Análise de genoma na linha de comando

Neste documento, fornecemos instruções para uma análise inicial, incluindo a criação de uma estrutura filogenética, de sequências de agentes patogénicos bacterianos. Todas estas análises, e muitas outras, requerem que primeiro se gerem e montem genomas de consenso a partir de dados de sequenciação primárias.

Notas importantes para acompanhar este tutorial:

- O texto com um fundo cinzento em tipo de letra `monoespaçado` representa comandos a escrever. Geralmente, os comandos têm uma linha, no entanto, neste documento, os comandos podem passar para a linha seguinte visualmente. Adicionaremos uma linha em branco entre os comandos para indicar quando existem vários comandos.
- O texto negritado rodeado por < > é algo que terá de substituir pelo seu próprio nome de ficheiro, percurso ou amostra.
- Este tutorial pressupõe que você configurou o diretório do pipeline de bioinformática no seu computador e compilou sequências de consenso no formato FASTA. Se este não for o caso e tiver dados de sequenciação primários não compilados, siga as instruções Configuração da linha de comandos e Configuração do genoma na linha de comandos antes de continuar.
- A análise filogenética requer uma série de antecedentes de genomas para comparar as suas sequências. Os genomas anteriores podem ajudá-lo a identificar a que linhagens pertencem as suas sequências e a determinar se as sequências recentemente geradas são semelhantes às anteriormente publicadas. Deve ser dada uma atenção especial à inclusão de genomas que tenham uma diversidade temporal, geográfica e genética adequada.

Este tutorial irá guiá-lo através de uma análise inicial de sequências de genomas de agentes patogénicos. O processo é composto por duas etapas principais:

1. Digitar as sequências recém-geradas.
2. Criar uma estrutura filogenética de sequências de agentes patogénicos recentemente geradas e publicamente disponíveis.

Este tutorial baseia-se no pipeline bioinformático BACPAGE disponível em https://github.com/CholGen/bacpage. As instruções para configurar este pipeline e um diretório de projeto podem ser encontradas em Configuração da linha de comando e Montagem do genoma na linha de comando. Todos os comandos neste tutorial devem ser executados a partir do diretório do pipeline de bioinformática (normalmente em ~/bacpage) e fazer referência ao arquivo YAML de configuração editado no Montagem do genoma na linha de comando (geralmente localizado em `<**project-path**>/config.yaml`).

PASSO 0: Configuração do diretório do projeto

Esta análise pressupõe que você concluiu a Montagem do Genoma na Linha de Comando e criou um diretório de projeto dentro do diretório do pipeline. Além disso, você deve ter gerado sequências de consenso a partir de leituras de sequenciamento primário para cada uma das suas amostras.

1. Navegue até ao diretório de projeto:

```
cd ~/bacpage/<project-path>
```

2. Confirme que existem ficheiros FASTQ desmultiplexados para cada amostra (ou seja, dois ficheiros FASTQ por amostra) no seu diretório de projeto, visualizando o conteúdo do diretório input/ do seu diretório de projeto.

```
ls input
```

3. Confirme a existência de sequências de consenso para cada amostra (ou seja, um ficheiro FASTA por amostra) no diretório do projeto, visualizando o conteúdo do diretório `results/consensus_sequences/`:

```
ls results/consensus_sequences/
```

PASSO 1: Tipagem bacteriana

Um primeiro passo para muitos protocolos de análise é caraterizar amplamente os isolados utilizando a sua informação molecular. Para tal, utilizaremos o método de tipagem de sequências multi-locus (MLST), que detecta variantes alélicas de sete genes de manutenção ubíquos, e a serotipagem, que detecta a presença ou ausência de seis genes de factores de virulência.

Nota: Por predefinição, o pipeline do bacpage está configurado para analisar genomas de Vibrio cholerae. Se estiver a estudar outro agente patogénico, pode alterar os parâmetros "mlst_profiling/scheme" em <**project-path**>`/config.yaml` para qualquer esquema de tipagem publicado pelo PubMLST (execute mlst -list para ver os esquemas de tipagem disponíveis).

Além disso, é possível fornecer seu próprio diretório de genes de fator de virulência no formato FASTA em vez dos genes específicos do Vibrio cholerae fornecidos pelo pipeline. Para fazer isso, altere o valor de "reference_genes" em <**project-path**>`/config.yaml` para o diretório que contém arquivos FASTA para esses genes.

1. Navegue até o diretório do pipeline de bioinformática:

```
cd ~/bacpage
```

2. Para efetuar o MLST, execute o seguinte comando:

```
snakemake --configfile <project-path>/config.yaml --cores
<number-of-processors> --keep-going --until mlst_profiling
```

Este comando irá gerar um único arquivo `<project-path>/results/reports/mlst_types.csv`. O arquivo é um CSV que contém uma linha para cada amostra, descrevendo quais variantes alélicas ela possui para cada um dos genes de manutenção (adk, gyrB, mdh, metE, pntA, purM, pyrC). Utilizando estes alelos, o MLST atribui um tipo de sequência a cada amostra.

3. Os isolados também podem ser classificados pela presença ou ausência de cinco genes de factores de virulência. Execute o seguinte comando para detetar estes genes para cada uma das suas amostras:

```
snakemake --configfile <project-path>/config.yaml \
--cores <number-of-processors> --keep-going --until
virulence_factor_profiling
```

Esse comando gerará um único arquivo `<project-path>/results/reports/typing_information.csv`, que será parecido com o exemplo abaixo para três genomas de cólera. Cada linha relata a fração do gene (indicada pela coluna) que é coberta pelas leituras de uma amostra individual.

sample	ctxA	tcpA_ clássico	tcpA_eltor	toxR	wbeO1	wbfO139
Amostra 1	1.00	0.32	1.00	0.98	1.00	0.22
Amostra 2	1.00	0.31	1.00	0.98	1.00	0.22
Amostra 3	1.00	0.33	1.00	0.98	1.00	0.22

Os genes encontrados numa amostra terão uma fração coberta de cerca de 1,00, indicando que as leituras mapeiam todo o gene. Os genes não encontrados nesta amostra terão uma fração de cobertura inferior. No exemplo acima, podemos dizer que a primeira amostra contém os genes ctxA, wbeO1 e tcpA El Tor, mas não os genes wbfO139 ou tcpA clássico. Podemos, portanto, dizer que se trata de Vibrio cholerae O1, e que provavelmente pertence ao biótipo El Tor.

PASSO 2: Alinhamento de Sequências Múltiplas

Vamos agora gerar uma filogenia incluindo os seus genomas recém-gerados. Embora às vezes possa ser útil fazer uma árvore filogenética usando apenas as sequências recém-geradas, geralmente é mais útil combinar as sequências recém-geradas com um conjunto de sequências publicadas anteriormente, chamado de "conjunto de dados anteriores".

A análise filogenética requer um alinhamento de sequências múltiplas como entrada. Uma vez que todas as suas sequências recém-geradas e o conjunto de dados histórico foram ambos montados através do alinhamento com uma referência comum, podemos facilmente gerar um alinhamento concatenando os seus ficheiros individuais. Se todas as suas sequências (incluindo quaisquer sequências previamente publicadas que esteja a incluir como parte do seu conjunto de dados históricos) não foram montadas através do alinhamento com o mesmo genoma de referência, terá de utilizar um software de alinhamento computacionalmente intensivo para alinhar as sequências umas com as outras.

Nota: Recomendamos que apenas as sequências que cobrem pelo menos 90% do genoma sejam incluídas na inferência filogenética. Por padrão, o pipeline só incluirá sequências no alinhamento se elas atingirem esse limite de cobertura. Se desejar um rigor diferente, altere o valor "*tree_building/required_coverage*" em <**project-path**>/config.yaml para o valor desejado. Além disso, você pode considerar o uso de resultados de digitação para informar quais sequências incluir (por exemplo, incluir apenas sequências classificadas como um determinado serótipo).

1. Localize o conjunto de dados histórico que pretende comparar com as sequências recentemente geradas. Este conjunto de dados deve ser um único ficheiro FASTA contendo entradas separadas para cada sequência do conjunto de dados anteriores, também designado por "multi-FASTA". Recomendamos que coloque o ficheiro multi-FASTA no diretório de recursos/ do pipeline de bioinformática (normalmente ~/bacpage/resources). Pode fazê-lo utilizando a linha de comandos ou movendo simplesmente o ficheiro utilizando o navegador de ficheiros do seu computador.

2. Depois de ter colocado o conjunto de dados de fundo FASTA no diretório de recursos, determine o percurso absoluto do conjunto de dados anteriores. Se foi colocado na localização recomendada, pode encontrar o percurso absoluto navegando para o diretório de recursos e verificando a saída de pwd:

   ```
   cd ~/bacpage/resources

   pwd
   ```

   O percurso absoluto do conjunto de dados anteriores será o resultado de pwd mais o nome do ficheiro do conjunto de dados existente (terminando com .fasta).

3. Navegue de volta para o diretório do pipeline bioinformático:

   ```
   cd ~/bacpage
   ```

4. Adicione o percurso absoluto do seu conjunto de dados anteriores a <**project-path**>/config.yaml. Abra o arquivo de configuração em um editor de texto, altere o valor de <**background-dataset-path**> na linha 8 (veja abaixo), altere o valor de "*generate/phylogeny*" na linha 15 para "True" e salve o arquivo.

   ```
   background_dataset: "<background-dataset-path>"
   ```

5. Gere um alinhamento de sequência múltipla executando o seguinte comando:

```
snakemake --configfile <project-path>/config.yaml --cores
<number-of-processors> --keep-going --until concatenate_sequences
```

Este passo simplesmente concatena seus genomas recém-gerados, e o conjunto de dados anteriores caso esteja presente, em um multi-FASTA localizado em: <**project-path**>/intermediates/illumina/phylogeny/complete_alignment.fasta

PASSO 3: Criar uma árvore filogenética

Tendo gerado o alinhamento múltiplo de sequências, podemos prosseguir com a inferência de uma árvore filogenética.

1. Execute o passo de inferência filogenética do pipeline executando o seguinte comando:

```
snakemake --configfile <project-path>/config.yaml \
--cores <number-of-processors> --keep-going --until generate_tree
```

Resumidamente, este passo gera uma filogenia usando iqtree. Por padrão, esta etapa usará o modelo de substituição GTR e calculará o suporte do ramo conduzindo 1000 bootstraps. Essas opções podem ser alteradas modificando o valor de "*tree_building/iqtree_parameters*" em <**project-path**>/config.yaml. Esse processo é computacionalmente intensivo e pode levar algumas horas para ser concluído, dependendo do tamanho e do número de sequências que estão sendo analisadas, bem como da velocidade do seu computador.

A saída deste comando é uma filogenia no formato Newick, <**project-path**>/results/<project-directory-name>.ml.tree

2. Embora o ficheiro de árvore seja um ficheiro de texto que pode ser aberto e lido num editor de texto, geralmente não é interpretável neste formato. Utilizaremos um visualizador de árvores GUI chamado FigTree para visualizar ficheiros de árvores. No diretório de aplicações do seu computador, abra o FigTree. Clique em File -> Open, e seleccione a filogenia recentemente gerada no navegador de ficheiros que se abre. Alternativamente, você pode abrir o arquivo <**project-directory-name**>.ml.tree diretamente do navegador de arquivos.

3. A filogenia deve agora aparecer na janela principal do FigTree. Procure na árvore pelas suas amostras, determine quais são as amostras mais próximas..