Infraestrutura de Tecnologia de Informação para Projetos de Sequenciamento de Genomas de Fungos

Rudinei Zacaria, Scheila de Avila e Silva

Área do Conhecimento de Ciências Exatas e Engenharias
Campus Universitário da Região dos Vinhedos (CARVI)
Universidade de Caxias do Sul (UCS)
Alameda João Dal Sasso, 800 – 95705-266 – Bento Gonçalves – RS – Brasil

{rudi.zacaria@gmail.com, sasilva6@ucs.br}

Resumo. *A Bioinformática tem adquirido uma importância crescente na manipulação de dados biológicos. Através da combinação de procedimentos e técnicas auxilia biólogos com a complexidade de ferramentas tanto de hardware quanto de software, otimizando o fluxo de trabalho em um ambiente distribuído e reduzindo a sobrecarga do movimento de dados entre programas. Este trabalho tem por objetivo propor um workflow de configuração de servidores para pesquisas de genoma de fungo. Baseado na Revisão Sistemática da Literatura, buscou-se identificar, selecionar, avaliar e sintetizar evidencias relevantes disponíveis, apresentando informações referentes a softwares e configurações de hardware destinados a sequenciamento e anotação genômica de fungos.*

Palavras-chave: *sequenciamento de fungo, software de bioinformática, hardware.*

1.Introdução

Atualmente a bioinformática é imprescindível para a manipulação dos dados biológicos. Esta área do conhecimento abrange todos os aspectos relacionados à aquisição, processamento, armazenamento, distribuição, análise e interpretação da informação biológica. Através da combinação de procedimentos e técnicas da matemática, estatística e ciência da computação, são desenvolvidas ferramentas que auxiliam a compreender o significado biológico representado nos dados genômicos (BORÉM, 2001).

O sequenciamento de um genoma disponibiliza a ordem que as informações genômicas estão em um organismo. Neste sentido o sequenciamento permite obter informações sobre a linha evolutiva dos organismos, resultar em novos métodos de diagnóstico, formular novos medicamentos, vacinas, na prevenção e tratamentos mais eficazes contra doenças ou pragas. Essas informações são obtidas por meio da anotação:

> A anotação genômica é um processo computacional, de vários passos classificados em três categorias básicas: a anotação em nível de nucleotídeos, anotação em nível de proteínas e anotação em nível de processo. Na anotação em nível de nucleotídeos procura-se encontrar a localização física das sequências de DNA e descobrir onde estão os genes, RNAs, elementos repetitivos, etc. Na anotação em nível proteico procura-se descobrir a provável função dos genes, identificando quais são aqueles que determinado organismo possui e quais ele não possui. Já a anotação em nível de processo procura identificar as vias e processos nos quais diferentes genes interagem, montando uma anotação funcional eficiente (STEIN, 2001).

O primeiro passo para a configuração do ambiente computacional é a definição da plataforma de aplicativos e sistema operacional a ser utilizado. O passo seguinte consiste na procura do equipamento propriamente dito. Traduzir as demandas de um projeto para as especificações de um servidor (*hardware*) é fundamental para dimensionar as necessidades atuais. Com base nos aplicativos escolhidos é importante realizar uma análise de volume de processamento, utilização de memória e consumo espaço em disco. Vários fatores influenciam o desempenho de novos montadores de genoma: cobertura de leitura, conteúdo de GC (guanina e citosina), fração de repetições, dentre outros ao genoma a ser sequenciado.

Do ponto de vista computacional uma questão determinante para a montagem de genoma é o tipo de dado e a cobertura desejada, para criar um ponto de partida. *Assemblers* são programas responsáveis pelas montagens de genomas, existem vários fatores que contribuem para o uso de memória e processamento, alguns dos quais são óbvios, como o número de fragmentos gerados conhecidos por *reads* e outros que são mais difíceis de estimar, como a complexidade do genoma. A maioria dos *assemblers* de genoma, como Velvet, requer memória e processamento de acesso aleatório adequado para dados de sequência de cluster, em um conjunto de segmentos de DNA denominado *contigs*. Por exemplo, de acordo com a empresa DNA STAR[1] especializada em sequenciamento genômico, o organismo *Saccharomyces cerevisiae*, possui um comprimento de gemona aproximadamente de 15MBases. Para o sequenciamento deste organismo a utilização de memória RAM aproxima-se de 20 Gb, o recomendado é de pelo menos 1 GB de RAM por MBase de comprimento do genoma.

Os requisitos de memória para montagens de sequenciamento de genoma de novo aumentam em proporção ao número de sequências e *contigs* montados. Diante desta realidade, o servidor deve ser configurado adequadamente para essas atividades. Adicionalmente, a instalação de ferramentas de *software* também representa um desafio para usuários com menor conhecimento computacional. Os resultados produzidos por uma ferramenta não estão sempre em um formato que pode ser usado pela próxima ferramenta em um fluxo de trabalho. Suprindo as demandas que foram impostas, o trabalho tem por objetivo realizar o levantamento das dificuldades apresentadas no processo de configuração de servidores para pesquisas de anotação genômica de fungo.

O presente artigo aborda alguns aspectos sobre Infraestrutura da Tecnologia da Informação para projetos de sequenciamento de genomas de fungos, mostrando as melhores condutas a serem aplicadas ao propor um *workflow* de configuração de servidores para pesquisas de anotação genômica de fungo. Através da Revisão Sistemática da Literatura, a qual analisou projetos semelhantes ao do *Penicillium echinulatum* já realizados, a fim de buscar relações na forma de sequenciamento, montagem e anotação.

Tendo a seguinte questão de pesquisa: “Qual a configuração de um servidor para atender as necessidades de um projeto de anotação genômica de fungo?”

2. Revisão bibliográfica

2.1. revisão sistemática da literatura

[1] https://www.dnastar.com/

A revisão sistemática, trata de um tipo de investigação dedicada a uma questão bem definida, que visa identificar, selecionar, avaliar e sintetizar as evidências relevantes disponíveis. (GALVÃO; PEREIRA, 2014)

Os métodos para elaboração de revisões sistemáticas preveem: (1) elaboração da pergunta de pesquisa; (2) busca na literatura; (3) seleção dos artigos; (4) extração dos dados; (5) avaliação da qualidade metodológica; (6) síntese dos dados (meta-análise); (7) avaliação da qualidade das evidências; e (8) redação e publicação dos resultados.

Dessa forma, foram elaboradas duas questões de pesquisa para serem respondidas após a obtenção dos estudos primários. São elas: (I) Quais as ferramentas de *software* utilizadas no sequenciamento, montagem e anotação de genoma de fungo; (II) Quais as configurações de *hardware* utilizados no sequenciamento, montagem e anotação de genoma de fungo.

2.2. estratégias de pesquisa e seleção

Elaboradas as questões da pesquisa, o próximo passo consistiu na definição dos termos de busca (*strings* de busca) que foram elaboradas a partir da necessidade de pesquisa sobre sequenciamento e anotação genômica, em conjunto com temas relacionados a *software*, *hardware*, *tool* e *annotation*. Por fim, foram realizadas combinações com palavras chaves, *fungi*, *penicillium* e *trichoderma*.

Assim sendo, foram construídas *strings* de busca com termos em inglês.

Tabela 1. *STRINGS* de busca

Pergunta I	**Pergunta II**
fungi AND annotation AND software *fungi AND annotation AND tool* *fungi AND sequencing AND annotation* *fungi AND sequencing AND software* *fungi AND sequencing AND tool* *penicillium AND sequencing AND annotation* *trichoderma AND sequencing AND annotation*	*fungi AND annotation AND hardware* *fungi AND sequencing AND hardware*

Fonte: elaborado pelo autor.

2.3. seleção das bases de dados

Os termos de busca definidos foram aplicados nas bases de dados estabelecidas no protocolo da pesquisa. A seleção das bases se deu a partir do reconhecimento acadêmico em âmbito nacional e internacional. Com isso as bases selecionadas para a pesquisa foram: Pubmed e ScienceDirect.

2.4. estratégias de inclusão e exclusão

O procedimento para a inclusão e exclusão de documentos que retornaram após a aplicação dos argumentos de pesquisa (*strings* de busca) nas bases de dados se deu em três momentos. No primeiro momento (Filtragem 1), foram definidos no protocolo de pesquisa como critérios de inclusão: título (*tittle*), ser gratuito e a delimitação do tempo de publica-

ção de 5 anos, ou seja, publicados entre 2012 a 2017. Com estes parâmetros, a pesquisa retornou 777 documentos.

No segundo momento (Filtragem 2), foi realizada uma análise do título e resumo com finalidade de reduzir a enorme gama de resultados que havia sido retornada após a aplicação dos argumentos. Ao término dessa análise restaram 376 documentos. Após em um terceiro momento (Filtragem 3), eles foram lidos para confirmar se estavam de acordo com o que foi estabelecido nos critérios de inclusão. Por fim foram aceitos os documentos que tratavam de sequenciamento de fungos, restando 49 documentos relevantes que foram utilizados para responder as questões propostas pela revisão, conforme a Tabela 2.

Tabela 2. Quantidade de artigos relacionados a sequenciamento de fungos

String	**Base de Dados**	**Filtragem 1**	**Filtragem 2**	**Filtragem 3**
fungi AND annotation AND software	*Pubmed*	4	2	2
	ScienceDirect	65	34	6
fungi AND sequencing AND software	*Pubmed*	12	8	1
	ScienceDirect	223	100	10
fungi AND sequencing AND tool	*Pubmed*	35	16	6
	ScienceDirect	231	104	11
fungi AND annotation AND tool	*Pubmed*	6	4	1
	ScienceDirect	66	34	3
fungi AND sequencing AND annotation	*Pubmed*	31	17	4
	ScienceDirect	85	33	2
penicillium AND sequencing AND annotation	*Pubmed*	2	1	-
	ScienceDirect	6	6	2
trichoderma AND sequencing AND annotation	*Pubmed*	1	-	-
	ScienceDirect	2	2	1
fungi AND annotation AND hardware	*Pubmed*	-	-	-
	ScienceDirect	14	14	-
fungi AND sequencing AND hardware	*Pubmed*	-	-	-
	ScienceDirect	1	1	-
		777	376	49

Fonte: elaborado pelo autor.

2.5. análise e interpretação dos artigos

A seguir são apresentadas as respostas paras as questões de pesquisa tendo como base os aspectos citados nos artigos selecionados.

2.5.1. quais as ferramentas de *software* utilizadas no sequenciamento, montagem e anotação de genoma de fungo?

Com o desenvolvimento da tecnologia de análise genômica, a obtenção de uma grande quantidade de dados de sequenciamento de DNA é hoje uma realidade. Estes dados, associados às ferramentas computacionais desenvolvidas para sua análise, permitem o acesso às informações sobre genomas de diversos organismos, através da montagem de suas sequências parciais ou completas, bem como sua caracterização estrutural e funcional (ALLENDORF, 2010).

Conforme Tabela 3, foram identificados cerca de 70 *softwares* destinados ao sequenciamento, montagem e anotação de fungos. Estes *softwares* realizam algumas funções, entre elas pode-se citar: isolamento do DNA genômico, sequenciamento do genoma, análise de qualidade, filtragem de qualidade, montagem do genoma, predição e anotação, extração da sequência, análise e comparação dos elementos, visualização de montagens.

Tabela 3. Quantidade de citações de *softwares* relacionados a sequenciamento de fungos

FERRAMENTAS/PIPELINE	FUNÇÃO	QTD DE CITAÇÕES
GENEid	Predição de genes	2
FGENESH		6
TWINSCAN		1
Augustus	Predição e anotação de genes	7
SNAP		4
SignalP	Predição de peptídeo	6
Phobius		1
TMHMM	Predição de helices	3
MAKER	Anotação de genes	3
GeneMark-es		5
OrthoFiller		1
CEGMA		3
BUSCO		1
BLaST/BLAST2GO	Comparação de sequências	30
MaSuRCA		1
Genewise		1
BioEdit	Alinhamento de sequência	3
CustalX/ClustaW		14
HMMER		4
MUSCLE		6
SAMtools		1
Apollo	Visualização de sequências	3
EMBOSSTranseq	Tradução de sequências	2
PANDAseq	Montagem de genoma	1
AbySS		1
SOAP/SOAPdenovo2		6
Trinity		3
Velvet		1
GS de Novo Assembler		1
DNAMAN	Detecção de sequências quiméricas	5
RepeatMasker		1
MEGA		16
QIIME		2
MOTHUR	Corte e filtragem de qualidade	2
TRIM-GALORE		2
Seqclean		2
Trimmomatic		1
FastQC		2
InterProScan	Análise e avaliação de sequências	1
Bioconductor	Análise e compreensão de dados	1

Fonte: elaborado pelo autor. (versão completa em material suplementar)

Visando uma maneira simples de fazer algumas verificações de controle de qualidade em dados de sequência bruta provenientes de pipeline de sequenciamento, o *software* FastQC fornece um conjunto modular de análises que pode ser usado para verificar problemas dos quais o pesquisador deve estar ciente antes de fazer qualquer outra análise.

Ferramentas relacionadas à montagens e alinhamentos de novos genomas podem ser realizadas por um pacote de algoritmo denominado Velvet, que foi projetado para lidar com a montagem e alinhamentos de sequencias de leitura curta. Isto é conseguido manipulando gráficos de Bruijn, através da simplificação e compressão, que permite a remoção de erros e a simplificação de regiões repetidas. O Velvet também foi implementado dentro de pacotes comerciais, como o Geneious e MaSuRCA. Este último usado em comparações de montagens, permite montar genomas com conjuntos de dados que contém apenas leituras do sequenciamento Illumina ou uma mistura de *reads* curtas e longas.

Para comparações de montagens, além do MaSuRCA, existe a popular família BLAST (*Basic Local Alignment Search Tool*), formada de programas desenvolvidos para comparar sequências (DNA, RNA ou proteína), com sequências de uma base de dados e calcular a significância estatística dos alinhamentos, com rapidez e sem perder sensibilidade. O programa BLAST pode ser usado para inferir relações evolutivas e funcionais entre as sequências, assim como para ajudar a identificar os membros de famílias proteicas. (ALTSCHUL, 1990).

Outra ferramenta utilizada o Augustus, tem como função a predição e anotação de genes supervisionados *ab initio* com maior frequência e melhor desempenho, oferece parametrizações para várias espécies de fungos. Para espécies que não possuem uma parametrização proporcionada, é necessária uma entrada manual considerável para obter essa parametrização específica. O que acaba limitando sua aplicabilidade apenas às espécies para as quais a parametrização já está disponível. Para genes não supervisionados *ab intio*, GeneMark-ES é uma alternativa que se treina interativamente com a sequência do genoma de entrada, porém é relatado como impreciso na predição de genes de exão simples.

Baseando-se no levantamento destes *softwares* é importante considerar a utilização de mais de uma ferramenta para realizar a mesma função, com intuito de avaliar a qualidade das informações permitindo ao usuário obter um comparativo referente aos dados de saída, que se enquadre na solução que lhe for mais adequada para tal situação. Também pode ser combinado *softwares* de diferentes funções, de modo que a saída de cada elemento seja a entrada do próximo, formando assim um pipeline, que permite que cada *software* realize sua função de forma conjunta com a informação já processada do *software* anterior, automatizando o processo de sequenciamento para a obtenção do resultado final.

A maioria dos *softwares* destinados a sequenciamento, montagem e anotação de fungos listados possuem suas particularidades e dependências de outros *softwares*. Em sua maioria, são de distribuição Unix, o qual possuem interface web e são *softwares* livres o que torna um sistema atrativo, pelo fato do alto custo que o sequenciamento genômico possui na atualidade.

2.5.2. quais as configurações de *hardware* utilizados no sequenciamento e anotação de genoma de fungo?

A estrutura computacional para decifrar códigos genéticos é extremamente complexa. Inicialmente é preciso um aparato de *hardware* robusto, formando por servidores multiprocessados (dois processadores ou mais), discos rígidos de centenas de gigabytes e memória RAM disponível. A montagem do genoma requer acesso a servidores de computação ou a um computador com 120–250 Gb de memória. Para a montagem do gráfico de Bruijn, maiores escalas de memória podem ser necessárias para fungos maiores que 100 Mb (VRIES, 2018). Esta infra explica-se devido ao fato de que as tecnologias de sequenciamento de genoma, de alto rendimento geram um número excessivo de pequenos segmentos de DNA, chamadas de *short reads* que causam carga computacional significativa.

Partindo disso, o ideal é que as configurações sejam baseadas em uma arquitetura de *hardware* de 64 bits, que permite uma maior utilização de memória RAM, obtendo ainda mais qualidade de processamento de dados. É importante considerar que todo o dimensionamento de *hardware* seja configurado de forma uniforme, ou seja, não convém uma máquina ter um tipo de sistema e outro tipo de processador. Uma configuração associada a um processamento de alto desempenho, combinado com um dimensionamento de memória baixo, acarretará na perda de desempenho, pelo fato da memória não acompanhar o rendimento do processador.

Na prática, para exemplificar cita-se o *software* de análise de sequências InterProScan, que faz uso intensivo de memória e processamento. Uma especificação mínima recomendada é de 2 núcleos e 4 GB de RAM, o que permitirá a análise de 5 a 10 sequências de uma só vez. Ainda, para o organismo *Saccharomyces cerevisiae*, que possui um comprimento de genoma aproximadamente de 15 MBases, a utilização de memória RAM aproxima-se de 20 GB, o recomendado é de pelo menos 1 GB de RAM por 1 MBase de comprimento do genoma.

Conforme já mencionado, um fator limitante à montagem de genomas grandes e complexos é o alto consumo de memória ocasionado pelos métodos de grafos de Bruijn. Embora esta seja uma estrutura de dados compacta, todas as implementações fazem uso de algum tipo de estrutura de dados auxiliar ao núcleo do grafo de Bruijn, com o objetivo de mapear as leituras ao grafo. Com isso, muitas implementações que funcionam bem em menor escala (genoma menor do que 50 Megabases), chegam a necessitar de 2 Terabytes de disco rígido para a montagem de um genoma complexo. Soluções como ABySS e SOAPdenovo, no entanto, buscam minimizar o problema. ABySS faz uso de uma estratégia comunicação de dados paralela, ao passo que SOAP de novo emprega múltiplas passagens por estruturas de dados compactas, as quais podem ser mantidas em disco para o tratamento de grandes volumes. (FLICEK; BIRNEY,2009).

Esses requisitos de memória para montagens de sequenciamento de genoma de novo, aumentam em proporção ao número de sequências e *contings* montados. Acrescenta-se também a possibilidade de estar executando mais de um *software* ao mesmo tempo, sendo que, nestes casos aumenta-se a requisição de memória e processamento, fazendo-se neces-

sário mensurar este consumo no momento de dimensionar as configurações de *hardware*, de acordo com as necessidades do projeto a ser realizado.

3. Metodologia

A metodologia consistiu no levantamento de um *workflow* de instalação e configuração de servidor para sequenciamento, montagem e anotação genômica de fungos. O organismo que será utilizado para validação do *workflow* proposto foi o *Penicillium echinulatum.*

O fungo filamentoso *Penicillium echinulatum* tem sido uma alternativa para a produção de celulases visando à obtenção de novos processos que apresentem maior eficiência na hidrólise de substratos celulósicos. Fungos do gênero *Penicillium* tem sido considerado como potenciais alternativas para produção de biocombustíveis de segunda geração (GUSAKOV, 2011). As linhagens mutantes do fungo filamentoso *Penicillium echinulatum* estão entre os microrganismos que apresentam excelente potencial biotecnológico para a secreção de enzimas visando a conversão enzimática eficiente da biomassa lignocelulósica em etanol de segunda geração.

3.1. ferramentas para montagem e anotação de genomas de fungo

A escolha das ferramentas foi baseada na revisão sistemática da literatura e na experiência da equipe envolvida com a pesquisa. As ferramentas escolhidas foram:

- Funannotate - É um pacote de *software* de previsão, anotação e comparação do genoma;
- Eggnog-mapper - É uma ferramenta para anotação funcional rápida de novas sequências;
- Interproscan - É um pacote de *software* que permite que sequências (proteínas e nucleicos) sejam analisadas e avaliadas em relação às assinaturas da InterPro.

3.2. recursos de *hardware* e sistema operacional

Para a elaboração da proposta do *workflow* de configuração, os recursos de *hardware* utilizados possuem a seguinte configuração:

- Processador: Intel® Xeon® CPU x5650 @2.67 x 12;
- Memória instalada (RAM): 6.00GB;
- Tipo de sistema: Sistema de 64 bits, processador com base em X64.
- Sistema Operacional: Ubuntu 16.04.4 server.

3.3. *workflow*

Primeiro foi necessário definir o ambiente de trabalho, se as configurações seriam realizadas em uma máquina física ou virtual. Após foi realizada a instalação do sistema operacional, configuração do ambiente de rede, firewall e permissões de contas de usuário.

A próxima etapa foi realizar a escolha dos *softwares* listados na Tabela 3. Esta definição necessita a compreensão do *software*, para decisão de qual ferramenta atende do objetivo do pesquisador. Pesquisadores de projetos genômicos possuem fundamental impor-

tância nesta etapa, devido as experiências vivenciadas em outros projetos e atuaram de forma facilitadora opinando qual ferramenta apresenta um melhor resultado para função necessitada.

Dando sequência, foram realizadas as instalações e configurações dos *softwares* para o sequenciamento, juntamente com as permissões necessárias para poder operar de forma plena. Com os *softwares* devidamente instalados pode-se atuar no gerenciamento de memória e processamento dedicado a cada um, finalizando a configuração da estrutura necessária para realizar as pesquisas referentes ao projeto de genoma de fungos. O fluxo das etapas acontece conforme ilustrado na Figura 1.

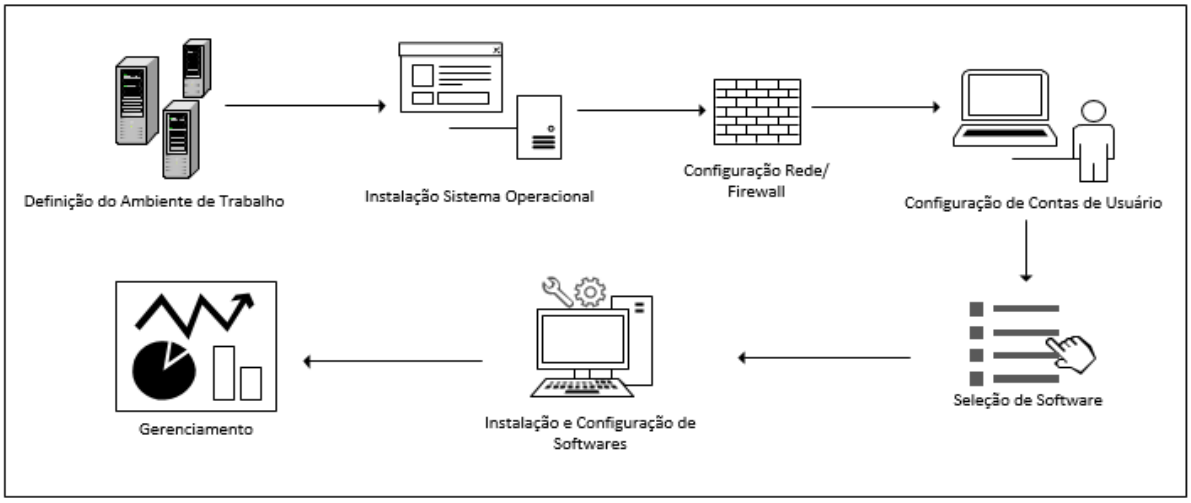


Figura 1. Etapas de configuração (elaborado pelo autor).

4. Resultados

Nesta seção são listadas as operações a serem realizadas.

4.1. etapas de instalação dos *softwares* escolhidos

Para a instalação dos *softwares* selecionados sugere-se seguir a ordem abaixo, devido as suas dependências.

OBS: Os caminhos utilizados como exemplos, podem variar dependendo do local que cada usuário definir sua instalação.

A - LINUXBREW: Gerenciador de pacotes para o Linux, utilizado para realizar o *download* dos *softwares*.

> Pode ser instalado em seu diretório pessoal e não requer acesso root. A instalação do Linuxbrew pode ser realizada em: /home/linuxbrew/.linuxbrew
>
> Cole o seguinte comando em um terminal: (irá realizar o download da biblioteca responsável pela instalação do linuxbrew.)
> # sh -c "$(curl -fsSL
> https://raw.githubusercontent.com/Linuxbrew/install/master/install.sh)"

Criando variável de ambiente:

Adicionar o Linuxbrew ao seu PATH e ao seu script de perfil bash shell, ~ / .profile. Para as aplicações tornarem-se visíveis de qualquer diretório do sistema.

```
# test -d ~/.linuxbrew &&
PATH="$HOME/.linuxbrew/bin:$HOME/.linuxbrew/sbin:$PATH"
# test -d /home/linuxbrew/.linuxbrew &&
PATH="/home/linuxbrew/.linuxbrew/bin:/home/linuxbrew/.linuxbrew/sbin:$PATH"
# test -r ~/.bash_profile && echo "export PATH='$(brew --prefix)/bin:$(brew --
prefix)/sbin'":'"$PATH"' >>~/.bash_profile
# echo "export PATH='$(brew --prefix)/bin:$(brew --prefix)/sbin'":'"$PATH"'
>>~/.profile
```

Dependências:
sudo apt-get install build-essential curl file git
Biblioteca complementar para a versão básica do linuxbrew.

Testar instalação:
brew install hello

B - FUNANNOTATE Versão 1.3.3 – Figura 2

Setup homebrew taps:
brew tap brewsci/bio && brew tap brewsci/science && brew tap nextgenusfs/tap && brew update
Dependências do homebrew para a instalação do Funannotate.

Instalação cpanminus:
brew install cpanminus
Script para obter, construir e instalar módulos, recurso que será utilizado na instalação dos módulos do Perl na sequência.

Instalação módulos do Perl:
cpanm Getopt::Long Pod::Usage File::Basename threads threads::shared Thread::Queue Carp Data::Dumper YAML Hash::Merge Logger::Simple Parallel::ForkManager DBI Text::Soundex Scalar::Util::Numeric Clone JSON LWP::UserAgent DBD::mysql URI::Escape

Instalação funannotate:
brew install funannotate
Isso instalará automaticamente a maioria das dependências, bem como a versão mais atual do funannotate. (Figura 3 – Dependências Funannotate).

Baixe / instale o GeneMark-ES / ET: (gmes_petap.pl deve estar no *PATH*)
Download: http://topaz.gatech.edu/GeneMark/license_download.cgi
Renomeia a chave para gm_key
Copia para o diretório que o GenMark foi extraído exemplo (usr/local/bin)
cp gm_key /.gm_key

Instalar os módulos python via PIP:

pip install -U numpy biopython natsort psutil goatools fisher pandas matplotlib seaborn scikit-learn ete3. Instalar inicialmente o módulo numpy, é pré-requisito para os demais módulos.

Instalar as bibliotecas do RepeatMasker :
wget –user (informe usuário) --password (informe senha) http://www.girinst.org/server/RepBase/protected/repeatmaskerlibraries/RepBaseRepeatMaskerEdition-20170127.tar.gz.

Após baixar a biblioteca, copiar a pasta compactada para dentro do diretório onde o RepeatMasker foi instalado. Ex:
/home/linuxbrew/.linuxbrew/opt/repeatmasker/libexec

Descompactar:
tar -vzxf RepBaseRepeatMaskerEdition-20170127.tar.gz

Configuração:
perl ./configure

Tecle ENTER para iniciar a configuração:
Inserir o caminho do perl. Se o caminho não for preenchido automaticamente, é necessário informar. Exemplo: [/usr/bin/perl]

Em seguida será solicitado o caminho do RepeatMasker. Se o caminho não for preenchido automaticamente, é necessário informar.
Exemplo: [home/linuxbrew/.linuxbrew/opt/repeatmasker]

Após inserir o caminho do programa TRF. Exemplo:
[/home/linuxbrew/.linuxbrew/opt/trf/bin]
No menu apresentado escolha a opção número 2 para configurar o rmblast

Inserir o caminho do rmblast. Exemplo:
[/home/linuxbrew/.linuxbrew/opt/rmblast/bin/]

Após selecionar opção 5 e verificar se o RepBase foi combinado com o Dfam.

Configuração do RepeatModeler:
Acessar diretório de instalação.
Exemplo:#cd /home/linuxbrew/.linuxbrew/opt/repeatmodeler/

Comando para iniciar a configuração:
perl ./configure

Inserir o caminho do perl. Se o caminho não for preenchido automaticamente, é necessário informar. Exemplo: [/usr/bin/perl]

Em seguida será solicitado o caminho do RepeatModeler. Se o caminho não for preenchido automaticamente, é necessário informar.
Exemplo: [home/linuxbrew/.linuxbrew/opt/repeatmodeler]

Informar o caminho do RepeatMasker.
Exemplo:[/home/linuxbrew/.linuxbrew/opt/repeatmasker/libexec]

Informar o caminhdo do Recon. Exemplo:
[/home/linuxbrew/.linuxbrew/opt/recon/bin]

Informar caminho do Repeatscout. Exemplo
[/home/linuxbrew/.linuxbrew/opt/repeatscout]

Informar caminho do TRF. Exemplo: [/home/linuxbrew/.linuxbrew/opt/trf/bin]

No menu apresentado escolha a opção número 1 para configurar o rmblast
Inserir o caminho do rmblast. Exemplo:
[/home/linuxbrew/.linuxbrew/opt/rmblast/bin/]

Selecione a opção número 3, para concluir a configuração das bibliotecas.

Configurar bancos de dados do funannotate:
funannotate setup -d /path/to/DB

Exporte as variáveis ENV necessárias:
#export EVM_HOME=#{HOMEBREW_PREFIX}/opt/evidencemodeler
#export AUGUSTUS_CONFIG_PATH=#{HOMEBREW_PREFIX}/opt/augustus
/libexec/config
#export BAMTOOLS_PATH=#{HOMEBREW_PREFIX}/opt/bamtools/bin
#export GENEMARK_PATH=/path/to/gmes_petap.pl
#export FUNANNOTATE_DB=/path/to/DB

C - EGGNOG-MAPPER Versão 2.4 – Figura 4

Realizar o *download* do *software*s via git:
git clone https://github.com/jhcepas/eggnog-mapper.git

Após mover diretório para local de preferência e acessa-lo via terminal, executar o comando:
./download_eggnog_data.py ascNOG; Este comando irá baixar as bases de dados referentes a ascNOG.

Adicionar Eggnog-mapper nas variáveis de ambiente:
Exemplo: #export EGGNOGMAPPER=”$HOME/linuxbrew/.linuxbrew/bin/eggnog-mapper”

Executar o programa de dentro do diretório instalado, através do comando:
#python emapper.py combinado com a sintaxe pretendida.

D – INTERPROSCAN Versão 5.29-68 – Figura 5

Realizar o download do *software* via wget:
wget ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/5/5.29-68.0/interproscan-5.29-68.0-64-bit.tar.gz

Necessita instalação do java 8 oracle, não funciona no java versão 9.

Descompactar Arquivo:
tar -vzxf interproscan-5.29-68.0-64-bit.tar.gz

Criar variável de ambiente:
#export INTERPROSCAN5="$HOME/.linuxbrew/bin/interproscan-5.29-68.0"

Executar o programa de dentro do diretório instalado, através do comando:
./interproscan.sh. Combinado com a sintaxe pretendida.

4.2. considerações sobre as instalações

Na bioinformática diversos *software*s são destinados ao propósito de sequenciamento, montagem e anotação de fungos. Durante o estudo realizado foi possível levantar algumas problemáticas as quais impediram a validação do *workflow*.

A validação do projeto se deu apenas na parte da instalação dos *softwares*, devido à complexidade a qual exigiu uma demanda de tempo e conhecimento, não sendo possível realizar a execução das aplicações, com o intuito de mensurar o consumo de recursos e apresentar uma análise para o dimensionamento de uma configuração necessária, para atender os requisitos do projeto. Contudo foi possível compreender que as ferramentas de sequenciamento, montagem e anotação genômica de fungos demandam de um elevado poder de máquina, consumindo memória, processamento e espaço em disco.

No decorrer das instalações foi visto que, grande parte do material disponibilizado em forma de tutoriais estão erroneamente formulados. Faltando informações de etapas das instalações, comandos que não atendem a função especificada. Entretanto servem como referências na busca para realização das instalações.

Ao iniciar uma instalação que venha a falhar mais adiante, surge outra dificuldade, pelo fato de não ser possível reverter os erros gerados, necessitando até mesmo a formatação da máquina. Recomenda então, realizar as instalações em uma máquina virtual, onde é possível configurar snapshots, ou seja, gerar uma imagem do sistema que pode ser utilizada como ponto de restauração, diante de instalações malsucedidas.

Observar as possíveis dependências que cada *software* necessita é importante para a definição de uma sequência de instalação, esta sequência facilita de modo que ao configurar uma aplicação, que pode não listar uma dependência, mas possuir, seja configurada de forma correta pelo fato da dependência ter sido instalada pela aplicação anterior. A cada nova versão de *software* disponibilizada possivelmente serão geradas novas dependências ou incompatibilidades, não encontradas em uma versão anterior.

5. Conclusão

Através do estudo realizado na primeira etapa do trabalho, baseado na revisão sistemática da literatura, onde foi sugerido a elaboração de um *workflow* de instalação e configuração de servidor para sequenciamento, montagem e anotação genômica de fungos, além da validação, por meio da instalação e configuração dos *softwares*.

Foi possível compreender a complexidade dos *softwares* voltados a bioinformática, estes possuem dependências de bibliotecas, que interferem diretamente nas instalações

pretendidas, dificultando a elaboração de um *workflow* de configuração genérica para aplicação em projetos de genomas. Acrescenta-se as dificuldades técnicas, relacionadas a configuração do ambiente de trabalho de um projeto de genoma de fungo.

Por isso fica claro a ambivalência de dados presente nesse estudo, onde, a ferramenta proposta hora contribui para o pesquisador e hora mostra ser inviável de forma que apresenta situações imprevisíveis dificultando a validade do *workflow* de configuração.

Como trabalhos futuros sugere-se o desenvolvimento de uma pipeline que realize a instalação automática das ferramentas de sequenciamento, montagem e anotação genômica de fungos. Recomenda elaborar uma forma de analisar as dependências necessárias na integração de um *software* com outro.

6. Referências

Allendorf, F. W.; Hohenlohe, a. P.; Luikart, G. Genomics and the future of conservation genetics. Nature Reviews Genetics, London, (2010).

Altschul, S. F., Gish, W., Miller, W., Myers, e. W., and Lipmanl, D. J. Basic local alignment search tool. J. Mol. Biol, (1990).

Borém, Aluízio; Santos, Fabrício Rodrigues. Biotecnologia Simplificada. Ed. Suprema. Viçosa, MG, (2001).

Cooper, Necia Grant. The Human Genome Project. Univ. Science Books, Mill Valey, CA, (1994).

De Vries, Ronald P.; Tsang, Adrian; Grigoriev, Igor V. Fungal Genomics Methods and Protocols, (2018).

Elmasri, Rames; Navathe, Shamkant B. Sistemas de banco de dados: fundamentos e aplicações. 6º. Ed. São Paulo: Addison Wesley, (2011).

Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. Nat. Methods (2009).

Galvão, Tais Freire; Pereira, Mauricio Gomes. Revisão sistemática de literatura: passos para sua elaboração. Epidemiologia e Serviço de Saúde, (2014).
Disponível em: <http://scielo.iec.pa.gov.br/pdf/ess/v23n1/v23n1a18.-pdf>
Acesso em 30 de Outubro de 2017.

Gorla, Narasimhaiah; Chiravuri, Ananth. Information systems outsourcing success: a review. In: 2010 International Conference on E-business, Management and Economics, (2011).

Gusakov, Alexander V. Alternatives Trichoderma reesei in biofuel production. Trends Biotechnol, (2011).

Insidedna. Benchmarking seven most popular genome assemblers. Disponível em:

<https://insidedna.me/tutorials/view/benchmark-seven-popular-genome-assemblers>. Acesso em 25 de Agosto de 2017.

Marx, Vivien. Biology: The big challenges of big data, Nature, (2013).

Prosdocimi, Francisco et al. Bioinformática: Manual do usuário n. 29. Revista: Biotecnologia Ciência & Desenvolvimento, Brasília, Brazil, (2001).

Stein, Lincoln. Genome annotation: from sequence to biology. Nature Reviews Genetics, (2001).

Thompson, J. D.; Higgins, D. G.; Gibson, T. J. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, (1994).

Yin, Zekun et al. Computing Platforms for Big Biological Data Analytics: Perspectives and Challenges. Computational and Structural Biotechnology Journal, (2017).

MATERIAL SUPLEMENTAR

Tabela 3. Quantidade de citações de *softwares* relacionados a sequenciamento de fungos

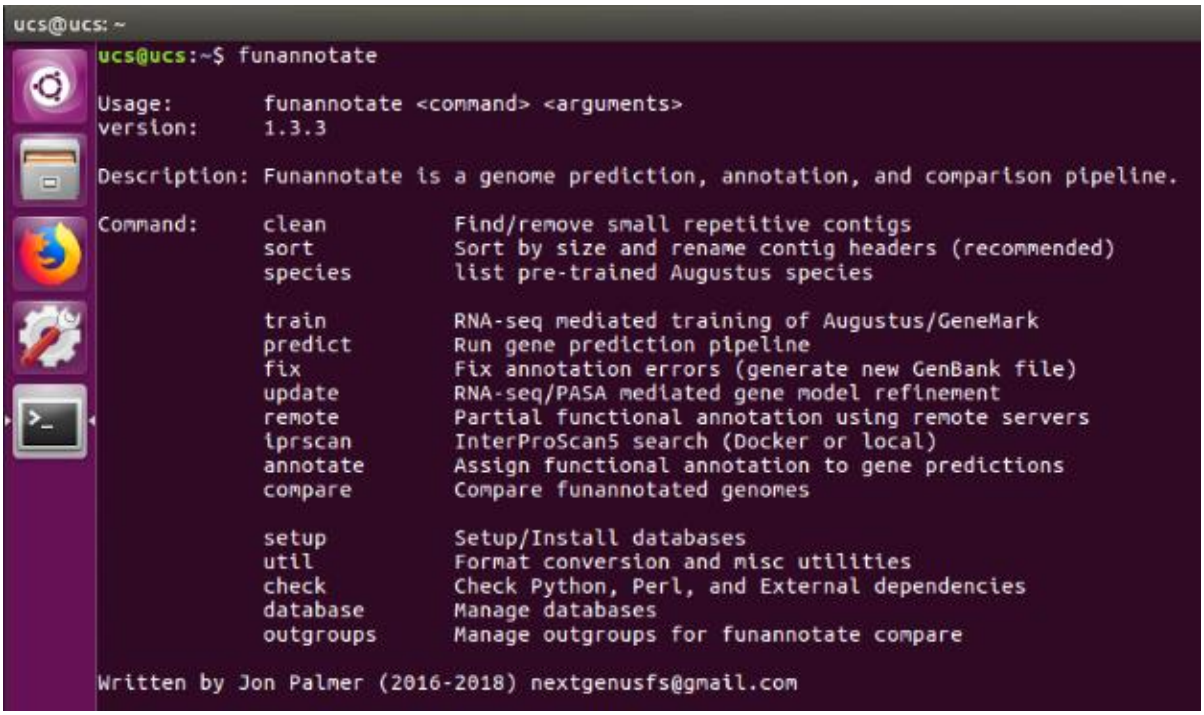
FERRAMENTAS/PIPELINE	FUNÇÃO	BANCO DE DADOS	QTD DE CITAÇÕES
GENEid	Predição de genes	Broad Institute	1
		Repbase	1
FGENESH		Broad Institute	1
		NCBI/GENBANK	5
TWINSCAN		NCBI/GENBANK	1
PSIPRED		NCBI/GENBANK	1
TargetP		NCBI/GENBANK	1
NetSurfP		NCBI/GENBANK	1
Phyre2		NCBI/GENBANK	1
MIREAP		NCBI/GENBANK	1
WoLF PSORT		Repbase	1
PredGPI		Repbase	1
Augustus	Predição e anotação de genes	Broad Institute	1
		NCBI/GENBANK	4
		Repbase	1
		KEGG	1
SNAP		Broad Institute	1
		NCBI/GENBANK	1
		Repbase	1
		KEGG	1
SignalP	Predição de peptídeo	NCBI/GENBANK	5
		Repbase	1
Phobius		Repbase	1
TMHMM	Predição de helices	NCBI/GENBANK	2
		Repbase	1
MAKER	Anotação de genes	Broad Institute	1
		Repbase	1

FERRAMENTAS/PIPELINE	FUNÇÃO	BANCO DE DADOS	QTD DE CITAÇÕES
		NCBI/GENBANK	1
GeneMark-es		Broad Institute	1
		Repbase	1
		NCBI/GENBANK	2
		KEGG	1
CLC Genomic Workbench		Broad Institute	1
		NCBI/GENBANK	1
OrthoFiller		NCBI/GENBANK	1
CEGMA		Repbase	1
		NCBI/GENBANK	1
		KEGG	1
BUSCO		Repbase	1
BALST/BLAST2GO	Comparação de sequências	Broad Institute	1
		NCBI/GENBANK	24
		Repbase	1
		KEGG	2
		InterPro	1
		UNIQUE	1
MaSuRCA		Repbase	1
Genewise		KEGG	1
BioEdit	Alinhamento de sequência	Broad Institute	1
		NCBI/GENBANK	2
CustalX/ClustaW		Broad Institute	1
		NCBI/GENBANK	11
		KEGG	1
		InterPro	1
HMMER		NCBI/GENBANK	2
		KEGG	2
SINA		NCBI/GENBANK	1
MUSCLE		NCBI/GENBANK	5
		Repbase	1
Bwa		NCBI/GENBANK	1
SAMtools		NCBI/GENBANK	1
HISAT2		Repbase	1
Tophat2		Repbase	1
Apollo	Visualização de sequências	Broad Institute	1
		Repbase	1
		KEGG	1
EMBOSSTranseq	Tradução de sequências	NCBI/GENBANK	2
PANDAseq	Montagem de genoma	NCBI/GENBANK	1
AbySS		NCBI/GENBANK	1
SOAP/SOAPdenovo2		NCBI/GENBANK	5
		KEGG	1
Trinity		Repbase	1
		KEGG	2
Geneious		NCBI/GENBANK	1
Velvet		NCBI/GENBANK	1
GS de Novo Assembler		InterPro	1
UCHIME	Detecção de sequências quiméricas	NCBI/GENBANK	1

FERRAMENTAS/PIPELINE	FUNÇÃO	BANCO DE DADOS	QTD DE CITAÇÕES
USEARCH	Análise de sequência	NCBI/GENBANK	1
DNAMAN		NCBI/GENBANK	5
SortMeRNA		Repbase	1
RepeatMasker		Repbase	1
PS-Scan		Repbase	1
WEGO		KEGG	1
CoreAlyze		KEGG	1
MEGA		NCBI/GENBANK	15
		InterPro	1
LEFSe		KEGG	1
ITScan		NCBI/GENBANK	1
MG-RAST		KEGG	1
QIIME		UNIQUE	1
		KEGG	1
MOTHUR	Corte e filtragem de qualidade	NCBI/GENBANK	2
TRIM-GALORE		NCBI/GENBANK	2
Seqclean		KEGG	1
		NCBI/GENBANK	1
Trimmomatic	Corte e filtragem de qualidade	Repbase	1
Phrap	Análise e controle de qualidade	NCBI/GENBANK	1
FastQC		Repbase	1
		NCBI/GENBANK	1
PRINSEQ-lite		Repbase	1
InterProScan	Análise e avaliação de sequências	Repbase	1
Bioconductor	Análise e compreensão de dados	NCBI/GENBANK	1
tRNAscan	Detecção de genes	Repbase	1
GlimmerHMM		KEGG	1
HPknotter	Detecção de pseudoknots	NCBI/GENBANK	1
Primer3	Detecção de sequência	NCBI/GENBANK	1
Pilon	Melhoria de montagens de rascunho	KEGG	1
PHYLIP	Pacote de inferência de filogenia	NCBI/GENBANK	1

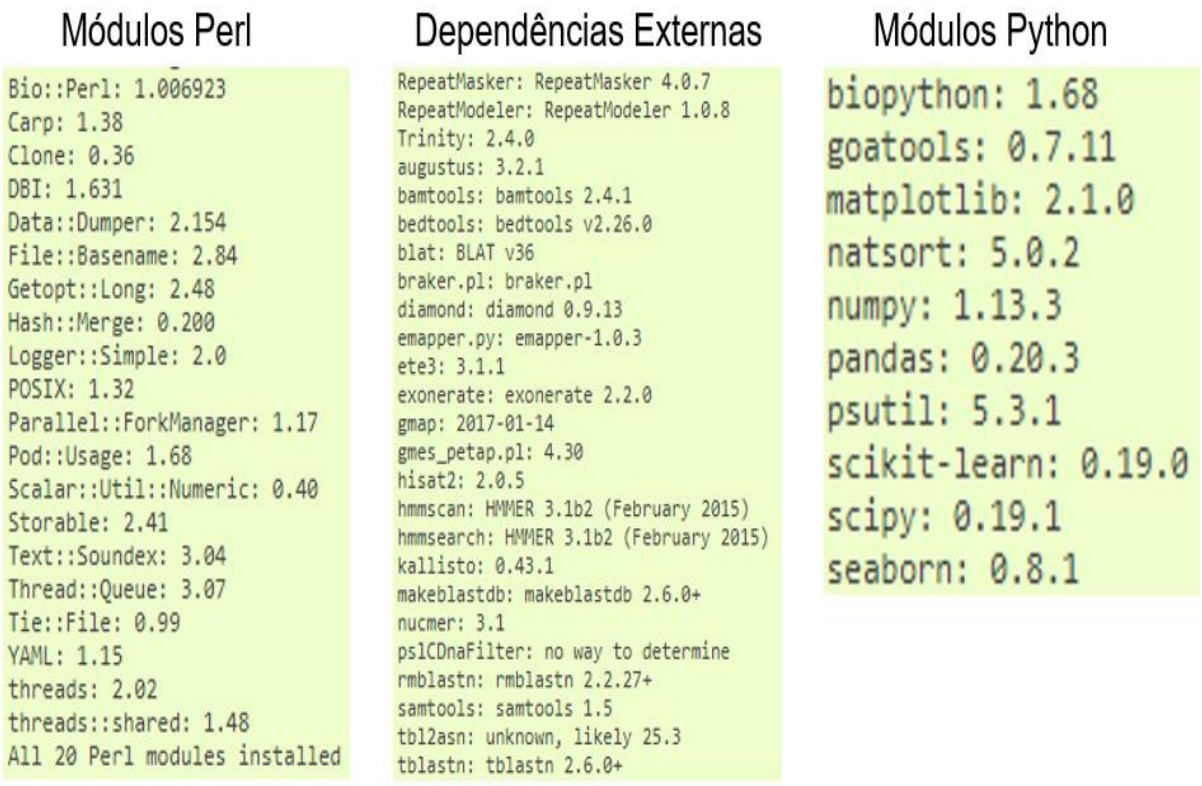
Fonte: Elaborado pelo Autor.

Figura 2 - FUNANNOTATE Versão 1.3.3



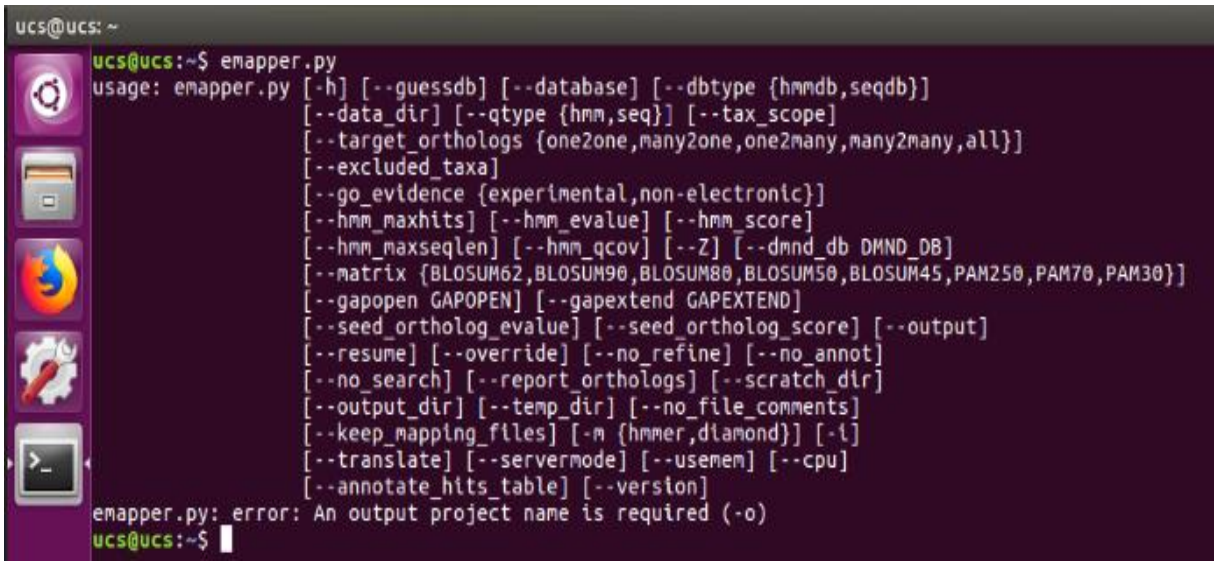
Fonte: Elaborado pelo Autor.

Figura 3 – Dependências Funannotate



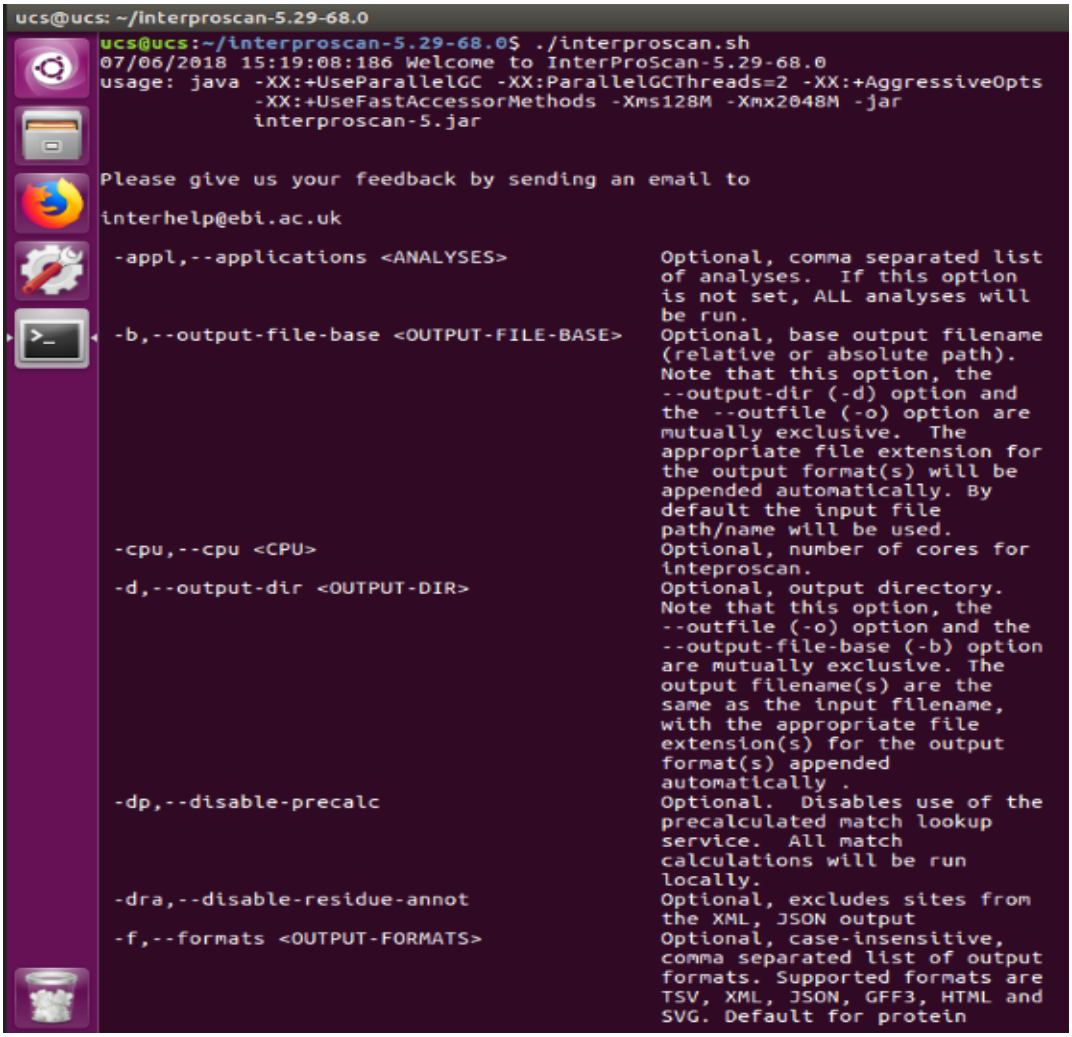
Fonte: Elaborado pelo Autor.

Figura 4 - EGGNOG-MAPPER Versão 2.4



Fonte: Elaborado pelo Autor.

Figura 5 - INTERPROSCAN Versão 5.29-68



Fonte: Elaborado pelo Autor.