Conceção do genoma bacteriano com base em referências na linha de comando

Neste documento, fornecemos instruções para efetuar uma montagem baseada em referência de Vibrio cholerae. Pode executar os passos com dados de sequenciação primárias que tenha gerado e copiado para o seu computador (ver instruções para configurar o seu diretório do projeto abaixo).

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 ficheiro, percurso ou nome de amostra.
- Este tutorial pressupõe que você tenha configurado o diretório do pipeline de bioinformática no seu computador Antes de iniciar o tutorial abaixo, verifique se o seu computador tem um ficheiro chamada bacpage/ no diretório inicial. Se não houver essa pasta no computador, conclua as instruções de Configuração do pipeline de bioinformática para instalar os arquivos e o software necessários antes de continuar.

Este tutorial irá guiá-lo através de montagens baseadas em referências a partir de dados de sequenciação do amplicon. O processo é composto por cinco etapas principais, incluindo uma etapa de avaliação da sequência:

1. Alinhar ficheiros FASTQ de extremidade emparelhada a um genoma de referência, gerando ficheiros BAM.
2. Chamar as variantes no ficheiro BAM relativamente à referência.
3. Gerar uma sequência de consenso a partir das variantes.
4. Mascarar regiões recombinantes conhecidas a partir da sequência de consenso
5. Avaliar a qualidade das leituras de sequenciação em bruto e do alinhamento.

Estes passos são efectuados utilizando a plataforma snakemake. Snakemake é uma ferramenta para criar pipelines bioinformáticos reprodutíveis e modulares. Cada tarefa de uma análise, incluindo as acima referidas, podem ser escritas como um passo individual de um pipeline. O Snakemake facilita a realização de análises, reduzindo o número de comandos que tem de escrever, paraleliza passos em todas as suas amostras e confirma que os passos foram concluídos com êxito.

Abaixo, encontrará instruções para utilizar o snakemake para executar todo o pipeline de uma só vez, bem como instruções (no Apêndice) que o orientam em cada um dos passos acima, um por um.

PASSO 0: Configuração do Diretório do Projeto [sempre necessário]

Antes de iniciar qualquer análise bioinformática, é boa prática criar um diretório especificamente para o conjunto de dados que vai analisar. Este diretório é onde irá executar as análises e guardar os resultados e é designado por diretório de projeto. É aconselhável configurar o diretório de projeto dentro do diretório bacpage/ no seu diretório INICIAL para garantir que todos os dados e resultados de sequenciamento estejam no mesmo lugar e possam usar as mesmas ferramentas e software de pipeline.

Referimo-nos ao percurso do arquivo bacpage/ como <**sequencing-path**>. Na maioria das máquinas, o <**sequencing-path**> será ~/bacpage.

1. Navegue até o ficheiro do pipeline de bioinformática como descrito acima:

   ```
   cd ~/bacpage
   ```

2. Na pasta bacpage/ existe uma subpasta chamada exemplo/. O primeiro passo é fazer uma cópia desta pasta. Será feita uma nova cópia sempre que executar uma nova sequenciação, garantindo assim que os ficheiros e pastas dentro do diretório de exemplo são configurados exatamente da mesma forma de cada vez. Para copiar esta pasta e dar-lhe um nome específico para o projeto, execute o comando abaixo. Recomendamos que dê ao seu diretório de projeto um nome informativo, como <data>_<nome da execução de sequências> (por exemplo: 20220609_cholera_run1).

   Escreva o seguinte na janela do terminal e, em seguida, pressione **Enter**:

   ```
   cp -R example/ <project-directory-name>
   ```

2. Navegue até ao diretório do seu projeto e registe o percurso absoluto utilizando.

   ```
   cd <project-directory-name>

   pwd
   ```

   Referimo-nos ao percurso absoluto para o seu diretório de projeto como <**project-path**> Você precisará do percurso absoluto para o seu diretório de projeto na etapa 6 abaixo.

3. Localize os dados de entrada que deseja usar no pipeline de montagem. Este pipeline requer arquivos FASTQ demultiplexados (ou seja, dois arquivos FASTQ por amostra). Utilizando o localizador de ficheiros do seu computador, mova estes dados para o input/ do diretório do seu projeto (ou seja, a pasta que acabou de criar acima).

   No diretório do seu projeto, deve existir um ficheiro chamado sample_data.csv, que contém as informações sobre as suas amostras. Abra este ficheiro no Excel ou num software de folha de cálculo semelhante e adicione a informação relativa a cada uma das suas amostras numa linha individual.

Na coluna da **sample**, registe o nome ou o identificador de cada amostra (estes devem ser únicos e não conter caracteres invulgares, como pontos "." ou barras "/"). Na coluna **read1**, registar o percurso absoluto do ficheiro FASTQ correspondente ao primeiro conjunto de leituras de uma amostra (geralmente com "R1" no nome do ficheiro). Na coluna **read2**, registar o percurso absoluto do ficheiro FASTQ correspondente ao primeiro conjunto de leituras de uma amostra (geralmente com "R2" no nome do ficheiro).

O percurso absoluto de um ficheiro pode ser determinado com o comando pwd. Se os seus ficheiros FASTQ desmultiplexados foram colocados no input/ do seu diretório de projeto, conforme descrito acima, pode encontrar o seu percurso absoluto executando o seguinte comando:

```
cd <project-path>/input

pwd
```

O percurso absoluto dos seus arquivos será a saída do pwd mais "/" mais o nome do arquivo. Um exemplo de ficheiro sample_data.csv completo é mostrado abaixo:

sample	**read1**	**read2**
amostra1	/home/user/bacpage/20220609_run1/input/sample1_S13_L001_R1_001.fastq.gz	/home/user/bacpage/20220609_run1/input/sample1_S13_L001_R2_001.fastq.gz
amostra2	/home/user/bacpage/20220609_run1/input/sample2_S3_L001_R1_001.fastq.gz	/home/user/bacpage/20220609_run1/input/sample2_S3_L001_R2_001.fastq.gz
amostra3	/home/user/bacpage/20220609_run1/input/sample3_S22_L001_R1_001.fastq.gz	/home/user/bacpage/20220609_run1/input/sample3_S22_L001_R2_001.fastq.gz

4. Guarde o ficheiro depois de introduzir os dados das suas amostras.

5. No diretório do seu projeto, abra o ficheiro de configuração chamado config.yaml com um editor de texto à sua escolha. Este ficheiro contém os parâmetros e as opções para a análise. Os parâmetros podem ser alterados para se adequarem à sua análise, mas as opções predefinidas devem ser apropriadas para a maioria das análises.

   Com o arquivo de configuração aberto em um editor de texto, substitua <**project-path**> e <**sequencing-path**> por seus percursos absolutos para os seis primeiros parâmetros.

```
# General parameters; point to input files.
run_type: "Illumina"
samples: "<project-path>/sample_data.csv"
output_directory: "<project-path>"
reference: "<sequencing-path>/resources/vc_reference.fasta"
reference_genes: "<sequencing-path>/resources/cholera_ref_genes/"
recombinant_mask: "<sequencing-path>/resources/cholera_mask.gff"
```

4. Por fim, determine quantos processadores estão disponíveis no seu computador, pois o uso de mais processadores fará com que o pipeline seja executado mais rapidamente. A saída do comando a seguir

indicará quantos processadores você tem disponíveis. Execute o comando abaixo e anote o resultado para mais tarde:

```
cat /proc/cpuinfo | grep processor
```

PASSO 1: Executar o pipeline completo

Se pretender executar todo o pipeline sem passar por cada um dos passos intermédios, pode gerar sequências de consenso e calcular métricas de qualidade com um único comando.

1. Navegue até à localização do pipeline bioinformático:

```
cd ~/bacpage
```

2. Execute o pipeline com o seguinte comando:

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

Isso irá gerar uma sequência de consenso no formato FASTA para cada uma das suas amostras e colocá-las em <**project-path**>/results/consensus_sequences/<**sample**>.masked.fasta. Um relatório HTML contendo alinhamento e métricas de qualidade para suas amostras pode ser encontrado em <**project-path**>/results/reports/qc_report.html.

Se um comando snakemake for concluído com sucesso, deverá ver algo como isto no ecrã:

```
[Tue Sep  5 12:09:40 2023]
Finished job 119.
253 of 253 steps (100%) done
Complete log: .snakemake/log/2023-09-05T120148.784555.snakemake.log
```

Nota: O número da tarefa, o passo e o total de passos dependerão do número de amostras que tiver.

Se um comando snakemake não tiver sido bem sucedido, verá o seguinte no final da saída:

```
Error in rule x:
    jobid: 1
    shell:
        command -options arguments
Shutting down, this might take some time.
Exiting because a job execution failed. Look above for error message
Complete log: .snakemake/log/2023-09-05T120148.784555.snakemake.log
```

Registe a regra falhada, a amostra que está a ser processada e o ficheiro de saída esperado da regra falhada. Pode tentar repetir o passo numa amostra específica, executando o seguinte comando:

Configuração do genoma na linha de comando

```
snakemake --configfile <project-path>/config.yaml --cores
<number-of-processors> --keep-going <expected-output-file>
```

PASSO 2: Avaliar a qualidade da compilação do genoma

Um componente fundamental da geração de montagens de genomas é a avaliação da sua qualidade. O pipeline de bioinformática gera um relatório para inspecionar visualmente a qualidade, a cobertura e a confiança que temos nas sequências de consenso resultantes.

1. Para revisar as métricas de qualidade de suas amostras, abra <**project-path**>/results/reports/qc_report.html com um navegador da Web.

Apêndice à Montagem do Genoma na Linha de Comando

As instruções seguintes descrevem os passos intermédios da geração das sequências de consenso. Este processo pode ser feito opcionalmente em vez de executar o pipeline completo usando as instruções acima.

PASSO 1: Alinhar ficheiros FASTQ de extremidades emparelhadas a um genoma de referência

O primeiro passo da montagem do genoma consiste em pegar nos ficheiros FASTQ de extremidade emparelhada produzidos por uma máquina de sequenciação e alinhá-los com um genoma de referência. Isto irá gerar um ficheiro BAM para cada amostra que é necessário para os passos futuros do pipeline.

Nota: por predefinição, o pipeline alinhará as leituras com uma referência Vibrio cholerae. Se você estiver estudando outro patógeno, poderá alterar os parâmetros de "referência" em <**project-path**>/config.yaml para outra sequência de referência.

1. Navegue no seu diretório de trabalho atual para a localização do pipeline de bioinformática (geralmente ~/bacpage). Todos os passos seguintes devem ser realizados neste diretório. Escreva o seguinte na janela do terminal e pressione **Enter**:

```
cd ~/bacpage
```

2. Execute o passo alignment_bwa do pipeline utilizando o seguinte comando.

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

   Este comando alinha individualmente cada amostra indicada no arquivo sample_data.csv e levará alguns minutos, dependendo do número de amostras.

   Um arquivo BAM será gerado para cada amostra com o formato <**sample**>.sorted.bam. Os arquivos BAM para cada amostra podem ser encontrados em <**project-path**>/intermediates/illumina/alignments/.

PASSO 2: Chamar Variantes Relativas ao Genoma de Referência

O próximo passo na montagem do genoma é determinar as diferenças entre as leituras sequenciadas e o genoma de referência. Estas diferenças são designadas por variantes e descrevem a evolução da amostra em relação ao genoma de referência (normalmente o genoma mais antigo de um organismo).

1. Execute o passo de identificação de variantes do pipeline utilizando o seguinte comando:

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

Esse comando chama variantes para cada amostra indicada no arquivo sample_data.csv, filtra variantes de baixa qualidade e sem suporte e normaliza inserções e exclusões. Um arquivo de formato de chamada de variante (arquivo VCF) será gerado para cada amostra com o formato `<sample>.filt.norm.vcf.gz` (consulte https://en.wikipedia.org/wiki/Variant_Call_Format para obter uma descrição do formato VCF). Os ficheiros VCF para cada amostra podem ser encontrados em `<project-path>/intermediates/illumina/variants/`

PASSO 3: Gerar uma sequência de consenso

Iremos agora gerar uma sequência genómica para cada amostra, aplicando as variantes à nossa sequência de referência. Uma vez que este genoma é um resumo de muitas leituras de sequenciação, chamamos a esta sequência uma sequência de consenso.

1. Execute o passo de chamada de consenso do pipeline utilizando o seguinte comando:

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

Esse comando gera uma sequência de consenso para cada amostra indicada no arquivo sample_data.csv. Um arquivo FASTA contendo a sequência de consenso será gerado para cada amostra com o formato `<sample>.consensus.fasta`. Os ficheiros FASTA para cada amostra podem ser encontrados em `<project-path>/intermediates/illumina/consensus/`.

PASSO 4: Mascarar a sequência de consenso

Recomendamos mascarar as regiões do genoma que têm baixa cobertura e/ou são totalmente recombinantes. O mascaramento destas regiões evita resultados erróneos das análises a jusante, incluindo a tipagem e a inferência filogenética. As regiões de baixa cobertura podem ser determinadas diretamente a partir do ficheiro BAM gerado para cada amostra, enquanto as regiões totalmente recombinantes requerem um conhecimento prévio do organismo em estudo. Como parte do pipeline, fornecemos um ficheiro que indica as regiões totalmente recombinantes do genoma da cólera. Se estiver a estudar outro organismo, terá de atualizar o parâmetro recombinant_mask no ficheiro de configuração para outra máscara específica do organismo.

1. Para executar o passo de mascaramento do pipeline, utilize o seguinte comando:

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

Este comando mascara a sequência de consenso para cada uma das suas amostras. Um arquivo FASTA contendo a sequência de consenso mascarada será gerado para cada amostra com o formato <**sample**>.masked.fasta. Os arquivos FASTA mascarados para cada amostra podem ser encontrados em: <**project-path**>/intermediates/illumina/consensus/

PASSO 5: Avaliar a qualidade da montagem

Um componente chave da conceção de genomas é a avaliação da sua qualidade. Iremos gerar relatórios para inspecionar visualmente a qualidade, cobertura e confiança que temos nas sequências de consenso resultantes.

1. Gerar os relatórios de controlo de qualidade utilizando o seguinte comando:

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

Esse comando gera um único relatório HTML contendo métricas de qualidade para todas as amostras. O relatório HTML pode ser encontrado em <**project-path**>/results/reports/qc_report.html

2. Abra esse arquivo com um navegador da Website.