MoDIFI User Guide

Example Data

Download the required omics datasets from **ENCODE** (https://www.encodeproject.org/) and place them into the appropriate folders under the resources directory. Alternatively, you may download the packaged example dataset from the following link: *(insert link here)*.

RNA-seq

Cell Line	Source URL	Destination Directory
IMR90 Replicate 1	https://www.encodeproject.org/files/ENCFF019KLP/@@download/ENCFF019KLP.tsv	./resources/imr90/rna_seq/rep1/
IMR90 Replicate 2	https://www.encodeproject.org/files/ENCFF268IJX/@@download/ENCFF268IJX.tsv	./resources/imr90/rna_seq/rep2/
GM12878 Replicate 1	https://www.encodeproject.org/files/ENCFF014JRQ/@@download/ENCFF014JRQ.tsv	./resources/gm12878/rna_seq/rep1/

ATAC-seq

Cell Line	Source URL	Destination Directory
IMR90 Replicate 1 (BAM)	https://www.encodeproject.org/files/ENCFF848XMR/@@download/ENCFF848XMR.bam	./resources/imr90/atac_seq/bam/rep1/
IMR90 Replicate 2 (BAM)	https://www.encodeproject.org/files/ENCFF715NAV/@@download/ENCFF715NAV.bam	./resources/imr90/atac_seq/bam/rep2/
IMR90 Peak Calls (BED.gz)	https://www.encodeproject.org/files/ENCFF114GDS/@@download/ENCFF830JMV.bed.gz	./resources/imr90/atac_seq/
GM12878 Replicate 1 (BAM)	https://www.encodeproject.org/files/ENCFF646NWY/@@download/ENCFF646NWY.bam	./resources/gm12878/atac_seq/bam/rep1/
GM12878 Replicate 2 (BAM)	https://www.encodeproject.org/files/ENCFF415FEC/@@download/ENCFF415FEC.bam	./resources/gm12878/atac_seq/bam/rep2/
GM12878 Peak Calls (BED.gz)	https://www.encodeproject.org/files/ENCFF114GDS/@@download/ENCFF478ZAW.bed.gz	./resources/gm12878/atac_seq/

Hi-C

Cell Line	Source URL	Destination Directory
IMR90	https://www.encodeproject.org/files/ENCFF267VBT/@@download/ENCFF267VBT.bedpe.gz	./resources/imr90/hic/
GM12878	https://www.encodeproject.org/files/ENCFF661SAZ/@@download/ENCFF661SAZ.bedpe.gz	./resources/gm12878/hic/

Configuration Settings

MoDIFI Container Image

To run the MoDIFI Nextflow pipeline, you will need the container image file (modifi.sif), which provides the full execution environment with all required Python packages, R packages, and command-line sequencing tools (e.g., samtools, bcftools and bedtools, etc.) pre-installed. This image file can be downloaded from the provided link.

Replicate Numbers

Modify the number of replicates in **modifi_example.config**:

RepNUM_atac = ['GM12878': 2, 'IMR90': 2]

RepNUM_rna = ['GM12878': 1, 'IMR90': 2]

Prior Information

Specify the prior information in the configuration file. The prior file can be downloaded from the provided link and referenced as:

prior_file = "${resources_dir}/Gnocchi.tsv"

Sample Pairing

SamplePair.tsv defines target vs. reference samples for MoDIFI calculations:

Target	Reference	Check
IMR90	GM12878	PASS

Set the path in the config:

SamplePair = "${resources_dir}/SamplePair.tsv"

Promoter Annotation

Promoter genes are defined as **protein-coding genes with canonical transcripts** and should be stored at:

./resources/Promoter/Promoter.tsv

Output Directory

Set the output folder using:

output_dir = "${projectDir}/output/MoDIFI_Example"

The output files generated from running the example dataset can also be downloaded from the provided link if you would like to review the expected results.