

PIPELINE

MAIN.NF Estamos verificando qual etapa da pipeline foi solicitada (via parametro step) pelo usuário.

SE(step == 1): exibe as configurações de basecalling e alinhamento

SE(step == 2): exibe as configurações de filtragem e controle de qualidade.

SE(step == 3): exibe as configurações de metilação e relatório multiQC.

Import Workflows:

Importa workflows que foram definidos em outros arquivos dentro do projeto.

arquivos iniciais:

Definindo os arquivos iniciais e os canais de dados com base no valor do parâmetro --step

Organizar os dados e distribuir eles para os processos

workflow

Toda a lógica de fluxo de trabalho será executada conforme for definida aqui.

SE(step == 1): chama basecalling, passa os parâmetros.

SE(2_from_step_1"): chama a filtragem, com os dados do step 1.

SE(2_from_minknow) chama a filtragem, com os dados de MinKNOW

SE(step == 3): análise e visualização dos dados



Modulos:

Basecalling: converter sinais brutos de sequenciamento em sequências de DNA legíveis

process FAST5_to_POD5 {

conversão de arquivos FAST5 para o formato POD5.}

BASECALL_CPU {Esse processo usa a ferramenta Dorado para fazer o basecalling . Dorado:transforma os dados "crus" em leituras legíveis de nucleotídeos}

BASECALL_CPU_DEMUX

BASECALL_GPU

.BASECALL_GPU_DEMUX

*Todos aqui são processos.

CPU/ GPU porque o Dorado pode rodar em ambos os tipos de hardware, mas com diferenças significativas de desempenho e requisitos.

dorado só aceita POD5

Entendendo comandos:

o bloco **script:** define o que será executado no shell (terminal)

echo no contexto do script: (e em Bash em geral) é usada para imprimir mensagens no terminal ou log

Modulos:

ALCULATE_COVERAGE: Calcular a cobertura média por cromossomo de um arquivo BAM (sequenciamento alinhado).

MERGE_COVERAGE: Juntar todos os arquivos de cobertura de múltiplas amostras em um único arquivo e formatar para o MultiQC.

*Todos aqui são processos.

Modulos:

CONVERT_INPUT_FROM_MINKNOW convertem dados de saída do MinKNOW (software de sequenciamento da Oxford Nanopore) para um formato padronizado (arquivos .bam e .txt) que o pipeline possa processar.

MinKNOW (software de sequenciamento da Oxford Nanopore)

*Todos aqui são processos.

Modulos:

FILTER_BAM: realiza o filtro e a qualificação de arquivos BAM gerados por sequenciamento, com base em critérios como MAPQ (qualidade de mapeamento) e número mínimo de leituras mapeadas. Ele também gera estatísticas para análise em MultiQC.

****O samtools é uma ferramenta para manipular arquivos de alinhamento de sequenciamento, como arquivos SAM (Sequence Alignment/Map) e BAM (a versão binária compactada do SAM). Ele permite que você visualize, filtre, converta, ordene, indexe e gere estatísticas desses arquivos.**

Modulos:

MODKIT : processo para analisar dados de metilação de DNA extraídos de arquivos BAM (sequências alinhadas). A ferramenta principal usada é o Modkit, que é especializada em trabalhar com modificações epigenéticas, como a metilação.

*Todos aqui são processos.

Modulos:

MULTIQC: executa o MultiQC, que serve para agregar e resumir resultados de várias ferramentas de bioinformática em um único relatório HTML.

*Todos aqui são processos.

Modulos:

NUM_READS_REPORT

sses dois processos do Nextflow, MAKE_QC_REPORT e MERGE_QC_REPORT, fazem parte de uma etapa de avaliação de qualidade (QC) de dados de sequenciamento. Eles geram e consolidam tabelas de métricas de qualidade por amostra que depois podem ser incluídas num relatório MultiQC.

MAKE_QC_REPORT Extrai estatísticas por amostra (nº de leituras, tamanhos, qualidade)

MERGE_QC_REPORT Junta esses arquivos em tabelas compatíveis com MultiQC

*Todos aqui são processos.

Modulos:

PYCOQC_NO_FILTER e PYCOQC_FILTER: são responsáveis por rodar a ferramenta PycoQC, que gera relatórios de qualidade para dados de sequenciamento Nanopore com base nos arquivos .bam e nos resumos de sequenciamento (sequencing_summary.txt).

*Todos aqui são processos.

Modulos:

PYCOQC
PYCOQC_NO_FILTER e PYCOQC_FILTER. Ambos são responsáveis por rodar a ferramenta PycoQC, que gera relatórios de qualidade para dados de sequenciamento Nanopore com base nos arquivos .bam e nos resumos de sequenciamento (sequencing_summary.txt).

*Todos aqui são processos.

sub_workflows - Workflow

BASECALLING.NF

O código define um chamado BASECALLING

Converte arquivos de sequenciamento Nanopore (.fast5) em um formato mais eficiente (.pod5)
Executa basecalling (transforma sinais brutos em sequências de DNA)
Usa CPU ou GPU, com ou sem demultiplexação (separação por barcodes), dependendo dos parâmetros fornecidos
Coleta e organiza as saídas (arquivos BAM e TXT) para etapas seguintes do pipeline

sub_workflows - Workflow

FFILTERING_AND_QC_FROM_MINKNOW,

É usado para filtrar dados de sequenciamento e gerar relatórios de qualidade. O processo segue uma sequência de passos com base nas entradas fornecidas, usando módulos pré-definidos.

sub_workflows - Workflow

FILTERING_AND_QC_FROM_STEP_1

Filtra arquivos BAM com base em qualidade (mapq).
Gera dois relatórios de qualidade com a ferramenta PycoQC:
Um para os dados não filtrados.
Outro para os dados filtrados.
Gera um relatório final resumindo número de leituras, comprimentos e limiares de qualidade.

sub_workflows - Workflow

MODKIT_AND_MULTIQC,

Executa etapas finais de análise de qualidade e modificação epigenética em dados de sequenciamento filtrados.

CALCULATE_COVERAGE Calcula a cobertura dos arquivos BAM.
MERGE_COVERAGE Junta os arquivos de cobertura em um só.
MERGE_QC_REPORT Agrupa e formata relatórios de qualidade anteriores.
MULTIQC Gera um relatório visual integrando todos os dados de QC.
MODKIT Extrai e analisa modificações epigenéticas nos BAMs.