## Download and Installation

This *Discriminant-Cut* analysis tool does not require installation. Simply unpack and run follow the instructions as below.

## Prepare dataset file

All dataset files should be placed under folder “data” in the same directory where the application executable is located.

Each dataset file is a MATLAB “.mat” file, which includes the following mandatory fields.

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Data Type** | **Description** |
| **X** | *M* x *N1* matrix | Data value for the 1st group, where *M* denotes the number of features and *N1* denotes the number of samples. |
| **Y** | *M* x *N2* matrix | Data value for the 2nd group, where *M* denotes the number of features and *N2* denotes the number of samples. |
| **paired** | Logical | 0 if samples in two groups are unpaired; 1 if paired. |
| **probes** | *M* x 1 cell vector | Feature symbols. |
| **descs** | *M* x 1 cell vector | Feature descriptions. |
| **trueAlt** | *m* x 1 vector | The indexes of features that are truly significant. Will be used to calculate true FDR. If the ground truths are unknown, simply put a zero vector and do not trust the “true FDR” provided in the output. |

## Run *Discriminant-Cut* analysis tool using default settings

Once the dataset file is ready, you can start analyzing this dataset with the *Discriminant-Cut* analysis tool.

First go to command line, move to the directory where the executable is located.

Type in the executable file name followed by the dataset file name to analysis, i.e. “DiscriminantCut [dataset file name]”, and then press Enter. The *Discriminant-Cut* analysis tool will automatically analyze the given dataset using the default parameter settings. To specify parameters other than the default settings, please refer to the “**Specify parameters**” chapter, or simply run “DiscriminantCut” without providing the dataset file name.

A folder called “stats” will be created automatically by the analysis tool to store intermediate results. A corresponding intermediate file named “processed\_[dataset file name]” will be created in the middle of the analysis process. If the intermediate result is successfully generated, the analysis tool will start with the intermediate result rather than the dataset file. To force the analysis tool to generate new intermediate result, just delete the existing one.

A folder called “result” will be also created automatically. This is where your final output will be placed. For each analysis, the analysis tool will generate two result files. One is named “result\_[dataset file name].mat”. This file can be opened by MATLAB and it stores all final output variables; The other one is called “result\_[dataset file name].txt”, which is a text report. For details about how to interpret these two analysis result files, please refer to the “**Interpret analysis result files**” chapter.

## Specify parameters

The full usage of the *Discriminant-Cut* is

“DiscriminantCut dataFile [attList = "voom|ranksum|DESeq2"] [targetFDR=0.05] [nPerm=100] [resolution=10] [FPmethod=median]”.

The details are listed in the following table.

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Type** | **Default Value** |
| **dataFile** | String | **(required)** |
| Required. Full file name including extension under folder "datasets" | | |
| **attList** | String | voom|ranksum|DESeq2 |
| Basic attribute list, up to 3, delimited by a pipe character "|"  Format: "att1", "att1|att2", "att1|att2|att3"  Supported basic attributes:   1. ttest: Corrected t-statistic of SAM 2. logttest: Corrected t-statistic of SAM on logarithmic values 3. ranksum: Corrected ranksum statistic of SAM 4. \*limma: Moderated statistic of limma (require R + limma package) 5. \*voom: Moderated statistic of voom (require R + limma package) 6. \*DESeq2: Wald statistic for NB-based differential expression test from DESeq2 (require R + DESeq2 package) | | |
| **targetFDR** | Float | 0.05 |
| The desired FDR cutoff  Require a float number between 0 and 1 | | |
| **nPerm** | Integer | 100 |
| Number of permutations for FDR estimation  Require an integer number between 100 and 1000 | | |
| **resolution** | Integer | 10 |
| The empirical resolution for the weights of linear integration function  Require an integer number between 2 and 20  All weights will be limited to the integer multiple of 1 / resolution  Example: if resolution = 10, then all weights will be chosen from [0, 0.1, 0.2, ..., 0.9, 1] | | |
| **FPmethod** | String | median |
| The method to estimate the expected false positive among permutations  Must be chosen from the following set:   1. median: Expected number of false positives is estimated as the median of observed false positives among all permutations 2. mean: the mean of observed false positives among all permutations 3. ceil\_mean: the ceiling of the mean 4. max\_median\_or\_mean: the median or the mean, whichever is larger 5. max\_median\_or\_ceil\_mean: the median or the ceiling of mean, whichever is larger | | |

\*Note: if “voom”, “limma” or “DESeq2” are specified in attList, DiscriminatCut will call the corresponding R package (limma or DESeq2 package) to calculate basic attributes. In such case, it is required to have either/both package installed on R and the absolute path to the R binary correctly specified in “Rpath.txt”, e.g., “C:\Program Files\R\R-3.1.2\bin\x64\Rscript.exe”.

## Interpret analysis result files

The *Discriminant-Cut* analysis tool exports two types of result files. One is the text report (.txt) and the other is the MATLAB data file (.mat).

The MATLAB data file can be viewed in MATLAB. The result of each individual method is packed in a MATLAB struct object. In each struct object, the features called significant are stored in an array called “feature”.

The text report file can be viewed in any text editor. At the beginning of the report, the parameters used in this analysis will be listed as follows:

*============================= Parameter settings ==========================*

*Dataset name: example.txt*

*Number of genomic features: 16173*

*Number of sample in group 1: 8*

*Number of sample in group 2: 8*

*Target FDR: 0.050000*

*Function resolution: 10*

*Number of permutation: 100*

*Sampling rate: 1.000000*

*False positive estimation method: median*

*Number of basic attributes: 2*

*Basic attributes: ranksum ttest*

After the parameter setting, the result summary will be shown. Start with up-regulated features, followed by down-regulated features and then the overall summary:

*====================== Result summary (Up-regulated) ======================*

*Pi0: 0.708836*

*Number of detected DE genomic features: 252*

*Discriminant value cutoff: 1.335473*

*Estimated FDR: 0.049225*

*Integration function: 0.600000 x ranksum + 0.400000 x ttest*

*##Note: Only trust the numbers below if the ground truths are reliable##*

*Number of true positives: 233*

*Number of false positives: 19*

*True FDR: 0.075397*

*===================== Result summary (Down-regulated) =====================*

*Pi0: 1.000000*

*Number of detected DE genomic features: 374*

*Discriminant value cutoff: -1.286507*

*Estimated FDR: 0.048128*

*Integration function: 1.000000 x ranksum*

*##Note: Only trust the numbers below if the ground truths are reliable##*

*Number of true positives: 368*

*Number of false positives: 6*

*True FDR: 0.016043*

*======================== Result summary (Overall) =========================*

*Number of detected DE genomic features: 626*

*Estimated FDR: 0.040198*

*##Note: Only trust the numbers below if the ground truths are reliable##*

*Number of true positives: 601*

*Number of false positives: 25*

*True FDR: 0.039936*

After that, the identified significant features of each method will be listed one after the other. Below is an example for *Discriminant-Cut*.

*============== Detected DE genomic features (Up-regulated) ================*

*Name Description Truth Discriminant value BasicAttribute#1 BasicAttribute#2*

*6553 SLC9A5 1 1.335473 0.997700 1.842132*

*9275 BCL7B 0 1.338700 0.997700 1.850200*

*116843 SLC18B1 0 1.341587 1.128976 1.660505*

*5660 PSAP 1 1.345181 0.840168 2.102701*

*9412 MED21 0 1.361701 1.102721 1.750171*

*…*

*============= Detected DE genomic features (Down-regulated) ===============*

*Name Description Truth Discriminant value BasicAttribute#1 BasicAttribute#2*

*100 ADA 1 -1.680336 -1.680336 -5.426287*

*10024 TROAP 1 -1.680336 -1.680336 -3.554450*

*10058 ABCB6 1 -1.680336 -1.680336 -5.614968*

*10105 PPIF 1 -1.680336 -1.680336 -5.382401*

*10112 KIF20A 1 -1.680336 -1.680336 -3.401188*

*…*