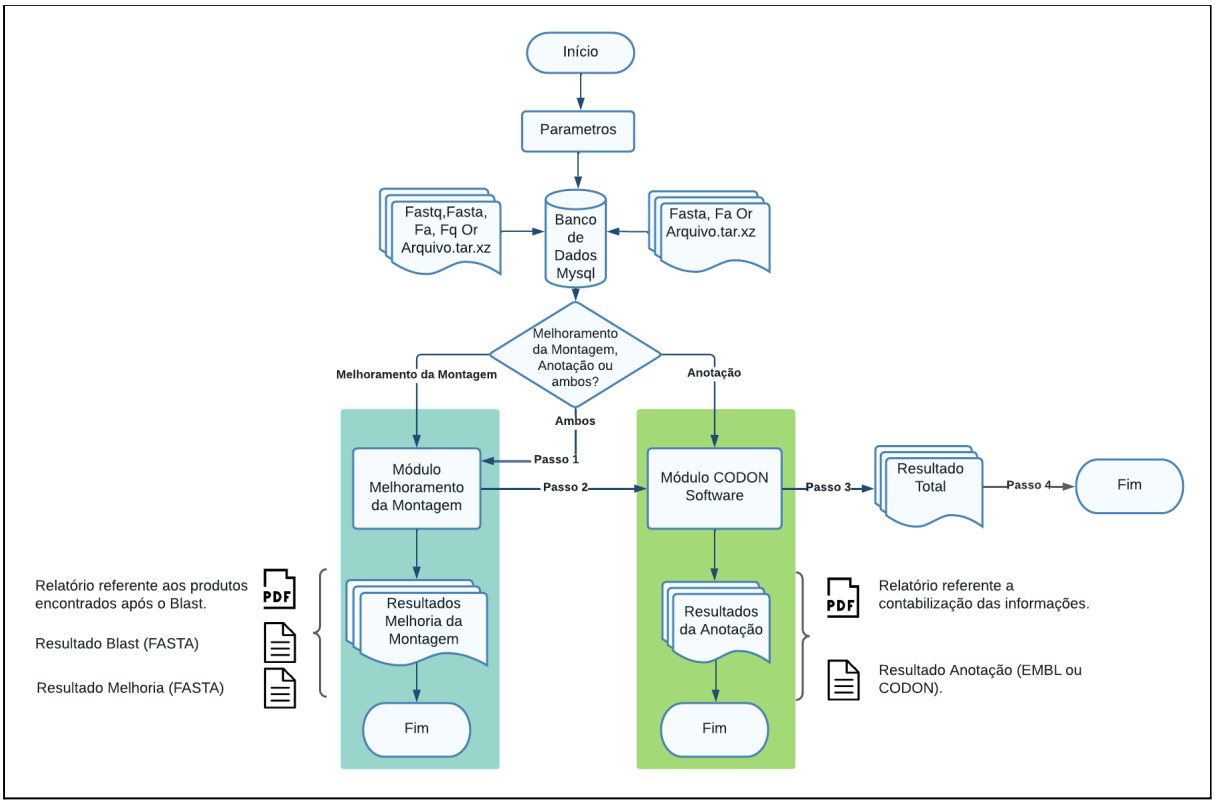

Figura 01. Pipeline ReNoteWeb: Destacado na cor azul o Módulo de melhoria da montagem e na cor verde o Módulo de anotação. As setas com a indicação dos passos sinalizam a execução do pipeline completo.

Validação da ferramenta

A validação da ferramenta foi realizada através de vinte organismos de diferentes espécies, listados na tabela 1, disponíveis no banco de dados do National Center for Biotechnology Information NCBI (https://www.ncbi.nlm.nih.gov/).

Tabela 1. Organismos utilizados para a validação da ferramenta

NOME DO ORGANISMO	SRA NÚMERO	BIBLIOTECA
Acinetobacter baumannii ATCC 17978	SRR4298913	Paired
Bacteroides fragilis NCTC 9343	SRR11462747	Paired
Corynebacterium diphtheriae NCTC 13129	SRR5481222	Single

Corynebacterium doosanense CAU 212 = DSM 45436	SRR3947906	Paired
Corynebacterium glutamicum strain ATCC 13032	SRR638977	Single
Escherichia coli O103:H2 str. 12009	ERR351260	Paired
Escherichia coli O111:H- str. 11128	ERR351258	Paired
Escherichia coli O26:H11 str. 11368	ERR351259	Paired
Escherichia coli str. K-12 substr. MG1655	SRR13093884	Paired
Escherichia coli strain P15-385	SRR13209798	Paired
Escherichia coli strain ST865	SRR15293247	Paired
Escherichia coli strain ZH193	SRR933455	Paired
Klebsiella pneumoniae SMKP03	DRR223366	Paired
Klebsiella pneumoniae strain ABFPV	SRR8778550	Paired
Mycobacterium bovis BCG Pasteur 1173P2	SRR14563279	Paired
Mycobacterium tuberculosis H37Rv	SRR12062794	Paired
Rhodococcus erythropolis strain X5	SRR10323925	Paired
Rhodococcus jostii RHA1	SRR8843224	Paired
Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150	ERR5192587	Paired
Salmonella enterica subsp. enterica serovar Typhimurium str. 14028S	ERR4147199	Paired

A validação foi dividida em três etapas. A primeira, consiste na comparação entre os resultados do ReNoteWeb, RAST e as informações de anotação depositadas no NCBI. Na segunda etapa, é realizada a comparação dos resultados obtidos pelo ReNoteWeb com e sem o uso do módulo de melhoria da montagem. E na terceira etapa, foi comparado os resultados obtidos no ReNoteWeb e no RAST, contudo, os resultados desta etapa são acrescidos dos produtos identificados no processo de melhoria da montagem realizada pelo ReNoteWeb.

O software Artemis (Rutherford et al., 2000) foi utilizado para realizar a contagem das informações relacionadas à quantidade de sequências de codificações (CDS), produtos, produtos com sigla de gene, rRNA (RNA ribossomal), tRNA (RNA transportador) e proteínas hipotéticas de cada organismo presente na análise.

RESULTADOS E DISCUSSÕES

A Tabela 2 lista os resultados obtidos no processo de anotação com o ReNoteWeb em comparação a anotação do RAST e a depositada no NCBI. A análise demonstra que o número de CDS identificadas no ReNoteWeb é maior, fato que se repete nos demais resultados, o

qual houve também, na maioria dos organismos, um aumento na quantidade de produtos contendo sigla de gene e na quantidade de proteínas hipotéticas. Vale destacar que, a anotação produzida pela ferramenta RAST não fornece uma sigla para gene, portanto não foi possível realizar a verificação do número total de produtos com a sigla de um gene.

Tabela 2. Análise dos resultados da anotação: Na primeira coluna são exibidos os SRA (sequence read archive), seguido das ferramentas utilizadas e as demais colunas a quantidade de produtos gênicos, produtos com sigla de gene e quantidade de proteínas hipotéticas.

SRA NÚMERO	ANOTAÇÃO	QUANTIDADE DE CDS	QUANTIDADE DE PRODUTOS	NÚMERO DE PRODUTOS NOMEADOS COM A SIGLA GENE	TOTAL DE PROTEÍNAS HIPOTÉTICAS
SRR4298913	NCBI	3725	2183	892	650
	RAST	3787	2915		872
	RENOTEWEB	3905	955	2097	853
SRR11462747	NCBI	4260	2305	634	1321
	RAST	4557	2872		1685
	RENOTEWEB	5034	2058	2014	962
SRR5481222	NCBI	2320	1169	536	615
	RAST	2415	1924		491
	RENOTEWEB	2375	1055	1067	253
SRR3947906	NCBI	2470	1747	43	680
	RAST	2612	1674		938
	RENOTEWEB	2501	1482	641	378
SRR638977	NCBI	3058	2118	237	703
	RAST	3211	2361		850
	RENOTEWEB	3250	1615	777	858
ERR351260	NCBI	5264	1062	3786	416
	RAST	5572	4912		660
	RENOTEWEB	6662	682	4459	1521
ERR351258	NCBI	5264	1131	3797	336
	RAST	5549	4949		600
	RENOTEWEB	6591	571	4565	1455
ERR351259	NCBI	5609	1322	3802	485
	RAST	5947	5222		725
	RENOTEWEB	7060	827	4554	1679
SRR13093884	NCBI	4489	2744	632	1113

	RAST	4557	4288		269
	RENOTEWEB	5544	164	4286	1094
SRR13209798	NCBI	4690	632	3825	233
	RAST	4885	4444		441
	RENOTEWEB	6011	461	4216	1334
SRR15293247	NCBI	4495	491	3801	203
	RAST	4637	4278		359
	RENOTEWEB	5742	346	4103	1293
SRR933455	NCBI	4801	852	3689	260
	RAST	4994	4482		512
	RENOTEWEB	6090	457	4229	1404
DRR223366	NCBI	4727	2554	1513	660
	RAST	4946	4435		511
	RENOTEWEB	5905	643	3948	1314
SRR8778550	NCBI	5202	3186	1534	482
	RAST	5173	4559		614
	RENOTEWEB	6182	530	4219	1433
SRR14563279	NCBI	3988	1042	1676	1270
	RAST	4251	3382		869
	RENOTEWEB	5045	903	2508	1634
SRR12062794	NCBI	4031	1430	1953	648
	RAST	4299	3427		872
	RENOTEWEB	5155	1551	2605	999
SRR10323925	NCBI	5918	4466	704	748
	RAST	6183	4051		2132
	RENOTEWEB	6332	2776	2245	1311
SRR8843224	NCBI	7211	4053	694	2464
	RAST	7613	5300		2313
	RENOTEWEB	7503	4195	1372	1936
ERR5192587	NCBI	4093	43	3194	856
	RAST	4658	4163		495
	RENOTEWEB	5406	564	3602	1240
ERR4147199	NCBI	5372	857	3257	1258
	RAST	4860	4338		522
	RENOTEWEB	5827	438	4079	1310

A Figura 02 exibe a comparação entre a quantidade de produtos gênicos que possuem sigla de gene identificados com a anotação usando o ReNoteWeb e a anotação depositada no NCBI. Com a análise do gráfico pode-se inferir que a identificação de um número maior de produtos com sigla de gene é devido a acurácia das informações obtidas no banco de dados do UniProt, juntamente com o método alternativo de identificação de CDS utilizados no processo de anotação.

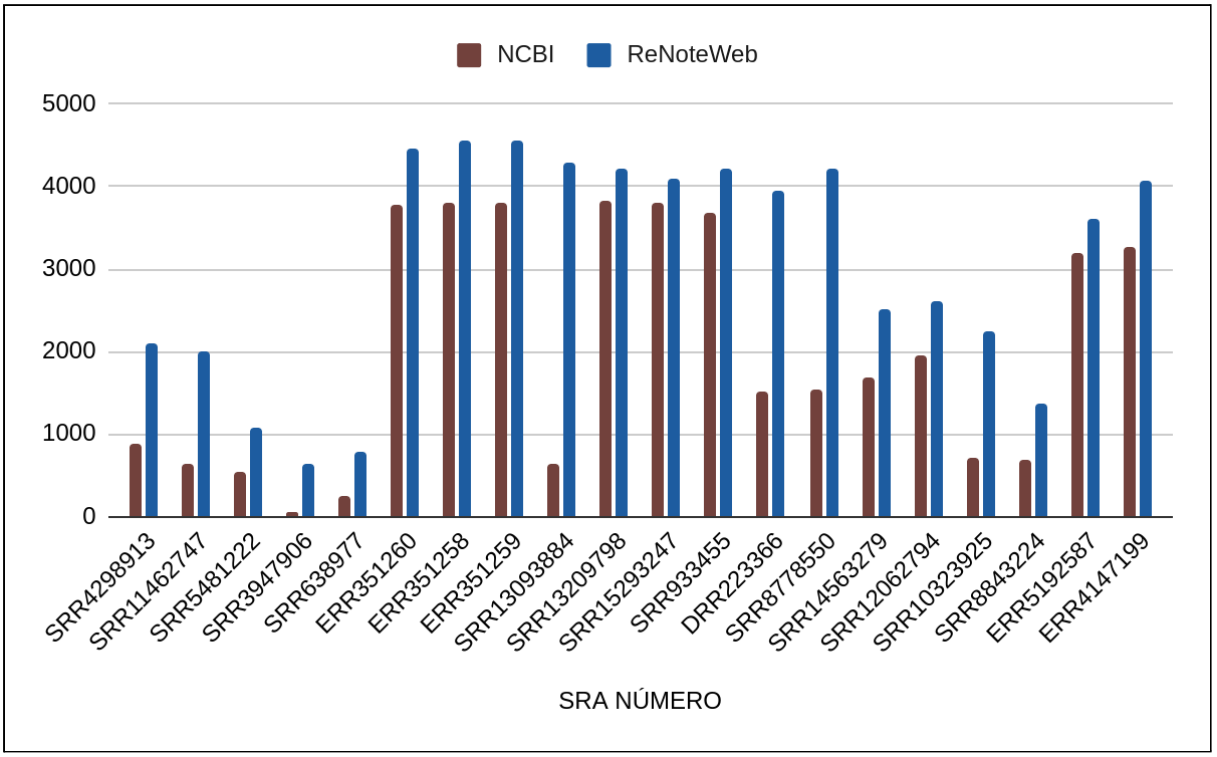

Figura 02. Quantidade total de produtos com sigla genes identificados: Análise comparativa entre as quantidades de produtos com sigla de gene identificados nas anotações do ReNoteWeb e a anotação depositada no NCBI.

A Figura 03 demonstra a comparação sobre a quantidade de proteínas hipotéticas identificadas nas anotações feitas na plataforma Web RAST, ReNoteWeb e anotação depositada no NCBI. A análise demonstra um número maior de proteínas hipotéticas identificadas na anotação realizada pelo ReNoteWeb, este comportamento é atribuído a abordagem utilizada no processo de identificação das ORF's utilizada na engine de anotação do ReNoteWeb, vale salientar a necessidade de posterior curadoria manual e análise destes produtos que podem conter algo de interesse biotecnológico.

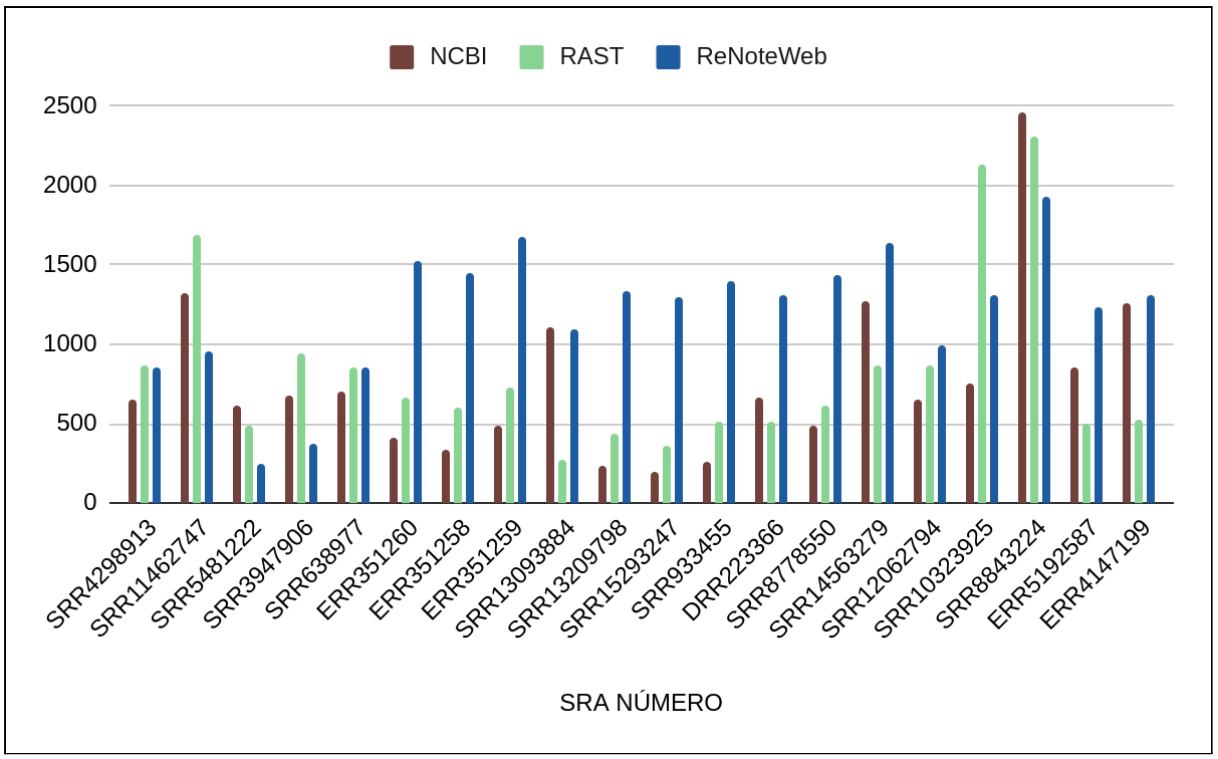

Figura03. Comparativo de quantidade total de proteínas hipotéticas: Comparação realizada sobre a quantidade total de proteínas hipotéticas identificadas nas anotações do RAST e ReNoteWeb em comparação a anotação depositada no NCBI.

A Tabela 3 lista de forma comparativa o resultado sobre a quantidade de rRNA e tRNA identificados nas anotações do ReNoteWeb, RAST e a depositada no NCBI. Na maioria dos organismos é possível observar que a quantidade de tRNA e rRNA identificados na anotação do ReNoteWeb foram iguais aos identificados na anotação depositada no NCBI.

Tabela 3. Quantidade de tRNA e rRNA encontrados para cada organismo

SRA NÚMERO	FERRAMENTA	rRNA	tRNA	SRA NÚMERO	FERRAMENTA	rRNA	tRNA
SRR4298913	NCBI	18	72	SRR15293247	NCBI	22	88
	RAST	18	72		RAST	22	90
	RENOTEWEB	18	72		RENOTEWEB	9	86
SRR11462747	NCBI	19	73	SRR933455	NCBI	22	86
	RAST	19	71		RAST	22	87
	RENOTEWEB	19	73		RENOTEWEB	9	87
SRR5481222	NCBI	15	54	DRR223366	NCBI	25	86
	RAST	15	54		RAST	25	86
	RENOTEWEB	15	54		RENOTEWEB	10	85
SRR3947906	NCBI	12	53	SRR8778550	NCBI	25	88
	RAST	8	53		RAST	25	88

	RENOTEWEB	12	54		RENOTEWEB	11	86
SRR638977	NCBI	18	60	SRR14563279	NCBI	3	47
	RAST	18	60		RAST	3	46
	RENOTEWEB	17	59		RENOTEWEB	3	47
ERR351260	NCBI	22	98	SRR12062794	NCBI	3	45
	RAST	22	98		RAST	3	44
	RENOTEWEB	22	90		RENOTEWEB	3	55
ERR351258	NCBI	22	107	SRR10323925	NCBI	15	53
	RAST	22	107		RAST	15	53
	RENOTEWEB	22	95		RENOTEWEB	6	37
ERR351259	NCBI	22	101	SRR8843224	NCBI	12	50
	RAST	22	101		RAST	12	50
	RENOTEWEB	22	92		RENOTEWEB	12	50
SRR13093884	NCBI	22	87	ERR5192587	NCBI	22	82
	RAST	22	88		RAST	22	82
	RENOTEWEB	22	88		RENOTEWEB	22	76
SRR13209798	NCBI	22	96	ERR4147199	NCBI	22	85
	RAST	22	97		RAST	22	86
	RENOTEWEB	10	92		RENOTEWEB	14	82

Para mensurar o impacto da etapa de melhoria da montagem foi realizada uma comparação do processo de anotação com e sem o uso do módulo de melhoria da montagem, o resultado é listado na Tabela 4. É possível observar um crescimento em todas as variáveis observadas. Confirmando que a utilização da etapa de melhoria da montagem de fato trouxe melhores resultados ao processo de anotação.

Tabela 4. Comparativo entre resultados de anotação do ReNoteWeb, com e sem o processo de identificação de produtos ausentes.

SRA NÚMERO	MÉTODO	QUANTIDADE DE CDS	QUANTIDADE DE PRODUTOS	NÚMERO DE PRODUTOS NOMEADOS COM A SIGLA GENE	TOTAL DE PROTEÍNAS HIPOTÉTICAS
SRR4298913	ReNoteWeb	3905	955	2097	853
	ReNoteWeb + Melhoramento	4062	981	2141	940
SRR11462747	ReNoteWeb	5034	2058	2014	962
	ReNoteWeb + Melhoramento	5109	1473	1936	1700

SRR5481222	ReNoteWeb	2375	1055	1067	253
	ReNoteWeb + Melhoramento	2401	1081	857	463
SRR3947906	ReNoteWeb	2501	1482	641	378
	ReNoteWeb + Melhoramento	2535	1527	464	544
SRR638977	ReNoteWeb	3250	1615	777	858
	ReNoteWeb + Melhoramento	3271	1625	784	862
ERR351260	ReNoteWeb	6662	682	4459	1521
	ReNoteWeb + Melhoramento	6758	746	4455	1557
ERR351258	ReNoteWeb	6591	571	4565	1455
	ReNoteWeb + Melhoramento	6983	585	4867	1531
ERR351259	ReNoteWeb	7060	827	4554	1679
	ReNoteWeb + Melhoramento	7221	857	4632	1732
SRR13093884	ReNoteWeb	5544	164	4286	1094
	ReNoteWeb + Melhoramento	5542	165	4290	1087
SRR13209798	ReNoteWeb	6011	461	4216	1334
	ReNoteWeb + Melhoramento	6203	449	4324	1430
SRR15293247	ReNoteWeb	5742	346	4103	1293
	ReNoteWeb + Melhoramento	5856	377	4170	1309
SRR933455	ReNoteWeb	6090	457	4229	1404
	ReNoteWeb + Melhoramento	6664	581	4457	1626
DRR223366	ReNoteWeb	5905	643	3948	1314
	ReNoteWeb + Melhoramento	6163	666	4075	1422
SRR8778550	ReNoteWeb	6182	530	4219	1433
	ReNoteWeb + Melhoramento	6404	583	4336	1485
SRR14563279	ReNoteWeb	5045	903	2508	1634
	ReNoteWeb + Melhoramento	5094	913	2512	1669
SRR12062794	ReNoteWeb	5155	1551	2605	999

	ReNoteWeb + Melhoramento	5279	926	2570	1783
SRR10323925	ReNoteWeb	6332	2776	2245	1311
	ReNoteWeb + Melhoramento	6630	2982	2257	1391
SRR8843224	ReNoteWeb	7503	4195	1372	1936
	ReNoteWeb + Melhoramento	8091	4457	1413	2221
ERR5192587	ReNoteWeb	5406	564	3602	1240
	ReNoteWeb + Melhoramento	5836	740	3671	1425
ERR4147199	ReNoteWeb	5827	438	4079	1310
	ReNoteWeb + Melhoramento	5944	455	4160	1329

A Figura 04 representa a análise referente aos produtos identificados com a sigla gene nas anotações sem e com o uso do módulo de melhoria da montagem. Pode-se inferir que o uso do módulo de melhoria da montagem proporcionou a identificação de mais produtos com sigla de gene e que após a efetiva análise através do processo de curadoria, podem representar algum alvo biotecnológico de interesse.

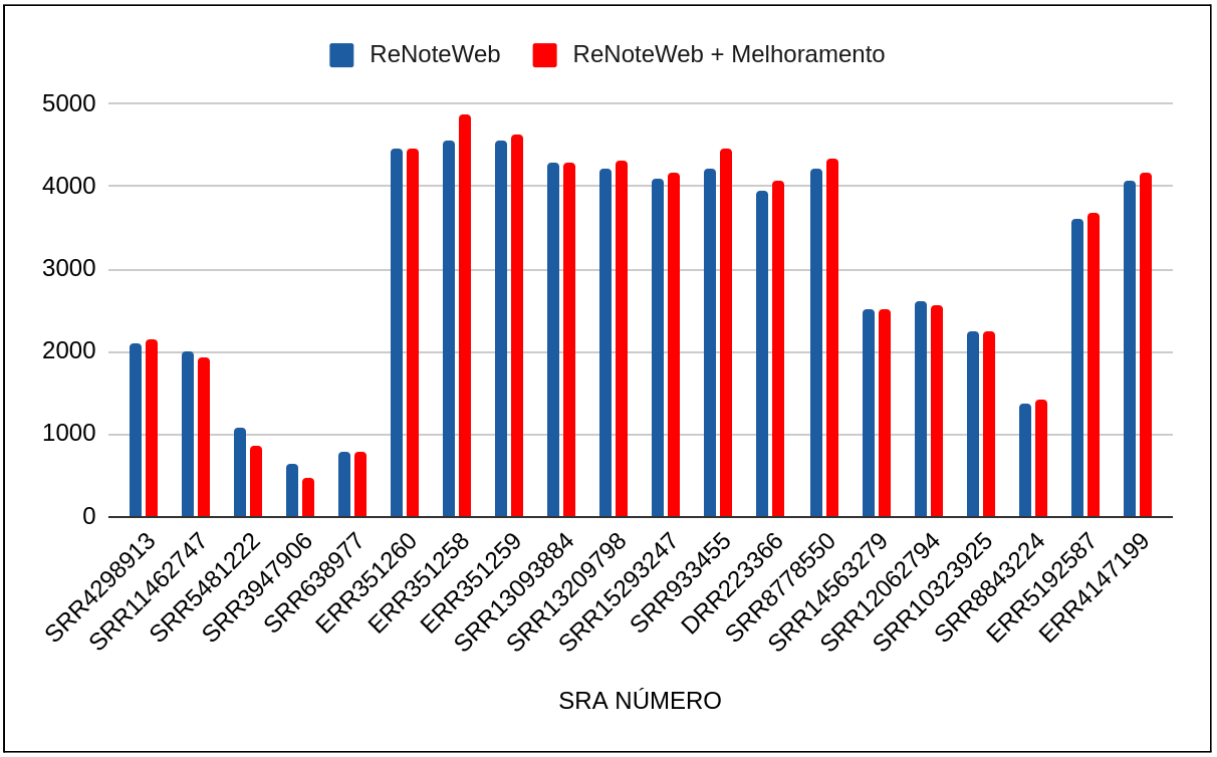

Figura 04. Quantidade total de produtos com sigla de gene: Comparação realizada sobre a quantidade de produtos com sigla de gene identificados na anotação do ReNoteWeb sem o uso do módulo de melhoria da montagem e após o uso do referido módulo.

A Tabela 5 lista o resultado da análise realizada nas anotações feitas pela plataforma RAST e ReNoteWeb, ambas utilizando o módulo de melhoria da montagem como etapa prévia à anotação. A análise do resultado demonstra que houve crescimento no número de CDS presentes na anotação do RAST, isto se deve a atualização da sequência FASTA de entrada no módulo de melhoria da montagem. Ademais, os resultados deste presente trabalho são promissores, visto que através da melhoria da montagem e anotação, o programa possibilitou, em alguns casos, identificar mais produtos gênicos em relação ao que estava depositado no NCBI e que foi predito pelo RAST. Dessa forma, a utilização do software pode impactar diretamente em análises que buscam identificar alvos com potencial biotecnológico, por utilizar informações de alta acurácia obtidas do Uniprot. E também pode aprofundar o conhecimento sobre um gene alvo que ainda não havia sido identificado em determinados organismos por não estar representado no genoma, colaborando com diversas análises genômicas.

Tabela 5: Resultados apresentados após realização da anotação dos organismos com melhoria na montagem.

SRA NÚMERO	FERRAMENTA	QUANTIDADE DE CDS	QUANTIDADE DE PRODUTOS	NÚMERO DE PRODUTOS NOMEADOS COM A SIGLA GENE	TOTAL DE PROTEÍNAS HIPOTÉTICAS
SRR4298913	Rast	4113	3028		1085
	ReNoteWeb	4062	981	2141	940
SRR11462747	Rast	4691	2883		1808
	ReNoteWeb	5109	1473	1936	1700
SRR5481222	Rast	2415	1925		490
	ReNoteWeb	2401	1081	857	463
SRR3947906	Rast	2662	1698		964
	ReNoteWeb	2535	1527	464	544
SRR638977	Rast	3639	2750		889
	ReNoteWeb	3271	1625	784	862
ERR351260	Rast	5769	5063		706

	ReNoteWeb	6758	746	4455	1557
ERR351258	Rast	5962	5152		810
	ReNoteWeb	6983	585	4867	1531
ERR351259	Rast	6082	5312		770
	ReNoteWeb	7221	857	4632	1732
SRR13093884	Rast	4845	4295		550
	ReNoteWeb	5542	165	4290	1087
SRR13209798	Rast	5091	4548		543
	ReNoteWeb	6203	449	4324	1430
SRR15293247	Rast	4786	4365		421
	ReNoteWeb	5856	377	4170	1309
SRR933455	Rast	5666	4852		814
	ReNoteWeb	6664	581	4457	1626
DRR223366	Rast	5204	4560		644
	ReNoteWeb	6163	666	4075	1422
SRR8778550	Rast	5408	4716		692
	ReNoteWeb	6404	583	4336	1485
SRR14563279	Rast	7556	3429		4127
	ReNoteWeb	5094	913	2512	1669
SRR12062794	Rast	4640	3458		1182
	ReNoteWeb	5279	926	2570	1783
SRR10323925	Rast	6575	4214		2361
	ReNoteWeb	6630	2982	2257	1391
SRR8843224	Rast	8710	5657		3053
	ReNoteWeb	8091	4457	1413	2221
ERR5192587	Rast	5104	4405		699
	ReNoteWeb	5836	740	3671	1425
ERR4147199	Rast	4974	4418		556
	ReNoteWeb	5944	455	4160	1329

Na Tabela 6 destacam-se algumas funções desempenhadas no pipeline do ReNoteWeb em comparação com as ferramentas RAST e PROKKA que executam tarefas de anotação. Demonstrando assim que o ReNoteWeb apresenta benefícios a mais ao usuário em relação aos programas que foram comparados, como a melhoria da montagem e o relatório ao final do processamento. Além do mais, a interface WEB, disponibiliza diversas opções de

download e ausência de instalação do programa permitem que o ReNoteWeb seja um programa de fácil manuseio e intuitivo, principalmente para usuários sem conhecimento computacional aprofundado.

Tabela 6: Resumo das tarefas realizadas pelas ferramentas RAST, PROKKA e ReNoteWeb

DESCRIÇÃO DAS TAREFAS	RAST	PROKKA	RENOTEWEB
Anotação	X	X	X
Melhoria da montagem			X
Interface WEB	X		X
Opções diversas para Download	X		X
Relatório no fim do processamento			X
Passagem de parâmetros de acordo com a necessidade	X	X	X
Não requer conhecimento prévio de computação e instalação de software	X		X

CONCLUSÃO

A ferramenta RenoteWeb provou ser eficiente para melhorar o resultado de montagem e anotação de genomas procariotos, a qual fornece a identificação de um número significativo de produtos que estavam anteriormente ausentes da fita genômica original.

Entretanto, a necessidade de realizar o processo de curadoria manual posterior não está descartada, o que é comum com o resultado de todas as ferramentas utilizadas nesse trabalho, visto que esses produtos adicionais que foram identificados dentre as análises podem conter informações relevantes, que pode contribuir para diversas análises, tais como análises evolutivas e proteômicas. Com a integração realizada entre o ReNoteWeb e o software CODON, os resultados referentes à anotação obtidos no ReNoteWeb podem ser exportados no formato de projeto CODON para curação manual posterior.

Dado o acima, pode-se inferir que o ReNoteWeb é uma alternativa para realizar o processo de melhoramento de montagem e anotação, o qual produz resultados com um alto nível de acurácia por meio de informações do banco de dados Uniprot, o que pode contribuir

para diversas análises posteriores e auxiliar no entendimento biológico de microrganismos, principalmente organismos não-modelo, contribuindo para a identificação de possíveis alvos com aplicabilidade biotecnológica.

Além do foco do ReNoteWeb, o qual é o processo de anotação e melhoramento de resultados de montagem, destaca-se que essa ferramenta pode ser de grande interesse para a comunidade científica e grupos de pesquisa, especialmente aqueles que não possuem recursos humanos especializados em bioinformática ou recursos computacionais necessários, pois é uma ferramenta computacional web simples, robusta, capaz de realizar esse tipo de análise. ReNoteWeb é uma ferramenta que não necessita de instalação e configuração de dependências de software, tornando-a muito interessante para usuários com ou sem experiência em computação.

AGRADECIMENTOS

Essa pesquisa recebeu financiamento do Conselho Brasileiro de Pesquisa (CNPq), número de concessão:405245/2018-1. Graças a Universidade Federal do Pará, esse trabalho recebeu ajuda da PROPESP/UFPA. Esse trabalho faz parte da pesquisa desenvolvida pelo BIOD (grupo de pesquisa Bioinformática, Ômicas e Desenvolvimento - www.biod.ufpa.br). AAOV agradece a Universidade Federal do Pará (UFPA), PHCGS agradece a PVTA341-2020 da Universidade Federal Rural da Amazônia (UFRA) e JTCA agradece a Universidade do Estado do Pará (UEPA).

DECLARAÇÃO DE CONFLITO DE INTERESSES

Os autores declaram não haver interesses conflitantes.

REFERÊNCIAS

Aziz, R. K., Bartels, D., Best, A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., … Zagnitko, O. (2008). The RAST Server: Rapid annotations using subsystems technology. BMC Genomics, 9. https://doi.org/10.1186/1471-2164-9-75

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology, 19(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Bye-AJee, H., Cowley, A., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Castro, L. G., Figueira, L., … Zhang, J. (2017). UniProt: The universal protein knowledgebase. Nucleic Acids Research, 45(D1), D158–D169. https://doi.org/10.1093/nar/gkw1099

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A. S., & Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research, 18(1), 188–196. https://doi.org/10.1101/gr.6743907

Haft, D. H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., Li, W., Chitsaz, F., Derbyshire, M. K., Gonzales, N. R., Gwadz, M., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Zheng, C., Thibaud-Nissen, F., Geer, L. Y., … Pruitt, K. D. (2018). RefSeq: An update on prokaryotic genome annotation and curation. Nucleic Acids Research, 46(D1), D851–D860. https://doi.org/10.1093/nar/gkx1068

Kremer, F. S., McBride, A. J. A., & Pinto, L. da S. (2017). Approaches for in silico finishing of microbial genome sequences. In Genetics and Molecular Biology (Vol. 40, Issue 3, pp. 553–576). Brazilian Journal of Genetics. https://doi.org/10.1590/1678-4685-gmb-2016-0230

Koren, S., Treangen, T. J., Hill, C. M., Pop, M., & Phillippy, A. M. (2014). Automated ensemble assembly and validation of microbial genomes. BMC Bioinformatics, 15(1), 1–9. https://doi.org/10.1186/1471-2105-15-126

Lantz, H., Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J. F., Vlasova, A., Leskosek, B. L., Soler, L., & Binzer-Panchal, M. (2018). Ten steps to get started in Genome Assembly and Annotation. F1000Research, 7. https://doi.org/10.12688/f1000research.13598.1

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics, 31(10), 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Merlin, B., Alves, J. T. C., de Sá, P. H. C. G., de Oliveira, M. S., Dias, L. M., da Silva Moia, G., dos Santos, V. C., & de Oliveira Veras, A. A. (2021). CODON-Software to manual curation of prokaryotic genomes. PLoS Computational Biology, 17(3). https://doi.org/10.1371/JOURNAL.PCBI.1008797

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. Genomics, 95(6), 315–327. https://doi.org/10.1016/j.ygeno.2010.03.001

Odell, S. G., Lazo, G. R., Woodhouse, M. R., Hane, D. L., & Sen, T. Z. (2017). The art of curation at a biological database: Principles and application. Current Plant Biology, 11–12, 2–11. https://doi.org/10.1016/j.cpb.2017.11.001

Otto, T. D., Dillon, G. P., Degrave, W. S., & Berriman, M. (2011). RATT: Rapid Annotation Transfer Tool. Nucleic Acids Research, 39(9). https://doi.org/10.1093/nar/gkq1268

Pfeiffer, F., & Oesterhelt, D. (2015). A manual curation strategy to improve genome annotation: Application to a set of haloarchael genomes. Life, 5(2), 1427–1444. https://doi.org/10.3390/life5021427

Richardson, E. J., & Watson, M. (2013). The automatic annotation of bacterial genomes. Briefings in Bioinformatics, 14(1), 1–12. https://doi.org/10.1093/bib/bbs007

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Ad`, M.-A., Rajandream, A., & Barrell, B. (2000). Artemis: sequence visualization and annotation. In BIOINFORMATICS APPLICATIONS NOTE (Vol. 16, Issue 10). http://www.acedb.org/

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. Bioinformatics, 30(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Silva de Oliveira, M., Thyeska Castro Alves, J., Henrique Caracciolo Gomes de Sá, P., & Veras, A. A. de O. (2021). PAN2HGENE-tool for comparative analysis and identifying new gene products. PloS One, 16(5), e0252414. https://doi.org/10.1371/journal.pone.0252414

Tanizawa, Y., Fujisawa, T., & Nakamura, Y. (2018). DFAST: A flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics, 34(6), 1037–1039. https://doi.org/10.1093/bioinformatics/btx713

Zerbino, D. R. (2010). Using the Velvet de novo assembler for short-read sequencing technologies. In Current Protocols in Bioinformatics (Issue SUPPL. 31). https://doi.org/10.1002/0471250953.bi1105s31