# Code Usage Notes

**Part I :**preparation work.

1. git clone --recursive <https://github.com/alcindor819/FDI_code_MATLAB.git>

2. cd FDI\_code\_MATLAB/FDI/Basic\_info/

3. wget -c https://zenodo.org/record/3928546/files/GC.zip

wget -c <https://zenodo.org/record/3928546/files/mappability.zip>

(If this command fails, please manually download the file from the link and upload it to Basic\_info/.)

4.unzip GC.zip

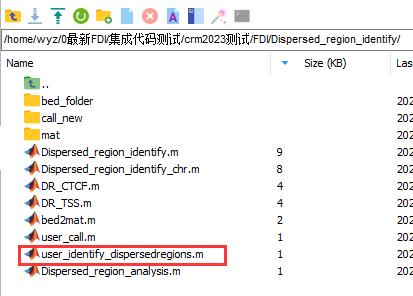
unzip mappability.zip

**Part II :**Notes for identifying dispersed regions.

There is a sample bed file in this folder.



Need to call the function **user\_identify\_dispersedregions.m** in the folder ‘/home/wyz/0latest\_FDI/integration\_code\_tests/crm2023\_tests/FDI/Dispersed\_region\_identify/’.



The parameters to be passed to the function user\_identify\_dispersedregions are as follows:

The first line is the name of the parameter, the second line is an example of the parameter, and the third line is a comment on the parameter.

bed\_folder =

'/home/wyz/0最新FDI/集成代码测试/crm2023测试/FDI/Dispersed\_region\_identify/bed\_folder/';

%This folder contains the .bed files.

bed\_name =

'SRR16574631';

%This is the .bed filename prefix for the samples needed to identify the dispersed region.

mat\_path =

'D:\wyzwork\精简版代码\FDI\Dispersed\_region\_identify\mat\';

%This is the address used to store the intermediate .mat file.

distribution =

'Beta';

%This is the type of distribution, and you can choose a distribution such as normal depending on the actual distribution of the data.

P =

0.01;

%Global and local P-value.

FDR =

0.05;

%FDR threshold.

res\_path =

'D:\wyzwork\精简版代码\FDI\Dispersed\_region\_identify\call\_new\';

%Folders for dispersed regions and TSS,CTCF results.

X =

0.5;

%A parameter of FDI, default is 0.5, range is (0-1].

Y =

10;

%A parameter of FDI, default is 10, range is [10-100].

chr\_n =

1:22;

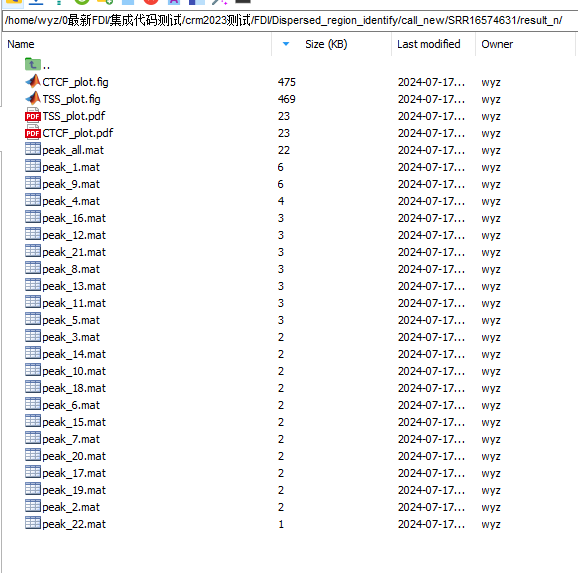
%Identify chromosomes from 1 to 22.

numWorkers =

2;

%The number of cores used in parallel is recommended to be no higher than 2 for an average local computer.

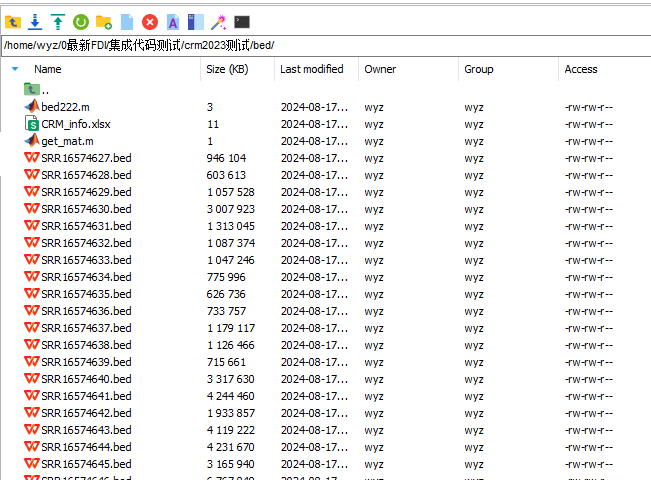
When it finishes running, it generates a folder named bed\_name in res\_path, where the results are stored.



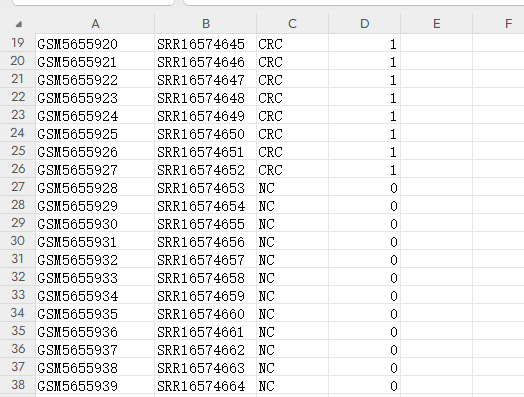
Dispersed regions and total dispersed regions corresponding to the 22 chromosomes, and also the coverage of these discrete regions to TSS and CTCF.

It is recommended to use BH01 samples for testing, because BH01 has enough sequencing depth, and you need to download the data and process it into a bed file by yourself and then call this tool.

**Part III :**Notes for diagnostic.

The bed file for each sample to be classified is stored in (‘/home/wyz/0latest-FDI/integration-code-test/crm2023-test/bed/’).

This is part of the information about these samples, the information we need is the name of the sample (second column) and the status label of the sample (fourth column).

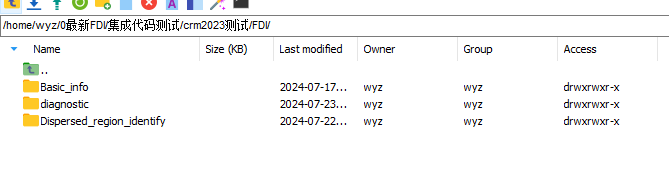


This is the code folder, which contains 3 sections.

Basic\_info stores some basic genomic information about human hg19 required for computation.

Diagnostic is the code module for cross validation.

Dispersed\_region\_identify is a module for identifying and evaluating dispersed regions.



When you need to execute the classification module, you can directly call the **classify\_Kfold** function in the Diagnostic folder.

The input parameters and examples are: user can change bed\_folder,sample\_info\_path,mat\_path to their own absolute path.

bed\_folder =

'/home/wyz/0最新FDI/集成代码测试/crm2023测试/bed/';

%Folder for .bed files.

sample\_info\_path =

'/home/wyz/0最新FDI/集成代码测试/crm2023测试/bed/CRM\_info.xlsx';

%Full path to where the .xlsx file is stored.

bed\_col =

2;

%Columns in the xlsx file that hold the prefix name of the bed file.

pos\_neg\_col =

4;

%xlsx file holding the columns where the sample is cancer (1) or healthy (0).

fold =

10;

%The k-fold cross-validation of k is fold.

X =

0.5;

%A parameter of FDI, default is 0.5, range is (0-1].

Y =

10;

%A parameter of FDI, default is 10, range is [10-100].

readlen =

200;

%Length of each scanning region.

buchang =

20;

%Interval between each scanning region.

gp =

0.05;

%Global and local P-value.

fdr =

0.05;

%FDR threshold.

mat\_path =

bed\_folder;

%Path to the .mat file (this is an intermediate file generated for ease of handling, by default it is the same path as bed\_folder).

This is a parallel MATLAB program that ends with the results of each fold of the k-fold cross validation, the figure shows the 10-fold process, the intermediate data and results are in the folders of the respective folds, and the total AUC and ROC curves are returned to the current folder after the k-folds are finished.

