EpiRegNet: Constructing epigenetic regulatory network from high throughput gene expression data for humans
Lily Yan Wang, Panwen Wang, Mulin Jun Li, Jing Qin, Xiaowo Wang, Michael Q. Zhang & Junwen Wang
Published online: 01 Dec 2011.

To cite this article: Lily Yan Wang, Panwen Wang, Mulin Jun Li, Jing Qin, Xiaowo Wang, Michael Q. Zhang & Junwen Wang (2011) EpiRegNet: Constructing epigenetic regulatory network from high throughput gene expression data for humans, Epigenetics, 6:12, 1505-1512, DOI: 10.4161/epi.6.12.18176

To link to this article: http://dx.doi.org/10.4161/epi.6.12.18176

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions
EpiRegNet
Constructing epigenetic regulatory network from high
to throughput gene expression data for humans

Lily Yan Wang,1,† Panwen Wang,1,† Mulin Jun Li,1 Jing Qin,1 Xiaowo Wang,2 Michael Q. Zhang2,3 and Junwen Wang1,*

1Department of Biochemistry; LKS Faculty of Medicine; The University of Hong Kong; Hong Kong; 2Bioinformatics Division; TNLST; Tsinghua University; Beijing, China; 3Department of Molecular and Cell Biology; Center for Systems Biology; The University of Texas at Dallas; Dallas, TX USA

†These authors contributed equally to this work.

Key words: epigenetics, network, histone modification, gene regulation, microarray, RNA-seq

Abbreviations: TF, transcription factor; TFBS, transcription factor binding site; TSS, transcription start site; HMM, histone modification mark

High throughput profiling methods, such as microarray gene profiling and RNA-seq, have generated thousands of data sets that measure gene expression changes at the whole genome level. However, from gene expression changes, we could only observe the phenomenon of the biological process. To decipher the underlying mechanisms of these phenomena, we have to understand the factors regulating gene expression. Understanding gene regulation is key to unraveling the mechanism of many biological processes, including cell development, lineage commitment and differentiation, and pathogenesis of cancers. Since gene regulation is controlled at both genetic and epigenetic levels, it is necessary to construct both genetic and epigenetic regulatory networks for these differentially expressed genes.

At the genetic level, the interaction between transcription factors (TFs) and their binding sites (TFBSs) in the promoter or enhancer regions determines the time, location and the level at which target genes are expressed. Extensive studies have been done to pinpoint the genetic determinants of gene regulation, with both experimental1 and computational2,3 approaches. Interactive applications and web services have been developed to infer regulatory networks based on transcription factor binding sites enriched in the gene promoter regions, based on either experimental ChIP-Seq data4,5 or computational predicted data.6

At the epigenetic level, histone modifications and DNA methylation can affect the chromatin structure and thus activate or repress gene expression. Genome wide profiles of histone modifications have been studied extensively and their effects on gene regulation have been reported. For example, Barski et al.7 and Wang et al.8 performed genome wide histone acetylation and methylation profiling in CD4+ T cells and discovered strong associations between gene expression level and histone modification modules in the promoter regions. Karlic et al.9

www.landesbioscience.com

Epigenetics 1505

DOI: 10.4161/epi.6.12.18176

*Correspondence to: Junwen Wang; Email: junwen@uw.edu
Submitted: 08/10/11; Revised: 09/16/11; Accepted: 09/21/11
and Cheng et al.\textsuperscript{10} were able to predict gene expression based on histone modification profiles in worm and human, respectively, with computational models. These studies indicated that regulation of a gene was also affected by its epigenetic status.

However, there is no web server reported so far that is able to construct epigenetic regulatory networks from gene expression data. Furthermore, as the Roadmap Epigenomics Project is producing large amount of genome wide epigenetic data,\textsuperscript{11} we will have abundant data sets to link epigenetic landscapes to their target genes in different cell types. Here we present an interactive web, named EpiRegNet, to fit this purpose. EpiRegNet can construct epigenetic regulatory networks for gene expression data by incorporating either our locally curated epigenetic and TFBS data or data provided by the users.

Given different categories of genes, we hypothesize that there are epigenetic factors responsible for gene expression differences among the categories. The server therefore aims to search for the promoter/enhancer region of these genes and find the epigenetic factors that most likely contribute to the differences, and then construct an epigenetic regulatory network and visualizes it in a straightforward graphic interface. The server also provides information on transcription factors that co-occur with the epigenetic factors. Furthermore, it can draw a heatmap to demonstrate the cooperative or competitive relationship among these factors in activating or repressing their target gene's expression. Finally, we present three case studies to show how our system could be applied to discover the mechanisms of transcriptional regulation involving DNA-protein interaction and chromatin structure.

Results

Implementation of EpiRegNet. EpiRegNet takes a list of gene names with their labels as initial input (Fig. 1A). These gene names can be in several formats, including RefSeq id, official gene symbol, and microarray probe id from Affymetrix, Agilent or Illunima. Labels are decided by users, depending on how they categorize the genes. For example, for differentially expressed genes, labels could be tagged as “1” for upregulated genes, “-1” for downregulated and “0” for not differentially expressed genes (as in case study 1). For genes expressed in different levels in one cell line, labels could be tagged as “1” for highly expressed genes, “0” for moderately expressed genes and “-1” for lowly expressed genes (as in case study 2). In the case that the user only inputs one category of genes or genes are unlabeled, the server will treat all these genes as one category and randomly pick same number of genes from all other human genes as the other category and continue the analysis.

Since epigenetic regulations are cell specific, users need to choose a cell type for their analysis. User can either choose one from the 12 cell types annotated in our database or create a self-defined cell type. For the annotated cell type, user can select histone marks from our database; for the user specified cell type, user has to upload histone modification data from their own experiments, such as ChIP-seq. It is possible the uploaded mark name is among the marks of the same cell line curated in the web server and when it is incorporated in the network it could be analyzed with the corresponding curated mark at the same time. For user-defined marks, we add an underscore as prefix to the mark name to denote it is defined by user and the data are newly uploaded (as shown in case study 3 resulting network and heatmap figure). It is also required to select the region of promoter to be analyzed. The region ranges are defined by the location relative to transcription start site (TSS). Three options are provided: -500 bp to +100 bp, -2,000 bp to +500 bp and -8,000 bp to +2,000 bp. Finally, user needs to set parameters that determine the number of epigenetic marks and genes to be displayed in the network. “Mark selection p value” determines the p level of following statistical tests and the “Gene selection” determines how many genes will be displayed for each mark (Fig. 1A).

With inputs from the user, the server will utilize our local database to find the (epi)genetic factors that potentially regulate these genes. Currently, our database contains histone modification marks and TFBSs for 12 cell types. It includes ChIP-seq binding sites for 170 epigenetic marks and 73 TFs collected from public available experimental data and computational predicted TFBSs for 419 TFs based on the procedures described previously in reference 12 and 13. We hypothesize that if the sets of genes are regulated by a certain factor, the frequency/strength of this factor in the promoter/enhancer region of these genes will be significantly different from one category to another. For the epigenetic factors that regulate the inputted genes with p value less than user specified α level (see Materials and Methods for details), the server will draw the epigenetic regulatory network in the “network” page (Fig. 1B). This network is composed of selected histone modification marks and up to ten genes with highest signals for each mark. Marks are connected with lines, the thickness of which demonstrates their degree of correlation; each mark is connected to genes for which their promoter regions are highly occupied by this particular mark. When a gene is clicked in the network, a pop-up window will show mark signals and all detected TFBSs around this gene’s promoter.

In addition, user may be interested in how the marks colocalize with each other to regulate genes. The information is displayed in the “heatmap” page (Fig. 1C). Pairs of marks could function cooperatively or competitively in the network, depending on whether they have a significant correlation coefficient among their binding sites (see Materials and Methods for details). In both “network” and “heatmap,” by clicking any histone modification mark a pop-up window will display functional transcription factors and their correlative relations with the specific histone modification mark (Fig. 1D).

The above-mentioned functions focus on the static data for multiple histone modification marks and are implemented in the “Static” page of our web server. However, users may want to analyze dynamic change of a particular histone modification mark under different cell conditions; therefore, we have built an additional “Dynamic” page for this purpose. Given the epigenetic status at two cell conditions and their corresponding gene expression change, the server aims to find how the epigenetic mark pattern is changed in the promoter region of the gene, and how it consequently changes the target gene’s expression level. In addition, the server can find intermediate TFs that are regulated
has been shown that to change a cell’s function and type, the intrinsic change of gene profiles is crucial. It was also found that histone modification profiles were changed with cell type changing. Because histone modification is involved in transcription regulation, it is reasonable to believe that these histone modifications are tightly correlated with cell differentiation. We chose 142 genes that have similar expressions in ESC (with RPKM values between three and five), but different expressions in CD4+ T cell, and then divided them into two groups, named group “U” and group “D,” based on whether their expression levels are upregulated or downregulated in CD4+ T cells. With 0.005 as $\alpha$ level, our analysis showed that no histone modification enrichment was significantly different in ESC (Fig. 2A), but more than 10 histone modifications were significantly different in CD4+ T cell (Fig. 2B), including

by the mark, and at the same time regulating other differentially expressed genes. The server will present the direct and indirect target genes, as well as the histone modification mark and intermediate TFs in an integrative network. The network will be displayed in an user friendly interface, with the histone modification mark as the hub, and target genes as leafs. The meaning of each element in the network is explained on the website.

**Case Study**

Gene expression profiles are altered dramatically from ESC to differentiated cell partly due to change in histone modification pattern. It has long been known that tissue specificity of the cell is determined to a large extent by the unique gene expression profile. In various multiple tissues studies, it

![Figure 1. Graphical interface of EpiRegNet and the result pages. (A) Input gene list and set parameters. (B) Network in result page. Blue nodes are top ten functional histone modification marks (HMMs) with p values less than $\alpha$. The size is inversely proportional to its p value from enrichment test. Yellow nodes are genes which are most enriched with HMMs. The size is proportional to the number of HMMs which regulate it. Blue lines between HMM nodes stand for their correlative relations. They are shown only if the coefficients have absolute values greater than 0.8. Red lines between HMM and gene nodes denote this gene is highly regulated by the mark. Lines with deeper color and more thickness denote stronger relations. (C) Heatmap in result page. Marks in pink are functional ones with p values less than $\alpha$, while white ones are nonfunctional. Marks in red are arbitrarily selected by user. Squares in the matrix explain the correlative relation between a pair of HMMs. When the mouse hovers on one square, the corresponding marks would be highlighted with cyan color in row names column and the p value and correlation coefficient would be displayed. Deeper color stands for stronger relation (red for positive correlation, while blue for negative). (D) Pop-up window in result page showing functional transcription factors. In the table, in second column the p values are for TFs, to denote whether the TF contributes to gene expression differentiation. In third column, p values and coefficients from the correlation tests between these TFs and the chosen HMM are shown (the coefficient is not given if p value is greater than $\alpha$ level). Deeper color stands for stronger relation (red for positive correlation, while blue for negative).](https://www.landesbioscience.com/epigenetics/1507-01.png)
We also performed one-sided Wilcoxon tests on H3K9me1/2/3, another reported group of repressive histone modification. Similar to H3K27 methylation, we found that H3K9me2/3 were significantly deprived in genes of group U compared with those of group D (p < 10^{-3}), while H3K9me1 tended to be more enriched in group U genes (p < 10^{-3}). In summary, our system could effectively detect the histone modification marks that are reported to have activation or repression roles in gene expression. Besides, we discovered novel insights that indicate that, for the repressive histone modification marks H3K4 and H3K27 methylation, the number of methyl groups added to histone tails should be taken into consideration when exploring their specific roles in regulation.

The role of “modification backbone” in regulating gene expression. A set of histone marks, termed “modification backbone,” which is composed of 16 histone modifications and H2A.Z, is present in most of the gene promoters. According to

![Figure 2. Result for case study one (A) is the result of all ESC histone modifications, and (B) is that of signals in CD4+ T cell (here only 23 out of 38 histone modifications are shown). Histone modifications are colored pink if they contribute to gene expression differentiation based on enrichment test results.](image-url)
Wang et al.\(^8\) genes with this backbone are expressed in a higher level compared with genes without this backbone. To confirm this observation, we used our system to analyze the expression level of genes based on the RNA-seq data,\(^6\) and found that the marks in this “backbone” were all of great importance in transcriptional regulation. We calculated the expression levels of around 30,000 RefSeq genes according to the RPKM score calculation method,\(^{18}\) and separated the genes into three equally sized groups: highly, lowly and medially expressed genes. From each of the three groups, we selected 100 genes randomly and analyzed all 38 histone modification marks in CD4\(^+\) T cells using our server. The marks were ranked by ascending p values and all the “backbone” histone modification marks were on top of the list, showing significance in determining gene expression levels (Fig. 3A and \(p < 10^{-3}\), Kruskal-Wallis tests). Similar results were observed by adjusting sample size to 900 and 1,500 genes. In addition, we found that some histone modifications that are not “backbone” component, including H3K27me3, were shown as being significant in repressing gene expression (Fig. 3B). The finding was in accordance with Wang et al.’s conclusion that genes were not only activated by the backbone modules but also repressed by other non “backbone” marks, which may act cooperatively with “backbone” marks to affect the chromatin state, and consequently affect gene expression.\(^8\) As a control test, we shuffled these 300 genes and regrouped them into three groups for the analysis. We found that none of the histone modifications was significant (\(\alpha = 0.01\)), which confirmed that our observation was not by chance.

**Histone modifications in human pancreatic islets analyzed with user-uploaded ChIP-seq data.** Our server also accepts...
The resulting network (Fig. 4B). In the core promoter (-500 bp to +100 bp), H3K4me1 did not show significant difference in enrichment between highly expressed genes and lowly expressed genes, while when larger promoter regions were used (-2,000 bp to +500 bp and -8,000 bp to +2,000 bp), H3K4me1 differences were significant in distinguishing the two groups. This observation suggested that H3K4me1 occurred more in the distant regulatory elements rather than core promoters. Third, co-occurrence between H3K4me2 and H3K4me3 was constantly observed in the heatmap page (Fig. 4A and B). These two histone modification marks had strong correlation (p < 0.05, Pearson's correlation coefficient r > 0.7), regardless of the type of promoter region selected. In contrast, H3K4me1 was less correlated with either of them under all circumstances. In addition, we found that H3K27me3 has either negative correlation with H3K4 methylations or no correlation. In summary, we were able to confirm all conclusions in Bhandare et al.'s paper. Our server is therefore uploaded histone modification data by the user and performs the analysis in the same pipeline. In this case study, we uploaded four types of pancreatic histone modification data from Bhandare et al. into our system. We selected 100 highly expressed and 100 lowly expressed genes from the Microarray results of Bhandare et al. to form the gene list. Through the resulting network and heatmap pages, we observed several phenomena that were consistent with those reported in Bhandare et al. First, all the four histone modification marks were significant in regulating gene expression (p < 0.05, Fig. 4A). As expected, H3K4 methylations tended to activate gene expression as we observed genes with strongest H3K4 methylation signals were highly expressed. On the contrary, H3K27me3 was suggested to repress gene expression as H3K27me3 occurred more in the lowly expressed genes. Second, H3K4me1 was correlated with enhancers rather than promoters. When we chose different ranges of promoter region, H3K4me1 was not always picked out as significant mark in the resulting network (Fig. 4B).
effective and efficient in exploration of histone modification and gene expression data to build epigenetic regulatory networks.

Discussion

Many studies have been conducted on the regulatory network composed of transcription factors. However, the majority of the research either focuses on a small number of transcription factors in a certain type of tissue, or emphasizes a large number of transcription factor binding motifs obtained from computational predictions without considering tissue specificity. Different from transcription factors, histone modification data are tissue specific. In our system, we have curated histone modification profiles in 12 cell lines. The resulting transcriptional regulatory network based on these data would be quite elaborate without missing any important histone modifications that contribute to the network to a great extent. Besides, the majority of the data were generated under similar circumstances and processed using the same protocol, since in each whole genome mapping experiment in a specific cell line, all possible histone modification profiles were produced simultaneously. This is different from transcription factor binding sites detection experiments, in which the number of transcription factors studied per experiment is limited. Thus, we believe that the data on histone modifications used in this system are homogeneous, given that they were derived from the same cell line at the same stage of development. In addition, data were normalized using the same method.

Histone modification has received increasing attention in multiple fields due to its wide effects in cellular processes, among which transcriptional regulation is most intensively studied. Genome wide mapping of certain types of histone modification has produced many available data for various cell lines.\(^1\,7\,8\,16\,17\,20\,22\) Meanwhile, visualization of these signals in the region of an individual gene could be realized with the help of tools such as UCSC genome browser. Unfortunately, for gene expression data with information for hundreds or thousands of genes, no application is available to analyze the role histone modification plays in the transcriptional regulation and differential gene expression. Our system makes up this spot to provide researchers a tool to explain their results in a meaningful way. On the other hand, the analytical method, which is based on enrichment statistical test, is not employed for the first time in specifying the correlation between histone modification signals in a gene’s promoter region and the gene’s expression level, and some common conclusions have been reached, including that in various tissues of different animals, any histone acetylation or H3K4 methylation is believed to play an upregulating activity,\(^16\,16\) and H3K27 methylation is suggested to be downregulating transcription.\(^16\,17\) Therefore, this application, which has integrated classical analytical methods and user-friendly interactive interface is reliable and practical for users.

Histone modification pattern is quite dynamic due to the fact that it is sensitive to temporal and spatial change.\(^23\,24\) In the main page of our web server, we solved the problem of how spatial specific (tissue specific) histone modification patterns affect gene expression profiles. We also provide a “Dynamic” page addressing the problem of how temporal elements affect genes’ differential expression. With the increase of studies in this field, we hope users could utilize this platform in their research and assist us in improving its function.

Materials and Methods

EpiRegNet is implemented in perl, PHP and MySQL with all major browsers supported. Currently, the system has collected whole genome wide histone modifications profiles for 12 cell lines in human, and will expand, as more data are available. The data for ESC and IMR90 cells were downloaded from Roadmap Epigenomics Project\(^11\) and data for CD4+ T cells were downloaded from Barski et al.\(^7\) and Wang et al.\(^8\) ChIP-seq data for 73 TFBS and histone modification in other nine cell lines were from ENCODE.\(^1\) Computationally predicted TFBSs of 419 transcription factors were obtained using the algorithm introduced in Levy et al.\(^2\) with TFBS p values calculated by FastPval.\(^23\) The TSS annotations were obtained from GRCh37/hg19 version of RefSeq genes from the UCSC Genome Browser Database.\(^26\)

With the hypothesis that if the set of genes are regulated by a certain factor, the frequency/strength of this factor in the promoter/enhancer region of these genes will be significantly different from one category to another, statistic tests are applied to histone modification marks or transcription factors to obtain the p-value, which denotes whether this factor contributes to the different gene expression profiles. For epigenetic modification data, we measure the read (peak) counts for each gene, and use Wilcoxon test to compare the values for two categories of genes and use Kruskal-Wallis test for multiple categories of genes. For transcription factors, we assign either 1 or 0 for each gene depending on whether the TF has binding sites in target gene’s promoter region or not. We then use chi-square test or fisher’s exact test to calculate the significance.

In the “heatmap” page, corellelative relations between each pair of marks are used to construct a matrix of histone modification marks. Spearman’s correlation test is applied to every pair of factors on their corresponding enrichment scores in promoter regions of genes of interest. In “network” page, the correlation is also shown by lines connecting pairs of histone modification marks.

For the “Dynamic” page of EpiRegNet, the initial parameter settings are similar to those of the main page. Given the user uploaded data for histone modification distribution patterns in two states, we normalize these two data sets so that they have equal number of total reads. With a list of differentially expressed gene names and two processed histone modification mark signal files, our system calculates the difference of signal strength in promoter region (distance to TSS is defined by the user) between two states for each gene and then ranks these genes by this value of difference in a descending way. Genes on the top of the rank are defined as direct targets of the particular histone modification mark. For these top genes, if they are TFs, we search for their targets among all genes in the uploaded gene list. If the TFBS to this TF can be detected in a gene’s promoter region, we define this gene as a target of the TF, and as an indirect target of
the histone modification mark because the histone modification mark regulates expression of the TF.
A detailed manual on how to utilize EpiRegNet is available on the website and, in each step, there is explanation on how to set parameters and interpret results.

Conclusion
Combining both epigenetics and genetics factors, we present this web-based application to facilitate the detection of functioning histone modification marks and transcription factors for researchers who study transcriptional regulation. The system provides swift and flexible service for biologists to process their gene expression data to obtain further information. It is available for academic use at http://jwlanglab.org/EpiRegNet.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Acknowledgments
We thank Sam Bevan for editing the manuscript. The project was supported by University Postgraduate Fellowships (to L.Y.W. and J.Q.) from the University of Hong Kong, the Research Grants Council of Hong Kong (N_HKU752/10, 781511M, 778609M, AoE M04/04), and Food and Health Bureau of Hong Kong (10091262).

References

Downloaded by [UT Southwestern Medical Center at 13:06 01 June 2015]