D178–D183 *Nucleic Acids Research, 2014, Vol. 42, Database issue* *Published online 2 December 2013*
doi:10.1093/nar/gkt1254

DPRP: a database of phenotype-specific regulatory programs derived from transcription factor binding data

David T. W. Tzeng[1], Yu-Ting Tseng[1,2], Matthew Ung[3], I-En Liao[2], Chun-Chi Liu[1,4,*] and Chao Cheng[3,5,6,*]

[1]Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung 402, Taiwan, [2]Department of Computer Science and Engineering, National Chung Hsing University, Taichung 402, Taiwan, [3]Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, NH, USA, [4]Agricultural Biotechnology Center, National Chung Hsing University, Taichung 402, Taiwan, [5]Institute for Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA and [6]Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

Received August 22, 2013; Revised November 8, 2013; Accepted November 11, 2013

ABSTRACT

Gene expression profiling has been extensively used in the past decades, resulting in an enormous amount of expression data available in public databases. These data sets are informative in elucidating transcriptional regulation of genes underlying various biological and clinical conditions. However, it is usually difficult to identify transcription factors (TFs) responsible for gene expression changes directly from their own expression, as TF activity is often regulated at the posttranscriptional level. In recent years, technical advances have made it possible to systematically determine the target genes of TFs by ChIP-seq experiments. To identify the regulatory programs underlying gene expression profiles, we constructed a database of phenotype-specific regulatory programs (DPRP, http://syslab.nchu.edu.tw/DPRP/) derived from the integrative analysis of TF binding data and gene expression data. DPRP provides three methods: the Fisher's Exact Test, the Kolmogorov–Smirnov test and the BASE algorithm to facilitate the application of gene expression data for generating new hypotheses on transcriptional regulatory programs in biological and clinical studies.

INTRODUCTION

In the past decade, gene expression profiling by microarray and more recently by RNA-seq experiments has been extensively used to study transcriptional regulation, resulting in a plethora of expression data available in public databases such as the Gene Expression Omnibus (1). These data sets are informative in elucidating transcriptional regulation under various biological and clinical conditions. For example, a comparison of gene expression between breast cancer and normal breast tissues identifies differentially expressed genes (DEGs) that are presumably critical for carcinogenesis. Such gene expression alterations in response to conditional changes are programmed by a set of transcription factors (TFs). Unfortunately, TF activity is often regulated by phosphorylation/dephosphorylation and other posttranscriptional mechanisms and can be modified by mutation. Thus, it is usually difficult to identify TFs responsible for gene expression changes solely based on the expression of TFs (2–4).

In principle, the activity of TFs can be reflected by the expression changes of their target genes: on TF activation, the expression of a TF's target genes are more likely to be upregulated in the case of a transcriptional activator, and downregulated in the case of a transcriptional repressor; the opposite would be expected if a TF is deactivated. For example, we cannot consistently detect the expression change of the malfunctional p53 with a point mutation that abolishes the tumor suppressor's transcriptional regulatory activity in tumor samples. However, the p53 gene targets are more likely to be differentially expressed in the tumor sample with respect to a normal sample. Based on this rationale, several methods have been proposed to infer the regulatory activity of TFs based on the expression change of their target genes (5–7). These methods

*To whom correspondence should be addressed. Tel: +886 4 22840338 (ext. 7031); Fax: +886 4 22859329; Email: jimliu@nchu.edu.tw
Correspondence may also be addressed to Chao Cheng. Tel: +1 603 650 1220; Fax: +1 603 650 1188; Email: chao.cheng@dartmouth.edu

The authors wish to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

© The Author(s) 2013. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://nar.oxfordjournals.org/ at National Chung Hsing University Library on January 1, 2014

have achieved substantial success in yeast because the target genes for most yeast TFs have been determined by ChIP-chip experiments (8). Regardless, the application of these methods in human is limited by incomplete knowledge of TF–gene interactions. In fact, to systematically identify human TF–gene interactions, previous studies have attempted to predict TF targets based on the existing TF binding motifs in DNA regions upstream of genes, albeit with a high false-positive rate (5,9).

In recent years, technical advances have made it possible to systematically determine the target genes of TFs by ChIP-chip and ChIP-seq experiments (10,11). In fact, a large number of ChIP-chip and ChIP-seq data have been generated by large-scale projects or individual laboratories. For instance, the ENCODE consortium (Encyclopedia of DNA Elements) has generated 424 ChIP-seq profiles, including >120 human TFs with various cell lines (12). Additionally, enormous amounts of gene expression data have accumulated over the past decade from studies addressing biological and clinical questions. The increasing availability of ChIP-seq data sets provides us with an unprecedented opportunity to reanalyze these gene expression data to further understand and dissect the regulatory networks underlying these expression profiles. In previous reports, the ChEA databases collected large-scale ChIP-seq data and provided the integrative analysis of both ChIP-seq and gene expression data (13); the ChIP-array web server can integrate ChIP-seq data and gene expression profiles to construct regulatory networks (14). Both databases only support Fisher's Exact test but database of phenotype-specific regulatory program (DPRP) provided three algorithms, suggesting better functionality and flexibility.

In this study, we established a second-level database, named the Database of Phenotype-specific Regulatory Programs, to facilitate the search and application of gene expression data for generating new hypotheses on transcriptional regulatory programs under diverse biological and clinical contexts. In the database, we have collected 984 gene expression data sets, which include 29 744 samples. Each data set has several phenotype-specific subsets, and each subset is a group of samples. To study DEGs between two subsets, we defined the subset pair as two subsets within a data set. In the database, we have collected 984 gene expression data sets, which include 29 744 samples and 3754 subset pairs. It contains a wide range of phenotypes such as disease, drug treatment and tissue type. Meanwhile, we have defined a collection TF–gene regulatory relationships containing 424 TF binding profiles derived from the ENCODE ChIP-seq data. We applied three different methods, the Fisher's exact test, the Kolmogorov–Smirnov test (KS test) and the Binding Association with Sorted Expression (BASE) algorithm we previously developed (7), to integrate gene expression profiles and the TF–gene interaction data to infer regulatory networks underlying each expression data set. DPRP provides a user-friendly interface for generating testable hypotheses on transcriptional regulation underlying a wide range of biological and clinical phenomena. DPRP is freely available at http://syslab.nchu.edu.tw/DPRP/.

MATERIALS AND METHODS

Database construction

Gene expression data

We collected 984 gene expression data sets, which include 29 744 samples. These data sets were originally generated to explore differential gene expressions under various conditions or treatments, e.g. gene expression changes during development; differential gene expression between different subtypes of breast cancer. Thus, each data set has several phenotype-specific subsets and each subset has a group of samples. To identify DEGs for each data set, we selected the subsets with at least three samples, and then performed t-test between each pair of subsets without overlapping samples. We obtained the DEGs (significantly upregulated or downregulated genes) for 3754 subset pairs representing a wide range of biological contexts (Supplementary Table S1).

Phenotype annotation of gene expression data

To systematically annotate gene expression data and address synonymous issues, we used the Unified Medical Language System (UMLS) technology that provides a comprehensive catalog of medical concepts (15). The UMLS includes Metathesaurus, semantic network and lexical resources. To concentrate on human disease study, we limited the UMLS concepts to three disease-related semantic types: 'Pathologic Function', 'Injury or Poisoning' and 'Anatomical Abnormality'. To obtain the UMLS concepts for each data set, we used the UMLS natural language processing tool, MetaMap program (15), to process the summary description and the Medical Subject Headings of the PubMed record of the data set. It resulted in 4162 data set-concept relations including 757 distinct UMLS concepts (Supplementary Table S2). The phenotype annotation facilitates users to search specific biological or clinical concepts in enormous gene expression data.

ChIP-seq data

We downloaded 424 ChIP-seq track files from the ENCODE project (16), which represent the binding profiles of >120 human TFs in different cell lines. Based on ChIP-seq data, we applied a method called Target identification from profiles (TIP) algorithm (17) to calculate the binding affinity of each TF with all human RefSeq genes (18), resulting in a matrix containing binding affinities for all TF–gene pairs. TIP is a probabilistic model for the identification of TF target genes. Moreover, TIP calculated the *P* value and the Q value for each TF–gene pair, allowing us to define the target gene set for each TF profile (a TF under a specific cell line).

Inference of phenotype-specific regulatory programs

We applied three different methods to integrate gene expression data with TF binding data to infer the regulatory programs underlying expression profiles. Given a subset pair (e.g. estrogen treated versus untreated MCF7 cell lines), we inferred the regulatory programs responsible for the DEGs. We connected the significant TFs based on ChIP-seq data to construct a regulatory network, in which

Downloaded from http://nar.oxfordjournals.org/ at National Chung Hsing University Library on January 1, 2014

the TF→TF interactions were identified by the TIP algorithm ($P < 0.01$) and indicated that one regulates the transcription of the other. A brief description of these methods is as follows:

Fisher's exact test

Given a subset pair, we select the upregulated and the downregulated DEGs with $P < 0.01$. In case that the number of DEGs with $P < 0.01$ is <500, we instead select the top 500 significant genes to ensure enough genes are included for stable results in subsequent statistical analyses. To estimate the significance of differential TF activity, we performed Fisher's exact tests to examine the overlap between the up-/downregulated gene set and TF target genes. This method requires two cutoff values: one is used to define the up- and the downregulated genes, and the other is used to define the TF target genes. A more detailed description of applying Fisher's exact test for TF activity inference can be found in previous studies (5,19).

KS test

Given a subset pair, we calculated the t-scores for all genes by comparing their expression levels between the two subsets. For each TF we performed KS test to compare the distributions of the t-scores between target genes and nontarget genes. To define the target gene set of a TF, we set the cutoff value as $P < 0.01$. If the number of target genes with $P < 0.01$ is <500, we select the top 500 significant target genes for the regulatory program analysis. For each TF, the KS test resulted in a P value, indicating the significance of its activity change, and a D value, indicating the direction of its activity change. A positive D value indicates that target genes of a TF have significantly higher expression levels than nontarget genes, and a negative D-value indicates the reverse. A similar KS test-based method has been proposed by Tsai *et al.* (6) to identify cell cycle–related TFs in yeast.

BASE algorithm

The cutoff values for defining TF target genes and DEGs are usually arbitrary and hard to determine in advance. Comparing with the Fisher's exact test and the KS-test, BASE is a nonparametric algorithm that requires no cutoff setting for TF target genes or DEGs (20). First, we calculated the t-scores for all genes by comparing their expression levels between a pair of subsets, and sorted them in the decreasing order to obtain a ranked gene list. Each gene in the list is associated with a t_i, the t-score for this gene, and a b_i, the binding affinity of a TF to this gene calculated by TIP algorithm (17). Then we calculated a cumulative distribution function by aggregating $|t_i \cdot b_i|$ and a reference function by aggregating $|t_i|$. Finally, we calculated the maximum deviation between the functions and applied a permutation-based method to normalize the score and to estimate its significance. The normalized score is called regulatory activity score (RAS), which indicates the direction of the activity change of a TF. For a transcriptional activator, a positive/negative RAS indicates enhanced/reduced activity of the TF, while for a transcriptional repressor, the reverse is true. A more detailed description about BASE can be found in (20).

WEB INTERFACES

We integrated gene expression data sets, phenotype information and ChIP-seq data sets to construct the DPRP database with a user-friendly web interface (Figure 1). Users can search a disease concept to discover all related gene expression data sets, choose the interested data set and then select a subset pair within the data set for TF regulatory program analysis. For example, a user can type in 'Breast Carcinoma' as a keyword to obtain a list of data sets related to breast cancer. To facilitate user-friendly text search, we adopted the jQuery AutoComplete technique to guide the user for keyword selection (http://jqueryui.com/autocomplete/). When a specific data set is selected, the database will list a number of subset pairs (e.g. breast cancer versus normal) for investigating regulatory activity of TFs.

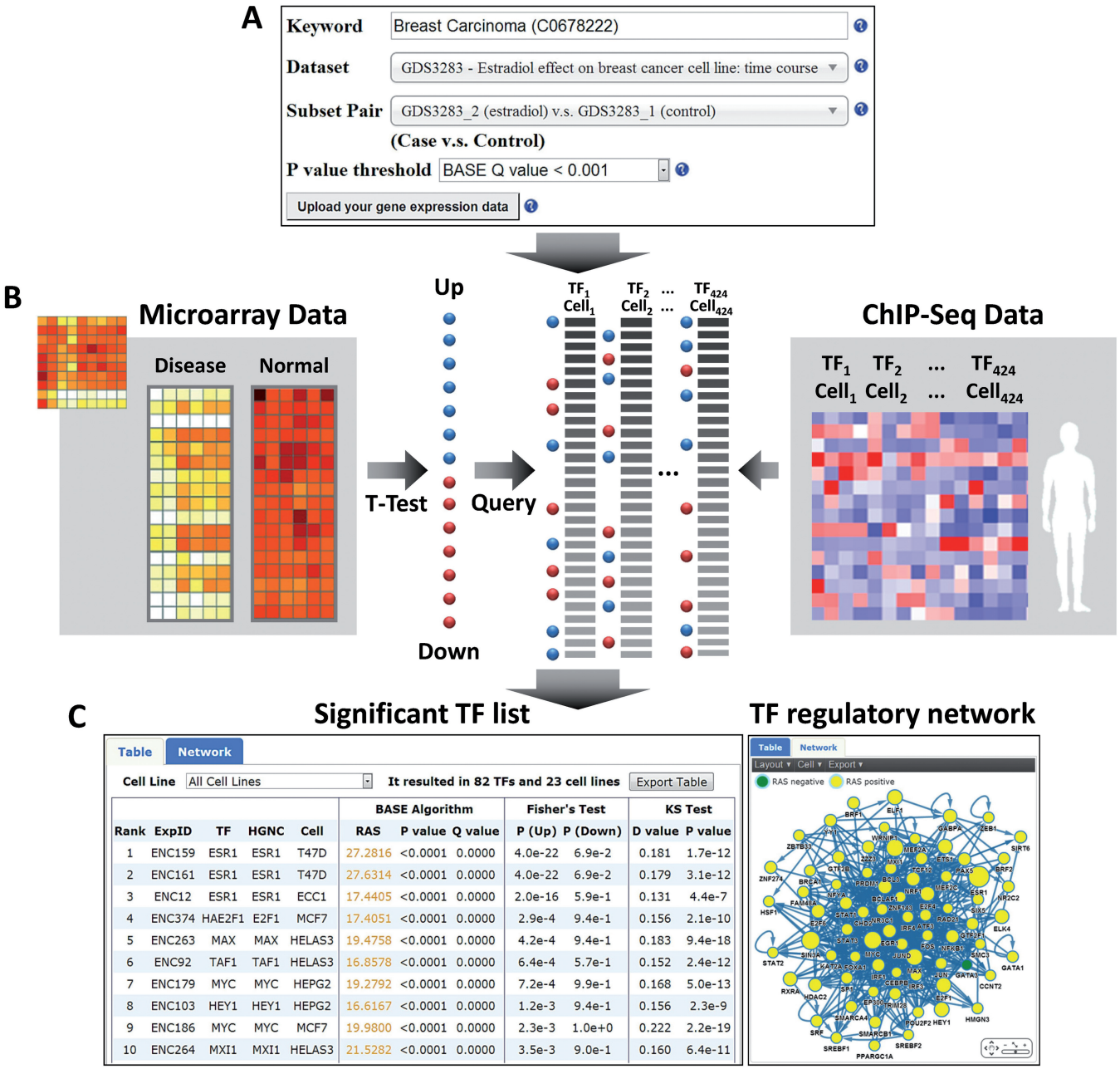
Given a subset pair, DPRP will generate a list of TFs with significant activity changes. To visualize the TF regulatory program, the web server draws a regulatory TF network consisting of all significant TFs, in which the TF→TF interaction indicates that one regulates the transcription of the other, which is identified by the TIP algorithm ($P < 0.01$) from ChIP-seq data. The ChIP-seq data support TF-gene regulations in different cell lines. Users can select a specific cell line to display a cell line–specific network or use all cell lines to display an integrated network. Some cell lines only have a few ChIP-seq experiments, which is not sufficient for TF network construction. Thus, the web interface only allows users to select cell lines with at least 12 ChIP-seq experiments. Moreover, users can upload their own gene expression data onto the database, and then DPRP will perform Fisher's Exact test, KS test and BASE algorithm in the data.

EXAMPLE APPLICATIONS

To demonstrate the biological importance of DPRP, we used GDS3283 and GDS3044 as examples to show the cell-specific regulatory TF networks (Figure 2). GDS3283 is a gene expression data set with estradiol treatment using MCF7 breast epithelial cancer cells (21). The BASE algorithm identifies 74 TFs with significantly differential activity ($Q < 0.001$), in which the most significant TF is estrogen receptor alpha (ESR1) based on ChIP-seq experiments carried out in T47D cells (Figure 2A). Obviously, this result is consistent with our knowledge that the estradiol treatment significantly induces the activity of ESR1 in breast cancer cells. Interestingly, when the expression levels of ESR1 are compared between estradiol treatment and control samples, we cannot detect significant expression change of ESR1 in GDS3283. Thus, the BASE algorithm identified the key regulator that cannot be discovered by differential expression analysis.

Since the GDS3283 is a breast cancer data set, we selected the ChIP-seq experiments from breast cancer cell lines (T47D and MCF7 cells) to generate the regulatory TF network with $Q < 0.001$ (Figure 2B), which contained five significant TFs: ESR1, GATA3, FOXA1, MYC and E2F1. In this regulatory network, ESR1,

Downloaded from http://nar.oxfordjournals.org/ at National Chung Hsing University Library on January 1, 2014



					BASE Algorithm			Fisher's Test		KS Test	
Rank	ExpID	TF	HGNC	Cell	RAS	P value	Q value	P (Up)	P (Down)	D value	P value
1	ENC159	ESR1	ESR1	T47D	27.2816	<0.0001	0.0000	4.0e-22	6.9e-2	0.181	1.7e-12
2	ENC161	ESR1	ESR1	T47D	27.6314	<0.0001	0.0000	4.0e-22	6.9e-2	0.179	3.1e-12
3	ENC12	ESR1	ESR1	ECC1	17.4405	<0.0001	0.0000	2.0e-16	5.9e-1	0.131	4.4e-7
4	ENC374	HAE2F1	E2F1	MCF7	17.4051	<0.0001	0.0000	2.9e-4	9.4e-1	0.156	2.1e-10
5	ENC263	MAX	MAX	HELAS3	19.4758	<0.0001	0.0000	4.2e-4	9.4e-1	0.183	9.4e-18
6	ENC92	TAF1	TAF1	HELAS3	16.8578	<0.0001	0.0000	6.4e-4	5.7e-1	0.152	2.4e-12
7	ENC179	MYC	MYC	HEPG2	19.2792	<0.0001	0.0000	7.2e-4	9.9e-1	0.168	5.0e-13
8	ENC103	HEY1	HEY1	HEPG2	16.6167	<0.0001	0.0000	1.2e-3	9.4e-1	0.156	2.3e-9
9	ENC186	MYC	MYC	MCF7	19.9800	<0.0001	0.0000	2.3e-3	1.0e+0	0.222	2.2e-19
10	ENC264	MXI1	MXI1	HELAS3	21.5282	<0.0001	0.0000	3.5e-3	9.0e-1	0.160	6.4e-11

Figure 1. An overview of the DPRP web interface. (**A**) Users can perform a query by the following procedures: (i) Users can input a disease name in the auto-completed keyword field, which provides a list of partially matched UMLS concepts for selection. Alternatively, users can also input a data set ID in the keyword field to select a specific data set. (ii) After UMLS concept selection, the data sets associated with the selected concept will be shown in a data set list, from which the user can select the data set of interest. (iii) Given a specific data set, the subset pairs from the selected data set will be displayed in a subset list, and then the user can select the subset pair to search TF regulatory programs. DPRP provides three different methods to rank the potential TFs, in which users can determine which ranking guidelines to use. In addition, users can upload their own gene expression data with gene list and t-value of t-test or log ratios between two subsets. (**B**) The database integrated gene expression data and ChIP-seq TF binding data to identify the regulatory programs underlying a selected phenotype pair. (**C**) The output web pages: DPRP generates a list of the TFs and ranks them by their *P* values or Q values. In TF table view, users can export the table of candidate TFs as a text file. Based on the ranked TF list, DPRP generates a regulatory network consisting of all significant TFs, in which users can export the TF network as a png, svg or xml file.

FOXA1 and GATA3 formed a tight regulatory module. These results are consistent with the finding by Kong *et al.* (22) that FOXA1 and GATA3 are essential co-regulators in estrogen response pathway and that ESR1, FOXA1 and GATA3 formed an enhanceosome in breast cancer cells. Activation of MYC and E2F1 may indicate that estrogen treatment can promote cell proliferation of MCF7 cells.

DPRP also provides insights into drug mechanisms. GDS3044 is a gene expression data set in K562 (the leukemia cell line) cells treated with imatinib. Figure 2C shows that the imatinib treatment significantly increases

Downloaded from http://nar.oxfordjournals.org/ at National Chung Hsing University Library on January 1, 2014

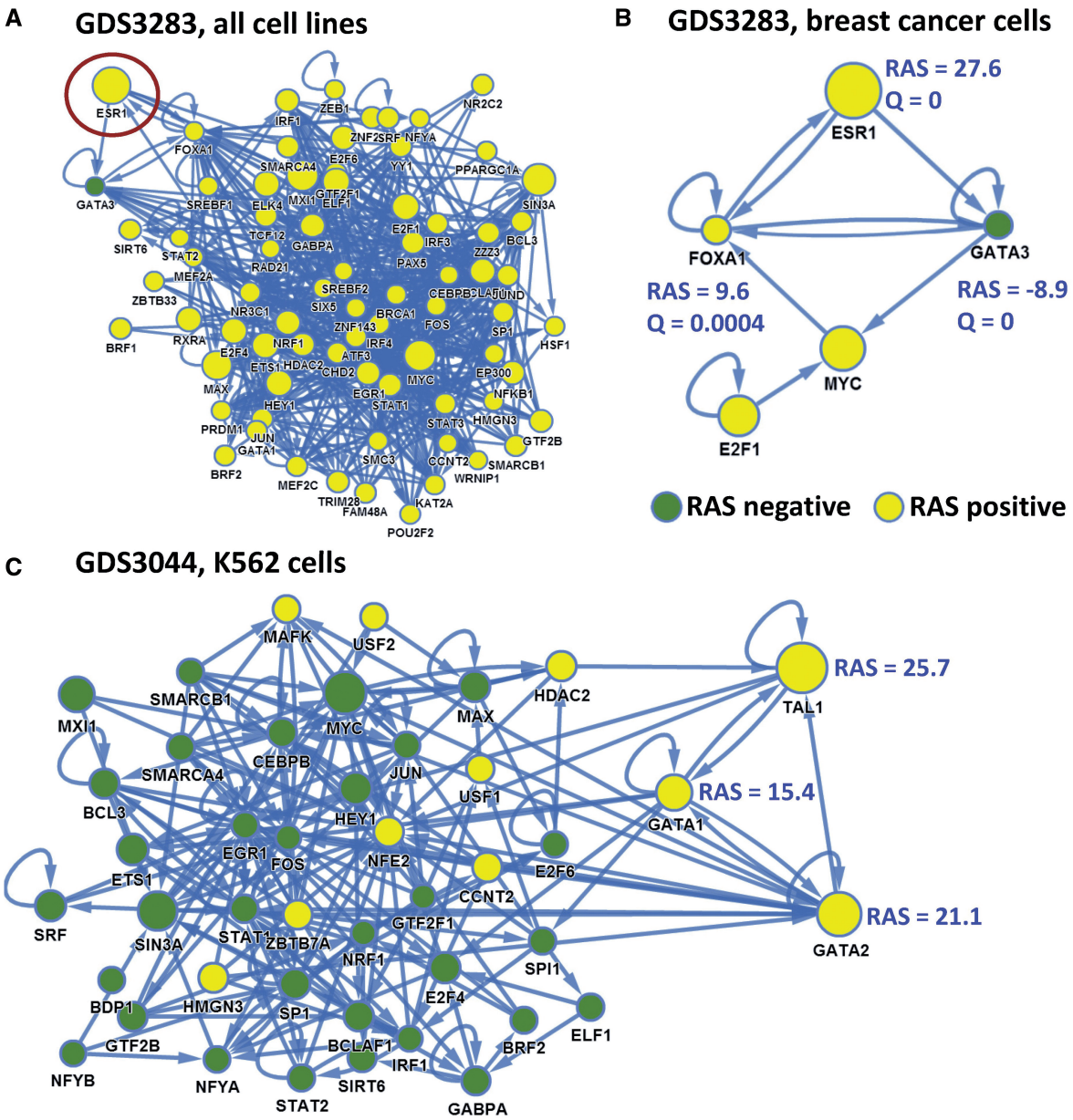


Figure 2. Example applications. (**A**) The complete regulatory TF network associated with estradiol treatment in MCF7 cells from the GDS3283 data set. The network contains 74 significant TFs identified from ChIP-seq data in all cell lines, in which the most significant TF is ESR1. This is the regulatory network output by the BASE method with $Q < 0.001$, when users select the GDS3283 data set and subset pair 'estradiol treatment versus control'. (**B**) The regulatory TF network specific for T47D + MCF7 cell lines. In the network, only the significant TFs with ChIP-seq data from T47D and MCF7 are displayed. (**C**) The regulatory TF network associated with imatinib treatment in K562 cells from the GDS3044 data set. This is the output by the BASE method ($Q < 0.001$), when users select the GDS3044 data set with subset pair 'imatinib treatment versus control', and then select the K562-specific TF network. The network contains 43 significant TFs, in which the most significant TF is TAL1.

the activity of three TFs: GATA1, GATA2 and TAL1 in K562 cells. It is known that imatinib inhibits the kinase activity of BCR-ABL protein, which is the pathophysiologic cause of chronic myelogenous leukemia. Previous studies have shown that BCR-ABL suppresses the GATA1 activity, and thus explains why we observed an increased activity of GATA1 in response to imatinib treatment (16,23). In addition, TAL1, the T-cell acute lymphocytic leukemia protein 1, is specifically expressed in early erythroid cells and interacts with GATA1 (24), which also supports our result.

DISCUSSION

In this study, we applied three different methods to infer the regulatory programs underlying given gene expression profiles. To apply the Fisher's exact test, DEGs have to be defined based on the gene expression data. At the same significance level, the numbers of DEGs vary substantially in different gene expression data sets, depending on the quality of the data and the sample size. As a consequence, we expect variability in statistical power and robustness of Fisher's exact test. Similarly, the effectiveness of this

Downloaded from http://nar.oxfordjournals.org/ at National Chung Hsing University Library on January 1, 2014

method is also influenced by the number of TF target genes. TFs with more target genes are more likely to be identified as significant TFs. The KS-test does not require the up-/downregulated gene sets, but it still requires the predefined TF target genes. The BASE method requires neither a differential gene set nor a target gene set, and thus is more convenient in practice and does not have the bias issue. However, it estimates the significance of the TF RAS using the permutation of gene expression profiles (shuffle all genes in the profile), which often overestimates their significance. Because none of these methods is perfect, we provide the results from all three methods in the DPRP database. This allows users to determine the stringency level and make decisions according to their own requirements, e.g. selecting the significant TFs identified by all methods to obtain a TF list of high confidence.

Currently, we have included the ChIP-seq data generated by the ENCODE project in our database. There are many ChIP-seq and ChIP-chip data sets that have been generated by other large-scale projects or by individual laboratories that will be included in the database. Moreover, we anticipate that an increasing number of TF ChIP-seq data will be generated in the near future. We will maintain our database with routine updates to ensure that we maintain a comprehensive list of TFs. We believe DPRP will be a useful database and resource for biological and clinical studies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

American Cancer Society Research [IRG-82-003-27]; start-up funding package provided by the Geisel School of Medicine at Dartmouth College (to C.C.); National Science Council grants [NSC99-2320-B-005-008-MY3 and NSC101-2627-B-005-002 to C.C.L.]. Funding for open access charge: American Cancer Society Research Grant, [#IRG-82-003-27].

Conflict of interest statement. None declared.

REFERENCES

1. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al*. (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*., **41**, D991–D995.
2. Calkhoven,C.F. and Ab,G. (1996) Multiple steps in the regulation of transcription-factor level and activity. *Biochem. J*., **317(Pt 2)**, 329–342.
3. Boulikas,T. (1995) Phosphorylation of transcription factors and control of the cell cycle. *Crit. Rev. Eukaryot. Gene Expr*., **5**, 1–77.
4. Ouyang,J., Valin,A. and Gill,G. (2009) Regulation of transcription factor activity by SUMO modification. *Methods Mol. Biol*., **497**, 141–152.
5. Rhodes,D.R., Kalyana-Sundaram,S., Mahavisno,V., Barrette,T.R., Ghosh,D. and Chinnaiyan,A.M. (2005) Mining for regulatory programs in the cancer transcriptome. *Nat. Genet*., **37**, 579–583.
6. Tsai,H.K., Lu,H.H. and Li,W.H. (2005) Statistical methods for identifying yeast cell cycle transcription factors. *Proc. Natl Acad. Sci. USA*, **102**, 13532–13537.
7. Cheng,C., Yan,X., Sun,F. and Li,L.M. (2007) Inferring activity changes of transcription factors by binding association with sorted expression profiles. *BMC Bioinformatics*, **8**, 452.
8. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B., Yoo,J. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
9. Cheng,C., Li,L.M., Alves,P. and Gerstein,M. (2009) Systematic identification of transcription factors associated with patient survival in cancers. *BMC Genomics*, **10**, 225.
10. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
11. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
12. Gerstein,M.B., Kundaje,A., Hariharan,M., Landt,S.G., Yan,K.K., Cheng,C., Mu,X.J., Khurana,E., Rozowsky,J., Alexander,R. *et al*. (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91–100.
13. 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.
14. Qin,J., Li,M.J., Wang,P., Zhang,M.Q. and Wang,J. (2011) ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. *Nucleic Acids Res*., **39**, W430–W436.
15. Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*., **32**, D267–D270.
16. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al*. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*., **22**, 1760–1774.
17. Cheng,C., Min,R. and Gerstein,M. (2011) TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics*, **27**, 3221–3227.
18. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res*., **40**, D130–D135.
19. Essaghir,A., Toffalini,F., Knoops,L., Kallin,A., van Helden,J. and Demoulin,J.B. (2010) Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. *Nucleic Acids Res*., **38**, e120.
20. Zhu,M., Liu,C.C. and Cheng,C. (2013) REACTIN: regulatory activity inference of transcription factors underlying human diseases with application to breast cancer. *BMC Genomics*, **14**, 504.
21. Lin,Z., Reierstad,S., Huang,C.C. and Bulun,S.E. (2007) Novel estrogen receptor-alpha binding sites and estradiol target genes identified by chromatin immunoprecipitation cloning in breast cancer. *Cancer Res*., **67**, 5017–5024.
22. Kong,S.L., Li,G., Loh,S.L., Sung,W.K. and Liu,E.T. (2011) Cellular reprogramming by the conjoint action of ERalpha, FOXA1, and GATA3 to a ligand-inducible growth state. *Mol. Syst. Biol*., **7**, 526.
23. Xu,C., Fu,H., Gao,L., Wang,L., Wang,W., Li,J., Li,Y., Dou,L., Gao,X., Luo,X. *et al*. (2012) BCR-ABL/GATA1/miR-138 mini circuitry contributes to the leukemogenesis of chronic myeloid leukemia. *Oncogene*, Dec 3 (doi: 10.1038/onc.2012.557; epub ahead of print).
24. Wilkinson-White,L., Gamsjaeger,R., Dastmalchi,S., Wienert,B., Stokes,P.H., Crossley,M., Mackay,J.P. and Matthews,J.M. (2011) Structural basis of simultaneous recruitment of the transcriptional regulators LMO2 and FOG1/ZFPM1 by the transcription factor GATA1. *Proc. Natl Acad. Sci. USA*, **108**, 14443–14448.

Downloaded from http://nar.oxfordjournals.org/ at National Chung Hsing University Library on January 1, 2014