Published in final edited form as:
Nat Protoc. 2013 December ; 8(12): 2502–2515. doi:10.1038/nprot.2013.150.

Target analysis by integration of transcriptome and ChIP-seq data with BETA

Su Wang[1], **Hanfei Sun**[1], **Jian Ma**[1], **Chongzhi Zang**[2], **Chenfei Wang**[1], **Juan Wang**[1], **Qianzi Tang**[1], **Clifford A Meyer**[2], **Yong Zhang**[1], and **X Shirley Liu**[2]

[1]Department of Bioinformatics, School of Life Science and Technology, Tongji University, Shanghai, China

[2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, Massachusetts, USA

Abstract

The combination of ChIP-seq and transcriptome analysis is a compelling approach to unravel the regulation of gene expression. Several recently published methods combine transcription factor (TF) binding and gene expression for target prediction, but few of them provide an efficient software package for the community. Binding and expression target analysis (BETA) is a software package that integrates ChIP-seq of TFs or chromatin regulators with differential gene expression data to infer direct target genes. BETA has three functions: (i) to predict whether the factor has activating or repressive function; (ii) to infer the factor's target genes; and (iii) to identify the motif of the factor and its collaborators, which might modulate the factor's activating or repressive function. Here we describe the implementation and features of BETA to demonstrate its application to several data sets. BETA requires ~1 GB of RAM, and the procedure takes 20 min to complete. BETA is available open source at http://cistrome.org/BETA/.

INTRODUCTION

Gene expression is regulated through multiple mechanisms, two of which include the binding of TFs and chromatin regulators. TFs bind DNA and interact with transcriptional machinery to activate or repress the expression of target genes. In contrast, chromatin regulators bind to or catalyze histone modifications to affect chromatin structure and function. *In vivo* binding of both TFs and chromatin regulators (hereafter referred to collectively as factors) can be discovered by chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). In addition, the influence of factor binding on gene expression can be investigated by using transcriptome data obtained from conditions that contrast between the bound and unbound states.

Correspondence should be addressed to Y.Z. (yzhang@tongji.edu.cn) and X.S.L. (xsliu@jimmy.harvard.edu).

AUTHOR CONTRIBUTIONS S.W., C.A.M., Q.T. and X.S.L. designed the method; S.W., H.S., J.M., C.W., C.Z., J.W. and X.S.L. implemented the algorithm; S.W. performed the data analysis; and S.W. and X.S.L. wrote the initial manuscript. All authors contributed to the discussion and writing of the final manuscript.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

However, in mammalian experimental systems, the concordance between gene expression changes and TF binding is often difficult to interpret. First, factor-binding sites and target genes usually lack a one-to-one relationship. The same factor could bind anywhere between the proximal promoter to hundreds of kilobases downstream to regulate gene expression. Alternatively, the same binding site could regulate multiple genes by interacting with different promoters in different subpopulations of cells. Second, not all factor-binding sites found in a ChIP-seq experiment are functional, potentially owing to the lack of collaborating factors or conditions favorable to their function. Finally, the binding of one factor may cause secondary effects owing to transcriptional changes of its direct targets. Addressing these issues requires making general working assumptions about gene regulation combined with robust statistical analyses on available ChIP-seq and transcriptome data. Although several target gene prediction methods have been published, few of these provide a user-friendly algorithm package for target gene detection. The GREAT target analysis tool provides several *ad hoc* options for designating target genes, which it subsequently analyzes for annotation enrichment[1]. Databases such as TRED[2] provide target genes for a selection of factors on the basis of motif analysis or public ChIP-seq data, but they cannot infer targets specific to user-defined factors or conditions[3,4].

Development of the protocol

We developed BETA as an integrated software package for analyzing factor binding and differential expression in mammalian genomes. It is available open source at http://cistrome.org/BETA, and it can be run as a web tool directly from http://cistrome.org/ap/. The program has three main functions: (i) to predict whether a factor has activating or repressive function; (ii) to infer the factor's target genes; and (iii) to identify the binding motif of the factor and its collaborators, which might modulate the factor's activating or repressive function. Figure 1 illustrates the main operational stages of BETA. Instead of assigning one-to-one mapping between binding sites and genes, BETA models the influence of a binding site on the expression of a gene with a monotonically decreasing function that is based on the distance between the binding site and transcription start site. The regulatory potential of a gene is scored as the sum of the contribution of individual sites[5]. However, genes with promoters in a repressive chromatin environment, or those lacking prerequisite collaborating factors may not respond to factor binding despite a high regulatory potential. In these cases, gene expression changes associated with factor binding can give better confidence that a gene is a direct target. To take this into account, BETA ranks genes on the basis of both regulatory potential of factor binding and differential expression upon factor binding, and then it calculates the rank product[6] of the two to predict direct targets. To determine whether a factor has overall activating and/or repressive functions, a nonparametric statistical test contrasts regulatory potentials for genes that are differentially expressed with genes that are statically expressed in the factor perturbation experiment. The activating or repressive functions of factors are often modulated by other collaborating factors, some of which also directly bind DNA. BETA conducts sequence motif analysis on binding sites near upregulated or downregulated targets to identify putative collaborating factors.

BETA contains three subprotocols: BETA-basic, BETA-plus and BETA-minus. BETA-basic can be used to predict whether a factor has activating or repressive function and detect direct target genes. BETA-plus can be used to predict whether a factor has activating or repressive function, whether it can detect direct target genes and whether it can analyze sequence motifs in target regions. Both binding and differential expression data are required for BETA-basic and BETA-plus, whereas BETA-minus is used when only binding data are available to predict target genes.

Application of BETA

The purpose of this protocol is to predict genes that are the direct targets of TFs or chromatin regulators. Once the factor's set of target genes is known, further analysis can be done by using gene ontology–based tools such as DAVID[7] to link functions to this set. The motif analysis function of BETA identifies motifs that are associated with candidate *cis*-regulatory regions relative to regions that are nonregulatory, which enables the identification of potential cofactors.

Comparison with other methods

Various methods and software process ChIP-seq data and analyze TF target genes with different strategies. Model-based analysis of ChIP-seq (MACS)[8], CisGenome[9] and SICER[10] are some of the peak caller tools globally used to identify precise TF-binding sites. With ChIP-seq pre-processed data, a simple peak-based way to identify targets is to assign the proximal nearest gene or the gene containing peaks in its promoter region[11,12]. With most TF ChIP-seq data, only a small percentage of binding is found at the promoters, and the use of nearest peaks to assign target genes is very unreliable. TIP[13], which builds a probabilistic model to identify the target genes by TF-binding profiles, does not consider gene expression data. In contrast, some earlier studies predict the targets on the basis of gene expression information only. Qian *et al.*[14] predict the target genes by identifying the relationship of gene expression with support vector machines (SVM); Honkela *et al.*[15] do that with time-series expression data by creating a linear activation model based on Gaussian process; and Redestig *et al.*[16] developed the CERMT algorithm by using multiple short expression time-series data with several treatments, defining the TF target candidates as the genes with similar responses to the TF in these treatments. The false-positive rate is reduced when using integrated binding and expression information to identify target genes compared to solely using either. We compared some simple methods in our previous study[5].

Many ChIP-seq experiments, together with expression profile experiments, are performed in conditions in which the expression of the factor of interest is perturbed (knockdown or overexpression). The expression changes of all other genes, especially for the regulatory targets of the factor, are considered as being the result of the perturbation. Several methods and web servers were released recently for TF target gene prediction by using the integrative analysis of binding and expression data. ChIP-Array[17] is a web server that identifies the direct targets by simply marking the genes that are both differentially expressed and binding-enriched as targets, but all binding peaks and the expression changes of each gene receive equal weight. EMBER[18], developed by the Dinner group, integrates the binding and expression data by an unsupervised machine learning with an expectation maximization

algorithm to detect the potential targets. It gives each gene an expression behavior but ignores the distance between binding sites, and it considers all genes 100 kb within the binding peaks as potential targets. BETA uses a distance-weighted measure to gauge the regulatory potential of all the binding sites of the factor within a certain distance to a target gene. In addition, when a factor's expression is perturbed, affected genes often include both upregulated and downregulated genes—one group might represent directly affected genes, whereas the other group might represent indirectly affected genes or squelching. BETA integrates differential expression with binding to evaluate whether the direct effect of factor binding is an activating or repressing expression, and it assigns direct up- or down-targets. ChIP-seq often yields tens of thousands of peaks in a single experiment, but only a few hundred direct targets. Because sometimes co-regulation of TFs is key to influencing gene expression at certain conditions, the differential motif finding function (e.g., peaks near differential genes versus peaks near nondifferential genes) provided by BETA will help find the correct co-regulators. Finally, some factors can both directly activate and repress gene expression, and it is often through interaction with different partners that their specific effects are determined; therefore, differential motif analysis comparing the peak associated with the up-target and down-target regions is also important.

Experimental design

To illustrate how this protocol works and to interpret its results, we use androgen receptor (*AR*) ChIP-chip data obtained in LNCaP cells in combination with microarray data of gene expression after 16 h of dihydrotestosterone (DHT) treatment. AR is a member of the nuclear receptor family, and it has a key role in gene regulation in normal prostate and in prostate cancer. Detection of AR target genes is important to understand its regulation and function. Here, we show how to use BETA to analyze AR regulation by integrating binding and transcriptome information. The details of this case study will be discussed at the ANTICIPATED RESULTS section. To show other features of BETA, we also provide examples involving the genes encoding human estrogen receptor-α (*ESR1*) and the mouse tet methylcytosine dioxygenase 1 (Tet1).

Activating/repressive function prediction—ChIP-seq data are often examined in the context of gene expression, and thus expression profiles are available for both the factor-bound and factor-unbound conditions. We used LIMMA[19] and Cuffdiff[20] to obtain differentially expressed genes from microarray or RNA-seq experiments, respectively (BETA also accepts expression data with other applied algorithms). We determined the list of differentially expressed genes in the DHT-induced AR system by using LIMMA[19], and we show top lines here as follows:

```
ID logFC AveExpr t P.Value adj.P.Val B
NR_045762_at 3.16711734 9.140369116 35.91057535 6.99e-11 4.18e-07 14.13456018
NM_001002231_at 3.214550493 9.169929883 35.32505807 8.07e-11 4.18e-07
14.05227211
NM_001256080_at 3.214550493 9.169929883 35.32505807 8.07e-11 4.18e-07
14.05227211
```

```
NM_005551_at 3.214550493 9.169929883 35.32505807 8.07e-11 4.18e-07
14.05227211
```

We divided genes into three groups according to their expression pattern, upregulated, downregulated or unchanged, after DHT treatment, and labeled them as UP, DOWN or NON in the workflow, respectively. BETA users can specify the number of genes in each group by count or by specific statistical measures such as false discovery rate (FDR) via the parameters --da and --df, respectively, in BETA basic and BETA plus.

For ChIP-seq data, we use Bowtie[21] to map sequencing reads to the reference genome and MACS[22] for peak calling. We use the peak-calling program model-based analysis of tiling array (MAT)[23] to identify the binding events from ChIP-chip data. Four lines of the output from this analysis of AR ChIP-chip data from Wang *et al.*[24] are as follows:

```
chr1 1208689 1209509 AR_LNCaP_2 51.58
chr1 1334246 1335348 AR_LNCaP_7 54.55
chr1 2179351 2180790 AR_LNCaP_9 257.72
chr1 2341577 2342737 AR_LNCaP_11 199.59
```

The regulatory potential, which is a gene's likelihood of being regulated by a factor, is estimated for each gene. The regulatory potential is calculated as $S_g = \sum_{i=1}^{k} e^{-(0.5+4\Delta_i)}$ (ref. 5). All binding sites (*k*) near the transcription start site of the gene (*g*) within a user specified range (100 kb as default) are considered. Δ is the exact distance between a binding site and the TSS proportional to 100 kb (Δ = 0.1 means the exact distance = 10 kb). BETA users can also specify the top number of binding sites by count or by specific statistical measures. BETA then generates a cumulative distribution function of the gene groups and uses a one-tailed Kolmogorov-Smirnov test[25] to determine whether the UP and DOWN groups differ significantly from the NON group. As shown in Figure 2a, the dotted line represents the background, the genes that are not differentially expressed, whereas the red and the blue lines represent the genes upregulated and downregulated, respectively. BETA sorts genes by the regulatory potential score from high to low. The *y* axis of Figure 2a represents the proportion of genes in a category that are ranked at or better than the *x*-axis value, which represents the rank on the basis of the regulatory potential score from high to low. The *P* value listed in the top left represents the significance of the UP or DOWN group relative to the NON group as determined by the Kolmogorov-Smirnov test. From the AR activating/repressive function prediction result, it is clear that the UP-regulated genes have a much higher regulatory potential score than the DOWN-regulated and the nonregulated genes. That is to say, the genes with a gain in gene expression after 16 h of DHT treatment tend to also have an enrichment of AR-binding sites.

Direct target prediction—BETA predicts factor target genes by combining the binding potential from ChIP-seq data with differential expression data. Each gene is assigned two ranks: one based on binding potential R_{gb} and one based on differential expression R_{ge}. Direct targets are then assigned on the basis of the rank product[6] of the two, and those with

more nearby binding and more differential expression are more likely to be called as real targets.

There is evidence that binding of the transcription factor CCCCTC-binding factor (CTCF) can form regulatory boundaries and that the expression patterns of genes within such regulatory blocks correlate better than genes in different blocks[26]. BETA provides a conserved CTCF boundary file that integrates all available Encyclopedia of Data Elements (ENCODE) data for humans (hg19 assembly) and mice (mm9). However, this block will have a big contribution only when the certain range is set via the parameter -d—when we use 100 kb (as default) or smaller distances, there is little difference.

BETA provides the target prediction file in a user-friendly format; the first six columns are in the standard BED format. In addition to refseq gene IDs, BETA also provides official gene symbols. With these results, users can easily perform further downstream analysis with GREAT[1] or DAVID[7]. This is an example of the top lines of that output for direct target gene prediction:

```
Chroms txStart txEnd refseqID rank product Strands GeneSymbol
chr19 51376688 51383823 NM_001256080 2.186e-07 + KLK2
chr19 51376688 51383823 NM_005551 2.186e-07 + KLK2
chr19 51376688 51383823 NR_045762 2.186e-07 + KLK2
chr19 51376688 51383823 NR_045763 2.186e-07 + KLK2
chr19 51376688 51383823 NM_001002231 2.186e-07 + KLK2
chr1 207191865 207206101 NM_023938 8.822e-07 - C1orf116
chr1 207191865 207206101 NM_001083924 8.822e-07 - C1orf116
chr21 42836477 42880085 NM_005656 1.033e-06 - TMPRSS2
chr21 42836477 42879992 NM_001135099 1.041e-06 - TMPRSS2
```

Suppose there are n genes (both differentially expressed and with regulatory potential >0, which means at least one binding event around it within the range defined by the parameter -d). Two ranks (R) are associated with each gene (g): one is based on decreasing regulatory potential (R_{gb}), such that is $R_{gb} = 1$ for the gene with the largest regulatory potential score, and the other is based on the increasing of the FDR or P value (R_{ge}), that is, $R_{ge} = 1$ for the most strongly differentially expressed gene. Then the rank product of the gene (g), $RP_g = (R_{gb}/n)*(R_{ge}/n)$. The RP can be interpreted as a P value[6], because it shows the probability that this gene has a regulatory potential rank $\leq R_{gb}$ and a differential expression changed rank $\leq R_{ge}$. On the basis of RP, users can judge the targets with a certain cutoff, for example, genes with an RP less than 10^{-3} will be more likely to be the true target genes of AR. Kallikrein-related peptidase 2 (*KLK2*), a gene that is highly expressed in prostate cancer, has been reported to be regulated by AR[27]. As a prognostic marker of prostate cancer, the function of *KLK2* is still unknown. As expected, BETA found *KLK2* to be upregulated by AR with the most significant RP value (RP = 2.186e–07). We also found chromosome 1 open reading frame 116 (*Clorf116*; also known as *SARG*), and transmembrane protease serine 2 (*TMPRSS2*) to have significant RP values (RP = 8.822e–07 and RP = 1.033e–06, separately) for upregulation by AR. In previous studies, it was already been proven that

Clorf116 is an AR-upregulated target gene[28] and that the *TMPRSS2*–ETS-related gene (*ERG*) fusion gene is upregulated by AR and is present in prostate cancer with a high frequency[29]. BETA outputs hundreds of target genes, which is helpful in understanding the function and regulation of factors and provides some genes with highly potential values.

Except for direct targets, BETA provides the target gene– associated peaks as well, which can be easily used for motif searching or meta-profiling. Here we show the associated peaks of AR-upregulated target genes as follows:

```
chrom pStart pEnd Refseq Symbol Distance Score
chr19 51354060 51354999 NM_001256080 KLK2 -22159 0.249983590819
chr19 51372841 51373704 NM_001256080 KLK2 -3416 0.529067106385
chr19 51392207 51393248 NM_001256080 KLK2 16039 0.319320493096
chr19 51354060 51354999 NM_005551 KLK2 -22159 0.249983590819
chr19 51372841 51373704 NM_005551 KLK2 -3416 0.529067106385
chr19 51392207 51393248 NM_005551 KLK2 16039 0.319320493096
chr19 51354060 51354999 NR_045762 KLK2 -22159 0.249983590819
```

The first three columns are the basic information about the peaks, the 4th and 5th columns are the target genes' RefSeq ID and gene symbol, and the 6th column is the distance from the peak center to the gene transcription start site (TSS) (a positive value represents the peak downstream of the gene and negative one is upstream). The regulatory potential score in the last column represents the contribution of each peak (the nearer the peak is, the higher the score). The user can upload the first three columns to IGV[30] or to the University of California Santa Cruz (UCSC) genome browser[31] to visualize the relationship between target genes and associated peaks.

If ChIP-seq data do not have corresponding differential expression data, users can apply BETA-minus, a simpler method that defines the targets as the genes with a high regulatory potential, derived only from TF binding within a specific region (100 kb default). In such cases, BETA is not able to predict whether the factor binding is activating or repressing gene expression.

Binding motif analysis—To identify factor-binding motifs associated with ChIP-seq and differential expression data, BETA conducts motif analysis on sites proximal to the targets. It calls the function 'model-based interval scanner with PSSM' (MISP) to search for enriched sequence motifs represented as position-specific scoring matrices (PSSM). MISP adopts the algorithm proposed in MOODS[32], which can scan hundreds of matrices to chromosome-sized sequences in a few seconds. BETA then compares the number of motifs near the ChIP-seq binding summits with that in flanking regions to detect motifs with marked summit enrichment (Fig. 2b). Requiring summit enrichment improves the specificity of reported motifs, and this approach has been adopted in the SeqPos algorithm in the Cistrome analysis pipeline[33].

A similar analysis could also be conducted in peaks near the UP and DOWN targets to identify differential motifs enriched over the NON targets or over each other. These differential motifs could provide important insight into collaborating factors or novel regulatory mechanisms. As AR has an activating function only in LNCaP cells, the motif analysis focuses only on the upregulated target gene regions. Our analysis summarizes all significant motif results in a web page, part of which is shown in Figure 2c; additional details from the original results can be found in text files in BETA deposited results folder, which can be defined by the BETA parameter -o. To remove redundant motifs from this summary, BETA classifies motifs into groups on the basis of similarity scores from Habib's method[34]. Other information, including the motif ID (our database ID), the official symbol of the factor and DNA-binding domains (integrated from the TFCat database[35]), detail each motif. *t* scores and corresponding *P* values measure the significance of motif enrichment. From the motif analysis on AR target regions (Fig. 2c), nuclear receptor subfamily 3, group C, member 1 (NR3C1) was detected as the strongest motif and grouped with NR3C2, AR and progesterone receptor (PGR). Forkhead box AI (FOXA1), as a pioneer factor of AR, was detected with a significant *P* value and summarized with other forkhead domain family members (Fig. 2c).

Limitations of BETA

BETA requires at least ChIP-seq data to identify putative target genes. For combined expression data analysis, BETA currently supports standard LIMMA and Cuffdiff outputs or a tab-delimited differential expression text file with BETA required information (see examples at http://cistrome.dfci.harvard.edu/BETA/). BETA recognizes both Refseq IDs and official gene symbols in differential expression data. Other types of gene identification should be converted to these formats before analysis with conversion tools such as the DAVID gene ID conversion tool[36].

MATERIALS

REAGENTS

- Data sets formatted as described below: factor-binding data and differential gene expression data

EQUIPMENT

- Computer: any computer running a Unix-like system with at least 2 GB of RAM can be used. A 64-bit machine running either Linux or Mac OS X (10.6 or later) with 4 GB or more of RAM is preferred
- Software: downloaded and installed as described below: Python2.6 or newer and the NumPy Python package; R 2.13.1 or newer; GNU Compiler Collection (GCC)

PROCEDURE

1| (Optional) *Install git*. Git-clone is widely used to make a copy of a project. To copy the Python NumPy module (Step 2), install git first. Follow option A for Mac OS X users, option B for Linux Ubuntu users, option C for Linux Fedora

users or option D for other Linux distribution users. If git is already installed, skip this step.

A. For Mac OS X users

i. Download the dmg file from http://code.google.com/p/git-osx-installer, and then double click the file to install it.

B. For LINUX Ubuntu users

i. Type the following code: $ apt-get install git

C. For LINUX Fedora users

i. Type the following code: $ yum install git

D. For other LINUX distribution users

i. Refer to http://git-scm.com/download/linux

?TROUBLESHOOTING

2| (Optional) *Install python module NumPy*. Linux and Mac OS X usually have Python built in. Ensure that the version is newer than Python 2.6. Next, install NumPy for multidimensional array analysis. If Python already has NumPy installed, skip this step.

```
$ git clone git://github.com/numpy/numpy.git numpy
$ cd numpy
$ python setup.py build
$ python setup.py install
```

? TROUBLESHOOTING

3| *Verify the installation of Numpy*. To verify the installation of NumPy, launch Python and type the following:

```
>>> import numpy
>>> numpy
>>> numpy.__version__
```

The NumPy installed directory and version information will be displayed in the screen if it is installed successfully. For example:

```
>>> numpy
>>> <module 'numpy' from
'/Library/Frameworks/Python.framework/Versions/2.7/lib/python
2.7/site-packages/numpy/__init__.pyc' >
```

```
>>> numpy.__version__
>>> 1.8.0.dev-ccbf5cf
```

? TROUBLESHOOTING

4| *Install R.* R is free software that is widely used for statistical computing and graphics. If R is already installed, ensure that the version is newer than R 2.13.1. Otherwise, perform option A for Linux and option B for Mac OS X to install R.

A. For Linux users

i. Type the following:

```
$ sudo apt-get update
$ sudo apt-get install r-base-core
```

B. For Mac OS X users

i. Choose the CRAN mirrors at http://www.r-project.org/.

ii. Download R for Mac OS X and choose the latest version http://cran.cnr.berkeley.edu/.

iii. Install R by double-clicking the package (e.g., R-3.0.1.pkg).

? TROUBLESHOOTING

5| *Download BETA.* Download the BETA source code from http://cistrome.dfci.harvard.edu/BETA/,

```
$ cd BETA
```

? TROUBLESHOOTING

6| *Install BETA.* Follow option A for global installation or option B for local installation.

A. For global installation, if the user is a root or an administrator of the machine

i. Type the following:

```
$ sudo python setup.py install
```

B. For local installation

i. Type the following:

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

In this case, you should modify PYTHONPATH by adding the following two lines to the .bashrc file in the home directory if necessary:

```
> export PATH=/your_directory/bin:$PATH
> export PYTHONPATH=/your_directory/lib/
python2.X/site- packages/:$PYTHONPATH
```

▲ **CRITICAL STEP** Do not install BETA in the source code directory.

? TROUBLESHOOTING

7| *Download the reference genome sequence data*. BETA requires reference genome data to perform motif analysis. Currently, BETA only supports reference genome data in FASTA format. It can be downloaded from the UCSC Genome Bioinformatics site at http://genome.ucsc.edu/ (ref. 37).

▲ **CRITICAL STEP** Ensure that the chromosome identification is ‘chr1’, ‘chr2’… instead of ‘chrI’, ‘chrII’.

8| *Format factor-binding data sets for BETA analysis*. Factor-binding data files should be in BED format. BETA only supports three-column (chrom, chromStart, chromEnd) or five-column (chrom, chromStart, chromEnd, name, score) BED files. To get the BED-format binding events, perform option A for ChIP-seq data sets or option B for ChIP-chip data sets.

A. For ChIP-seq data sets

i. Align the data to a reference sequence with an alignment tool, such as Bowtie[21].

ii. Analyze the alignment results with a peak-calling program, such as MACS[8].

B. For ChIP-chip data sets

i. Analyze TF-binding events with a peak-calling program for tiling arrays, such as MAT[23].

?TROUBLESHOOTING

9| *Format differential gene expression data sets for BETA analysis.* Differential expression data files should be tab-delimited text files from LIMMA (LIM), Cuffdiff (CUF), BETA-specific format (BSF) or other file types (O) with prescribed expression information (Box 1). Raw expression data should have both control and experimental conditions. To get the eligible differential gene expression data, perform option A for LIM format from microarray data; use option B for CUF format from RNA-seq data; use option C for BSF; or use option D for other types.

A. For LIM format from microarray data

i. Ensure that the experiment has at least two replicates.

ii. Download the custom CDF file[38] from BRAINARRAY (http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/CDF_download.asp).

iii. Run LIMMA[19] with R.

B. For CUF format from RNA-seq data

i. Align RNA-seq reads to the whole genome with TopHat[20].

ii. Obtain differentially expressed genes by using Cuffdiff[20] in the Cufflinks package.

C. For BSF

i. Convert to a tab-delimited text file.

ii. Ensure that the file consists of three columns: gene ID, expression change and *P* value or other statistical significance.

D. For other file types

i. Convert to a tab-delimited text file.

ii. Ensure that the file has three columns: gene ID, expression change and *P* value or other statistical significance. Specify the column number via --info.

▲ **CRITICAL STEP** All genes should be identified with Refseq IDs or official gene symbols.

? TROUBLESHOOTING

10| *Parameter selection for example data sets*. Example data sets (described in Table 1 and available at http://cistrome.org/BETA/#download) are used here to illustrate BETA analysis. Each data set is analyzed with BETA-basic, BETA-plus and BETA-minus. For data input, basic commands indicate the following parameters (additional, optional BETA parameters can be found in Box 2): -p specifies the name of factor-binding data; -e specifies the name of the corresponding differential expression data; -k specifies the format of the differential expression data—LIM, CUF, BSF or O; -d specifies a distance (in bp) within which peaks will be considered, default = 100,000 (100 kb); -g specifies the reference genome: hg19 for human or mm9 for mouse; for other genomes, see parameter -r in Box 3; -n specifies the prefix of the output files; and --da limits analysis to a specific number of differentially expressed genes in either direction (up and down).

▲ **CRITICAL STEP** Input of LIM, CUF, BSF or O under -k depends on the data set; ensure that the input file format is BETA supported.

? TROUBLESHOOTING

11| *BETA analysis of example data sets*. We use AR in LNCaP, Tet1 in mouse embryonic stem (ES) cells and ESR1 in MCF-7 cells to illustrate how the different BETA subprotocols work for different data sets. Perform option A for BETA-basic, option B for BETA-plus or option C for BETA-minus.

A. BETA-basic: TF activating and repressive function prediction and direct target detection

i. Predict *AR* function and direct targets in LNCap cells with 16 h of DHT treatment:

```
$ BETA basic –p 3656_peaks.bed –e
AR_diff_expr.xls –k LIM –g hg19 --da500 -o basic
```

ii. The screen output lists all arguments used in this procedure, reports the input file format checking status and shows warnings and progress. In the end, BETA also reports the total time of the procedure. An example of this screen information is shown:

```
[16:52:07] Argument List:
[16:52:07] Name = basic
[16:52:07] Peak File = 3656_peaks.bed
[16:52:07] Top Peaks Number = 10000
[16:52:07] Distance = 100000 bp
[16:52:07] Genome = hg19
[16:52:07] Expression File = AR_diff_expr.xls
[16:52:07] Expression Type = MicroArray, LIMMA
result
[16:52:07] Number of differential expressed
genes = 500.0
[16:52:07] Differential expressed gene FDR
Threshold = 1
[16:52:07] Up/Down Prediction Cutoff = 0.001000
[16:52:07] Check 3656_peaks.bed successfully!
[16:52:07] limma output file format successful
passed
[16:52:07] You do not like filter peak by CFCT
boundary, it will be filtered only
by the distance
[16:52:07] Read file <3656_peaks.bed> OK! All
```

```
<7059> peaks.
[16:52:20] Process <41168> genes
[16:52:30] Finished! result output to
<basic.txt>
[16:52:32] Prepare file for the Up/Down Test
null device
1
[16:52:40] Finished, Find the result in
NA_score.pdf
[16:52:40] Get the Rank Product of the
“upregulate” genes
[ ‘“upregulate”’ ]
[16:53:36] pick out the peaks 100000 bp around
the selected genes
[16:53:37] Finished: Find target gene
associated peaks in basic
total time: 0:1:29
```

iii. Predict Tet1 function and direct targets in mouse ES cells. Input Tet1 BED format peak file via -p and input the differential expression file via -e; set the genome assembly mm9 via -g:

```
$ BETA basic –p ../5795_peaks.bed -
e ../Tet1_diff_expr.xls -k LIM -g mm9 -o basic
--
da 500
```

iv. Predict ESR1 function and direct targets in MCF-7 cells with 12 h of E2 treatment:

```
$ BETA basic -p ../349_peaks.bed –
e ../ESR1_diff_expr.xls -k LIM -g hg19 -o basic
--
da 500
```

B. BETA-plus: TF activating and repressive function prediction, direct target detection and motif analysis

i. BETA-plus uses the same parameters as BETA-basic and additional parameters: --gs is required for motif scanning and it specifies a FASTA format reference genome; --bl is an optional parameter that is on when boundaries are considered. In these commands, the parameter –da was not set, meaning that the default top 50% of upregulated

genes and top 50% of downregulated genes were chosen as differentially expressed. This parameter can be specified depending on the data set. We turn --bl on and use the CTCF boundaries provided by BETA to ensure that the gene and associated peaks are in one CTCF block. This boundary file allows other data with at least three-column BED format, and it can be set via the parameters in Box 3.

ii. Integration analysis (function prediction, target detection and motif analysis) of AR in LNCap cells with 16 h of DHT treatment:

```
$ BETA plus –p 3656_peaks.bed –e
AR_diff_expr.xls
–k LIM –g hg19 --gs hg19.fa --bl
```

iii. Integration analysis of Tet1 in mouse embryonic stem cells:

```
$ BETA plus –p ../5795_peaks.bed –
e ../Tet1_diff_expr.xls -k LIM -g mm9 --gs
mm9.fa ––bl
```

iv. Integration analysis of ESR1 in MCF-7 cells with 12 h of E2 treatment:

```
$ BETA plus -p ../349_peaks.bed -
e ../ESR1_diff_expr.xls -k LIM -g hg19 --gs
hg19.fa ––bl
```

C. BETA-minus: target prediction based solely on binding events

▲ **CRITICAL STEP** BETA-minus requires only parameters –p and –g.

i. Regulatory potential–based target prediction for AR in LNCap cells with 16 h of DHT treatment:

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

ii. Regulatory potential–based target prediction for Tet1 in mouse ES cells:

```
$ BETA minus -p 5795_peaks.bed --bl -g mm9
```

iii. Regulatory potential–based target prediction for ESR1 in MCF-7 cells with 12 h of E2 treatment:

```
$ BETA minus -p 349_peaks.bed --bl -g hg19
```

? TROUBLESHOOTING

Box 1

Differential expression data file formats

- **LIMMA standard output (LIM)**[19]

  ID (optional), RefseqID, logFC, AveExpre, *T*, *P*-value, Adj. *P*-value, B

- **Cuffdiff standard output (CUF)** (http://cufflinks.cbcb.umd.edu/manual.html#gene_exp_diff)

  Test ID, gene ID, gene, locus, sample1, sample2, status, value1, value2, log2 (fold change), test stat, *P* value, *Q* value, significant

- **BETA-specific format (BSF)**

  Gene ID, regulatory status (value with + or -), statistical value (e.g., FDR or *P* value)

- **Other formats (O)**

  Gene ID, regulatory status, statistical value

Box 2

Optional BETA parameters

-n NAME, --name NAME	Name result file
-o OUTPUT, --output OUTPUT	Directory to store all output files
--gname2	If switched on, gene or transcript IDs in files given through -e will be considered official gene symbols DEFAULT=FALSE
--info EXPREINFO	Specify gene ID, up/down status and statistical values Columns? of expression data; DEFAULT: 2,5,7 for LIMMA; 2,10,13 for Cuffdiff; 1,2,3 for BETA-specific format
--pn PEAKNUMBER	Number of peaks contributing to regulatory potential score DEFAULT=10000
--df DIFF_FDR	Input a value 0–1 as a significance threshold for differentially expressed genes by statistical value DEFAULT=1 (all genes)
--da DIFF_AMOUNT	Most significant differentially expressed genes by proportion (0–1) or number (>1) and ranked by statistical value; for example, 2,000 will set the top 2,000 up genes and 2,000 down genes DEFAULT=0.5 (top 50% up genes and top 50% down genes)

-c CUTOFF, --cutoff CUTOFF	Input a value 0–1 as a threshold of one-tail Kolmogorov-Smirnov test to determine significant difference DEFAULT=1e-3

Box 3

Parameters for BETA-plus extension usage

- **Application of BETA beyond human and mouse data:**

-r REFERENCE, --reference REFERENCE	The gene annotation file is downloaded from http://genome.ucsc.edu/. Input only if the genome is neither hg19 nor mm9

- **Peaks and associated genes may be detected by some boundaries such as CTCF binding sites; BETA provides built-in CTCF-conserved binding sites integrated from ENCODE CTCF and DNase1 ChIP-seq data:**

--bl BOUNDARYLIMIT	Boolean value; whether or not to use a CTCF boundary to obtain a peak's associated gene, DEFAULT=FALSE
--bf BOUNDARYFILE	BED format boundary file; use only when --bl is set and genome is neither hg19 nor mm9

? TROUBLESHOOTING

Troubleshooting advice can be found in Table 2.

● TIMING

In general, it takes 2, 1 and 20 min to run BETA-basic, BETA-minus and BETA-plus, respectively. The run time is closely related to the number of binding events (users can set top binding sites by using the parameter --pn) and the number the differentially expressed genes (which depends on –da and --df).

ANTICIPATED RESULTS

BETA-basic

BETA-basic rapidly analyzes a factor's function and its direct targets. The resulting output files are listed here:

File name	Description	Function (PROCEDURE Steps)
NA_byscores.R	R script for function prediction	Up or down or both Steps 11A(i, iii and iv), 11B(ii, iii and iv)
NA_score.pdf	PDF file for function prediction	Up or down or both Steps 11A(i, iii and iv), 11B(ii, iii and iv)
NA_uptarget.txt	Upregulated direct targets	Up Steps 11A(i and iv), 11B(ii and iv)
NA_uptarget_associate_peaks.bed	Associated peaks of upregulated targets	Up Steps 11A(i and iv), 11B(ii and iv)
NA_downtarget.txt	Downregulated direct targets	Down Steps 11A(iii and iv), 11B(iii and iv)

File name	Description	Function (PROCEDURE Steps)
NA_downtarget_associate_peaks.bed	Associated peaks of downregulated targets	Down Steps 11A(iii and iv), 11B(iii and iv)

Functional prediction results are presented as a cumulative distribution function plot; direct target genes can be downloaded as a tab-delimited text file with the first six columns in standard BED format. In addition to these outputs, the file basic_score.pdf predicts whether the factor has an activating or repressive function or both (Figs. 2a and 3). Among the three data sets used in Step 11, BETA found AR to have an activating function in the prostate cancer cell line LNCaP (Fig. 2a), Tet1 to have a repressing function in mouse ES cells (Fig. 3a) and ESR1 to have both activating and repressing functions in the breast cancer cell line MCF-7 (Fig. 3b); these finding are consistent with the previous studies[5,39]. Direct upregulated targets or downregulated targets are listed in the NA_uptarget.txt or NA_downtarget.txt file, and the associated peaks named NA_uptarget_associated_peaks.bed or NA_downtarget_associated_peaks.bed will be output as well. All the output results have the same format with the *AR* output shown in the Experimental design.

BETA-plus

BETA-plus runs function prediction, target detection and binding motif analysis step by step. Motif results are deposited into a directory named ‘motifresult’, which is under the BETA results directory defined by the parameter -o. If a factor functions as both an activator and a repressor of gene expression (e.g., ESR1 in MCF7 cells), BETA will perform motif searches in both upregulated and downregulated target gene regions, enrichment analysis of motifs in up- or down-target over nontarget regions and identification of upregulation- and downregulation-specific motifs. The binding motif analysis file betamotif. html summarizes all significant binding motif results on a web page. The result of the ESR1 analysis are shown in Figure 4, with Figure 4a,b depicting binding motifs found in upregulated and downregulated target gene regions, respectively. The ESR1 binding motif was the most significant one in both up- and down-target regions. The retinoid X receptor alpha (RXRA) binding motif was found to have a negative *t* score, which represents RXRA enriched in downregulated genes. Motifs with a positive *t* score represent enrichment in upregulated genes (Fig. 4c). In addition, binding motifs found in upregulated and downregulated genes (compared with nontargeted genes) represent potential collaborating factors to ESR1 (Fig. 4d,e).

In addition to .html summarized files, the original tab-delimited text files include analysis of motifs in upregulated and downregulated gene regions: UP_MOTIFS.txt and DOWN_MOTIFS.txt. Additional files display the results of motif comparisons: UP_NON_MOTIFS.txt and DOWN_NON_MOTIFS.txt. BETA also summarizes the differential motifs between upregulated and downregulated gene regions in the UP_DOWN_DIFFERENTIAL_MOTIFS.txt file. The common format of these files is shown below (using ESR1 UP motifs as an example). Motif ID, species and DNA-binding domain provide basic information for the binding motif. PSSM (an example is shown below) can be used to draw a motif logo, to perform motif similarity comparisons or to get

the motif sequence for further analysis; *t* scores and *P* values represent the statistical values for the enrichment.

Results of motif analysis:

```
MotifID Species Symbol DNA BindDom PSSM Tscore Pvalue
MC00335 Homo sapiens ESR1 Hormone-nuclear Receptor Family PSSM 20.94 1.39e–81
MC00333 Homo sapiens ESR2 Hormone-nuclear Receptor Family PSSM 19.49 2.88e–73
MS00657 Homo sapiens NR2F1 Hormone-nuclear Receptor Family PSSM 15.64 3.59e–
51
MA0066 Homo sapiens PPARG Hormone-nuclear Receptor Family PSSM 12.12 3.61e–32
MS00081 Homo sapiens MEIS1 Homeodomain Family PSSM 11.45 2.96e–29
MS00829 Homo sapiens ESRRA Hormone-nuclear Receptor Family PSSM 10.88 1.03e–
26
```

Format of PSSM:

M00179: ATF2

0.143 0.143 0.714 0.551 0.01 0.01 0.01 0.97 0.01 0.01 0.01 0.286

0.143 0.286 0.143 0.01 0.01 0.97 0.01 0.01 0.97 0.01 0.571 0.418

0.286 0.561 0.01 0.429 0.143 0.01 0.97 0.01 0.01 0.01 0.286 0.286

0.428 0.01 0.133 0.01 0.837 0.01 0.01 0.01 0.01 0.97 0.133 0.01

BETA-minus

BETA-minus predicts factor target genes from binding data only, and provides as output two text files: a target file and a list of target-associated peaks. The target-associated peaks file has the same format as BETA-basic output files. A sample output for the target gene file is shown below, where score refers to the regulatory potential calculated with the same method we described above:

Argument List:

```
#Chromsome TSS TTS RefseqID Score Strand GeneSymbol
chr5 180630119 180632177 NM_033342 2.991 - TRIM7
chr5 180620923 180632177 NM_203293 2.991 - TRIM7
chr5 180620923 180631340 NM_203294 2.969 - TRIM7
chr5 180620923 180631340 NM_203296 2.969 - TRIM7
chr5 180620923 180631340 NM_203295 2.969 - TRIM7
chr5 180620923 180627930 NM_203297 2.916 - TRIM7
chr6 26538571 26547164 NM_006353 2.469 + HMGN4
chr5 180649565 180649633 NR_039781 2.446 - MIR4638
chr6 35541361 35696360 NM_001145775 2.406 - FKBP5
```

The three subprotocols provided by the BETA package have a wide applicability for the integration of ChIP-seq and transcriptome analysis. Target genes predicted by BETA and the prediction of their activating or repressing functions help researchers to understand the regulatory mechanisms of the analyzed factors. Furthermore, efficient binding motif analysis provides a new way to detect co-regulators.

Acknowledgments

This project was supported by the National Basic Research (973) Program of China (2010CB944904), the National Natural Science Foundation of China (31329003) and the US National Institutes of Health (HG4069 and U41 HG007000).

References

1. McLean CY, et al. GREAT improves functional interpretation of *cis*-regulatory regions. Nat Biotechnol. 2010; 28:495–501. [PubMed: 20436461]
2. Jiang C, Xuan Z, Zhao F, Zhang MQ. TRED: a transcriptional regulatory element database, new entries and other development. Nucleic Acids Res. 2007; 35:D137–D140. [PubMed: 17202159]
3. Buck MJ, Lieb JD. A chromatin-mediated mechanism for specification of conditional transcription factor targets. Nat Genet. 2006; 38:1446–1451. [PubMed: 17099712]
4. Palii CG, et al. Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages. EMBO J. 2010; 30:494–509. [PubMed: 21179004]
5. Tang Q, et al. A comprehensive view of nuclear receptor cancer cistromes. Cancer Res. 2011; 71:6940–6947. [PubMed: 21940749]
6. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett. 2004; 573:83–92. [PubMed: 15327980]
7. Sherman BT, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 2007; 35:W169–W175. [PubMed: 17576678]
8. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nat Protoc. 2012; 7:1728–1740. [PubMed: 22936215]
9. Ma W, Wong WH. The analysis of ChIP-seq data. Methods Enzymol. 2011; 497:51–73. [PubMed: 21601082]
10. Zang C, et al. A clustering approach for identification of enriched domains from histone modification ChIP-seq data. Bioinformatics. 2009; 25:1952–1958. [PubMed: 19505939]
11. Boyer LA, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell. 2005; 122:947–956. [PubMed: 16153702]
12. Rougemont J, Naef F. Computational analysis of protein-DNA interactions from ChIP-seq data. Methods Mol Biol. 2012; 786:263–273. [PubMed: 21938632]
13. Cheng C, Min R, Gerstein M. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. Bioinformatics. 2011; 27:3221–3227. [PubMed: 22039215]
14. Qian J, Lin J, Luscombe NM, Yu H, Gerstein M. Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. Bioinformatics. 2003; 19:1917–1926. [PubMed: 14555624]
15. Honkela A, et al. Model-based method for transcription factor target identification with limited data. Proc Natl Acad Sci. 2010; 107:7793–7798. [PubMed: 20385836]
16. Redestig H, Weicht D, Selbig J, Hannah MA. Transcription factor target prediction using multiple short expression time series from *Arabidopsis thaliana*. BMC Bioinformatics. 2007; 8:454. [PubMed: 18021423]
17. Qin J, Li MJ, Wang P, Zhang MQ, Wang J. ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. Nucleic Acids Res. 2011; 39:W430–W436. [PubMed: 21586587]

18. Maienschein-Cline M, Zhou J, White KP, Sciammas R, Dinner AR. Discovering transcription factor regulatory targets using gene expression and binding data. Bioinformatics. 2012; 28:206–213. [PubMed: 22084256]
19. Smyth, GK. Limma: linear models for microarray data. In: Gentleman, R.; Carey, VJ.; Huber, W.; Irizarry, RA.; Dudoit, S., editors. Bioinformatics and Computational Biology Solutions using R and Bioconductor. Springer; 2005. p. 397-420.
20. Trapnell C, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012; 7:562–578. [PubMed: 22383036]
21. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]
22. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9:R137. [PubMed: 18798982]
23. Johnson WE, et al. Model-based analysis of tiling-arrays for ChIP-chip. Proc Natl Acad Sci USA. 2006; 103:12457–12462. [PubMed: 16895995]
24. Wang Q, et al. A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth. Mol Cell. 2007; 27:380–392. [PubMed: 17679089]
25. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005; 102:15545–15550. [PubMed: 16199517]
26. Chan CS, Song JS. CCCTC-binding factor confines the distal action of estrogen receptor. Cancer Res. 2008; 68:9041–9049. [PubMed: 18974150]
27. Sun Z, Pan J, Balk SP. Androgen receptor-associated protein complex binds upstream of the androgen-responsive elements in the promoters of human prostate-specific antigen and kallikrein 2 genes. Nucleic Acids Res. 1997; 25:3318–3325. [PubMed: 9241247]
28. Steketee K, Ziel-van der Made AC, van der Korput HA, Houtsmuller AB, Trapman J. A bioinformatics-based functional analysis shows that the specifically androgen-regulated gene *SARG* contains an active direct repeat androgen response element in the first intron. J Mol Endocrinol. 2004; 33:477–491. [PubMed: 15525603]
29. Cai C, Wang H, Xu Y, Chen S, Balk SP. Reactivation of androgen receptor-regulated *TMPRSS2:ERG* gene expression in castration-resistant prostate cancer. Cancer Res. 2009; 69:6027–6032. [PubMed: 19584279]
30. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013; 14:178–192. [PubMed: 22517427]
31. Kent WJ, et al. The human genome browser at UCSC. Genome Res. 2002; 12:996–1006. [PubMed: 12045153]
32. Korhonen J, Martinmäki P, Pizzi C, Rastas P, Ukkonen E. MOODS: fast search for position weight matrix matches in DNA sequences. Bioinformatics. 2009; 25:3181–3182. [PubMed: 19773334]
33. Liu T, et al. Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol. 2011; 12:R83. [PubMed: 21859476]
34. Habib N, Kaplan T, Margalit H, Friedman N. A novel Bayesian DNA motif comparison method for clustering and retrieval. PLoS Comput Biol. 2008; 4:e1000010. [PubMed: 18463706]
35. Fulton DL, et al. TFCat: the curated catalog of mouse and human transcription factors. Genome Biol. 2009; 10:R29. [PubMed: 19284633]
36. Da Wei Huang BTS, Stephens R, Baseler MW, Lane HC, Lempicki RA. DAVID gene ID conversion tool. Bioinformation. 2008; 2:428. [PubMed: 18841237]
37. Fujita PA, et al. The UCSC genome browser database: update 2011. Nucleic Acids Res. 2011; 39:D876–D882. [PubMed: 20959295]
38. Dai M, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. 2005; 33:e175–e175. [PubMed: 16284200]
39. Williams K, et al. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature. 2011; 473:343–348. [PubMed: 21490601]

40. Hu M, Yu J, Taylor JM, Chinnaiyan AM, Qin ZS. On the detection and refinement of transcription factor binding sites using ChIP-seq data. Nucleic Acids Research. 2010; 38:2154–2167. [PubMed: 20056654]

41. Carroll JS, et al. Genome-wide analysis of estrogen receptor binding sites. Nat Genet. 2006; 38:1289–1297. [PubMed: 17013392]

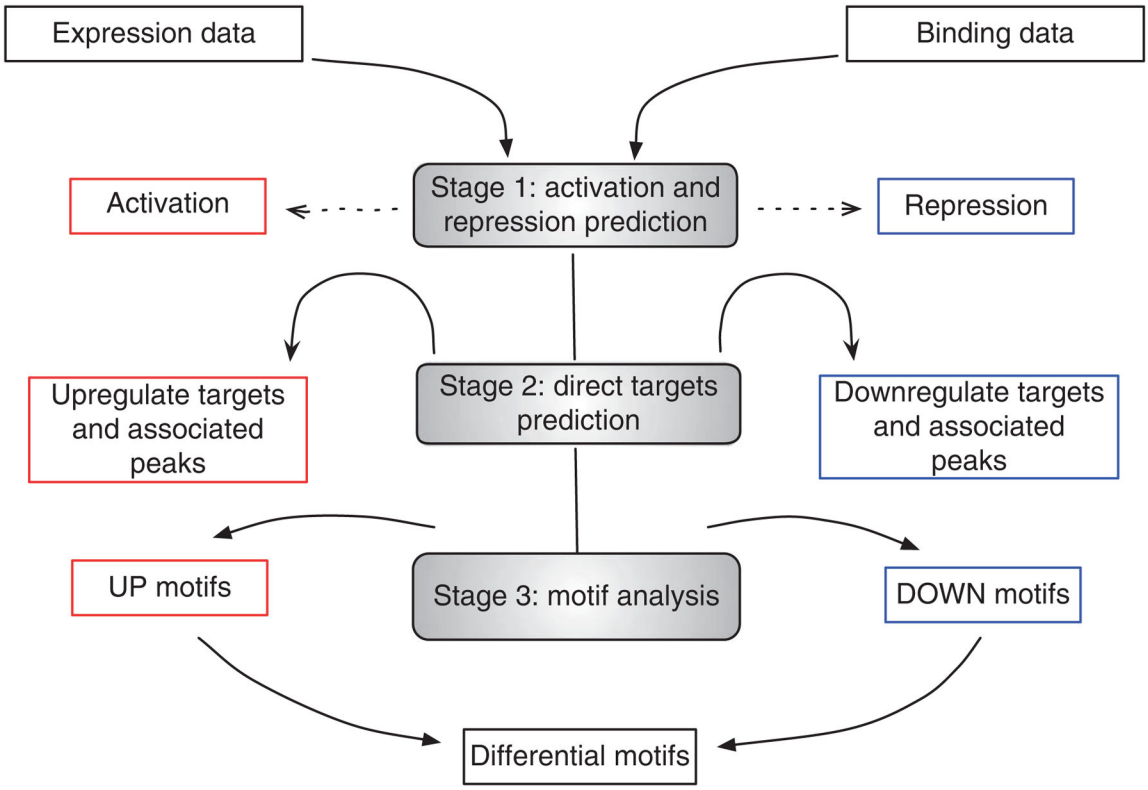

Figure 1.

BETA workflow. Stage 1 analyzes the differential expression and ChIP-seq binding data to predict whether a factor generally activates or represses gene expression. Stage 2 predicts direct target genes by their upregulation or downregulation. Stage 3 conducts motif analysis to identify putative collaborating factors that contribute to upregulation (UP) or downregulation (DOWN).

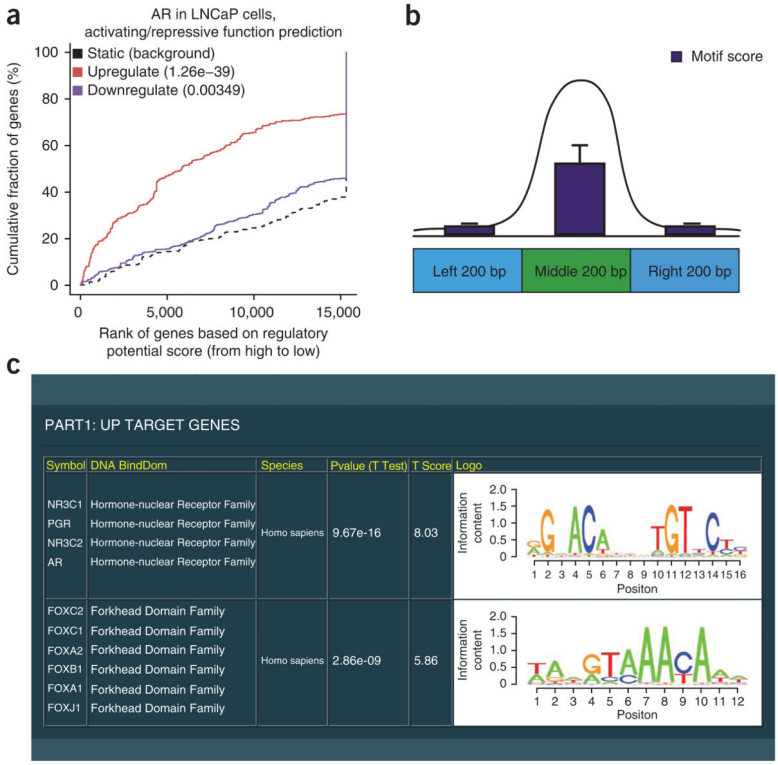

Figure 2.

BETA output of activating/repressive function prediction and motif analysis of AR. (**a**) BETA activating/repressive function prediction of the AR data set from the LNCaP prostate cancer cell line. The red and the purple lines represent the upregulated and downregulated genes, respectively. The dashed line indicates the nondifferentially expressed genes as background. Genes are cumulated by the rank on the basis of the regulatory potential score from high to low. *P* values that represent the significance of the UP or DOWN group distributions are compared with the NON group by the Kolmogorov-Smirnov test. (**b**) Motif scan algorithm. Motif scores in each binding peak are compared among three regions. The middle region consists of 200 bp centered on the peak summit; the left and right regions comprise 200 bp in either direction of the middle region. The significance of motif summit enrichment is measured by the *P* value from a one-tailed *t* test. (**c**) Screenshot of binding motif analysis on UP target regions of AR. Similar motifs are grouped together, and the motif logo of the most significant factor in the group is provided in the last column. The motif symbol, DNA-binding domain and species are shown in the first three columns; the *t* score and the *P* value from the *t* test are shown in the middle two columns.

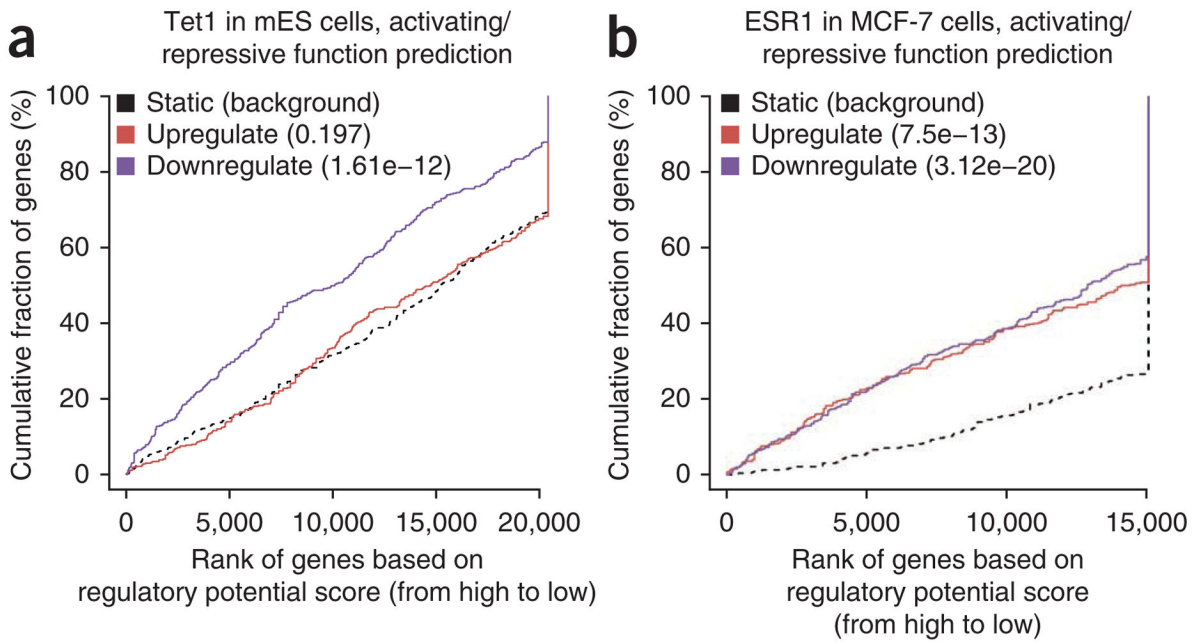

Figure 3.

Activating and repressive function prediction of Tet1 in mouse ES cells and ESR1 in MCF-7 cells. (**a**) BETA-basic analysis of the Tet1 binding and expression data sets from the mouse ES cell line identifies upregulated (red) and downregulated (purple) genes. The dashed line indicates the non-differentially-expressed (NON) genes as background. (**b**) BETA-basic analysis of *ESR1* targets in MCF-7 breast cancer cells. *P* values represent the significance of difference in the UP or DOWN groups compared with the NON group by the Kolmogorov-Smirnov test.

Figure 4.

Screenshots of summarized BETA-plus analysis of ESR1 motifs in html format. (**a**) BETA-plus combines factor-binding data and differential gene expression data to analyze sequence motifs in upregulated target genes (UP). ESR1, ESR2 and six other human estrogen receptor family members are classified into one group because of their high similarity scores. (**b**) Motifs in downregulated target genes (DOWN). (**c**) Differential motifs found in the UP and DOWN groups. A *t* score of >0 indicates motifs enriched in upregulated target genes, whereas a *t* score of <0 indicates enrichment in downregulated target genes. (**d**) Motif comparison between UP and NON regions. (**e**) Motif comparison between DOWN and NON regions.

TABLE 1

Example data sets.

Binding file name	Binding data resource	Data description	Peak no.	Expression file name	Expression data accession no.
3656_peaks.bed	Brown Laboratory[24]	*AR* in LNCaP cells	7095	AR_diff_expre.xls	GSE7868 (ref. 24)
5795_peaks.bed	GSE24841 (ref. 39)	*Tet1* in mES cells	35532	Tet1_diff_expre.xls	GSE24842 (ref. 39)
349_peaks.bed	GSE19013 (ref. 40)	*ESR1* in MCF-7 cells	7031	ESR1_diff_expr.xls	GSE11324 (ref. 41)

TABLE 2

Troubleshooting table.

Step	Problem	Possible reason	Solution
1–6	Installation failed	Variable problems (and see below)	Refer to the readme or detailed installation online http://cistrome.org/BETA/#inst
6	Setuptools error	Lack of the Python package 'setuptools'; multiple versions of Python installed	To install setuptools, input the following in the terminal: $ curl http://python-distribute.org/distribute_setup.py | sudo python For multiple python versions, specify the PYTHONPATH of 1 over 2.6 in environment variable: > export PYTHONPATH=PATH_TO_BETA_LIB
	Installation failed	Incorrect permission of the installation directory	Modify PYTHONPATH and reinstall BETA with: —prefix
8, 9	Error when checking input file format	Unsupported format of input data or incorrect [LIM/CUF/BSF/O] parameter	Check the input file, especially for differential expression data; see examples of test data in the BETA package
10, 11	No results output	Low data quality or unmatched binding and expression data	Use high-quality data or loosen some parameters