Nanopore Direct RNA Sequencing Pipeline Tutorial

This pipeline is designed to process, analyze, and interpret direct RNA sequencing data from Oxford Nanopore sequencing technology. The workflow includes steps from raw data processing through base modification analysis, motif searching, differential expression analysis, and gene set enrichment. Here’s an overview of each stage.

Overview of Steps

Rule: download\_dorado\_model (Download Required Models)

The pipeline begins with downloading base-calling and modification-calling models required for analyzing RNA signals. These models are obtained from the Oxford Nanopore Dorado repository.

Rule: get\_cdna\_transcriptome\_reference (Download Reference Transcriptome)

Retrieves the transcriptome reference file for Homo sapiens, needed for accurate alignment and annotation of the RNA sequences.

Rule: basecall (Base-calling the RNA Data; for each sample)

Base-calling converts raw ionic signal data into nucleotide sequences. Each sample is processed individually, producing sequence files for downstream analysis.

Rules: sort\_bam and index\_bam (Sort and Index BAM Files)

After base-calling, BAM files are sorted and indexed. Sorting arranges reads by genomic coordinates, facilitating efficient retrieval, and indexing enables quick access to specific genomic regions.

Rule: count\_modified\_bases (Count Base Modifications)

Counts specific modified bases (e.g., m6A) across the RNA sequences. This is crucial for understanding modifications in RNA molecules that may affect gene expression.

Rule: bed\_conversion (Convert BED Files to RData Format)

Converts BED files of modification counts into RData format for compatibility with downstream analysis in R. This conversion allows efficient loading and processing in R-based scripts.

Rule: filter\_modification\_sites (Filter Modification Sites)

Applies filters to modification sites, removing those with insufficient read coverage or confidence levels. This ensures only high-confidence modifications are retained for further analysis.

Rule: merge\_all\_samples (Merge All Samples)

Combines data from all samples, identifying common modification sites across conditions. This step is essential for comparison between control and treatment groups.

Rule: post\_processing\_mod\_sites (Post-process Modification Sites)

Conducts further processing on merged modification sites, such as calculating differential modification rates between conditions.

Rule: compute\_z\_statistics (Compute Z-statistics)

Computes statistical measures for modifications, including Z-scores, which help in identifying significantly modified sites.

Rule: generate\_count (Generate Expression Counts)

Generates count tables of RNA transcripts for each sample, providing data for gene expression analysis.

Rule: generate\_gene\_expression\_count (Generate Gene Expression Count Table)

Consolidates transcript counts into gene expression counts, yielding a summary table for RNA expression across genes.

Generate Fasta Subset for Gene Transcript

Rule: generate\_fasta\_subset\_gene\_transcript

Extracts relevant sequences from the reference fasta file, creating a subset containing only the transcripts of interest based on detected modifications.

Annotate Modified Sites

Rule: annotate\_modified\_sites

Annotates each modification site with gene and transcript information, providing context for each modification relative to genes.

Identify Consensus Motifs and Base Proportions

Rule: consensus\_motif\_and\_base\_proportion

Calculates consensus motifs and base compositions for each modification site, identifying characteristic motifs that may play functional roles.

Generate DRACH Signatures

Rule: generate\_DRACH\_signatures

Searches for DRACH motifs in the sequences, a well-known signature for m6A modifications, to validate the modification patterns observed.

Motif Search and Signature Detection

Rules: motif\_search and find\_signature\_motifs

These steps perform a targeted motif search to detect specific signature motifs that could indicate functional RNA elements or modifications.

Perform Differential Expression Analysis

Rule: perform\_differential\_expression\_analysis

Conducts a differential expression analysis to determine genes or transcripts with significant expression changes between control and treatment groups.

Identify Condition-characteristic Genes

Rule: identify\_condition\_characteristic\_genes

Identifies genes that are characteristic of each condition (control vs. treatment), potentially highlighting biomarkers or therapeutic targets.

Perform Gene Set Enrichment Analysis (GSEA) and Annotate Genes

Rules: GSEA\_and\_genes\_annotation and perform\_GSEA

The final steps perform GSEA to identify enriched biological pathways and annotate genes. This provides insights into the functional implications of observed modifications and differential expressions.

Compile Results

Rule: all

The all rule aggregates the outputs of each step, ensuring that all components of the analysis have completed successfully.