ChIP-Array 2: integrating multiple omics data to construct gene regulatory networks

Panwen Wang[1,2,†], Jing Qin[1,2,†], Yiming Qin[1], Yun Zhu[1,2], Lily Yan Wang[1,2], Mulin Jun Li[1,2], Michael Q. Zhang[3,4] and Junwen Wang[1,2,*]

[1]Centre for Genomic Sciences and Department of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China, [2]Shenzhen Institute of Research and Innovation, The University of Hong Kong, Shenzhen, Guangdong 518057, China, [3]Bioinformatics Division, TNLIST, Tsinghua University, Beijing 100084, China and [4]Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas at Dallas, Dallas, TX 75080, USA

ABSTRACT

Transcription factors (TFs) play an important role in gene regulation. The interconnections among TFs, chromatin interactions, epigenetic marks and cis-regulatory elements form a complex gene transcription apparatus. Our previous work, ChIP-Array, combined TF binding and transcriptome data to construct gene regulatory networks (GRNs). Here we present an enhanced version, ChIP-Array 2, to integrate additional types of omics data including long-range chromatin interaction, open chromatin region and histone modification data to dissect more comprehensive GRNs involving diverse regulatory components. Moreover, we substantially extended our motif database for human, mouse, rat, fruit fly, worm, yeast and *Arabidopsis*, and curated large amount of omics data for users to select as input or backend support. With ChIP-Array 2, we compiled a library containing regulatory networks of 18 TFs/chromatin modifiers in mouse embryonic stem cell (mESC). The web server and the mESC library are publicly free and accessible at http://jjwanglab.org/chip-array.

INTRODUCTION

Deciphering gene regulatory network (GRN) is crucial to understanding the mechanisms of various biological processes and the onset of diseases. Gene transcription can be regulated by multiple factors, such as transcription regulation, epigenetics and post-translational modifications. Previous version of ChIP-Array incorporates *in vivo* transcription factor (TF) binding and transcriptome data to construct a GRN for a TF (1). The TF's bindings are detected by chromatin immunoprecipitation followed by high-throughput technologies (ChIP-seq/chip/exo, collectively, ChIP-X), and the transcriptome data are obtained before and after the TF's perturbation, by either microarray or RNA-sequencing. However, TFs often collaborate with other factors for gene regulation. Recent studies showed that interplay among different TFs, chromatin modifiers and/or epigenetic marks comprehends the functional diversity of gene regulation (2–5). For example, super enhancers aggregating multiple TFs, transcription cofactors, chromatin regulators and polymerase are found to be associated with highly expressed genes that control cell identity and disease (6). Some enhancers can locate as far as 100 kb away from their target genes through chromatin loops, which is hard to identify with traditional technology. Fortunately, the development of advanced chromatin conformation capturing technologies, such as Hi-C and Chromatin Interaction Analysis by Paired-End Tag sequencing (ChIA-PET), enables us to detect long-range enhancers with less false positives (7). Thus, the omics data quantifying these chromatin events provide valuable information to improve GRN construction. Other information, for instance, genome-wide open chromatin regions are also helpful to improve the quality of TF-target identification, since TF binding can be impeded and dissociated by nucleosome (8), and most of TF binding sites (TFBSs) identified by the Encyclopedia of DNA Elements (ENCODE) consortium are located in high DNA-accessible regions (9).

In spite of the development of high-throughput technologies and the emergence of multiple omics data conducted on different regulatory factors, existing tools have not sufficiently integrated those data for GRN construction. Most of these tools either combine only ChIP-X and transcriptome data to reduce the false-positive rate (2,10–12), or require a large number of samples to build their models for the identification of interactions among different types of factors, such as genetic variation, DNA methylation and

miRNA (5,13,14) (Supplementary Section 4). Hence, we developed our enhanced version of ChIP-Array, which is able to incorporate not only ChIP-X and transcriptome data, but also long-range chromatin interaction, open chromatin region and histone modification data, to construct GRNs.

The new version can also construct regulatory networks triggered by a cellular event, such as a change in histone modification, or a cellular perturbation (knock in, knock out, drug treatment, or pathogen infection). To take advantages of public omics data, we also manually curated large volume of omics data of various cell types/lines for users to use in combination with their own data. The algorithm for target identification is also enhanced. In addition to directly determining the targets as the intersection of TFBS-enriched and differentially expressed genes (DEGs) in previous version, we offer another target detection method, 'Rank Product', which is based on relative peak position to transcription start site (TSS), peak intensity and gene expression change (2,15). Besides, we substantially extended our TF motif database to 6584 position weight matrices (PWMs) of 4727 TFs, and for more species including human, mouse, rat, fruit fly, worm, yeast and *Arabidopsis*. To provide better user experience, the web interface is more user-friendly. The events, including TF-binding, long-range chromatin interaction, open chromatin region and histone modification, around a TSS, are intuitively displayed in JBrowse (16). Furthermore, enriched Gene Ontology (GO) and Pathways from the resulting GRN are displayed in the result page. With this new web server, we constructed a network library containing regulatory networks of 18 TFs/chromatin modifiers in mouse embryonic stem cell (mESC) and made it freely accessible.

DESCRIPTION OF ChIP-ARRAY 2

Input data

The main input data contain three parts (Supplementary Table S2): (i) a list of genomic positions containing TF/chromatin modifier's binding peak regions, differential histone modification regions or differential DNA methylation regions. The centered regulator in the constructed GRN can be a TF, a type of histone modification or a biological perturbation, respectively; (ii) a list of differentially expressed genes under the perturbation of a TF/chromatin modifier, or other treatment; (iii) other types of omics data such as long-range chromatin interactions, Dnase hypersensitive sites (DHS) and various histone modifications. For TF binding data, users can either upload their own data, or use our curated ChIP-X data or TFBSs predicted by PWMs (17,18). Besides the general format 'BED' and 'GFF', the server supports peak files generated by popular peak calling tools including CisGenome (19) and MACS (20). To take into account the biological replicates, users can use the peak file generated by PePr (21). To consider differential TF binding or histone modification between two experimental conditions, users can use peak file generated by DBChIP (22) and DiffBind (23). For transcriptome data, ChIP-Array 2 is able to read the standard output of LIMMA (24) and Cuffdiff (25). The server also supports a simple format (gene ID, $\log_2$(fold change) and statistical value) for the essential information of gene expression changes. For convenience, in the transcriptome profile, most commonly-used array probe names and various gene IDs from different platforms and sources can be recognized by ChIP-Array 2. Users can either upload other types of omics data in the 'BED' format or select our curated data from the web server (Supplementary Section 2.1). They can also select the information of enhancers from VISTA database (26) and other experimentally-validated enhancers we curated from literature (Supplementary Section 2.2) for target detection. A sample result and a copy of sample data are available for users to browse before they run their own jobs.

Workflow

The main workflow of ChIP-Array 2 is shown in Figure 1A, ChIP-X and transcriptome data will be combined to detect direct targets, using either the 'Direct', by which the targets will be determined as intersection of ChIP peak-enriched genes and DEGs, or 'Rank Product' method (Supplementary Section 1.2). In the later method, genes are ranked by both peak abundance and expression changes, and scored by the rank product. When a cutoff is used, genes with higher peak abundance and higher expression change are chosen. Here higher peak abundance denotes more peaks with higher intensity and shorter distance to TSS. To detect TF-target relation involved in enhancer-promoter interaction, long-range chromatin interaction data is used to incorporate the TSSs with peaks located in the distal interacting regions. If the direct target is a TF, the same method will be applied to detect its downstream targets with our curated ChIP-X data, or putative TFBSs predicted by Position Weigh Matrix (PWM) method (Supplementary Section 1.1). When putative TFBSs are used, open chromatin regions and histone modifications are further used to filter out the inactive, non-tissue-specific TFBSs. Finally, the results will be presented in a web page, and all the intermediate and resultant files are downloadable for users to perform further analysis. Complement to single factor analysis, we offer an opportunity for users to study the synergy among regulatory factors by co-occupancy analysis (Supplementary Section 1.3).

Result page

As shown in Figure 1B, the result page consists of four main parts. The first part is a GRN shown in CytoscapeWeb (27) or rendered by HTML5 using cytoscape.js (http://js.cytoscape.org/) if flash is not supported by user's browser. By default, the number of nodes is limited to about 80 to avoid heavy burden of displaying the network. However, users can choose to show the full network by clicking the link on the right side. Each node or edge is clickable to show the details including the regulation relations, open chromatin regions, histone modifications and binding regions. These regions can be highlighted in next tab in JBrowse. If the TFBSs are detected by ChIP-X data, the motif enrichment analysis will be performed by MEME Suite (28), which can discover *de novo* motifs enriched in the surrounding region. MEME Suite will also compare the discovered motifs and the known motifs in our database, and find the potential co-occupied factors for the query factor. The
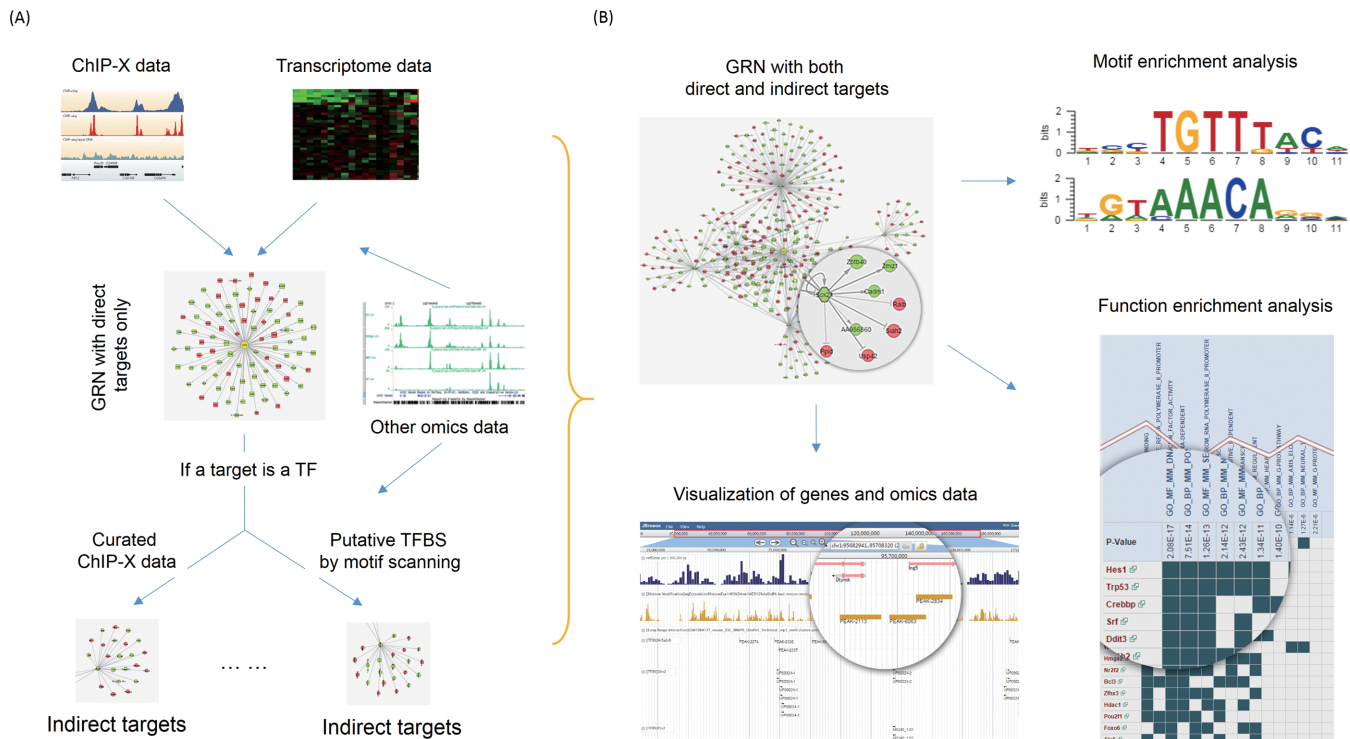
Figure 1. Workflow (A) and results (B) of ChIP-Array 2. (**A**) Direct targets are identified by combining ChIP-X and transcriptome data. Interplays between the TF of interest and other regulatory factors/target genes are supported by other omics data. Then indirect targets are detected by curated ChIP-X data or predicted TFBSs with the assistance of other omics data. (**B**) The results are composed of four parts: the resulting GRN shown in CytoscapeWeb, motif enrichment analysis by MEME Suite, functional enrichment analysis, and visualization in JBrowse.

server can also perform gene set enrichment analysis for the network inferred, against Gene Ontology (GO) terms and annotated pathways. A gene/gene-set overlap matrix is generated accordingly, in which a gene is highlighted if it is present in a particular GO term/pathway. For humans, the gene sets of GO and pathways are downloaded from MSigDB (29). The gene sets for other species are obtained from http://bioinformatics.sdstate.edu/gskb/db/, generated by bioinformatics research group of South Dakota State University.

EXAMPLES

With the omics data we collected from Gene Expression Omnibus (GEO) (Supplementary Table S3), we built a TF/chromatin modifier GRN library for mESC. It contains GRNs controlled by 18 TFs/chromatin modifiers. This library combines data from ChIP-seq of the TFs/chromatin modifiers, transcriptome profiling under their perturbations, together with ChIA-PET, DNase-seq and ChIP-seq of H3K4me3, H3K4me1, H3K27ac, H3K27me3 and H3K9me3. GRNs can be viewed in 'mESC Library'. Integration of diverse omics data significantly improves the target detection and our understanding of gene transcription regulation (Figure 2 and Supplementary Section 4). For example, with ChIA-PET data, ChIP-Array 2 is able to find gene clusters that are highly related to the functions of the TF of interest (Figure 2A). Pou5f1 (also known as Oct4, an important TF for the pluripotency of mESC), which binds on the promoters of Sox2 and Fgf4, is well known to reg-

ulate these two genes (30,31). Through long-range interaction, we found that both genes are connected to Ier2, whose promoter is also bound by Pou5f1. Ier2 is a member of immediate early response proteins that can be induced by the cytokine leukaemia inhibitory factor (LIF), and consistently shows significant down-regulation under the knockdown of Pou5f1 in mESC (32–34). Results from ChIP-Array 2 indicate the co-regulation of Ier2, Sox2 and Fgf4 by Pou5f1 via the aggregating evidence of Pou5f1 binding, down-regulation under Pou5f1 knockdown and the physical interactions among their promoters, which has not been reported by individual studies generating these omics data. Besides, long-range chromatin interaction links TF binding sites in distal elements to the target gene promoter.

ChIP-Array 2 also shows target genes that do not have significant TF binding sites in promoters, but are physically linked to a distal element with a strong TF binding signal (Figure 2B). For instance, although previous studies have reported high correlation between the expressions of Pou5f1 and Slc7a3 (5,32), the mechanism on how they are regulated was not clear. ChIP-Array 2 provides the direct evidence on how Pou5f1 regulates Slc7a3 through an inter-chromosomal interaction between an enhancer and the Slc7a3 promoter. These targets may be overlooked by previous methods in which long-range chromatin interactions were not considered. In addition, ChIP-Array 2 allows the study of interplays between TF and histone modifications (Figure 2C). Tcl1 and Cdx2 are two targets activated and repressed by Pou5f1, respectively. Promoters of both genes are bound by Pou5f1, however, Tcl1 promoter is
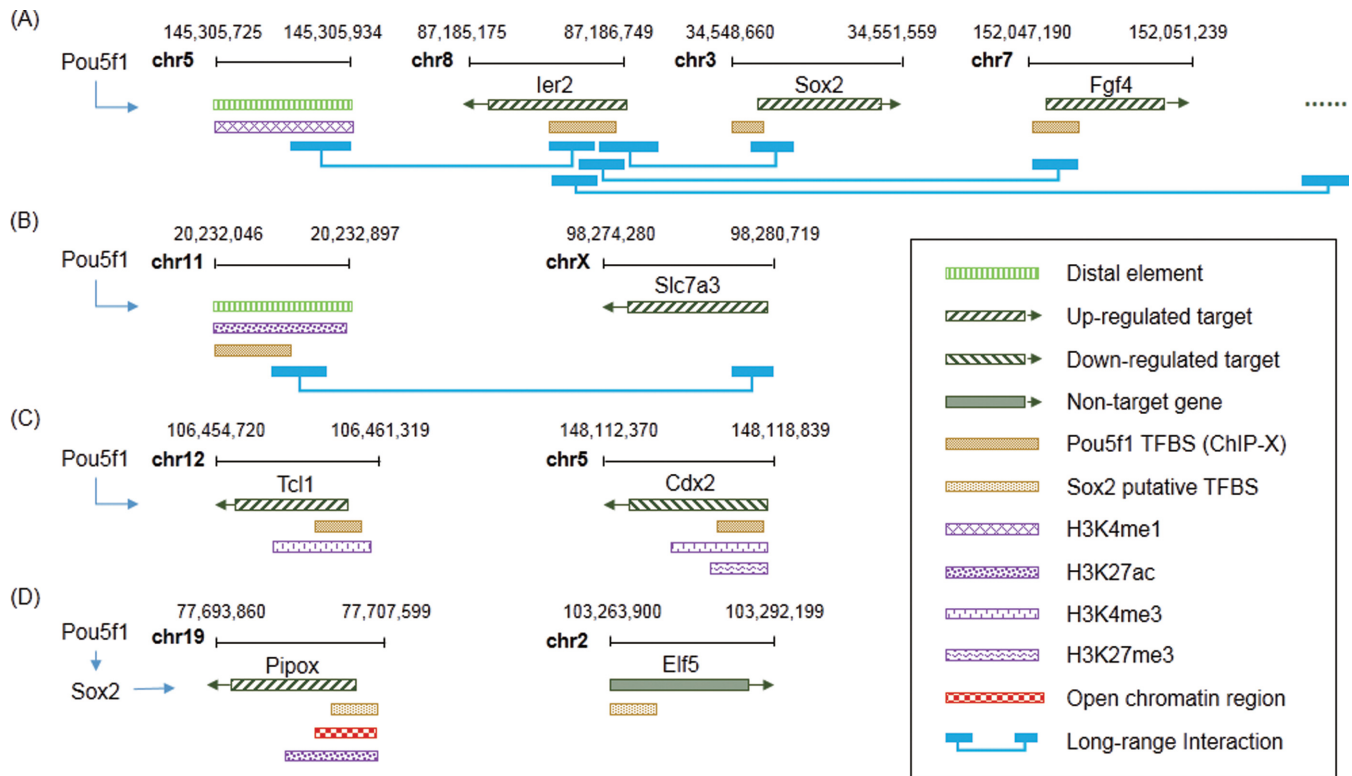
Figure 2. Example targets of Pou5f1 identified by ChIP-Array 2 with multiple omics data. (**A**) A group of genes and distal elements connect to each other through long-range chromatin interactions. Each of the genes has the Pou5f1 binding signals from ChIP-seq, and knockdown of Pou5f1 significantly changes the gene expression. (**B**) Distal elements containing ChIP-seq peaks or putative TFBSs may link to the target genes without significant binding sites. These targets may be overlooked by previous tools where the long-range chromatin interactions are not considered. (**C**) Target genes can be activated (Tcl1) or suppressed (Cdx2) by Pou5f1 when the histone modifications at their promoters are different. (**D**) Histone modification marks and open chromatin regions are used to filter the putative TFBSs for indirect target identification. Since they indicate the tissue-specific accessibility of the TFBSs, TFBSs located in these regions are more likely to be active than those beyond the regions. Thus, putative TFBSs outside of histone modification marks and open chromatin regions are not considered for indirect target inference.

modified with H3K4me3 but not H3K27me3, while Cdx2 promoter is bivalently modified by both marks in mESC. This implies that different histone modifications at Pou5f1 binding promoters may act differently, causing up- or down-regulation effects on the targets. Moreover, incorporating these histone modifications, as well as data identifying open chromatin regions, we successfully filtered inactive putative TFBSs predicted by sequence scanning when detecting indirect targets. Figure 2D shows two example genes. When searching indirect targets of Pou5f1, both Pipox and Elf5 contain predicted Sox2 binding sites on their promoters. They are significantly down- and up-regulated respectively when Pou5f1 is knocked down. Since Sox2 is a direct target of Pou5f1, both Pipox and Elf5 are identified as indirect targets of Pou5f1 by the old version of ChIP-Array. However, combining the histone modifications and open chromatin regions via ChIP-Array 2, which mark the chromatin accessibility of TFBSs, we found that only Pipox is the true target of Sox2. Elf5 is filtered out because of lacking any marks to support the activity of the predicted Sox2 binding site on its promoter. The direct regulation between Sox2 and Pipox is also confirmed by a ChIP-seq peak of Sox2 on its promoter and expression change when Sox2 is knocked down, while Elf5 shows no evidence in these studies (data not shown) (32,35,36). All these findings can be visualized

via JBrowse in the Pou5f1 network stored in ChIP-Array 2 mESC library by inputting the genomic positions of these genes and ticking the relevant data tracks in the full-screen mode of JBrowse.

DISCUSSION

ChIP-Array 2 integrates ChIP-X and transcriptome data, together with other omics data to construct GRNs. However, it does not require all the above data, and can run even without either ChIP-X or transcriptome data. If transcriptome data are not provided, only direct targets will be detected by ChIP-X data or predicted TFBSs, but no indirect targets will be inferred due to the high false-positive rate. Without ChIP-X data for direct target detection, TFBSs obtained by computational prediction will be utilized to infer the putative targets. To reduce the false positives, users are encouraged to use at least one type of omics data to detect active binding sites for the query factor. Unlike a standalone program, which may run hours for a job, ChIP-Array 2 aims to produce the results in tens of minutes, or even minutes. Some of the parameters are set in advance to reduce running time, but the flexibility may be lost accordingly. For instance, the putative TFBSs are scanned at several certain cutoffs, and then stored on the server, so that users can only

select the cutoffs that we used in the pre-scanning. For the motif analysis, we just use peaks in promoters of targets, ignoring other peaks since it may take up to hours for motif discovery. Users can download the job and extract all the peaks and perform the whole-genome motif analysis using standalone MEME Suite.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

REFERENCES

1. Qin,J., Li,M.L.J., Wang,P.W., Zhang,M.Q. and Wang,J.W. (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.
2. Wang,S., Sun,H., Ma,J., Zang,C., Wang,C., Wang,J., Tang,Q., Meyer,C.A., Zhang,Y. and Liu,X.S. (2013) Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat. Protoc.*, **8**, 2502–2515.
3. Wang,L.L.Y., Wang,P.W., Li,M.L.J., Qin,J., Wang,X.W., Zhang,M.Q. and Wang,J.W. (2011) EpiRegNet: constructing epigenetic regulatory network from high throughput gene expression data for humans. *Epigenetics*, **6**, 1505–1512.
4. Musselman,C.A., Lalonde,M.E., Cote,J. and Kutateladze,T.G. (2012) Perceiving the epigenetic landscape through histone readers. *Nat. Struct. Mol. Biol.*, **19**, 1218–1227.
5. Guan,D., Shao,J., Zhao,Z., Wang,P., Qin,J., Deng,Y., Boheler,K.R., Wang,J. and Yan,B. (2014) PTHGRN: unraveling post-translational hierarchical gene regulatory networks using PPI, ChIP-seq and gene expression data. *Nucleic Acids Res.*, **42**, W130–W136.
6. Hnisz,D., Abraham,B.J., Lee,T.I., Lau,A., Saint-Andre,V., Sigova,A.A., Hoke,H.A. and Young,R.A. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.
7. Dean,A. (2006) On a chromosome far, far away: LCRs and gene expression. *Trends Genet.*, **22**, 38–45.
8. Luo,Y., North,J.A., Rose,S.D. and Poirier,M.G. (2014) Nucleosomes accelerate transcription factor dissociation. *Nucleic Acids Res.*, **42**, 3017–3027.
9. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
10. Wu,G. and Ji,H. (2013) ChIPXpress: using publicly available gene expression data to improve ChIP-seq and ChIP-chip target gene ranking. *BMC Bioinformatics*, **14**, 188–189.
11. Maienschein-Cline,M., Zhou,J., White,K.P., Sciammas,R. and Dinner,A.R. (2012) Discovering transcription factor regulatory targets using gene expression and binding data. *Bioinformatics*, **28**, 206–213.
12. Qin,J., Hu,Y., Xu,F., Yalamanchili,H.K. and Wang,J. (2014) Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods*, **67**, 294–303.
13. Zhang,S., Liu,C.C., Li,W., Shen,H., Laird,P.W. and Zhou,X.J. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.*, **40**, 9379–9391.
14. Li,W., Zhang,S., Liu,C.C. and Zhou,X.J. (2012) Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, **28**, 2458–2466.
15. Breitling,R., Armengaud,P., Amtmann,A. and Herzyk,P. (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, **573**, 83–92.
16. 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.
17. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
18. Hannenhalli,S., Putt,M.E., Gilmore,J.M., Wang,J.W., Parmacek,M.S., Epstein,J.A., Morrisey,E.E., Margulies,K.B. and Cappola,T.P. (2006) Transcriptional genomics associates FOX transcription factors with human heart failure. *Circulation*, **114**, 1269–1276.
19. Jiang,H., Wang,F., Dyer,N.P. and Wong,W. (2010) CisGenome Browser: a flexible tool for genomic data visualization. *Bioinformatics*, **26**, 1781–1782.
20. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137–R138.
21. Zhang,Y., Lin,Y.-H.H., Johnson,T.D., Rozek,L.S. and Sartor,M.A. (2014) PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. *Bioinformatics*, **30**, 2568–2575.
22. Liang,K. and Keles,S. (2012) Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics*, **28**, 121–122.
23. Ross-Innes,C.S., Stark,R., Teschendorff,A.E., Holmes,K.A., Ali,H.R., Dunning,M.J., Brown,G.D., Gojis,O., Ellis,I.O., Green,A.R. *et al.* (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389–393.
24. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47–e48.
25. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
26. Visel,A., Minovitsky,S. and Dubchak,I. (2007) VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.
27. Lopes,C.T., Franz,M., Kazi,F., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.
28. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
29. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
30. Schoorlemmer,J. and Kruijer,W. (1991) Octamer-dependent regulation of the kFGF gene in embryonal carcinoma and embryonic stem cells. *Mech. Dev.*, **36**, 75–86.
31. Ding,L., Paszkowski-Rogacz,M., Nitzsche,A., Slabicki,M.M., Heninger,A.K., de Vries,I., Kittler,R., Junqueira,M., Shevchenko,A., Schulz,H. *et al.* (2009) A genome-scale RNAi screen for Oct4 modulators defines a role of the Paf1 complex for embryonic stem cell identity. *Cell. Stem Cell*, **4**, 403–415.
32. Nishiyama,A., Sharov,A.A., Piao,Y., Amano,M., Amano,T., Hoang,H.G., Binder,B.Y., Tapnio,R., Bassey,U., Malinou,J.N. *et al.* (2013) Systematic repression of transcription factors reveals limited patterns of gene expression changes in ES cells. *Sci. Rep.*, **3**, 1390.
33. Loh,Y.H., Wu,Q., Chew,J.L., Vega,V.B., Zhang,W., Chen,X., Bourque,G., George,J., Leong,B., Liu,J. *et al.* (2006) The Oct4 and

Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, **38**, 431–440.

34. Trouillas,M., Saucourt,C., Guillotin,B., Gauthereau,X., Ding,L., Buchholz,F., Doss,M.X., Sachinidis,A., Hescheler,J., Hummel,O. *et al.* (2009) Three LIF-dependent signatures and gene clusters with atypical expression profiles, identified by transcriptome studies in mouse ES cells and early derivatives. *BMC Genomics*, **10**, 73.

35. 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.

36. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.