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/gildas-python-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

2. python setup.py install --prefix=<your path>
3. Modify PYTHONPATH if necessary

See more from http://cistrome.org/BETA/#nst

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

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

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

-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 M or R.M for MicroArray, R for RNA-Seq

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

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

--gs=*GENOMESEQUENCE*

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. DEFAULT:2,5,7 for LIMMA; 2,10,13 for Cuffdiff and 1,2,3 for BETA specific format

-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

--pt=*PERMUTETIMES*

Permutation times. Give a reasonable value to get an exact FDR. Gene number and permute times decide the time it will take. DEFAULT=500

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 (3 or 5 colums)

  chroms start end name score [strand]

  If your bed don't have the name and score column, please fake one.

- Differential Expression File by Microarray: Result of Limma

  ID Refseq logFC AveExpre Tscore Pvalue adj.P.Value B

- Differential Expression File by RNAseq: Result of Cuffdiff

  Test_id gene_id gene locus sample_1 sample_2 status value_1 value_2 Log2(foldchange) test_stat p_value q_value significant

- CTCF conserved boundary file: BED format(at least 3 columns)

  chroms start end [name] [score] [strand]

  The conserved CTCF binding sites of all the cell lines in this species.

- Genome reference: Downloaded from UCSC

  refseqID chroms strand txstart txend genesymbol.

  We use that as a reference to get the gene information.

- Genome sequence data: The whole genome sequence data, fasta format

  The format is like: >chr1: xxxx-yyyyy

  ATCGGGACTTGACCC

!!! Make your input binding and expression file with header starts with'#' or just delete the header of it.

Output Files

- score.pdf A PDF figure to test the TF's funtion, Up or Down regulation.
- score.r The R script to draw the score.pdf figure
- uptarget.txt The uptarget genes, 4 column, Refseq, Gene Symbol, Rank Product, FDR
- 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)
  - UP_MOTIFS.txt
  - UP_NON_MOTIFS.txt
  - DOWN_MOTIFS.txt
  - DOWN_NON_MOTIFS.txt
  - UPVSDOWN_MOTIFS.txt
  - 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

*** See more on our tutorial online.