Published online 23 October 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D43–D50***
doi: 10.1093/nar/gkw965

ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data

Ke-Ren Zhou, Shun Liu, Wen-Ju Sun, Ling-Ling Zheng, Hui Zhou, Jian-Hua Yang[*] and Liang-Hu Qu[*]

Key Laboratory of Gene Engineering of the Ministry of Education, State Key Laboratory for Biocontrol, Sun Yat-sen University, Guangzhou 510275, P. R. China

Received August 19, 2016; Revised September 24, 2016; Editorial Decision October 06, 2016; Accepted October 12, 2016

ABSTRACT

The abnormal transcriptional regulation of non-coding RNAs (ncRNAs) and protein-coding genes (PCGs) is contributed to various biological processes and linked with human diseases, but the underlying mechanisms remain elusive. In this study, we developed ChIPBase v2.0 (http://rna.sysu.edu.cn/chipbase/) to explore the transcriptional regulatory networks of ncRNAs and PCGs. ChIPBase v2.0 has been expanded with ∼10 200 curated ChIP-seq datasets, which represent about 20 times expansion when comparing to the previous released version. We identified thousands of binding motif matrices and their binding sites from ChIP-seq data of DNA-binding proteins and predicted millions of transcriptional regulatory relationships between transcription factors (TFs) and genes. We constructed 'Regulator' module to predict hundreds of TFs and histone modifications that were involved in or affected transcription of ncRNAs and PCGs. Moreover, we built a web-based tool, Co-Expression, to explore the co-expression patterns between DNA-binding proteins and various types of genes by integrating the gene expression profiles of ∼10 000 tumor samples and ∼9100 normal tissues and cell lines. ChIPBase also provides a ChIP-Function tool and a genome browser to predict functions of diverse genes and visualize various ChIP-seq data. This study will greatly expand our understanding of the transcriptional regulations of ncRNAs and PCGs.

INTRODUCTION

Eukaryotic genomes encode thousands of protein-coding genes (PCGs) and non-coding RNAs (ncRNAs), such as microRNAs (miRNAs), long non-coding RNAs (lncRNAs), small nucleolar RNAs (snoRNAs) and pseudogenes (1–6). The dysregulation of these RNA molecules have been shown to contribute to developmental, physiological and pathological processes (5,7). However, how the majority of PCGs and ncRNA genes are transcriptionally regulated remains unknown.

The expression or transcription of genes is mainly governed by the specificity of transcription factors (TFs). Increasing evidences suggest that transcriptional regulatory circuitries involving in ncRNAs and TFs play important roles in controlling cellular differentiation, proliferation and embryonic stem (ES) cell identity (8–10). For example, miRNAs have been connected to the core transcriptional regulatory circuitry of ES cells that maintains ES cell identity (8). Interactions of dozens of lncRNAs and ES cell TFs control pluripotency and differentiation (9). However, deciphering the interactions between hundreds of TFs and thousands of PCGs and ncRNAs remain a daunting challenge.

Recent advances in chromatin immunoprecipitation followed by sequencing (ChIP-seq) have provided powerful ways to identify genome-wide profiling of DNA-binding proteins and histone modifications (11–13). The application of ChIP-seq methods has reliably discovered TF binding sites and histone modification sites (11–13). In fact, many more studies to date have been focused on understanding the transcriptional regulations of TF-PCG. We and others have used ChIP-seq data of some TFs to characterize transcriptional regulation of lncRNAs and miRNAs (8,9,14). However, with ChIP-seq technologies have been broadly used to identify the binding sites of thousands of TFs, transcription cofactors (TCFs) and chromatin-remodeling factors (CRFs), there is a great need to integrate these large-scale datasets to explore the transcriptional regulatory networks of diverse ncRNAs and PCGs and their roles in the human diseases.

[*]To whom correspondence should be addressed. Tel: +86 20 8411 2399; Fax: +86 20 8403 6551; Email: lssqlh@mail.sysu.edu.cn
Correspondence may also be addressed to Jian-Hua Yang. Tel: +86 20 8411 2517; Fax: +86 20 8403 6551; Email: yangjh7@mail.sysu.edu.cn

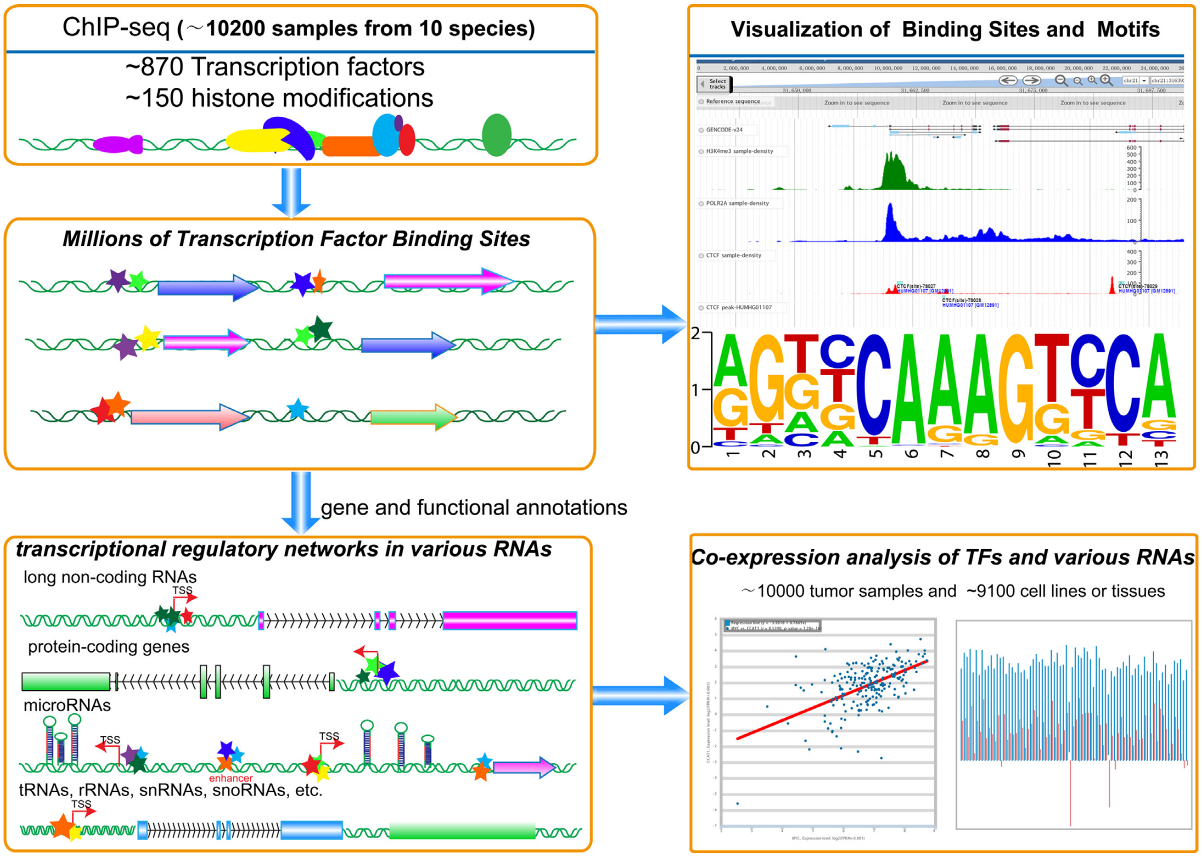
ChIP-seq (~10200 samples from 10 species)
~870 Transcription factors
~150 histone modifications
Millions of Transcription Factor Binding Sites
Visualization of Binding Sites and Motifs
gene and functional annotations
transcriptional regulatory networks in various RNAs
long non-coding RNAs
TSS
protein-coding genes
microRNAs
TSS
TSS
enhancer
tRNAs, rRNAs, snRNAs, snoRNAs, etc.
TSS
Co-expression analysis of TFs and various RNAs
~10000 tumor samples and ~9100 cell lines or tissues
2
1
0
1
2
3
4
5
6
7
8
9
10
11
12
13

Figure 1. System overview of ChIPBase v2.0 core framework. All results generated by ChIPBase v2.0 are deposited in MySQL relational databases and displayed in the visual browser and web page.

In ChIPBase v2.0, we integrated a large number of ChIP-seq peak datasets of trans-acting factors, including TFs, TCFs, CRFs, other DNA-binding proteins and histone modifications (Figure 1) to discover the interaction maps between trans-acting factors and various types of RNAs. Furthermore, by importing expression profiles of thousands of tumor samples from TCGA project, ChIPBase v2.0 can be used to illustrate the clinically relevant interactions between TFs and RNA molecules. With the integration of more than 10 000 ChIP-seq datasets from 10 species, ChIPBase v2.0 is expected to help the researchers to investigate the potential transcriptional regulatory mechanisms of ncRNAs and PCGs.

MATERIALS AND METHODS

Integration and exploration of public ChIP-seq datasets

We manually collected ∼10 200 peak datasets generated from ChIP-seq, ChIP-exo and MNChIP-seq. All our peak datasets were curated from NCBI GEO database (15), ENCODE project (16), modENCODE project (17,18) and NIH Roadmap Epigenomics Project (19) (Tables 1 and 2). These peak datasets were converted to the corresponding latest genome version by using liftOver tool (20), and the peaks that failed to be converted were discarded.

We calculated genome-wide experiment density of the binding sites of trans-acting factors by merging all their corresponding peak data and represented them as bigWig format files deposited in our genome browser (21). We *de novo* identified motifs of DNA-binding proteins and their corresponding peak regions by using HOMER program (22).

Integration of genome sequences, gene annotation sets and other metadata of 10 species

We downloaded genome sequences and annotation sets of 10 species (human, mouse, rat, chicken, *Xenopus tropicalis*, zebrafish, worm, fly, yeast and *Arabidopsis thaliana*) from UCSC genome browser, GENCODE project (23) and Ensembl database (24) (Supplementary Table S1). Annotations of miRNA primary transcripts were downloaded from the latest version of miRbase (25) (release 21) and converted to their corresponding latest genome version by using liftOver (20) (Supplementary Table S1). In addition, annotations of lncRNAs of chicken (galGal4), worm (ce10) and *X. tropicalis* (xenTro3) were obtained from our deepBase v2.0 (26).

To analyze transcriptional regulatory networks of ncRNAs and PCGs in 10 species, we extracted and classified genes into lncRNA, miRNA, other ncRNA and PCGs from their gene annotation sets according to gene biotypes defined by Ensembl (24). We also curated some metadata to ensure the annotation consistency and accuracy, including refSeq accessions, gene ontology (GO) terms and classification of DNA-binding proteins (15,20,24,27).

Table 1. The library statistics of ChIP-seq datasets in ChIPBase v2.0

Species	Total library	TF library	TCF library	CRF library	Other library	Histone library
human	5803	2498	433	192	214	2466
mouse	2500	1036	209	72	89	1094
worm	852	428	67	18	310	15
fruitfly	838	186	82	54	183	347
A.thaliana	54	51	/	/	/	3
yeast	52	52	/	/	/	/
rat	44	15	2	5	1	21
zebrafish	32	10	/	/	/	22
X. tropicalis	30	14	/	/	/	16
chicken	11	10	/	/	/	1

Library statistics indicating the numbers of sample library (ChIP-seq, ChIP-exo and MNChIP-seq), including TFs, TCFs, CRFs, other DNA-binding protein (other) and histone modifications (histone) in 10 species.

Table 2. The statistics of trans-acting factors in ChIPBase v2.0

Species	Total	TF	TCF	CRF	Other	Histone
human	711	480	51	43	89	48
mouse	302	189	22	18	42	31
worm	151	68	6	6	67	4
fruitfly	162	60	6	8	56	32
A.thaliana	29	26	/	/	/	3
yeast	15	15	/	/	/	/
rat	20	7	1	3	1	8
zebrafish	11	7	/	/	/	4
X. tropicalis	8	4	/	/	/	4
chicken	5	4	/	/	/	1

This statistics indicating the numbers of TFs, TCFs, CRFs, other DNA-binding protein (other) and histone modifications (histone) in 10 species.

Selection of examined transcriptional regulatory domains of ncRNA genes and PCGs

DNA-binding proteins like TFs might not almost exclusively bind at proximal promoters of ncRNA genes or PCGs, many of them can bind to enhancers region which can be up to tens of kb away from the genes and can be upstream or downstream from the transcription start sites (TSSs) (28–30). More than half of the observed binding events occurred on protein coding genes are distally binding (28). Similar to DNA-binding proteins, histone modifications both can occurred at promoter regions and enhancer regions (31–34). Therefore, we chose a 30-kb region upstream and 10-kb region downstream from TSSs as the examined regulatory domains of ncRNA genes and PCGs. Five-kb upstream region and 1-kb downstream region of each ncRNA and PCG were generally chosen as their promoter region (28). Then we intersected ChIP-seq peak regions of TFs with these regulatory domains to identify transcriptional regulatory relationships of TF-ncRNA and TF-PCG. The similar strategy was also applied to the identification of histone modification marks that occurred at upstream or downstream of ncRNAs and PCGs.

Development of co-expression patterns between TFs and various genes

Public gene expression data of RNA-seq and miRNA-seq were downloaded from UCSC Xena project (http://xena.ucsc.edu/) and EBI Expression Atlas (35), including ∼9900 miRNA-seq and ∼10 000 RNA-seq data of 32 carcinomas derived from TCGA project (36) and ∼7800 RNA-seq data of 31 normal tissues derived from GTEx project (37) in human (Table 3). All expression data were normalized with log2 (FPKM + 0.001) for RNA-seq or log2 (RPM + 0.001) for miRNA-seq.

Co-expression patterns between DNA-binding proteins and ncRNAs or PCGs were determined by Pearson correlation coefficient (Pearson' r) and *P*-value in t-test (student test). We visualized these expression data with scatter plots accompanied with regression lines, boxplots and histograms.

Construction of a web-based ChIP-Function tool and ChIPBase genome browser

Public GO terms data were downloaded from the Ensembl database (24). The genes did not have any GO terms were discarded. We used a hypergeometric test with false discovery rate (FDR) correction to determine the enrichment analysis of GO for TFs, TCFs, CRFs and other DNA-binding proteins.

To visualize our various ChIP-seq datasets, we used Jbrowse (38), which was a fast and embeddable genome browser built completely with JavaScript and HTML5, to construct the ChIPBase genome browser. We integrated all of our curated peak data, identified motifs data and calculated genome-wide experiment density of trans-acting factors and then displayed them in ChIPBase genome browser.

DATABASE CONTENT AND WEB INTERFACE

The comprehensive annotation and identification of transcriptional regulatory relationships of TF-ncRNA and TF-PCG

To investigate transcriptional regulatory networks and histone modification marks of ncRNAs and PCGs, we chose an examined transcriptional regulatory domain (from −30 to 10 kb away from TSSs) for these genes (see the 'Materials and Methods' section for details). On our website, we provided five web-based modules, including 'LncRNA', 'miRNA', 'OtherNcRNA', 'Protein' and 'Regulator', to explore transcriptional regulatory relationships. In the first

Table 3. The statistics of RNA-seq and miRNA-seq expression data used in ChIPBase v2.0

Species	Project name	Diseases or studies	Samples
human	TCGA Pan-Cancer (PANCAN, RNA-seq)	32	10 359
human	TCGA Pan-Cancer (PANCAN, miRNA-seq)	32	9966
human	Genotype-Tissue Expression (GTEx) project	31	7834
human	RNA-seq from the CCLE	20	780
human	RIKEN FANTOM5 project (human)	2	76
human	32 different tissues of human	1	32
mouse	RIKEN FANTOM5 project (mouse tissue)	10	156
mouse	RIKEN FANTOM5 project (mouse cell lines)	1	35
mouse	RNA-seq of mouse DBA/2J x C57BL/6J tissues	1	6
mouse	Individual-Th single cell RNA-Seq	1	91
rat	Strand-specific RNA-seq of nine rat tissues	3	27
worm	Developmental Stages, modENCODE	1	17
chicken	Strand-specific RNA-seq of nine chicken tissues	1	9
chicken	RNA-seq of poly-A enriched total RNA of tissue samples from chicken	1	5
X. tropicalis	RNA-seq of poly-A enriched total RNA of tissue samples from frog	1	5
A. Thaliana	RNA-seq during Arabidopsis meristem development from day 7 to 16 after germination	1	10
A. Thaliana	RNA-seq of coding RNA of Arabidopsis seedlings from 19 natural accessions	1	19
A. Thaliana	Transcriptomes for hybrids (F1s) between 18 *Arabidopsis thaliana* parents	1	9

This table contained the data sources and sample numbers of RNA-seq and miRNA-seq across seven species.

four modules, users can flexibly narrow or expand the regulatory domain or choose any items interested them to get the relationships of TF-gene, such as different cell lines, different experiments and whether there were any regions containing observed TF-binding motifs within the examined regulatory domains.

The Regulator module was developed to find the trans-acting factors bound upstream or downstream ncRNAs and PCGs. On the webpage of Regulator, users can search their interested genes to get how many TFs were bound around the promotor or enhancer regions of them. In human, we totally annotated and identified ∼4 658 040 transcriptional regulatory relationships of TF-ncRNA, including ∼2 147 760 TF-lncRNA and ∼273 760 TF-miRNA, as well as ∼3 438 820 TF-PCG (Supplementary Table S2).

De novo identification of binding motifs of DNA-binding proteins

As we all know, short DNA sequence motifs are important cis-regulatory elements recognized by DNA-binding proteins (e.g. TFs) (39–41), and are critical for elucidating transcriptional regulation of genes (41). We *de novo* identified ∼6200 motifs of DNA-binding proteins in different cell types, tissues and conditions across 10 species (Supplementary Table S3). All the identified position weight matrices (PWMs) and visualized motif logos of DNA-binding proteins are deposited at the web-based module named Motif. We also applied these motifs data to the annotation and identification of transcriptional regulatory relationships of TF-ncRNA and TF-PCG. These data have been integrated into most of the web-based modules and tools of our database, which will facilitate the analysis on transcriptional regulatory networks.

The co-expression patterns between DNA-binding proteins and ncRNA genes

We developed a web-based tool named Co-Expression to explore the co-expression patterns between DNA-binding proteins and ncRNA genes or any two genes (see the 'Materials and Methods' section for details). The Co-Expression can estimate the transcriptional relationships between TFs and their transcriptionally regulated genes. For example, miR-194-2 and miR-192 were both demonstrated to be upregulated by HNF1A and downregulated by TGFB1 (42,43), and their co-expression patterns drawn by our database also matched these correlations (Figure 2). In addition, CCAT1, a long non-coding RNA highly expressed in colon cancer, was found to be activated by MYC (44–46). In our database, the co-expression patterns of MYC-CCAT1 also showed a strong positive correlation in both colon adenocarcinoma and rectum adenocarcinoma (Supplementary Figure S1).

The Co-Expression also can be used to explore the transcriptional cooperativity between different TFs. For instance, YY1 bound with CTCF and formed a transcription complex around the genomic imprinting region (47,48), and we found that the co-expression patterns of CTCF and YY1 in our database were positively correlated in 32 cancer types (Supplementary Table S4).

Web interface of web-based modules and tools developed to the identification and exploration of transcriptional regulation of ncRNAs and PCGs

ChIPBase v2.0 consists of nine web-based modules and tools. The LncRNA, miRNA, OtherNcRNA, Protein and Regualtor modules are mainly developed to investigate the transcriptional regulation of ncRNAs and PCGs, while the Motif module is aimed to study the binding motifs of DNA-binding proteins. The ChIP-Function tool can be used to predict the functions of DNA-binding proteins. Additionally, the Co-Expression tool and ChIPBase browser assist users in exploring the co-expression patterns of TFs and various genes and visualizing various ChIP-seq data respectively.

Our database provides with multiple selectors and parameters to analyze the transcriptional regulatory relationships of TF-gene for all web-based modules and tools. In the most of these web-based modules and tools, there are some common selectors and parameters like 'Clade', 'Organism', 'Assembly', 'Factors', 'Experiment', 'Upstream', 'Downstream' and 'Motif'. Users can use the 'Clade', 'Organism' and 'Assembly' selectors to choose the interested organism and use the 'Factor' and 'Experiment' selectors

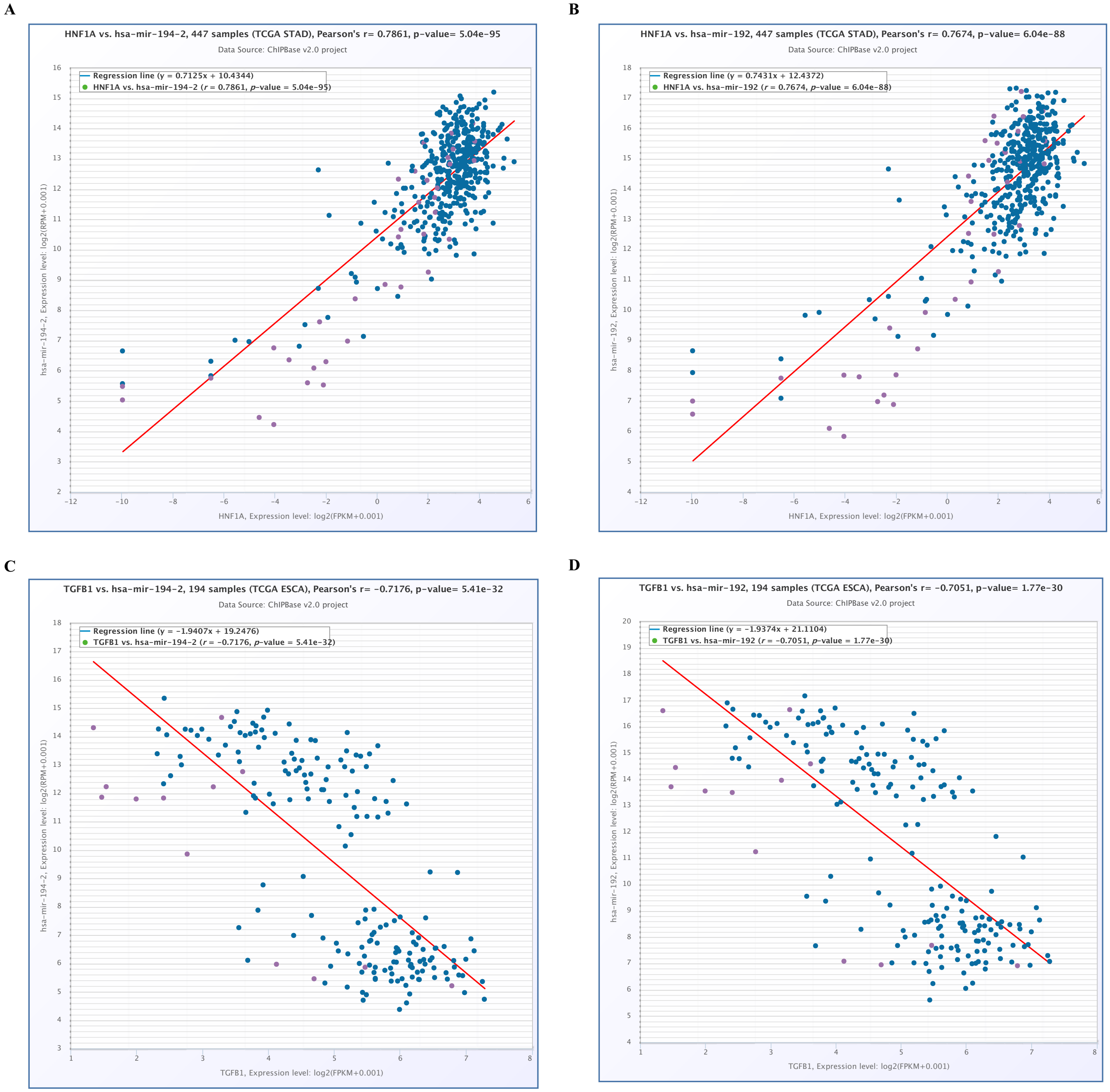

Figure 2. The co-expression patterns of HIF1A-miR194-2/192 and TGFB1-miR194-2/192 in stomach adenocarcinoma and esophageal carcinoma respectively. The dots in blue represent tumor samples, while the ones in purple represent normal samples. (**A**) The co-expression pattern of HIF1A-miR194-2 in 447 samples. (**B**) The co-expression pattern of HIF1A-miR192 in 447 samples. (**C**) The co-expression pattern of TGFB1-miR194-2 in 194 samples. (**D**) The co-expression pattern of TGFB1-miR192 in 194 samples.

to limit the research to the designated TF and treatment. The 'Upstream' and 'Downstream' parameters are used to narrow or expand the examined transcriptional regulatory domain of the selected TFs, while the 'Motif' selector is allowed users to restrict their analysis on the binding sites containing identified TF-binding motif of selected TFs.

Four web-based modules, LncRNA, miRNA, OtherNcRNA and Protein, shares the similar selectors mentioned above (Supplementary Figure S2). In the other module named Regulator, users can input a gene that they are interested in and analyzed results with some common selectors and a new selector called 'Type of regulators'. Users also can filter the results by multiple query criteria (Supplementary Figure S3).

On the webpage of Motif, users can choose TFs to browse their binding motifs in different cell types, tissues and conditions. All the identified results are displayed as PWMs, motif logos and their genomic binding regions, which esti-

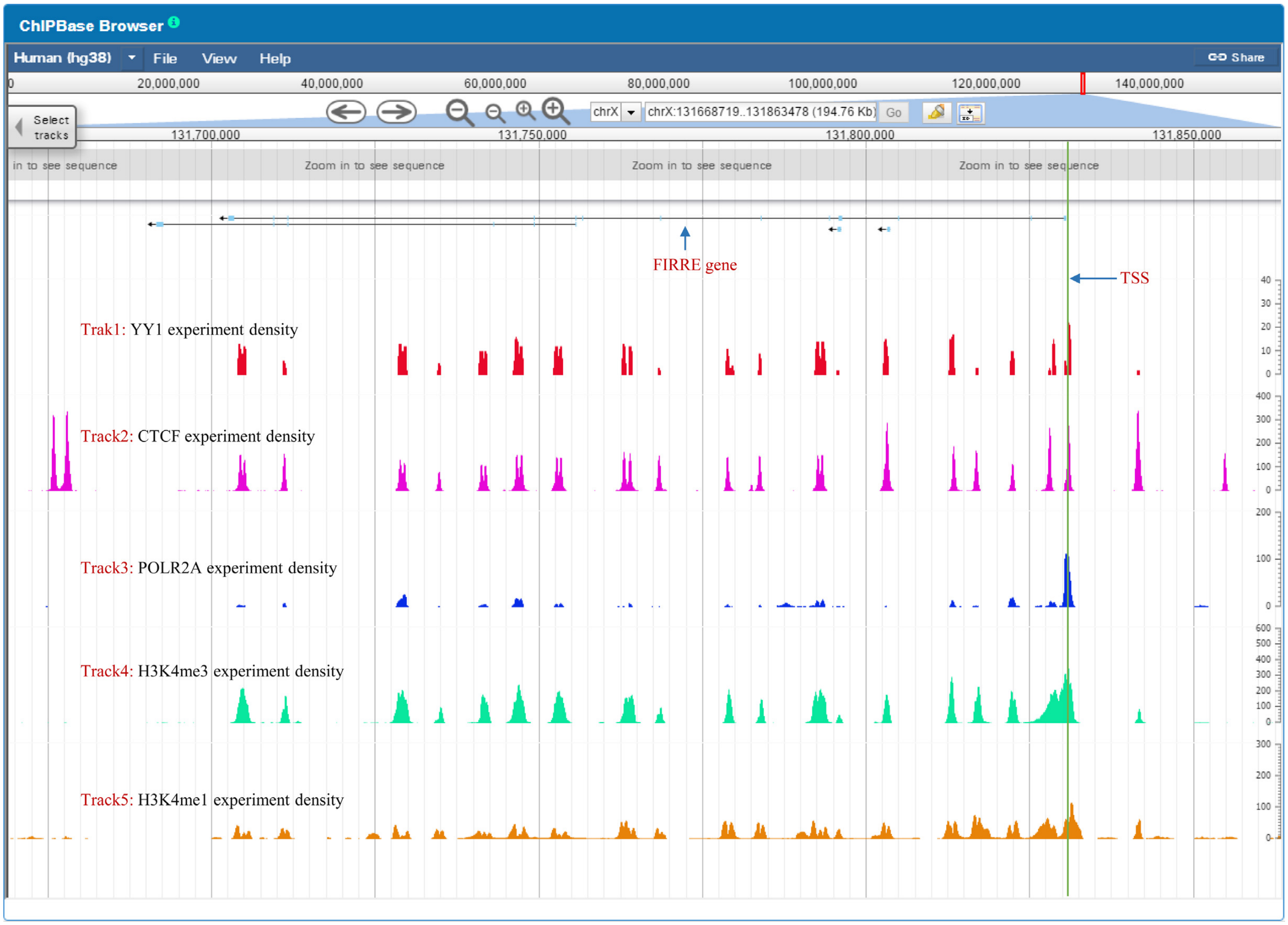

Figure 3. The binding patterns of YY1, CTCF, PORL2A, H3K4me3 and H3K4me1 across FIRRE gene in ChIPBase genome browser. The experiment density tracks of YY1, CTCF, PORL2A, H3K4me3 and H3K4me1 have the similar peak distribution across the FIRRE gene.

mated by *P*-value, percentage of targets and backgrounds, comparisons with known motifs and similar identification results (Supplementary Figure S4).

On the Co-Expression, users can choose the examined study or project (e.g. TCGA Pan-Cancer project) and input two genes in the input boxes, then click ‘Submit’ button to retrieve the query co-expression results (Figure 2 and Supplementary Figure S5). Another web-based tool, ChIP-Function, applies the GO enrichment analysis to predict the functions of TFs and other DNA-binding proteins from their transcriptional targets. In this web-based tool, there are two more parameters on the query page, ‘Adjusted *P*-value’ and ‘GO domain’ (Supplementary Figure S6).

Visualization of genomic binding features around ncRNAs and PCGs with ChIPBase genome browser

We developed ChIPBase genome browser that is built on JBrowse to facilitate visualization of the various ChIP-seq datasets and exploration of TF binding sites (Figure 3). In the main page of the browser, users can query one interested gene name or genomic region in the ‘Position/Search Term’ and select corresponding genome assembly to obtain an integrated view of various genomic features. Users can click the ‘+’ or ‘–’ button at the top to shrink or extend on the center of the annotation tracks window. Users can add tracks by clicking ‘Select Tracks’ button located in the upper-left corner and choose different types of ChIP-seq datasets.

With our genome browser, users can explore some new transcriptional features of genes. For example, the experiment density track of YY1, which was involved in repressing and activating diverse genes (49,50), showed that it discontinuously bound across the entire gene body and promoter region of FIRRE gene. Interestingly, the experiment density tracks of CTCF, POLR2A, H3K4me3 and H3K4me1 shared the similar binding patterns with YY1 (Figure 3).

DISCUSSION AND CONCLUSIONS

By integrating a large set of TF binding sites and their binding motifs derived from ChIP-seq methods and public resources, ChIPBase v2.0 reveals extensive and complex transcriptional regulatory network maps of ncRNAs and PCGs across 10 species.

The current resources for studying the transcriptional regulation of ncNRAs, including TransmiR (51) and CircuitsDB (52), only collected a limited amount of computationally predicted or experimentally supported data of TF-miRNA interactions. Compared with these resources and our previous release version (ChIPBase v1.0) (14), the advances and improvements of ChIPBase v2.0 are as listed follow: (i) ChIPBase v2.0 has been expanded with a large number of ChIP-seq datasets (20 times expansion), which covered thousands of TFs, TCFs, CRFs and histone modifications. (ii) ChIPBase v2.0 provides comprehensive annotations of transcriptional regulation of ncRNAs and PCGs by connecting TFs and their binding motifs to these RNA molecules. (iii) ChIPBase v2.0 allows researchers to get the available TFs that bound upstream and downstream their interested genes with web-based Regulator module. (iv) ChIPBase v2.0 provides genomic coordinates and PWMs of identified TF binding motifs across different cell types, tissues and conditions. These data will facilitate computational or experimental biologists to correlate their results. (v) ChIPBase v2.0 integrates a huge number of RNA-seq and miRNA-seq data across seven species and allows researchers to explore the co-expression relationships of TF-ncRNA and TF-PCG. (vi) ChIPBase v2.0 also can be used to explore the functions of TFs by performing GO analysis on their transcriptional targets. These enriched GO terms may provide valuable insights into the regulatory role and function of each TF. Overall, we have provided a variety of information and web interfaces to facilitate exploration of TF-ncRNA and TF-PCG interaction maps.

FUTURE DIRECTIONS

As the ChIP-seq technology has been applied to more and more DNA-binding proteins, cell types, tissues and conditions, there will be more and more ChIP-seq data in the future. We have developed an automatic pipeline that to automatically analyze ChIP-seq datasets, and then integrate these data into our local database. We will continually maintain and plan to update the database every two months. ChIPBase will continue to improve the computer server performance for storing and analyzing these new incoming data and dedicate to developing tools to decode the transcriptional regulatory networks of ncRNAs and PCGs.

AVAILABILITY

ChIPBase v2.0 is freely available at http://rna.sysu.edu.cn/chipbase/. All the data files in ChIPBase can be downloaded and used in accordance with the GNU Public License and the license of primary data sources.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Ministry of Science and Technology of China; National Basic Research Program [2011CB811300]; National Natural Science Foundation of China [31230042, 31370791, 31401975, 31471223, 91440110, 31671349]; Guangdong Province Funds [S2012010010510, S2013010012457]; The project of Science and Technology New Star in ZhuJiang Guangzhou city [2012J2200025]; Fundamental Research Funds for the Central Universities [2011330003161070, 14lgjc18]; China Postdoctoral Science Foundation [200902348]; Guangdong Province Key Laboratory of Computational Science (in part); Guangdong Province Computational Science Innovative Research Team (in part). Funding for open access charge: National Natural Science Foundation of China [91440110], Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase).
Conflict of interest statement. None declared.

REFERENCES

1. Ponting,C.P., Oliver,P.L. and Reik,W. (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.
2. Mercer,T.R., Dinger,M.E., Sunkin,S.M., Mehler,M.F. and Mattick,J.S. (2008) Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 716–721.
3. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
4. Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.
5. Fatica,A. and Bozzoni,I. (2014) Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.*, **15**, 7–21.
6. Bachellerie,J.P., Cavaille,J. and Huttenhofer,A. (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.
7. Krol,J., Loedige,I. and Filipowicz,W. (2010) The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.*, **11**, 597–610.
8. Marson,A., Levine,S.S., Cole,M.F., Frampton,G.M., Brambrink,T., Johnstone,S., Guenther,M.G., Johnston,W.K., Wernig,M., Newman,J. *et al.* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**, 521–533.
9. Guttman,M., Donaghey,J., Carey,B.W., Garber,M., Grenier,J.K., Munson,G., Young,G., Lucas,A.B., Ach,R., Bruhn,L. *et al.* (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, **477**, 295–300.
10. Lee,B.K., Bhinge,A.A., Battenhouse,A., McDaniell,R.M., Liu,Z., Song,L., Ni,Y., Birney,E., Lieb,J.D., Furey,T.S. *et al.* (2012) Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. *Genome Res.*, **22**, 9–24.
11. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
12. Visel,A., Blow,M.J., Li,Z.R., Zhang,T., Akiyama,J.A., Holt,A., Plajzer-Frick,I., Shoukry,M., Wright,C., Chen,F. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.
13. Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.
14. Yang,J.H., Li,J.H., Jiang,S., Zhou,H. and Qu,L.H. (2013) ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. *Nucleic Acids Res.*, **41**, D177–D187.
15. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets-10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
16. Myers,R.M., Stamatoyannopoulos,J., Snyder,M., Dunham,I., Hardison,R.C., Bernstein,B.E., Gingeras,T.R., Kent,W.J., Birney,E., Wold,B. *et al.* (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
17. Negre,N., Brown,C.D., Ma,L.J., Bristow,C.A., Miller,S.W., Wagner,U., Kheradpour,P., Eaton,M.L., Loriaux,P., Sealfon,R. *et al.* (2011) A cis-regulatory map of the Drosophila genome. *Nature*, **471**, 527–531.
18. Gerstein,M.B., Lu,Z.J., Van Nostrand,E.L., Cheng,C., Arshinoff,B.I., Liu,T., Yip,K.Y., Robilotto,R., Rechtsteiner,A., Ikegami,K. *et al.*

(2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, **330**, 1775–1787.

19. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
20. Speir,M.L., Zweig,A.S., Rosenbloom,K.R., Raney,B.J., Paten,B., Nejad,P., Lee,B.T., Learned,K., Karolchik,D., Hinrichs,A.S. *et al.* (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.*, **44**, D717–D725.
21. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
22. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
23. 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.
24. Yates,A., Akanni,W., Amode,M.R., Barrell,D., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P., Fitzgerald,S., Gil,L. *et al.* (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.
25. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.
26. Zheng,L.L., Li,J.H., Wu,J., Sun,W.J., Liu,S., Wang,Z.L., Zhou,H., Yang,J.H. and Qu,L.H. (2016) deepBase v2.0: identification, expression, evolution and function of small RNAs, LncRNAs and circular RNAs from deep-sequencing data. *Nucleic Acids Res.*, **44**, D196–D202.
27. Zhang,H.M., Liu,T., Liu,C.J., Song,S.Y., Zhang,X.T., Liu,W., Jia,H.B., Xue,Y. and Guo,A.Y. (2015) AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res.*, **43**, D76–D81.
28. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.
29. Pennacchio,L.A., Bickmore,W., Dean,A., Nobrega,M.A. and Bejerano,G. (2013) Enhancers: five essential questions. *Nat. Rev. Genet.*, **14**, 288–295.
30. Maston,G.A., Evans,S.K. and Green,M.R. (2006) Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, **7**, 29–59.
31. Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.
32. Wang,Z.B., Zang,C.Z., Rosenfeld,J.A., Schones,D.E., Barski,A., Cuddapah,S., Cui,K.R., Roh,T.Y., Peng,W.Q., Zhang,M.Q. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.
33. Seligson,D.B., Horvath,S., Shi,T., Yu,H., Tze,S., Grunstein,M. and Kurdistani,S.K. (2005) Global histone modification patterns predict risk of prostate cancer recurrence. *Nature*, **435**, 1262–1266.
34. Koch,C.M., Andrews,R.M., Flicek,P., Dillon,S.C., Karaoz,U., Clelland,G.K., Wilcox,S., Beare,D.M., Fowler,J.C., Couttet,P. *et al.* (2007) The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res.*, **17**, 691–707.
35. Petryszak,R., Keays,M., Tang,Y.A., Fonseca,N.A., Barrera,E., Burdett,T., Fullgrabe,A., Fuentes,A.M.P., Jupp,S., Koskinen,S. *et al.* (2016) Expression Atlas update-an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.*, **44**, D746–D752.
36. The Cancer Genome Atlas Research, N., Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R.M., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
37. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
38. Skinner,M.E., Uzilov,A.V., Stein,L.D., Mungall,C.J. and Holmes,I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.
39. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
40. Arvey,A., Agius,P., Noble,W.S. and Leslie,C. (2012) Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.*, **22**, 1723–1734.
41. Spitz,F. and Furlong,E.E. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**, 613–626.
42. Hino,K., Tsuchiya,K., Fukao,T., Kiga,K., Okamoto,R., Kanai,T. and Watanabe,M. (2008) Inducible expression of microRNA-194 is regulated by HNF-1 alpha during intestinal epithelial cell differentiation. *RNA*, **14**, 1433–1442.
43. Jenkins,R.H., Martin,J., Phillips,A.O., Bowen,T. and Fraser,D.J. (2012) Transforming growth factor beta 1 represses proximal tubular cell microRNA-192 expression through decreased hepatocyte nuclear factor DNA binding. *Biochem. J.*, **443**, 407–416.
44. Yang,F., Xue,X.C., Bi,J.W., Zheng,L.M., Zhi,K.K., Gu,Y. and Fang,G.E. (2013) Long noncoding RNA CCAT1, which could be activated by c-Myc, promotes the progression of gastric carcinoma. *J. Cancer Res. Clin.*, **139**, 437–445.
45. He,X.L., Tan,X.M., Wang,X., Jin,H.Y., Liu,L., Ma,L.M., Yu,H. and Fan,Z.N. (2014) C-Myc-activated long noncoding RNA CCAT1 promotes colon cancer cell proliferation and invasion. *Tumor Biol.*, **35**, 12181–12188.
46. Nissan,A., Stojadinovic,A., Mitrani-Rosenbaum,S., Halle,D., Grinbaum,R., Roistacher,M., Bochem,A., Dayanc,B.E., Ritter,G., Gomceli,I. *et al.* (2012) Colon cancer associated transcript-1: a novel RNA expressed in malignant and pre-malignant human tissues. *Int. J. Cancer*, **130**, 1598–1606.
47. Donohoe,M.E., Zhang,L.F., Xu,N., Shi,Y. and Lee,J.T. (2007) Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch. *Mol. Cell*, **25**, 43–56.
48. Moseley,S.C., Rizkallah,R., Tremblay,D.C., Anderson,B.R., Hurt,M.M. and Chadwick,B.P. (2012) YY1 associates with the macrosatellite DXZ4 on the inactive X chromosome and binds with CTCF to a hypomethylated form in some male carcinomas. *Nucleic Acids Res.*, **40**, 1596–1608.
49. Seto,E., Shi,Y. and Shenk,T. (1991) YY1 is an initiator sequence-binding protein that directs and activates transcription in vitro. *Nature*, **354**, 241–245.
50. Usheva,A. and Shenk,T. (1994) TATA-binding protein-independent initiation: YY1, TFIIB, and RNA polymerase II direct basal transcription on supercoiled template DNA. *Cell*, **76**, 1115–1121.
51. Wang,J., Lu,M., Qiu,C.X. and Cui,Q.H. (2010) TransmiR: a transcription factor-microRNA regulation database. *Nucleic Acids Res.*, **38**, D119–D122.
52. Friard,O., Re,A., Taverna,D., De Bortoli,M. and Cora,D. (2010) CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse. *BMC Bioinformatics*, **11**, 435.