

Dartmouth College

## Dartmouth Digital Commons

---

Open Dartmouth: Published works by  
Dartmouth faculty

Faculty Work

---

8-12-2016

### iTAR: A Web Server for Identifying Target Genes of Transcription Factors using ChIP-Seq or ChIP-Chip Data

Chia-Chun Yang

*National Chung Hsing University, Taichung*

Erik H. Andrews

*Dartmouth College*

Min-Hsuan Chen

*National Chung Hsing University, Taichung*

Wan-Yu Wang

*National Chung Hsing University, Taichung*

Follow this and additional works at: <https://digitalcommons.dartmouth.edu/facoa>



Part of the [Genetics and Genomics Commons](#)

---

#### Dartmouth Digital Commons Citation

Yang, Chia-Chun; Andrews, Erik H.; Chen, Min-Hsuan; and Wang, Wan-Yu, "iTAR: A Web Server for Identifying Target Genes of Transcription Factors using ChIP-Seq or ChIP-Chip Data" (2016). *Open Dartmouth: Published works by Dartmouth faculty*. 599.  
<https://digitalcommons.dartmouth.edu/facoa/599>

This Article is brought to you for free and open access by the Faculty Work at Dartmouth Digital Commons. It has been accepted for inclusion in Open Dartmouth: Published works by Dartmouth faculty by an authorized administrator of Dartmouth Digital Commons. For more information, please contact [dartmouthdigitalcommons@groups.dartmouth.edu](mailto:dartmouthdigitalcommons@groups.dartmouth.edu).

DATABASE

Open Access



# iTAR: a web server for identifying target genes of transcription factors using ChIP-seq or ChIP-chip data

Chia-Chun Yang<sup>1,2</sup>, Erik H. Andrews<sup>8</sup>, Min-Hsuan Chen<sup>2,3</sup>, Wan-Yu Wang<sup>2</sup>, Jeremy J. W. Chen<sup>1,3,4</sup>, Mark Gerstein<sup>5,6</sup>, Chun-Chi Liu<sup>2,3,4\*</sup> and Chao Cheng<sup>7,8,9\*</sup>

## Abstract

**Background:** Chromatin immunoprecipitation followed by massively parallel DNA sequencing (ChIP-seq) or microarray hybridization (ChIP-chip) has been widely used to determine the genomic occupation of transcription factors (TFs). We have previously developed a probabilistic method, called TIP (Target Identification from Profiles), to identify TF target genes using ChIP-seq/ChIP-chip data. To achieve high specificity, TIP applies a conservative method to estimate significance of target genes, with the trade-off being a relatively low sensitivity of target gene identification compared to other methods. Additionally, TIP's output does not render binding-peak locations or intensity, information highly useful for visualization and general experimental biological use, while the variability of ChIP-seq/ChIP-chip file formats has made input into TIP more difficult than desired.

**Description:** To improve upon these facets, here we present are fined TIP with key extensions. First, it implements a Gaussian mixture model for *p*-value estimation, increasing target gene identification sensitivity and more accurately capturing the shape of TF binding profile distributions. Second, it enables the incorporation of TF binding-peak data by identifying their locations in significant target gene promoter regions and quantifies their strengths. Finally, for full ease of implementation we have incorporated it into a web server (<http://syslab3.nchu.edu.tw/iTAR/>) that enables flexibility of input file format, can be used across multiple species and genome assembly versions, and is freely available for public use. The web server additionally performs GO enrichment analysis for the identified target genes to reveal the potential function of the corresponding TF.

**Conclusions:** The iTAR web server provides a user-friendly interface and supports target gene identification in seven species, ranging from yeast to human. To facilitate investigating the quality of ChIP-seq/ChIP-chip data, the web server generates the chart of the characteristic binding profiles and the density plot of normalized regulatory scores. The iTAR web server is a useful tool in identifying TF target genes from ChIP-seq/ChIP-chip data and discovering biological insights.

**Keywords:** Transcription factor, ChIP-seq, ChIP-chip, Gaussian mixture model, Gene ontology analysis

## Background

Transcription factors (TFs) constitute a family of proteins that play critical roles in regulating gene transcription [1, 2]. Mechanistically, they operate by recognizing and binding specific DNA sequences via DNA-binding domain(s) or by forming complexes with other

regulatory co-factors [3]. Owing to the development of technologies such as chromatin immunoprecipitation followed by massive parallel sequencing (ChIP-seq) or DNA hybridization (ChIP-chip), a large number of experiments have in recent years sought to identify genome-wide TF binding sites [4, 5], from which predictions about TF regulatory target genes can be made. Given the ChIP-seq/ChIP-chip data for a TF, several computational methods have been proposed to define these target genes [6–8]. Most of these methods apply a peak-based strategy: first, a peak-calling algorithm [9–11] is

\* Correspondence: jimliu@nchu.edu.tw; chao.cheng@dartmouth.edu

<sup>2</sup>Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung 402, Taiwan

<sup>7</sup>Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, NH 03755, USA

Full list of author information is available at the end of the article



employed to generate a list of enriched binding peaks of a TF, with target genes for that TF then defined by analyzing the intensity and location of the binding peaks to nearby genes. As signals from non-significant binding peaks are not considered, these methods are sensitive to the choice of peak-calling method. Indeed, a considerably different set of TF binding peaks and derived target genes are identified when different peak-calling methods or different parameter-settings are used, raising questions of validity and false positivity of this approach.

To address this issue, we have previously developed a probabilistic method, called TIP (Target Identification from Profiles), to identify TF target genes based on ChIP-seq or ChIP-chip data [12]. TIP requires no pre-defined binding peaks; rather, it considers raw experimental TF binding signals from all DNA regions of the genome. TIP works by characterizing a binding profile of a TF around the transcriptional start sites (TSS) of all genes and then uses this profile to weight the binding intensity of a TF in each gene's promoter region, yielding a continuous-valued binding score of a TF for each gene. TIP then estimates the significance of each gene as a regulatory target by comparing each gene's binding score to the distribution of all binding scores, using statistical methods that assume distributional normality.

As others and we have shown, TIP identifies TF targets with very high accuracy that can be readily used in downstream biological studies [13]. However, the following issues hinder the application of TIP. First, the assumption that binding scores follow a normal distribution makes the calculation of  $p$ -values exceedingly conservative, which results in a lower sensitivity of the method for identifying targets. This is because binding scores are generally not normally distributed and instead are either bimodal or positively skewed, reflecting the fact that the binding scores fundamentally encompass two different populations: genuine targets (with higher scores) and background genes (with lower scores). Thus, assuming distributional normality elides over this heterogeneity and loses statistical resolving power. Second, as TIP does not utilize binding peaks, it also does not output them, a shortcoming for downstream analyses as they are helpful for visualization and general experimental biology. Third, while a strength of TIP is that it uses all the track files from ChIP-seq and ChIP-chip experiments, in practice the file format variation of this data (Bed, BedGraph, Wiggle, and BigWig, among others) has made file input into TIP less straightforward than desired.

To address these limitations, here we extend TIP with key modifications. First, we revise the  $p$ -value calculation using a Gaussian mixture model, thereby more accurately taking into account the shape of the binding score distribution and improving target gene identification sensitivity. Second, we remedy the lack of binding peak

output by allowing TIP to optionally incorporate a binding peak file into its analysis and identify peak loci and intensity within the promoter regions of each significant target gene calculated by TIP. Finally, for ease of implementation we have created a web server (<http://syslab3.nchu.edu.tw/iTAR/>) for track files upload and end-to-end TIP processing. The web server currently accepts Wig, BigWig, and BedGraph compressed (.gz or.rar) formats from TF ChIP-seq/ChIP-chip experiments as the input, which contains the binding signals of a TF at each genomic position (e.g., raw/normalized read coverage or fold-change), and enables analysis for seven organisms (human, mouse, fly, worm, chicken, zebrafish and yeast) with multiple genome assembly support for human and mouse. With a user-friendly interface and extensible backend for future genome version support, the server is widely usable for TF target gene identification and ChIP-seq analysis going forward.

As a walkthrough of its capabilities we provide an example of its use for STAT3 ChIP-seq data. A validation of its output is conducted using the NFE2 ChIP-seq and gene expression dataset.

## Construction and content

### Server construction, supported organisms and genome assemblies

The iTAR web server runs on Linux and is implemented in JSP and Java. ChIP-seq data analysis is provided for seven different organisms: human, mouse, fly, worm, chicken, zebrafish and yeast, with multiple genome assemblies available for human (UCSC hg 18 and hg19) and mouse (UCSC mm8, mm9, and mm10). In addition, support for genome assemblies is extensible to enable further genome assembly availability going forward.

Annotation files for RefSeq genes are downloaded from the UCSC Genome Browser [14]. These files provide the chromosome, strand, transcriptional start site, transcription terminal sites, structures and other genomic information for all RefSeq genes for an organism, based on the corresponding genome assembly.

The server is freely available to the public and does not require login. Analysis of a ChIP-seq dataset for target gene identification using the server is straightforward and requires only three steps: (i) uploading input files, (ii) choosing parameters, and (iii) receiving the output files.

### Input files

The iTAR web server takes as input a ChIP-seq/ChIP-chip signal track file and optionally a binding peak file for a TF. The signal track file contains the binding signal of a TF at each nucleotide position of the genome as measured by sequencing in ChIP-seq or hybridization in ChIP-chip data. The server supports three signal track

formats: BedGraph, Wiggle, and BigWig. We note that each format may have multiple variants. For example, Wiggle files have two main formatting options—fixedStep and variableStep, designed for data with regular and irregular intervals between data points, respectively. The server contains a module to process different formats of signal track files to ensure the best usability.

The optional peak file is in bed format, containing the chromosome, start position, end position and other information of TF binding peaks. Users can run a peak calling program, e.g. MACS or PeakSeq [9, 11], using the uploaded ChIP-seq data as input to generate this peak file.

To improve the uploading time, the signal track files must be compressed into either.rar or.gz format before uploading. Alternatively, user can support download links for track file and peak file, and then iTAR will download the input files automatically.

### Parameter settings

After the input file(s) have been uploaded, users need to select the organism and the genome assembly version from the drop-down list (see Fig. 1a). The selected organism/assembly must match with the uploaded signal track file and an optional peak file to ensure accurate results in subsequent analysis. In addition, users need to specify three other parameters: the length of considering binding signals around TSS, the FDR (false discovery rate) threshold for TF targets and the  $P$ -value threshold for GO enrichment analysis. The length of promoter region will be used to select the regions of the putative promoters, the FDR threshold will be used to select a subset of target genes of the TF for subsequent GO enrichment analysis, and the  $P$ -value threshold will be used to select significantly enriched GO categories for displaying in the output webpage. After all parameters have been specified, users can click the “Submit” button to initiate online data analysis. A waiting page will be displayed and updated every 5 s to report the progress of the data analysis (Fig. 1b). Depending on the organism and the signal track file format, analysis may take 10–25 min to finish.

### Output files

After data analysis has completed, the results will be displayed in an output webpage. The main output consists of five panels that include (i) a table summarizing the parameter and input settings, (ii) a characteristic binding profile of the TF, (iii) the distribution of normalized regulatory scores for all genes, (iv) a list of significant genes and (optionally) their associated TF peak binding sites, and (v) a list of significantly enriched GO terms (see Fig. 1 for an example). Significant genes in (iv) are sorted by decreasing order of their regulatory scores. For each gene, the  $p$ -value and the multiple-

testing corrected FDR value are calculated, using both single normal and mixture normal models. To enable parameter setting adjustment and iterative analyses based on the output results, a “Rerun the program” option is included that accepts parameter setting modifications (to change statistical stringency requirements for GO analysis, for example) without the need for re-uploading input files.

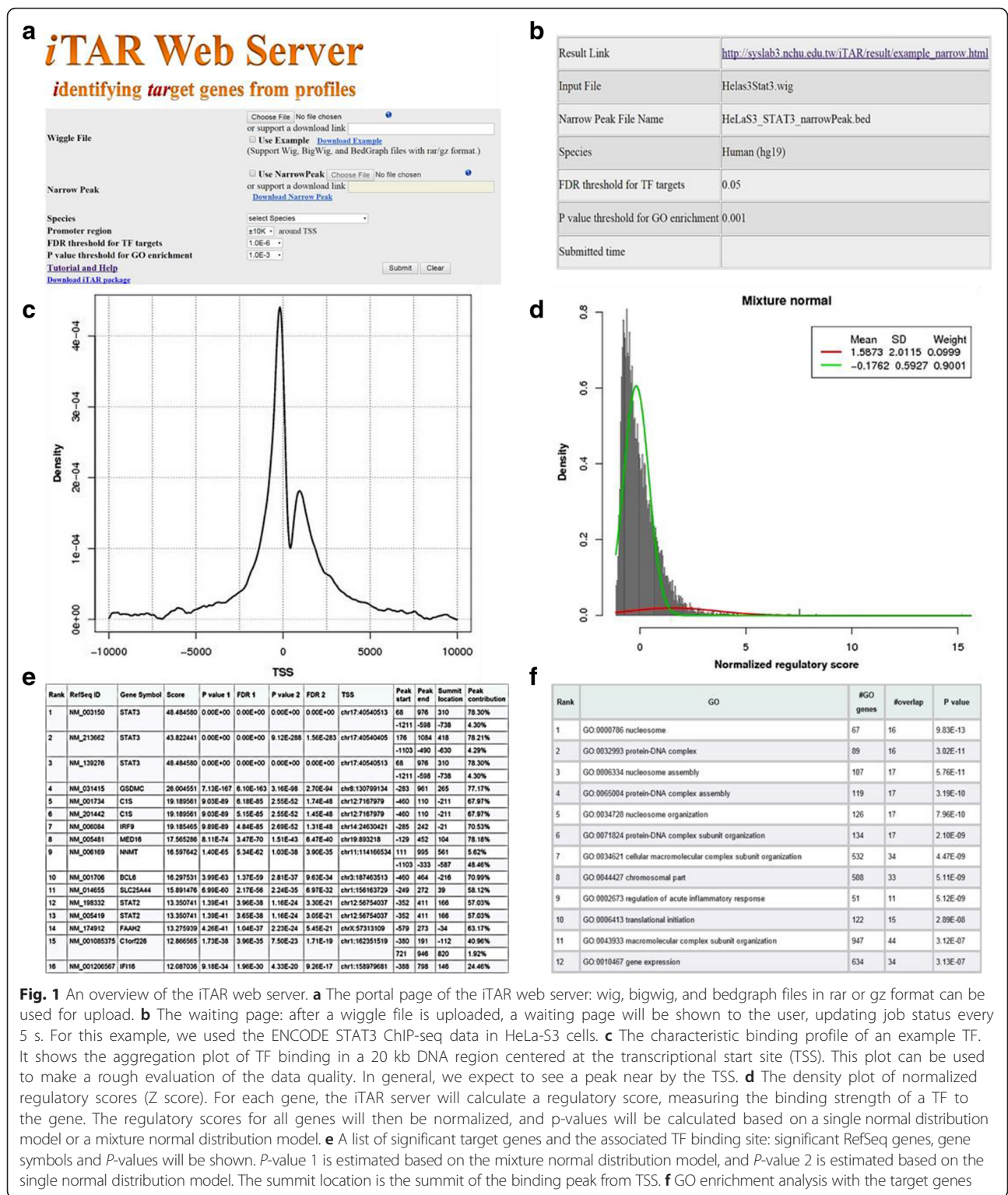
### TIP extensions

As mentioned above and previously described, TIP builds a characteristic, averaged binding profile for a TF around the TSS of all genes and then uses this profile to weight the sites associated with a given gene, providing a continuous-valued regulatory score of a TF for each gene. It then normalizes the regulatory scores into  $z$ -scores and estimates their  $p$ -values by referring to a standard normal distribution [12]. However, using a standard normal distribution to estimate the significance of regulatory scores is highly conservative, giving rise to a confident but small target gene set. This is because the distribution of the regulatory scores is typically not normal but positively skewed (Fig. 1d) or bimodal (Additional file 1: Figure S1), reflecting the fact that binding scores encompass two distinct groups: background, non-target genes with low regulatory scores, and genuine target genes with higher regulatory scores. By not taking this non-normal distributional shape into account, statistical resolution is lost and  $p$ -value estimates are conservative.

Motivated by these observations, here we refine TIP’s  $p$ -value calculation by applying a two-component Gaussian mixture model to estimate the significance of normalized regulatory scores. Our approach is as follows: suppose that each regulatory score  $y_1, \dots, y_n$  is instead derived from a two-component (i.e., non-target and genuine target genes, respectively) Gaussian mixture distribution. The log-likelihood for data consisting of  $n$  observations  $y = (y_1, \dots, y_n)$ , assuming a normal mixture model with two components, is then given by

$$\ell(\Theta|y) = \sum_{i=1}^n \log \left( \sum_{k=1}^2 w_k N(y_i | \mu_k, \sigma_k) \right) \quad (1)$$

where  $\Theta = (w, \Psi)$  represents all unknown parameters and  $N(\cdot | \mu_k, \sigma_k)$  denotes a Gaussian density function with mean  $\mu$  and standard deviation  $\sigma$ . Here the vector  $w = (w_1, w_2)$  consists of the mixing proportions and subjects to  $\sum_{k=1}^2 w_k = 1$ ,  $\Psi = (\mu_1, \mu_2, \sigma_1, \sigma_2)$ . The maximum likelihood estimation of  $\Theta$  can then be solved by



**Fig. 1** An overview of the iTAR web server. **a** The portal page of the iTAR web server: wig, bigwig, and bedgraph files in rar or gz format can be used for upload. **b** The waiting page: after a wiggle file is uploaded, a waiting page will be shown to the user, updating job status every 5 s. For this example, we used the ENCODE STAT3 ChIP-seq data in HeLa-S3 cells. **c** The characteristic binding profile of an example TF. It shows the aggregation plot of TF binding in a 20 kb DNA region centered at the transcriptional start site (TSS). This plot can be used to make a rough evaluation of the data quality. In general, we expect to see a peak near by the TSS. **d** The density plot of normalized regulatory scores (Z score). For each gene, the iTAR server will calculate a regulatory score, measuring the binding strength of a TF to the gene. The regulatory scores for all genes will then be normalized, and p-values will be calculated based on a single normal distribution model or a mixture normal distribution model. **e** A list of significant target genes and the associated TF binding site: significant RefSeq genes, gene symbols and P-values will be shown. P-value 1 is estimated based on the mixture normal distribution model, and P-value 2 is estimated based on the single normal distribution model. The summit location is the summit of the binding peak from TSS. **f** GO enrichment analysis with the target genes

$$\hat{\Theta} = \arg \max_{\Theta} \ell(\Theta|y) \tag{2}$$

To choose the best mixture model with optimal parameters of the two Gaussian distributions, we use an

expectation-maximization (EM) algorithm [15] (See Fig. 1d). EM is an iterative method for finding maximum likelihood estimation of parameters by using the following iterative algorithm,



$$\hat{w}_k = \frac{\sum_{i=1}^n \hat{z}_{ik}}{n}, \hat{\mu}_k = \frac{\sum_{i=1}^n \hat{z}_{ik} Y_i}{\sum_{i=1}^n \hat{z}_{ik}}, \quad (3)$$

$$\hat{\sigma}_k^2 = \sum_{i=1}^n \frac{\hat{z}_{ik} (Y_i - \hat{\mu}_k)^2}{\sum_{i=1}^n \hat{z}_{ik}}$$

with the posterior probabilities

$$\hat{z}_{ik} = \frac{\hat{w}_k N(y_i | \hat{\mu}_k, \hat{\sigma}_k)}{\sum_{h=1}^2 \hat{w}_h N(y_i | \hat{\mu}_h, \hat{\sigma}_h)} \quad (4)$$

Once the optimal parameters are determined, we estimate *p*-values by comparing each z-score to the left Gaussian distribution (with the smaller mean value).

### Gene ontology enrichment analysis

Given the target genes, we performed enrichment analysis with GO terms using the Fisher exact test based on a hypergeometric distribution [16]. To exclude non-informative general GO terms, we restricted our analysis to those with gene numbers < 1000. The web server shows significant GO terms ranked by *p*-value and provides the predicted target genes from them.

## Utility and discussion

### STAT3 ChIP-seq data

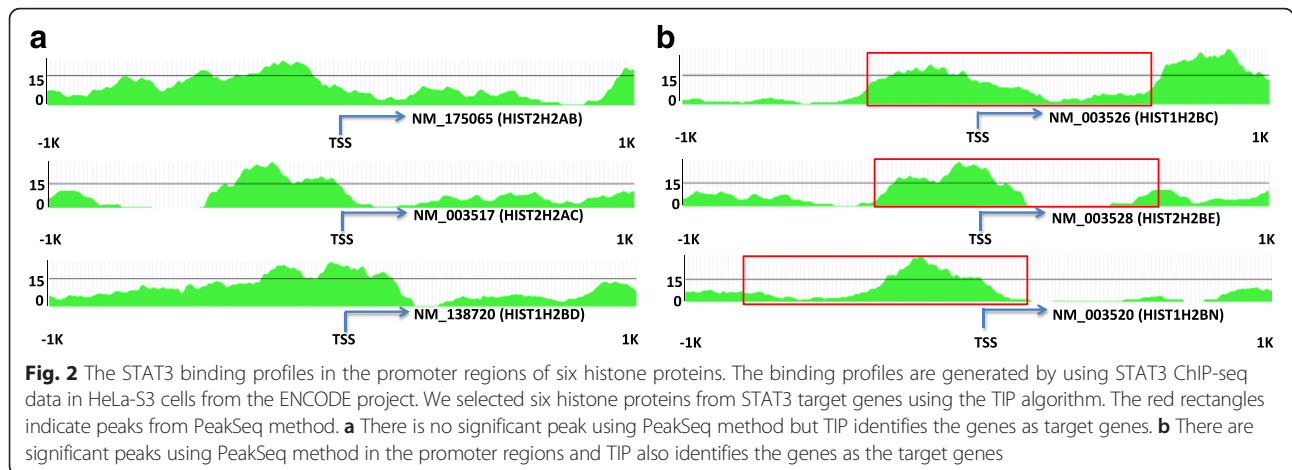
The ENCODE project contains STAT3 ChIP-seq data for experiments run in HeLa-S3 cells [5]. By checking the “Use Example” box and then clicking “Submit” in the interface page of the iTAR server (Fig. 1a), ChIP-seq data for STAT3 is processed and the results displayed in an output page. The output page contains five panels (Fig. 1b-f). The first panel shows a table that summarizes the data analysis including organism, version of genome assembly, parameter settings and submission time (Fig. 1b). The second panel shows the characteristic binding profile for STAT3, also referred to as an aggregation plot, which shows the average binding signals of STAT3 across all RefSeq genes in the +/- 10 kb DNA regions centered at the TSS of each gene (Fig. 1c). As shown, the profile displays a sharp peak around the TSS, suggesting that STAT3 shows a strong binding preference to the TSS proximal regions. While in general Fig. 1c’s purpose is to enable a visual check of binding profile characteristics, the location hints at underlying biology. Although most TFs show enriched binding signals around the TSS of genes, their characteristic binding profiles vary. This suggests that (i) different TFs tend to bind at different locations relative to TSSs, which could affect their transcriptional regulation of their target genes – signals closer to TSS may contribute more to their target gene regulation [8, 17]; (ii) the binding signals of different TFs at the same location might

influence nearby gene transcription differently – some TFs may exert their influence over long distances, while others exert only regional effects. Overall, Fig. 1c thus enables the study of overall trends of a TF’s regulation across the genome.

The third panel shows the distribution of normalized regulatory scores of STAT3 on all human RefSeq genes (Fig. 1d). The distribution shows a long tail to the right side, which can be decomposed into two separate distributions by using a two-component Gaussian mixture model (the red and the green curves). The main output of the iTAR web server is a ranked gene list as shown in the fourth panel (Fig. 1e), with genes sorted in decreasing order of their regulatory scores. *P*-values of all genes are calculated and then adjusted using the Benjamini–Hochberg multiple testing correction method (i.e. FDR) based on a single normal distribution as well as a mixture normal distribution. When the single normal distribution is used, a total of 241 STAT3 RefSeq target genes are identified at the 0.05 significance level (FDR < 0.05). In contrast, the method based on mixture normal distribution identifies 614 RefSeq target genes at the same significance level, highlighting its increased sensitivity of target gene prediction. By integrating Fig. 1d with 1e, users can analyze the binding profile distribution and decide which distribution model and *p*-value calculation works best for their data and application.

The fifth panel is a table containing results from GO enrichment analysis (Fig. 1f) of STAT3’s predicted target genes. Previous work has reported that unphosphorylated STAT3 (U-STAT3) influences gene transcription in response to cytokines [18]. In *Drosophila*, U-STAT92E is associated with HP1 and maintains heterochromatin stability. In addition, the U-STAT3-DNA interaction structure is important for chromatin organization [19]. Taken together, U-STAT3 is considered to function as a transcriptional activator of immune cells and a chromatin organizer. A total of 185 GO terms are significantly enriched in STAT3 target genes, in which the first, third, and fifth GO terms are associated with the nucleosome (GO:0000786, GO:0006334 and GO:0034728) and the ninth GO term is associated with the immune response (i.e. GO:0002673), consistent with STAT3’s known functionality. GO:0000786 nucleosome has 67 genes, among which 16 genes are STAT3 target genes (P = 1E-12) (Fig. 1f). To demonstrate the STAT3 binding signal in these 16 genes, Additional file 1: Figure S2 shows the STAT3 binding profiles in the promoter regions.

Among these 16 genes, some TF target genes have one or more strong binding peaks in their promoter regions (Fig. 2b), which can be identified by both TIP and peak-based methods. On the other hand, some TF targets are identified by TIP but not by peak-based methods (Fig. 2a). As shown, these target genes are also



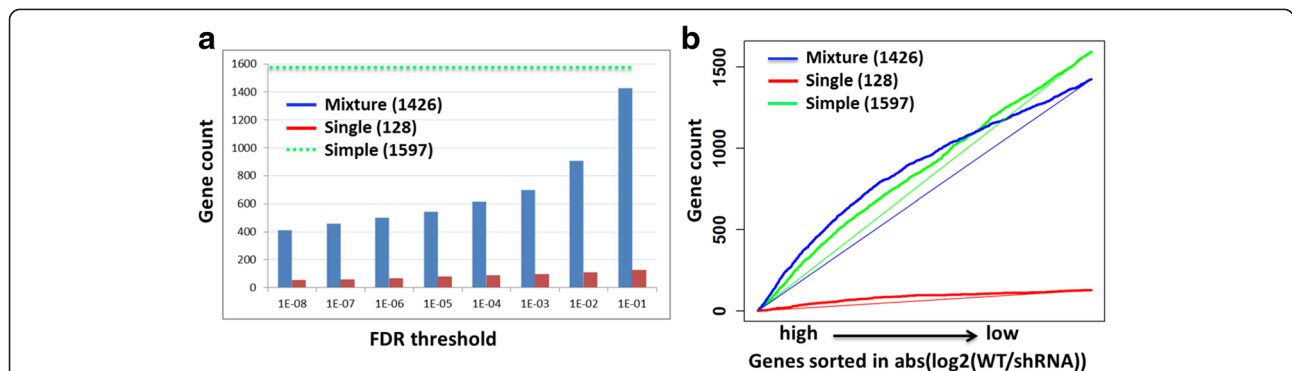
associated with one or more binding signals, but they are relatively weak and not identified as significant peaks by peak-calling methods. Thereby, TIP shows a higher sensitivity in detecting TF target genes with weak but considerable binding signals.

**The comparison of single normal distribution and mixture normal distribution**

To demonstrate the advantage of our extended TIP method with mixture normal distribution, we utilized the NFE2 ChIP-seq data and gene expression data treated by NFE2 shRNA in K562 cells from the ENCODE project [5]. We compared the identified NFE2 target genes using three methods as follows: simple method, TIP with single normal distribution and TIP with mixture normal distribution. In the simple method, we identified target genes using a conventional peak-based method: select target genes as those containing

one or more binding peaks in their proximal ( $\pm 1000$  bp of the TSS) promoter regions. The simple method identified 1597 target genes. A total of 128 target genes are identified for NFE2 by TIP at a FDR <0.1 significance threshold when employing a single normal distribution for significance testing. In contrast, the mixture normal based analysis identifies 1426 target genes at the same significance level. Figure 3a shows the number of significant target genes with various FDR thresholds of TIP algorithm. The TIP with single normal distribution is a stringent method and the numbers of significant target genes are stable but much less than the other methods, whereas the TIP with mixture normal distribution provides reasonable numbers of target genes.

For target validation, we calculated the expression changes (absolute value of log ratios) of all genes for the cells treated with shRNA as compared to the untreated cells, and then we sorted the genes in decreasing order



of absolute value of their log ratios. As shown in Fig. 3b, genes on the left have greater absolute value of log ratios (WT vs. shRNA) and are therefore more responsive to NFE2 regulation. Given a threshold of absolute log ratio ( $x$  axis), the  $y$  axis shows the number of target genes satisfied the absolute log ratio threshold for each methods. The area between the gene number curve and the straight line indicates accuracy of the target gene prediction. The area of the TIP with mixture normal distribution (the area between blue curve and blue straight line) has larger area than the other two methods. As shown, target genes identified by TIP are significantly more responsive to regulate by NFE2 than those identified by the conventional peak-calling method. Estimating  $p$ -value of genes based on mixture normal distributions increases the sensitivity of TIP, resulting in a larger target gene set than that based on the single normal distribution.

#### The comparison of read-coverage and fold-change signals of input data

To compare read-coverage and fold-change signals of input data, we downloaded the K562 Nfe2 ChIP-seq data and ChIP-seq control from the ENCODE project and then use bowtie to re-align the reads (parameters: `-m 1 -n 2 -S -a -best -strata`) [20]. Using bamCoverage and bamCompare tools of deepTools software (version 1.5.9.1) [21], we obtained signal files of the read coverage and fold change, respectively. 997 and 502 target genes were identified for read-coverage and fold-change signal files, respectively, by iTAR with mixture normal distribution (FDR <0.1). Additional file 1: Figure S3 shows the cumulative distribution of target genes identified by iTAR using read-coverage and fold-change signals. As shown, the target genes identified by iTAR using read-coverage signals is more responsive to NFE2 regulation.

#### Conclusions

With the rapidly increasing accumulations of ChIP-seq data, we refined the TIP algorithm to construct the web server, which offers a simple and user-friendly interface for experimental biologists. We believe that the iTAR web server will accelerate the research of gene regulation and help the experimental biologists to discover important biological mechanism. In addition, we provide a stand-alone JAVA package that users can choose to download and analyze their ChIP-seq data in their computers.

iTAR supports GO enrichment analysis for target gene list. In addition, users can download the predicted target gene list from iTAR, and submit to other interesting database.

#### Availability and requirements

Project name: iTAR

Project home page: <http://syslab3.nchu.edu.tw/iTAR/>

Operating system: Linux

Programming language: R and Java

Other requirements: R version  $\geq 2.10.0$ , R package: segmented, nor1mix (version  $\geq 1.2$ ) and mixtools.

Physical memory  $\geq 12$ GB

#### Additional file

**Additional file 1: Figure S1.** The density plot of normalized regulatory scores (Z score) using the ENCODE Stat1 ChIP-seq data in GM12878 cell line. **Figure S2.** The STAT3 binding profiles in the promoter regions of 16 histone proteins. **Figure S3.** Comparison of target genes identified by iTAR using read-coverage and fold-change signals. (DOCX 544 kb)

#### Abbreviations

ChIP-seq, chromatin immunoprecipitation followed by massively parallel DNA sequencing; FDR, false discovery rate; iTAR, a web server for identifying target genes of transcription factors using ChIP-seq or ChIP-chip data; TF, transcription factor; TIP, target identification from profiles; TSS, transcriptional start sites

#### Funding

This work is supported by the American Cancer Society Research grant IRG-82-003-27, the Centers of Biomedical Research Excellence (COBRE) grant GM103534, the start-up funding package from the Geisel School of Medicine at Dartmouth College (to C.C.), and the Taiwan Ministry of Science and Technology [MOST 104-2621-M-005-MY3 and MOST 104-2320-B-005-005] (to C.C.L.).

#### Availability of data and materials

The STAT3 ChIP-seq data in HeLa-S3 cells and NFE2 ChIP-seq data in K562 cells are downloaded from ENCODE project (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs>).

#### Authors' contributions

CC and CCL design the algorithm. CCY and EHA carried out the implementation of the algorithm. CCY, EHA, MG, CC and CCL prepared the manuscript. MHC and WYW participated in the design of the web server. JWC contributed to the literature study. All authors read and approved the final manuscript.

#### Competing interests

The authors declare that they have no competing interests.

#### Consent for publication

Not applicable.

#### Ethics approval and consent to participate

Not applicable.

#### Author details

<sup>1</sup>Institute of Molecular Biology, National Chung Hsing University, Taichung 402, Taiwan. <sup>2</sup>Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung 402, Taiwan. <sup>3</sup>Institute of Biomedical Sciences, National Chung Hsing University, Taichung 402, Taiwan. <sup>4</sup>Agricultural Biotechnology Center, National Chung Hsing University, Taichung 402, Taiwan. <sup>5</sup>Program in Computational Biology and Bioinformatics, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA. <sup>6</sup>Department of Molecular Biophysics and Biochemistry, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA. <sup>7</sup>Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, NH 03755, USA. <sup>8</sup>Geisel School of Medicine at Dartmouth, Institute for Quantitative Biomedical Sciences, Lebanon, NH 03766, USA. <sup>9</sup>Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Lebanon, NH 03766, USA.



Received: 26 June 2015 Accepted: 22 July 2016

Published online: 12 August 2016

**References**

- Butler JE, Kadonaga JT. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* 2002;16(20):2583–92.
- Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet.* 2006;7:29–59.
- Lemon B, Tjian R. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.* 2000;14(20):2551–69.
- ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science.* 2004;306(5696):636–640.
- Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489(7414):57–74.
- Sikora-Wohlfeld W, Ackermann M, Christodoulou EG, Singaravelu K, Beyer A. Assessing computational methods for transcription factor target gene identification based on ChIP-seq data. *PLoS Comput Biol.* 2013;9(11):e1003342.
- Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell.* 2008;133(6):1106–17.
- Ouyang Z, Zhou Q, Wong WH. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci U S A.* 2009;106(51):21521–6.
- Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008;9(9):R137.
- Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol.* 2008;26(12):1351–9.
- Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol.* 2009;27(1):66–75.
- Cheng C, Min R, Gerstein M. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics.* 2011;27(23):3221–7.
- Khaleel SS, Andrews EH, Ung M, DiRenzo J, Cheng C. E2F4 regulatory program predicts patient survival prognosis in breast cancer. *Breast Cancer Res.* 2014;16(6):486.
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res.* 2002;12(6):996–1006.
- Meng XL, Rubin DB. Maximum-Likelihood-Estimation Via the Ecm Algorithm - a General Framework. *Biometrika.* 1993;80(2):267–78.
- Liu Y, Li J, Sam L, Goh CS, Gerstein M, Lussier YA. An integrative genomic approach to uncover molecular mechanisms of prokaryotic traits. *PLoS Comput Biol.* 2006;2(11):e159.
- Maclsaac KD, Lo KA, Gordon W, Motola S, Mazor T, Fraenkel E. A quantitative model of transcriptional regulation reveals the influence of binding location on expression. *PLoS Comput Biol.* 2010;6(4):e1000773.
- Reich NC, Liu L. Tracking STAT nuclear traffic. *Nat Rev Immunol.* 2006;6(8):602–12.
- Timofeeva OA, Chasovskikh S, Lonskaya I, Tarasova NI, Khavrutskii L, Tarasov SG, Zhang X, Korostyshevskiy VR, Cheema A, Zhang L, et al. Mechanisms of unphosphorylated STAT3 transcription factor binding to DNA. *J Biol Chem.* 2012;287(17):14192–200.
- 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(3):R25.
- Ramirez F, Dunder F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* 2014;42(Web Server issue):W187–191.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

