Relatório de Análise

Coordenador: Wellington Pine Omori

Candidato à vaga: Giovanna Zuzarte Candido

Objetivos Gerais do estudo

1.1) Estes dados foram gerados com o intuito de compreender a variação da microbiota intestinal natural em resposta ao estado de saúde do hospedeiro;

1.2) Foram coletadas fezes frescas de camundongos durante 365 dias após a data do desmame. Foi realizado um sequenciamento de larga escala (MiSeq), usando primers do gene 16S rRNA para a região V4;

1.3) Durante os primeiros 150 dias após o desmame (dwp), nada foi feito com os camundongos, exceto permitir que se alimentassem, engordassem e se divertissem. Nos primeiros 10 dwp, observou-se um aumento brusco no peso dos camundongos, o que gerou dúvidas se a microbiota desses 10 dwp era diferente daquela observada entre os dias 140 e 150;

1.4) O arquivo metadata.csv oferecido neste desafio, contém a relação entre os arquivos fastq's (que estão no diretório fqs) e a idade de desmame dos camundongos. Os dados são parciais, muitos dos arquivos foram omitidos a fim de facilitar a execução dos desafios.

1.5) Junto dos dados a processar é disponibilizado um pequeno banco de sequências de referência FASTA (fasta_file.fasta) que contém sequências de bactérias com definição a nível de espécie.

1.6) Como dois dos três desafios dependem das análises dos dados descritos nos subitens 1.1 à 1.5, caso o candidato precise, estamos disponibilizamos uma OTU table e uma Tax table que podem ser usadas nos desafios subsequentes que dependeriam destes arquivos. Os arquivos estão dentro do diretório tables: otu_table_tax_amostras.tsv e tax_table_amostras.tsv.

Pré-desafio

Instalação e Criação do Docker

Inicialmente, foi utilizado o software ‘Visual Studio Code’ para o desenvolvimento. Foi realizado o download das extensões Docker (version 1.7.0) e Python (version 2020.10.332292344) e iniciado um novo arquivo. No shell do Linux foi também realizado o download do Docker.

No desenvolvimento do docker, no dockerfile foi utilizado a imagem do Ubuntu 18.04 como base e realizado o download dos softwares utilizados nas analises dos itens do desafio. Como demonstrado no arquivo README.TXT e no próprio Dockerfile.

O banco de dados disponibilizado para o desafio foi compactado e upado no DropBox, para sua utilização no docker em qualquer computador.

Competência básica para Bioinformática

Instruções Gerais

O objetivo deste desafio será o de gerar um script (recomendado utilizar a linguagem Python) que contemple os seguintes steps:

1.1) Efetuar trimagem dos dados, por qualidade, usando algum trimador de sua preferência;

1.2) Gerar reports da qualidade PHRED antes e após trimagem dos dados;

1.3) Fazer identificação taxonômicas das sequências que passaram pelo filtro de qualidade usando um banco de referência e um programa de sua escolha;

1.4) Gerar uma OTU table, onde as linhas serão taxonomias e as colunas os nomes das amostras. A intersecção de colunas e linhas devem mostrar as contagens da taxonomia na amostra em questão (vide arquivo em 'Desafio_Neoprospecta/tables');

1.5) Realizar todos os steps anteriores de forma automatizada

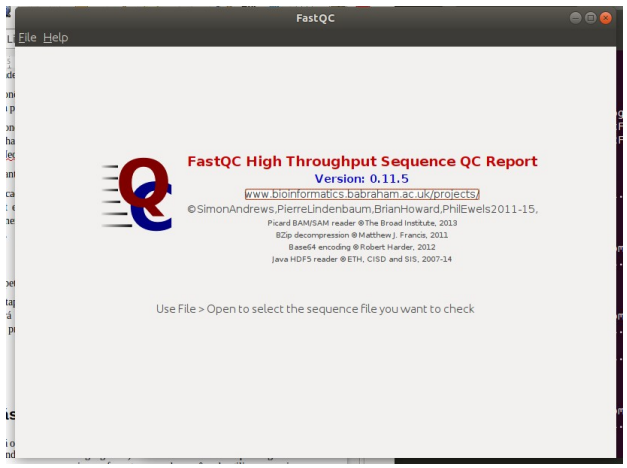
1.6) O código deve ser colocado em um container Docker, e depositado em uma conta do GitHub. Construa o requirements.txt e demais instruções para instalação. Descreva as instruções para a execução do pipeline. Recomendamos testar a instalação e execução completa do código do Docker antes de sua submissão final.

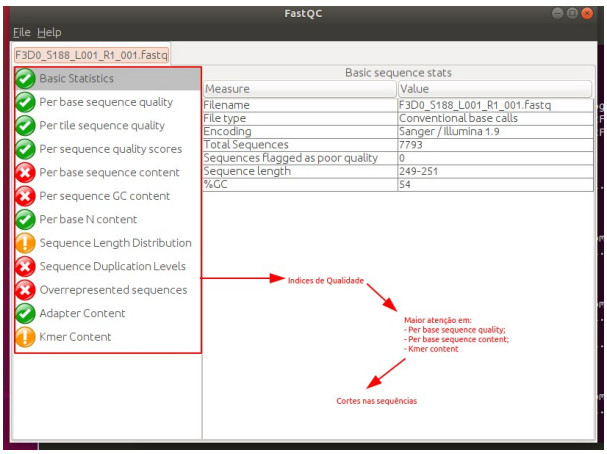
Resultados

Etapa 01: Para a primeira etapa da primeira competência é necessário fazer a execução do arquivo competencia01.py. Com o comando:

$ python3 competencia01.py

Este irá primeiramente fazer a criação das pastas para organização dos resultados para então abrir o primeiro software, o software para trimagem dos dados: FastQC. Para esta análise seguir os próximos passos manualmente:



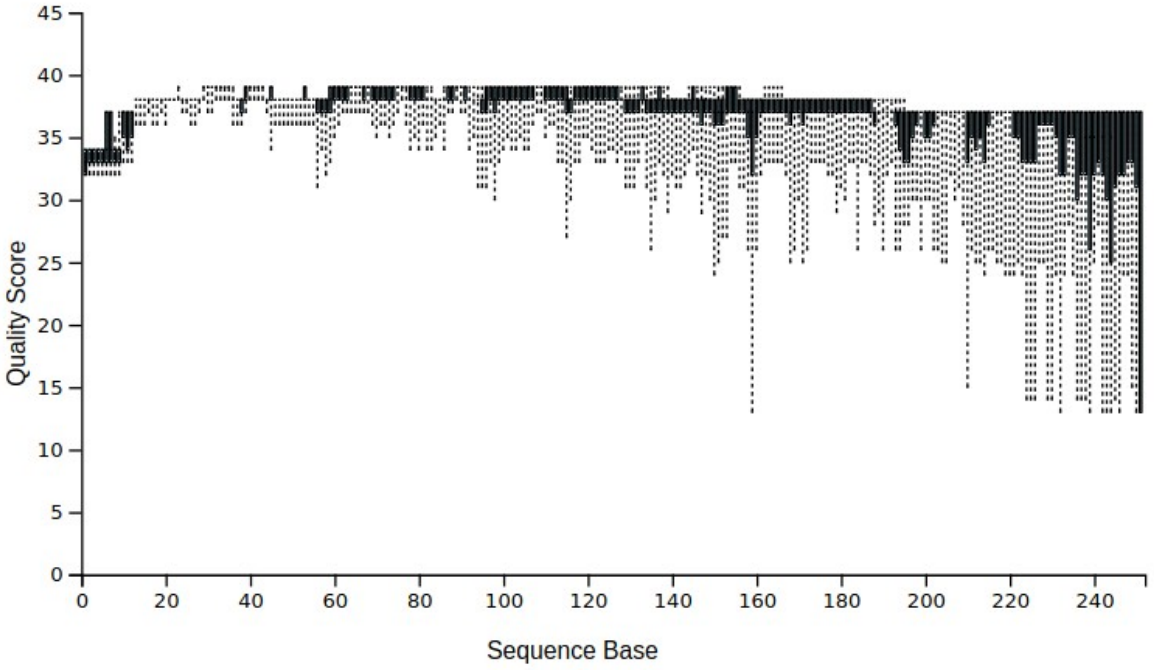


File > Open > caminho dos arquivos > verificações > File > Save Report > caminho da pasta de salvamento**

* O caminho para os arquivos é: desafio/testes/metadados/fqs

** O caminho para salvamento dos arquivos é: trimmagem

O script conduzirá as análises da base de dados, a única parte manual será a análise de qualidade utilizando o fastqc. Nesta analise podemos ver que as amostras apresentam resíduos dos adaptadores de squenciamento, justificando o corte que será realizado com o qiime2, presente na imagem a baixo, e ainda, o final destas sequências também apresenta diversos ‘ruídos’, então o corte final foi realizado na base 229, resultando em 215 pb.



A classificação taxonomica foi relizada também no qiime2 de forma automática utilizando o banco de dados SILVA 138, treinado para esta análise. Ja no caso da tabela foi realizado um barplot para a obtenção da tabela solicitada.

Competência básica para Análise de Dados

O objetivo deste desafio será o de gerar análises gráficas (descritivas e estatísticas) a partir de dados de sequenciamento (recomendado utilizar a linguagem R). Utilize como base os arquivos gerados na etapa 1.4. Caso não consiga gerar um arquivo no formato esperado, você pode utilizar o arquivo em 'Desafio_Neoprospecta/tables'.

OBS1: Tente automatizar as etapas quando possível, evitando etapas manuais.

OBS2: Lembre de disponibilizar o código usado para gerar os gráficos e informações/métricas, assim como descrever quaisquer etapas manuais usadas para manipular os dados com comentários no código (ex: '#a tabela de OTUs foi transposta no LibreOfficeCalc antes desta etapa').

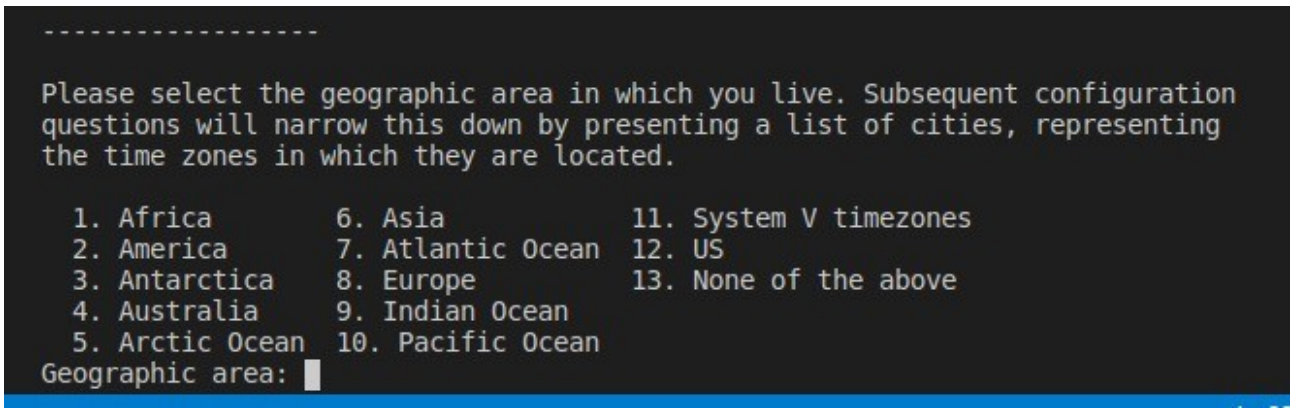
2.1) Plotar um gráfico de barras que mostre a contagem absoluta das 50 bactérias mais abundantes, agrupadas por tempo (dia após o desmame);

2.2) Plotar um gráfico de PCoA (Principal Coordinates Analysis) mostrando o perfil de agrupamento entre as amostras por dia após o desmame;

2.3) Usar alguma métrica que mostre as bactérias diferencialmente abundantes entre os dias de desmame (edegR ou DESeq2, por exemplo). Em um arquivo (PDF, HTML, DOC ou similar), descreva os resultados obtidos e explique quais foram os critérios de escolha dos métodos analíticos usados.

Resultados

Desejava-se que esta competência também fosse realizada de forma automática, dentro do Docker com a leitura de um arquivo em R. Porém, foi abortada a partir do momento que a instalação do pacote de dados do R fez com que o Docker tivesse que fazer questionamentos ao usuário referentes a instalação. Sendo assim, fez-se a utilização do RstudioCLoud no próprio Windows.



Conhecimentos de Bioinformática

Questões

Conhecimentos de bioinformática: questões dissertativas sobre montagem de genoma
3.1) Descreva todas as etapas que você usaria para realizar a montagem, anotação e verificação de qualidade de um genoma de um isolado bacteriano, tendo como base arquivos FASTQ de um sequenciamento Illumina MiSeq paired-end (arquivo R1 e R2), com 2 milhões de reads cada. Reads do arquivo R1 tem 306 pares de base, e reads do arquivo R2 tem 206 pares de base. Suponha que a amostra está bem isolada, contendo um único organismo.

3.2) Descreva como você faria a identificação taxonômica do genoma montado, considerando que a amostra realmente era um isolado bacteriano.

3.3) Os processos descritos em 3.1 e 3.2 são passíveis de automação? Seria possível montar um script que realize todo o processo, tendo como input apenas os arquivos FASTQ (R1 e R2) do sequenciamento? Discorra sobre a possibilidade disso, e, caso possível, como garantir (ou ao menos medir) a qualidade desta montagem/anotação/identificação. Comente sobre possíveis problemas.

3.4) Descreva como você faria a identificação taxonômica do genoma montado, considerando que a amostra não foi bem isolada, e pode conter mais de um organismo (considere que como houve uma tentativa de isolar, ela deve conter entre 1 e 5 organismos).

3.5) Você identificou a amostra e reconheceu que ela não estava bem isolada. Como você poderia solucionar este genoma? Descreva o que faria para separar os scaffolds em dois arquivos de genoma finais. Como poderia medir a qualidade da montagem final?

Discussão

3.1) Com a chegada do sequenciamento para este tipo (paired-end) seriam realizadas análises para qualidade das sequências, onde qualidades inferiores a phred 20 serão descartadas, vistas previamente no software FastQC e cortes realizados com o *Trimmomatic*. A montagem seria realizada com o software CLC Genome Workbench e a anotação funcional e possíveis correções, o software Artemis.

3.2) A identificação taxonômica poderia ocorrer com um simples blastn utilizando genes alvo, no caso de bactérias, o gene rRNA 16s. Para confirmação, seria feito uma árvore filogenética contendo organismos próximos e o output para confirmação em nível de espécie. Para estas últimas etapas poderão ser utilizados o software MEGA 7 e o bioedit.

3.3) Seriam passível de automatização em partes, pois os cortes de qualidade são “pessoais”, mesmo que utilizadas as mesmas métricas e o mesmo raciocínio, cada sequência precisa de uma análise única. Existem diversas problemáticas em questão aos arquivos gerados em imagens, não é impossível que sejam interpretados pela máquina, mas o machine learning envolvido deverá ser muito desenvolvido para que não hajam falsos positivos ou negativos.

3.4) A identificação taxonômica de amostras não muito bem isoladas pode ser realizadas com o sistema de barcodes para amostras possíveis. Se o genoma foi montado, porém não muito bem isolado, ou seja, podem haver genes de outros organismos, creio que somente um novo isolamento com melhor qualidade para a realização da montagem do genoma.

3.5) Não foi possível chegar a uma conclusão para a separação dos scaffolds em arquivos finais, podem haver misturas e a impossibilidade da separação.