**BETA: Binding and Expression Target Analysis**

**Introduction**

Binding and Expression Target Analysis (BETA) is a software package that integrates

ChIP-seq of transcription factors or chromatin regulators with differential gene

expression data to infer direct target genes.

**Python Version**

Python 2.6 or above is recommended.

pkg\_resources should be installed first, you can try it first

python

>>> from pkg\_resources import resource\_filename

Type curl http://python-distribute.org/distribute\_setup.py | python to do the installation

Python Numpy package should be installed first

Python

>>> import numpy

To install numpy, see more from http://www.iram.fr/IRAMFR/GILDAS/doc/html/gildaspython-

html/node38.html

**R Version**

R 2.13 or above is recommended

**Installation**

1. Install package dependencies:

a. numpy v. 1.3.0 or above

b. pkg\_resources if you don’t have

c. R v. 2.13.1 or above

2. As sudo, type: $sudo python setup.py install

(If you want to install it for your own)

Step 1 is the same with the above

2. python setup.py install --prefix=<your path>

3. Modify PYTHONPATH if necessary

See more from <http://cistrome.org/BETA/#inst>

**Command Line**

**Help**

**BETA Basic** will do the factor function prediction and direct target detecting

$ BETA basic –p 3656\_peaks.bed –e AR\_diff\_expr.xls –k LIM –g hg19 --da500 –n basic --info 1,2,6

**BETA Plus** will do TF active and repressive function prediction, direct targets detecting and motif analysis in target regions

$ BETA plus –p 3656\_peaks.bed –e AR\_diff\_expr.xls –k LIM –g hg19 --gs hg19.fa –bl – info 1,2,6

**BETA Minus** detect TF target genes based on regulatory potential score only by binding data

$ BETA minus -p 3656\_peaks.bed --bl -g hg19

**Main Arguments (refer to the Input file formats described below)**

**-p PEAKFILE**, --peakfile=PEAKFILE

The bed format peaks binding sites. (At least 5 column, CHROM,

START, END, NAME, SCORE)

**-e EXPREFILE**, --diff\_expr=EXPREFILE

The differential expression file get from limma for MicroArray data and

cuffdiff for RNAseq data

**-k KIND**, --kind=KIND

The kind of your differential expression data, this is required, it can be

LIM(Limma output), CUF(Cuffdiff output), BSF(BETA Specific output),and O (Other software output)

**-g GENOME**, --genome=GENOME

Select the species of your data, it can be hg39, hg19, hg18, mm10 or mm9. Other species can give the genome reference file via –r reference. DEFAULT=False

**--gs=GENOMESEQUENCE**

Whole genome reference data with fasta format, can be downloaded form UCSC table browser

**-r REFERENCE**, --reference=REFERENCE

Annotation file which contain the refgene info file downloaded from

UCSC, 6 columns (REFSEQID, CHROMS, STRAND, TSS, TTS,

NAME2 (GENE SYMBOL))

**Options**

--version Show program’s version number and exit

-h, --help Show this help message and exit

--pn=PEAKNUMBER

The number of peaks you want to consider, DEFAULT=10000

--gname2

If this switch is on, gene or transcript IDs in files given through -e will be considered as official gene symbols, DEFAULT=FALSE

-n NAME, --name=NAME

This Argument is used to name the result file. If not set, the peakfile name will be used instead.

**--info EXPREINFO**

specify the geneID, up/down status and statistical values column of your expression data. NOTE: use a comma as an connector. for example: 1,2,6 means geneID in the 1st column, logFC in 2nd column and FDR in 6th column. DEFAULT:1,2,6 for LIMMA; 2,10,13 for Cuffdiff and 1,2,3 for BETA specific format. You'd better set

it based on your exact expression file, it is required when –k=O.

-o OUTPUT, --output=OUTPUT

The directory to store all the output files, if you don't set this, files will be output into the BETA\_OUTPUT directory

-d DISTANCE, --distance=DISTANCE

Set a number which unit is ‘base’. It will get peaks within this distance from gene TSS. DEFAULT=100000(100kb)

--bl Weather or not use CTCF boundary to filter peaks around a gene, DEFAULT=FALSE

--bf=BOUNDARYFILE

CTCF conserved peaks bed file, use this only when you set --bl and the genome is neither hg19 nor mm9

--pn=PEAKNUMBER

The number of peaks you want to consider, DEFAULT=10000

-b BOUNDARYFILE, --boundaryfile=BOUNDARYFILE

Bed file of conserved CTCF binding sites in this species. Peaks be filtered consider this boundary if you set it. DEFAULT=False

--df=DIFF\_FDR Input a number 0~1 as a threshold to pick out the most significant differential expressed genes by FDR, DEFAULT = 1, that is select all genes

--da=DIFF\_AMOUNT

Input a number between 0-1, so that the script will pick out the differentially expressed genes by the rank. Input a number bigger than 1, for example, 2000, so that the script will only consider top 2000 genes as the differentially expressed genes. DEFAULT = 0.5, that is select top 25% genes. NOTE: if you want to use diff\_fdr, please set this parameter to 1, otherwise it will get the intersection of these two parameters

-c CUTOFF, --cutoff=CUTOFF

Input a number between 0~1 as a threshold to select the closer target gene list (up regulate or down regulate or both) with the p value was called by one side KS-Test, DEFAULT = 0.001

**Example**

BETA -p 2723\_peaks.bed -e gene\_exp.diff -k CUF -g hg19 --gs

/mnt/Storage/data/hg19.fa

**Input Files Format**

BETA will check the input file format first, the basic description of some

input files format are as follows

**• Peak File: BED format**

5 columns with (Chrom     Start    End   Name  Score) information

chr11 2086891 209509 AR\_LNCaP\_2 51.58

chr11 3342461 335348 AR\_LNCaP\_7 54.55

chr12 1793512 180790 AR\_LNCaP\_9 257.72

Or 3 columns with (Chrom     Start    End) information

chr11 2086891 209509

chr11 3342461 335348

chr12 1793512 180790

\*\*\* Note: Please do not contain the header in the bed file, and make sure it is tab delimitated.

**• Differential Expression File**

BETA supports **LIMMA output** differential expression format directly, which contains (ID logFC AveExpre Tscore Pvalue adj.P.Value B) informration

**LIM format (–k LIM)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| NM\_001548\_at | -6.945783684 | 9.632803007 | -138.2402671 | 6.92E-10 | 2.08E-05 | 11.83285762 |
| NM\_005409\_at | 6.11280866 | 6.322508161 | -117.5664651 | 1.51E-09 | 2.08E-05 | 11.57790488 |
| NM\_001565\_at | -6.352395593 | 7.838465214 | -113.6000902 | -113.6000902 | 2.08E-05 | 11.51589687 |

• **Cuffdiff output contains** (Test\_id gene\_id gene locus sample\_1 sample\_2 status value\_1 value\_2 Log2(foldchange) test\_stat p\_value q\_value significant) information.

**CUF format (-k CUF)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| NM\_000014 | NM\_000014 | - | chr12:9217772-9268558 | q1 | q2 | NOTEST | 0.102845 | 0.0820513 | -0.325878 | 0.498271 | 0.618293 | 1 | no |
| NM\_000015 | NM\_000015 | - | chr8:18248754-18258723 | q1 | q2 | NOTEST | 0.127358 | 0.30975 | 1.28221 | -1.32328 | 0.185744 | 1 | no |
| NM\_000016 | NM\_000016 | - | chr1:76190042-76229355 | q1 | q2 | NOTEST | 0 | 0 | 0 | 0 | 1 | 1 | no |
| NM\_000017 | NM\_000017 | - | chr12:121163570-121177811 | q1 | q2 | NOTEST | 3.47702 | 3.62422 | 0.0598207 | -0.195815 | 0.844755 | 1 | no |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |

• **BETA Specific format contains** (GeneID, Regulatory status (value with + or -), statistical value(e.g. FDR or Pvalue, the smaller value, the more significant it is)

) information.

**BSF format (-k BSF)**

|  |  |  |
| --- | --- | --- |
| NM\_000014 | -0.325878 | 0.618293 |
| NM\_000015 | 1.28221 | 0.185744 |
| NM\_000016 | 0 | 1 |
| NM\_000017 | 0.0598207 | 0.844755 |

• **Other format (-O, should contain the information described in BSF format, and –info is required)**

\*\*\* The differential expression file should contain all the genes in the genome, BETA will use all the info to get the static genes, and isolate the up regulated genes and down regulated genes based on the threshold you input.

\*\*\* Make sure your differential expression file do not have the header or add the ‘#’ in the front of your header line.

\*\*\* If your gene ID is the official gene symbol, please add the parameter --gname2

\*\*\* Although you can select the type of your differential expression format, in case to make sure BETA get the correct information, you would better set the columns information via --info except you have the same format with the above example.

See more from --info

**• boundary file (--bf)**: BED format(at least 3 columns)

chr1 521336 521779 3 0.986 +

chr1 839881 840447 19 0.986 +

chr1 919474 919976 36 1.0 +

chr1 968212 968748 48 0.986 +

**• Genome annotation (-r):** Downloaded from UCSC

BETA provides hg38, hg19, hg18, mm10, and mm9 annotation.

The annotation reference file should contain (refseqID chroms strand txstart txend genesymbol) information in order.

NM\_032291 chr1 + 66999824 67210768 SGIP1

NM\_001301823 chr1 + 33546729 33586132 AZIN2

NM\_013943 chr1 + 25071759 25170815 CLIC4

NM\_032785 chr1 - 48998526 50489626 AGBL4

**• Whole genome sequence data**: fasta format

The format is like:

>chr1: xxxx-yyyyy

ATCGGGACTTGACCC…

>chr2: xxxx-yyyyy

AGCGTGACTAGAGCC…

…

**Output Files**

• test.pdf A PDF figure to test the TF's funtion, Up or Down regulation.

• test.r The R script to draw the score.pdf figure

• uptarget.txt The uptarget genes, 7 columns, chroms, txStart, txEnd, refseqID, rank, product, Strands, GeneSymbol

• downtarget.txt The downtarget genes, the same format to uptargets

• Uptarget\_associated\_peaks.txt The peaks associated with up target genes

• Downtarget\_associated\_peaks.txt The peaks associated with down target genes

• Mitifresult (directory contain all the motif results)

o UP\_MOTIFS.txt

o UP\_NON\_MOTIFS.txt

o DOWN\_MOTIFS.txt

o DOWN\_NON\_MOTIFS.txt

o UPVSDOWN\_MOTIFS.txt

o betamotif.html

\*\*\* NOTE: Up or Down target file depends on the test result in the PDF file, it will be not produced unless it passed the threshold you set via -c –cutoff