W430–W436 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* *Published online 17 May 2011*
doi:10.1093/nar/gkr332

ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor

Jing Qin[1], Mulin Jun Li[1], Panwen Wang[1], Michael Q. Zhang[2,3] and Junwen Wang[1,*]

[1]Department of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong, 21 Sassoon Road, Hong Kong SAR, China, [2]Bioinformatics Division, TNLIST, Tsinghua University, Beijing 100084, China and [3]Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas at Dallas, Dallas, TX 75080, USA

Received February 7, 2011; Revised April 15, 2011; Accepted April 24, 2011

ABSTRACT

Chromatin immunoprecipitation (ChIP) coupled with high-throughput techniques (ChIP-X), such as next generation sequencing (ChIP-Seq) and microarray (ChIP–chip), has been successfully used to map active transcription factor binding sites (TFBS) of a transcription factor (TF). The targeted genes can be activated or suppressed by the TF, or are unresponsive to the TF. Microarray technology has been used to measure the actual expression changes of thousands of genes under the perturbation of a TF, but is unable to determine if the affected genes are direct or indirect targets of the TF. Furthermore, both ChIP-X and microarray methods produce a large number of false positives. Combining microarray expression profiling and ChIP-X data allows more effective TFBS analysis for studying the function of a TF. However, current web servers only provide tools to analyze either ChIP-X or expression data, but not both. Here, we present ChIP-Array, a web server that integrates ChIP-X and expression data from human, mouse, yeast, fruit fly and Arabidopsis. This server will assist biologists to detect direct and indirect target genes regulated by a TF of interest and to aid in the functional characterization of the TF. ChIP-Array is available at http://jjwanglab.org/ChIP-Array, with free access to academic users.

INTRODUCTION

Understanding the gene regulatory networks is critical to unraveling the complexity of various biological processes. A gene is regulated by transcription factors (TFs) which co-operatively interact at its regulatory region. Identification of the TF-target relationship is the first step in constructing a gene regulatory network. Several methods have been developed to study the TF-target relationship. For example, chromatin immunoprecipitation (ChIP) coupled with high-throughput techniques (ChIP-X), such as sequencing (ChIP-seq) or microarray (ChIP–chip), have been extensively used to map active TF binding sites (TFBSs) under specific conditions on a genome-wide scale. A known gene with TFBS around its promoter or enhancer region is usually considered as a direct target gene of the TF. However, TF knockdown studies show that a TF can activate, or suppress, or have no effect on the target genes (1). Although ChIP-X experiments can show active TFBS on the genes, this technique does not reveal the actual effect on the target gene's transcription. Moreover, the effect of a TF is usually manifested by changes in the gene expression, leading to a visible phenotypic change, after TF perturbation (knock-down/out or overexpression). Even though ChIP-X analysis reveals many potential direct targets, these targets are only a small portion of the genes affected by the TF. More methods are needed to map the TF-phenotype relationship, for example, finding the indirect targets of the TF.

Microarray technology has been widely used to measure the mRNA level changes of thousands of genes in the genome. The functions of a TF can be studied by observing expression changes of genes under the perturbation of the TF. Genes with expression changes that are considered to be caused by this perturbation can be direct or indirect targets of the TF. Web servers have been developed to analyze differentially expressed genes from microarray data. For example, CARRIE (2) can find TFs that have binding sites statistically overrepresented in the promoter regions of the differentially expressed genes. These statistical methods are based on the

*To whom correspondence should be addressed. Tel: +852 2819 2809; Fax: +852 2855 1254; Email: junwen@uw.edu

assumption that the co-expressed genes are regulated by a common TF or a set of TFs (3). In web servers using computational sequence-based methods (4–6), position weight matrices (PWMs) are used to scan the gene promoter for TFBS that are annotated in three popular databases: TRANSFAC (7), JASPAR (8) and UniPROBE (9). However, these web servers do not take into account TFBS binding in specific cell types and development stages.

ChIP-X technology has been used to identify the binding sites in an increasingly large number of TFs, and is more accurate compared to computational methods because it can detect TF binding in specific cell types and development stages. The binding not only depends on *cis* DNA sequence, but also on the chromatin structure of DNA and the expression level of the *trans* element under a specific cellular state. Lachmann *et al.* (3) compiled ChIP-X data from 87 publications into a TF-target interaction database, ChEA. Given a gene list, the ChEA database can find overrepresented TFs in their promoter regions. However, the number of TFs annotated in ChEA database is limited and the TFs are predefined. It does not provide a way to incorporate new ChIP-X and expression data, which restricts its usage on newly generated data, such as ChIP-X data of a new TF, or a TF in different cell types or conditions. Furthermore, ChEA analyses only the overrepresented TFs that directly regulate the differentially expressed genes. The relationships between perturbed TF and overrepresented TFs are unknown and differentially expressed genes could be regulated under regulatory networks through intermediate regulators (i.e. TFs). Methods are needed to find both direct and indirect target genes for a set of differentially expressed genes.

Both ChIP-X and microarray expression profiling methods experience high false positives. Due to their high throughput capability, each set of experiments usually generates over one thousand differentially expressed genes and thousands of ChIP-X target genes, the majority of which are false positives. Therefore, neither ChIP-X nor microarray expression profiling alone can reliably reveal the complex regulatory network of a TF. The combination of the two becomes a more effective tool for biologists to narrow down the list of true targets of a TF. One simple way to integrate the analysis is to use the intersection between the ChIP binding gene set and differentially expressed gene set. For example, the algorithm by Sharov *et al.* (10) identifies direct targets from this intersection, but it does not search for indirect targets, or considers sequence conservation of the binding sites, or provides binding information of co-occupied TFs, which restricts its general use for studying combinatory regulation and TF binding modules (5,11).

Currently available web servers for transcription regulatory network study only analyze either ChIP-X TFBS or expression profiling data, but not both. Here we present a web server, ChIP-Array, which can identify both direct and indirect targets based on both ChIP-X and microarray expression profiling data. Our ChIP-Array web server combines ChIP-X TFBS analysis, existing TFBS databases, TFBS conservation analysis and microarray expression profiling analysis. It provides biomedical researchers a convenient and efficient tool to construct the regulatory network of a TF and to obtain insights into the functions of the TF.

OVERVIEW OF CHIP-ARRAY

We developed ChIP-Array to help biologists analyze both ChIP-X and expression data together, and to construct a regulatory network controlled by a TF of interest. Data can be analyzed for five species: human, mouse, yeast, fruit fly and Arabidopsis. For ChIP-X and expression data of a given TF (or co-factor) X, the web server will: (i) find the direct targets of the TF by identifying the genes that are both differentially expressed and targeted by X; (ii) find indirect target Z by identifying an intermediate TF Y, which is both a putative regulator of Z and a target of X. The putative regulator of Z is identified by scanning all promoters in the genome with PWMs of all Ys from the three public accessible databases (JASPAR, UniPROBE and TRNASFAC derived TFBS database from UCSC genome browser); (iii) construct and display a regulatory network with all direct and indirect targets of TF X. Users can specify several parameters, such as the range of promoter regions, the cutoff P-values of the TFBS and its conservation score. Potential co-localized TFs in each target promoter can also be visualized in the network. Enrichment analysis can be done for the target genes of TFs in the network to assess the significance of their overlaps with the differentially expressed genes. Users can easily select a few targets identified by the web server for further experimental verification. Figure 1 describes the features of ChIP-Array web server.

DESCRIPTION OF CHIP-ARRAY

Input data

As shown in Figure 1a, the users need to input two sets of data, the 'binding location' from the ChIP-X data, and the 'differential expression' from the expression data. The interface also provides additional fields for users to select parameters, such as the definition of promoter regions, motif and conservation cutoff P-values, etc. The ChIP-X data can be in 'bed' or 'gff' formats, which shows the position of all identified peak regions or summits, or as peak files produced by the two popular CisGenome (12) and MACS (13) ChIP-X analysis softwares, or the input can simply be a list of genes with associated peaks. ChIP-Array generates a list of potential targets of the TF for the inputted peak files based on the distance between the transcription start site (TSS) of the closest gene (14) and the summit of peaks with cutoff distance specified by the users. Our system can cope with genome coordinates in the peak file of human genome assembly version hg19, mouse version mm9, yeast version sacCer2, fruit fly version dm3 and Arabidopsis version TAIR8. The expression data should be a list of genes that are differentially expressed. The data can be generated by high-throughput expression profiling experiments of genes under the perturbation of the TF, such as microarray or

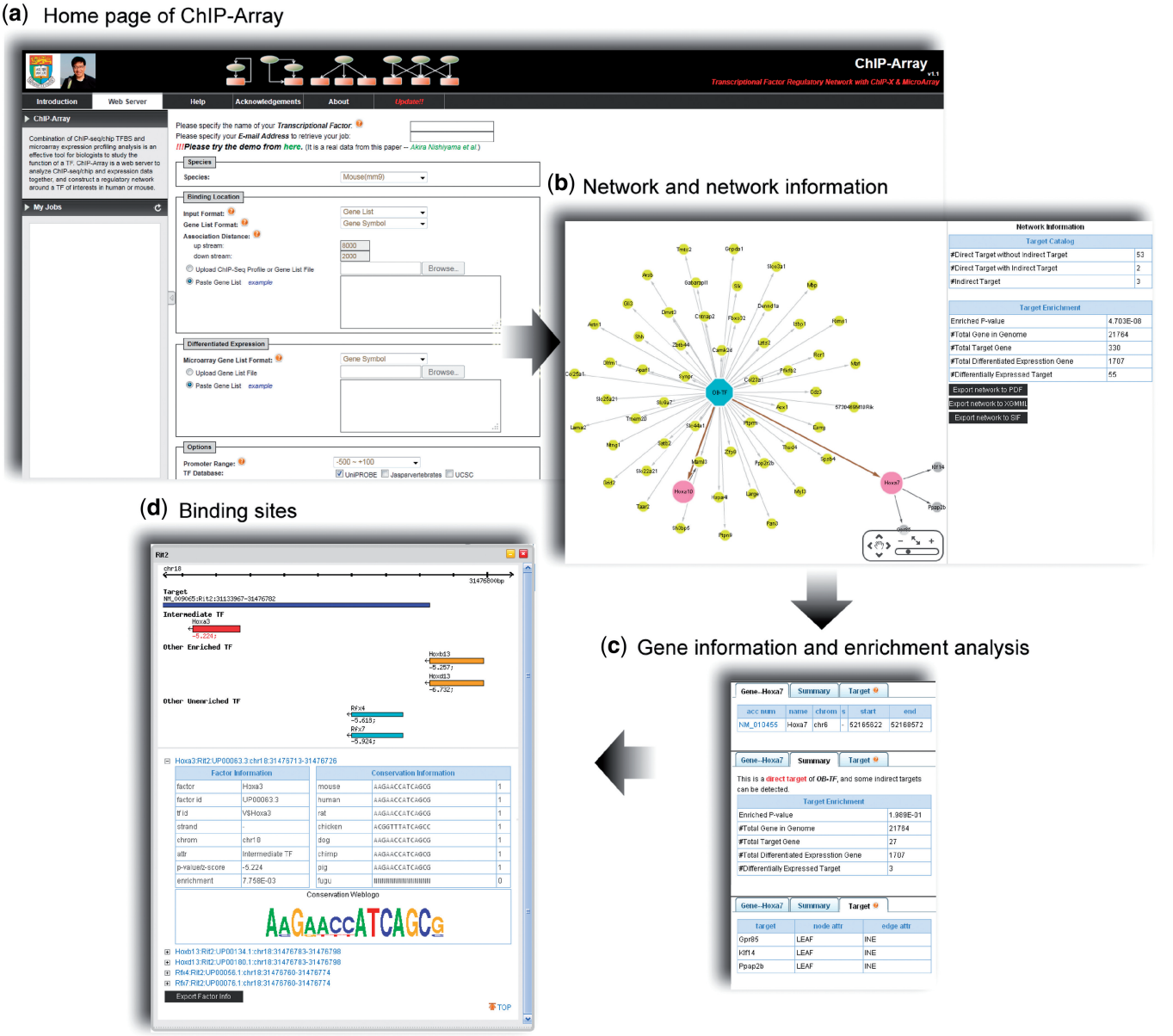

Figure 1. The overview of ChIP-Array web server. (**a**) Home page of ChIP-Array for data input and parameter selection, (**b**) Result page describing the transcriptional regulatory network, (**c**) Gene information including chromosome position, enrichment analysis and downstream targets and (**d**) Pop-up window showing the binding site information.

RNA-seq. The gene list can be obtained by bioinformatics analysis of these expression data, with tools such as 'affy' package in the *R* (15) and Bioconductor (16) programs. Several gene identifiers, such as gene symbol, accession number and microarray probe ID (from three major vendors, Affymetrix, Illumina and Agilent) are supported by the web server. The users should select the gene identifier and corresponding species. If the gene ID provided by the user does not match our database, the server will generate warning messages and allow user to remove, replace and modify the unmatched ID. If >20% of the IDs do not match, we determine that the type gene ID is not supported by the server. In such case, we recommend the users to use more sophisticated tools such as DAVID gene name conversion tool (17) to convert gene IDs into the gene type we supported before precede.

The user provides the name of the TF under investigation to be used for network visualization; otherwise the default name 'OB-TF' will be used. The users can adjust several other parameters to search for the intermediate target *Y*, for example, the range of promoter, which specifies the region around the TSS to be searched. The statistical significance of the putative motif can be specified by the *P*-value cutoff. The multiple species conservation was filtered by a conservation cutoff *P*-value, which measures the sequence identity of genomic

alignment for the putative TFBS compared to the random motif with the same length. Users can select from one or multiple available TF databases.

We have provided example data sets for users to test out the web server. The user can either click on 'try the demo here' to visualize the network generated from the example data, or click on 'example' in both 'Binding Location' and 'Differentiated Expression' fields to load the example data.

Running procedure

The web server queues the submitted job for processing, and the user can see how many jobs are waiting in the queue and the time the job has been run. A permanent link for the job is also provided so that the users can retrieve the result when the job is finished. The 'my jobs' on the left side panel provides a link to each of the jobs submitted by current user in the past 2 weeks. By clicking on the link, the user is able to retrieve the previously submitted jobs.

The server uses the inputted ChIP-X and expression data of a TF to find direct and indirect targets for the TF. The direct targets are identified as the intersection between the gene list from ChIP-X and the gene list from differential expression data. The indirect targets of TF *X*, for each intermediate TF *Y*, are identified by scanning the promoters of all differentially expressed genes for potential binding sites. Gene *Y* must be a TF and must have its PWM annotated in the selected database(s). For human and mouse, the JASPAR vertebrates and UniPROBE contain PWMs of 113 and 165 TFs or TF complexes respectively. The TFBS track of UCSC genomes browser provides potential conserved binding sites of 166 TFs or TF complexes in human genome (hg18). The binding sites are mapped to human genome version hg19 and mouse genome version mm9 by liftOver. JASPAR fungi, MacIsaac *et al.* (18) and SCPD (19) provide 177, 122 and 24 PWMs respectively for yeast. Fruit fly and Arabidopsis have 125 and 21 PWMs in JASPAR insects and plants respectively.

The binding site is scanned using PWMSCAN (20) and the significance of the binding site is measured by the *P*-value, which is calculated through a permutation-based method, FastPval (21). A sequence hit with *P*-value less than the user specified cutoff is considered as a putative binding site. Putative binding sites are further filtered based on their conservation scores, which is evaluated by another permutation-based test. Binding sites with a conservation *P*-value less than the user specified value are accepted as potential conserved binding sites of the TF *Y*. Genes that possess potential conserved binding sites of the TF *Y* in their promoter region are considered as the potential targets of *Y*. The sequence of each binding site is extracted from the corresponding genome. Alignments of the binding site with other species were extracted from multiple alignments of corresponding genomes. Multiple alignment files have been downloaded from the UCSC genome browser http://hgdownload.cse.ucsc.edu/downloads.html. The sequence identity between aligned sequences and the binding site sequence are calculated. This sequence is then compared with a random background to obtain a conservation *P*-value, which indicates the likelihood of finding a random site of identical length with equal or high sequence identity among the different species.

A network is then constructed of all the identified direct and indirect targets using Cytoscape Web (22), as shown in Figure 1b. The network visualizes the regulatory relationships among TF *X* and its direct and indirect targets. Each node is a gene and each edge is an arrow pointing out from a TF to its target, which indicates the regulatory function of the TF to its target. An intermediate TF *Y* can regulate itself when it has a putative binding site in its own promoter. An arrow from a TF pointing to itself shows a self-regulatory function. An arrow from an intermediate TF pointing to another direct target of TF *X* shows that it can regulate a direct target with a binding site co-localized in the TF *X* targeting promoter, indicating that both TF *X* and *Y* might regulate target *Z*. The user can adjust the visualization with the 'navigation' button in the lower-bottom corner of the graphic panel. The transcription regulatory network can be downloaded as a graph in PDF format, eXtensible Graph Markup and Modeling Language (XGMML), or in Simple Interaction Format (SIF).

The statistical significance of the overlap of TF targets and differentially expressed genes is calculated using a hypergeometric test. The estimated *P*-value is calculated for TF *X* and each intermediate TF *Y*. The statistical results are displayed in the right-upper panel (Figure 1b).

When the user clicks on any node (gene) on the graph, an information pop-up window (Figure 1d) will show all the putative binding sites in the promoter region of the gene and the right-upper panel will display general information about this gene including accession number and chromosome position (Figure 1c). If a TF has targets, a 'Summary' tag will give the number of targets and the results of enrichment analysis, and a 'Target' tag will show a list of the names of targets. The upper half of the pop-up window illustrates the gene's coordinate in the genome and location of all the putative binding sites. Clicking on a binding site will display detailed information about: (i) the binding site including the TF's name and ID, binding site *P*-value and enrichment *P*-value; (ii) multiple species conservation, which shows the sequences in other species and the sequence identities; and (iii) a conservation sequence logo based on multiple alignments of the binding sites. The user is also able to explore co-occupied TFs from this pop-up window.

Examples

The example we used is the network regulated by Cdx2. This TF is not expressed in embryonic stem cells (ESCs), but induces dramatic transcriptome perturbation and cell differentiation when it is overexpressed in ESCs (23). We have re-analyzed the results from the study by Nishiyama *et al.* The ChIP-Seq input data includes 22 219 peaks with at least nine tags, and peaks are located in 1680 gene promoters within 8000 bp upstream and 2000 bp downstream of TSS. The expression input data contains 2976 genes

whose expressions change at least 2-fold in 48 h after Cdx2 overexpression in ESCs. Selected parameters are −8000 to +2000 for the promoter range, 10E-5 for PWM scan *P*-value and 0.001 for conservation filtering *P*-value. All of the three databases were selected to include the maximum number of TFs. Of the 263 genes identified as direct targets, 16 TFs have indirect targets that are differentially expressed. Enrichment analyses show that the overlap between the targets of 13 TFs and the differentially expressed gene set is significant (P-value < 0.05 in hypergeometric test), which indicates that the overlapping does not happen randomly. Using ChIP-Array, Hox genes are identified as direct targets of Cdx2, which is consistent with the findings of Nishiyama *et al*. Among the 28 Hox genes differentially expressed under Cdx2 overexpression, 10 are direct targets and 14 are indirect targets in our network. Also, 9 of the 10 directly regulated Hox genes have significant downstream targets that overlap with the differential expression gene set. These results further confirm the importance of Hox genes as a transcription regulatory TF set that combines downstream of Cdx2. Of the 2976 differentially expressed genes, 931 (31%) genes are identified as direct or indirect targets, which demonstrate the power of ChIP-Array to explore the regulatory connection between TFs and targets from both ChIP-X and expression profiling data. The GO annotation enrichment analysis by DAVID (17) of direct and indirect targets explains the effect of Cdx2 overexpression on the cells. Using GO terms for the Cdx2 targets, transcription regulation and embryonic organ morphogenesis are enriched in direct targets, whereas cell fate commitment and specific tissues' development are revealed for indirect targets. This demonstrates the early effect of Cdx2 on TF transcription and embryonic development and its further effects on cell differentiations (Supplementary Data).

Server design

The ChIP-Array web server is implemented in Perl and based on the Catalyst web framework. It offers an intact Model-View-Controller design pattern with web development, which facilitates the expansibility and efficiency. Non-computational data is stored in a MySql database as additional information for the TFs and genes. We use Oracle Grid Engine to manage the submitted jobs. The finished job is stored on our server for 2 weeks, and there are three ways for users to retrieve their jobs by browser cookies, fixed links and email notifications. The program runs on a high-end computing cluster with powerful computational performance. ChIP-Array is freely available for academic use and there is no registration required.

DISCUSSION

By integrating ChIP-X and gene expression data, our web server can dissect the transcriptional regulation function of a TF by identifying its direct and indirect targets. Recently, the integration of wet lab experimental data and bioinformatics analysis is becoming a popular method to characterize the function of a TF. ChIP-Array provides a convenient and effective tool for biologists to interpret their high-throughput data. The combinatory analyses and two-layer target identification approach of ChIP-Array takes advantage of ChIP-X binding site identification, gene expression profiling analysis and existing TFBS databases to provide the user with a comprehensive understanding of the regulatory network targeted by a TF of interest.

The recent studies that focused on discovering direct targets of TFs using ChIP-X data and gene expression data could only characterize 6–17% of the differentially expressed genes in 12–72 h after TF perturbation as direct targets with a cutoff distance of <20 kb between TSSs and ChIP-X peaks (10,24,25). Generally, the expression profiles after 48 h of TF perturbation are considered to be the best for direct target detection (10), however, the proportion of direct targets identified is far less when compared to the total number of differentially expressed genes. This suggests that the majority of genes with expression changes were indirectly affected. Our ChIP-Array can more deeply analyze the data to disclose indirect regulation. Indirect target analysis not only uncovers more regulatory pathways of genes with expression changes, but also reveals the functions of many direct targets by their downstream targets. This two-layer target identification method provides a more comprehensive and clearer view of the regulatory network of a TF of interest.

ChIP-X analysis has been used in an increasing number of high-throughput genome-wide studies. Because ChIP-X data are development stage and cell type specific, integrating the mRNA expression data and ChIP-X data requires the respective datasets to be obtained under the same or at least similar experimental conditions. However, to cater for different biological researchers' demands would mean detecting the TFBSs of all TF in all conditions, which is an unachievable task. By contrast, computational sequence-based methods are not condition-specific but are less accurate than ChIP-X analysis. However, they do provide more comprehensive candidate TFBSs when condition-specific ChIP-X data are lacking. We implement both types of systems into our web server to create a general approach for the functional study of any TF in any condition. The inputted ChIP-X data needs to be generated in the same or at least similar conditions as those of expression profiling data. For indirect target searching, we apply computational sequence-based methods to meet the requirements for general use for a given specific condition where ChIP-X TFBS information of most TFs is lacking. This approach retains the condition specificity of ChIP-X data, while at the same time maximizing the TFBS information by using conditional non-specific methods. Although the computational predicted TFBS contains a relatively higher false discovery rate, restriction of indirect target genes within a differentially expressed gene set diminishes the unrelated TFBS, and the stringency of scanning and conservation filtering (can be controlled by the user) also reduces the false positives. In addition, the enrichment analysis for TFs evaluates how significant the TFs regulate those indirect targets.

In the future, we plan to collect more data from reported experiments integrating ChIP-X and mRNA expression profiling of genes under TF perturbation. Using our web server, the regulatory network regulated by each TF in a certain condition can then be constructed. The combination of TFs in a specific cell type or developmental stage, which is well-accepted to control the transcriptome, can easily be identified by expression profiling. The regulatory networks of the TFs involved in protein–protein interactions can be incorporated into our system to reveal the pathways of how the transcription of each gene is regulated.

In summary, the ChIP-Array web server provides biological researchers an efficient platform for TF regulatory network analysis, which will enable them to gain further insights into the TF function. Furthermore, accumulation of transcription regulation networks of different TFs and TF sets in the future will provide us with a global understanding of transcription regulation in biological processes and developments.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank our colleagues at The University of Hong Kong who tested the web server and provided us with invaluable suggestions.

FUNDING

University Postgraduate Fellowship [to J.Q.] from The University of Hong Kong; Centre for Reproduction Development and Growth of The University of Hong Kong [start up fund to J.W.]; General Research Fund [grant number 778609M to J.W.] from the Research Grants Council of Hong Kong; The National Natural Science Foundation of China and Research Grants Council of Hong Kong Joint Research Scheme [grant number N_HKU752/10 to J.W. and M.Q.Z.]. Funding for open access charge: General Research Fund [grant number 778609M to J.W.] from the Research Grants Council of Hong Kong.

Conflict of interest statement. None declared.

REFERENCES

1. Loh,Y.H., Wu,Q., Chew,J.L., Vega,V.B., Zhang,W., Chen,X., Bourque,G., George,J., Leong,B., Liu,J. *et al*. (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet*., **38**, 431–440.
2. Haverty,P.M., Frith,M.C. and Weng,Z. (2004) CARRIE web service: automated transcriptional regulatory network inference and interactive analysis. *Nucleic Acids Res*., **32**, W213–W216.
3. Lachmann,A., Xu,H., Krishnan,J., Berger,S.I., Mazloom,A.R. and Ma'ayan,A. (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, **26**, 2438–2444.
4. Hannenhalli,S. (2008) Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics*, **24**, 1325–1331.
5. Wang,J.W., Zhang,S.L., Schultz,R.M. and Tseng,H. (2006) Search for basonuclin target genes. *Biochem. Biophys. Res. Commun*., **348**, 1261–1271.
6. Wang,J.W. and Hannenhalli,S. (2006) A mammalian promoter model links cis elements to genetic networks. *Biochem. Biophys. Res. Commun*., **347**, 166–177.
7. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*., **31**, 374–378.
8. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*., **32**, D91–D94.
9. Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res*., **37**, D77–D82.
10. Sharov,A.A., Masui,S., Sharova,L.V., Piao,Y., Aiba,K., Matoba,R., Xin,L., Niwa,H. and Ko,M.S. (2008) Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. *BMC Genomics*, **9**, 269.
11. Vardhanabhuti,S., Wang,J.W. and Hannenhalli,S. (2007) Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res*., **35**, 3203–3213.
12. Jiang,H., Wang,F., Dyer,N.P. and Wong,W.H. (2010) CisGenome browser: a flexible tool for genomic data visualization. *Bioinformatics*, **26**, 1781–1782.
13. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al*. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol*., **9**, R137.
14. Wang,J.W., Ungar,L.H., Tseng,H. and Hannenhalli,S. (2007) MetaProm: a neural network based meta-predictor for alternative human promoter prediction. *BMC Genomics*, **8**, 374.
15. Gautier,L., Cope,L., Bolstad,B.M. and Irizarry,R.A. (2004) affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.
16. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.
17. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc*., **4**, 44–57.
18. MacIsaac,K.D., Wang,T., Gordon,D.B., Gifford,D.K., Stormo,G.D. and Fraenkel,E. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.
19. Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*, **15**, 607–611.
20. Levy,S. and Hannenhalli,S. (2002) Identification of transcription factor binding sites in the human genome sequence. *Mamm. Genome*, **13**, 510–514.
21. Li,M.J., Sham,P.C. and Wang,J. (2010) FastPval: a fast and memory efficient program to calculate very low P-values from empirical distribution. *Bioinformatics*, **26**, 2897–2899.
22. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.
23. Nishiyama,A., Xin,L., Sharov,A.A., Thomas,M., Mowrer,G., Meyers,E., Piao,Y., Mehta,S., Yee,S., Nakatake,Y. *et al*. (2009) Uncovering early response of gene regulatory networks in ESCs

by systematic induction of transcription factors. *Cell Stem Cell*, **5**, 420–433.
24. Fang,X., Yoon,J.G., Li,L., Yu,W., Shao,J., Hua,D., Zheng,S., Hood,L., Goodlett,D.R., Foltz,G. *et al.* (2011) The SOX2 response program in glioblastoma multiforme: an integrated ChIP-seq, expression microarray, and microRNA analysis. *BMC Genomics*, **12**, 11.
25. Parisi,S., Cozzuto,L., Tarantino,C., Passaro,F., Ciriello,S., Aloia,L., Antonini,D., De Simone,V., Pastore,L. and Russo,T. (2010) Direct targets of Klf5 transcription factor contribute to the maintenance of mouse embryonic stem cell undifferentiated state. *BMC Biol.*, **8**, 128.