Genet. Res., Camb. (2014), vol. 96, e005. © Cambridge University Press 2014
doi:10.1017/S0016672314000081

The comprehensive transcriptional analysis in *Caenorhabditis elegans* by integrating ChIP-seq and gene expression data

KAN HE[1]*, JIAOFANG SHAO[2], ZHONGYING ZHAO[2] AND DAHAI LIU[1]*
[1]*Center for Stem Cell and Translational Medicine, School of Life Sciences, Anhui University, Hefei City, People's Republic of China*
[2]*Department of Biology, Hong Kong Baptist University, Hong Kong, China*

(*Received 1 February 2014; revised 26 February 2014; accepted 27 February 2014*)

Summary

The fundamental step of learning transcriptional regulation mechanism is to identify the target genes regulated by transcription factors (TFs). Despite numerous target genes identified by chromatin immunoprecipitation followed by high-throughput sequencing technology (ChIP-seq) assays, it is not possible to infer function from binding alone *in vivo*. This is equally true in one of the best model systems, the nematode *Caenorhabditis elegans* (*C. elegans*), where regulation often occurs through diverse TF binding features of transcriptional networks identified in modENCODE. Here, we integrated ten ChIP-seq datasets with genome-wide expression data derived from tiling arrays, involved in six TFs (HLH-1, ELT-3, PQM-1, SKN-1, CEH-14 and LIN-11) with tissue-specific and four TFs (CEH-30, LIN-13, LIN-15B and MEP-1) with broad expression patterns. In common, TF bindings within 3 kb upstream of or within its target gene for these ten studies showed significantly elevated level of expression as opposed to that of non-target controls, indicated that these sites may be more likely to be functional through up-regulating its target genes. Intriguingly, expression of the target genes out of 5 kb upstream of their transcription start site also showed high levels, which was consistent with the results of following network component analysis. Our study has identified similar transcriptional regulation mechanisms of tissue-specific or broad expression TFs in *C. elegans* using ChIP-seq and gene expression data. It may also provide a novel insight into the mechanism of transcriptional regulation not only for simple organisms but also for more complex species.

1. Introduction

Nowadays, *Caenorhabditis elegans* (*C. elegans*) is regarded as a model biological system, in which dozens of transcription factors (TFs) have now been identified and characterized by a combination of genetics, molecular screens and the genome sequencing project. Sufficient information for those TFs is available to suggest the important roles in cell determination, differentiation or other functions of them. However, only a small fraction of *C. elegans* TFs have so far been identified and studied on the candidate target genes and their expression patterns.

As we all know, the most basic step in the study of gene regulation is to uncover the target genes regulated by TFs. Recently, with the extensive application of chromatin immunoprecipitation followed by high-throughput sequencing technology (ChIP-seq), genome-wide identification of TF binding events becomes more facilitated. Nowadays, the numerous genome-wide TF binding data are freely available from the databases of ENCODE for human and mouse as well as modENCODE for *C. elegans* and *Drosophila melanogaster*, which may provide useful biological insights for understanding the mechanisms of transcriptional regulation. Nevertheless, the binding sites alone are not sufficient to infer regulation. Lots of previous studies were focused on identifying the target genes controlled by TFs and dissecting the transcriptional regulatory networks underlying the relevant biological process through a simple peak-based method (Boyer *et al.*, 2005; Chen *et al.*, 2008; Kim *et al.*, 2008). In those studies, the genes with promoters overlapping with one or more peaks were identified as the targets in most cases, but less information on the genes of other regions in genome was reported, especially for the regions out of 2 kb transcription

* Corresponding author: Center for Stem Cell and Translational Medicine, School of Life Sciences, Anhui University, Hefei City, Anhui 230601, People's Republic of China. Tel: +86-0551-63861140. Fax: +86-0551-63861140. E-mail: hekan803@gmail.com and dliu2009@gmail.com

Table 1. *The summary of collected ChIP-seq and gene expression data*

TFs	Type	Expression pattern	Stage in modENCODE	Stage in expression arrays
HLH-1	bHLH	Body wall muscle	Emb	LE-bwm
ELT-3	GATA 4/5/6	Hypodermis	Fed L1	LE-hypodermis
PQM-1	Zn finger	Intestine	L4	L2-intestine
SKN-1	bZIP	pharynx	Fed L1	LE-PhM
CEH-14	Homeodomain	Head neurons	L2	L2-panneural
LIN-11	Homeodomain	Head neurons	L2	L2-panneural
CEH-30	Homeodomain	Broad expression	LE	LE
LIN-13	DRM (Zn finger)	Broad expression	Emb	Emb
LIN-15B	Ac family	Broad expression	L3	L3
MEP-1	NuRD (Zn finger)	Broad expression	Emb	Emb

start site (TSS) of a gene. In this case, some of the important long-range interactions were missed and it is unable to distinguish functional issues from non-functional binding sites. Furthermore, some of TF networks using ChIP-seq and gene expression data based on matrix decomposition have been constructed to unravel transcriptional regulatory programmes in complex systems (Yan *et al.*, 2013; Guan *et al.*, 2014).

Meanwhile, gene expression profiles of more than 30 different cells and developmental stages as well as whole-animal RNA isolated from seven different developmental stages using tiling arrays have been generated to produce a spatial and temporal map of *C. elegans* gene expression (Spencer *et al.*, 2011). It provides a basis for establishing the roles of individual genes in cellular differentiation by measuring native transcripts from a broad array. The datasets of relative gene expression levels across tissue types and developmental stages are available online for further study.

It will be clear that the integration of the genome sequencing information and gene expression data in genome wide using *C. elegans* as a model biological system will result in uncovering the comprehensive transcriptional mechanisms. We hypothesized that a particular gene set regulated by a TF should behave in some measurably consistent manner. In this study, we tested the hypothesis by integrating TF binding and gene expression data of *C. elegans* in genome wide to further explore the mechanism of transcriptional regulation, which can be applied to other species.

2. Materials and methods

The summary of collected ChIP-seq and gene expression data was shown in Table 1. The workflow of gene expression and ChIP-seq data collection and integration analysis was shown in Fig. 1 and described as follows.

(i) *ChIP-seq datasets collection*

The ten ChIP-seq datasets of six TFs (HLH-1, ELT-3, PQM-1, SKN-1, CEH-14 and LIN-11) with tissue-specific and four TFs (CEH-30, LIN-13, LIN-15B and MEP-1) with broad expression patterns were collected from modENCODE database (Gerstein *et al.*, 2010; Niu *et al.*, 2011; Landt *et al.*, 2012). Briefly, the raw sequences were aligned to *C. elegans* genome (WS232) using ELAND allowing up to two mismatches per read. And peaks were called using PeakSeq with the threshold of *P*-value as 0·05 (Rozowsky *et al.*, 2009).

(ii) *Target gene assignment and classification*

To identify candidate target genes possibly regulated by the TFs involved in the ChIP experiments, 5 kb regions both up- and downstream of a peak summit were first searched for transcriptional start (TSS) and end (TES) sites in the *C. elegans* genome (WS232) as described previously (Zhong *et al.*, 2010). If no TSS was annotated for a gene, the TSSs were arbitrarily set to the positions 150 bases upstream of the translational start sites. A total of 31 640 protein-coding transcripts from 20 553 genes were used for target identifications. The targets of each ChIP-seq study were assigned to unique gene and the target genes were classified into 15 classes based on their positions relative to the peak summits. Those with peak summit within the gene region were defined as the class of C0; those with peak summit located 0–5 kb upstream the TSS were defined as C1–5 for each 1 kb interval, respectively, whereas 0–5 kb downstream the TES as C11–15, respectively (Fig. 2(*a*)). For those peaks with no candidate genes in the 5-kb flanking region, the two nearest genes were found. The nearest gene upstream the peak was defined as C21, and the second nearest was C22. Similarly, the nearest gene downstream the peak was defined as C23, and the second nearest was C24 (Fig. 2(*b*)). If the same gene could be assigned to several classes due to the different peaks nearby, it was assigned into a class with the smallest number (Fig. 2(*c*)). For example, if one gene could belong to both C1 and C4, it was defined as C1 target finally.

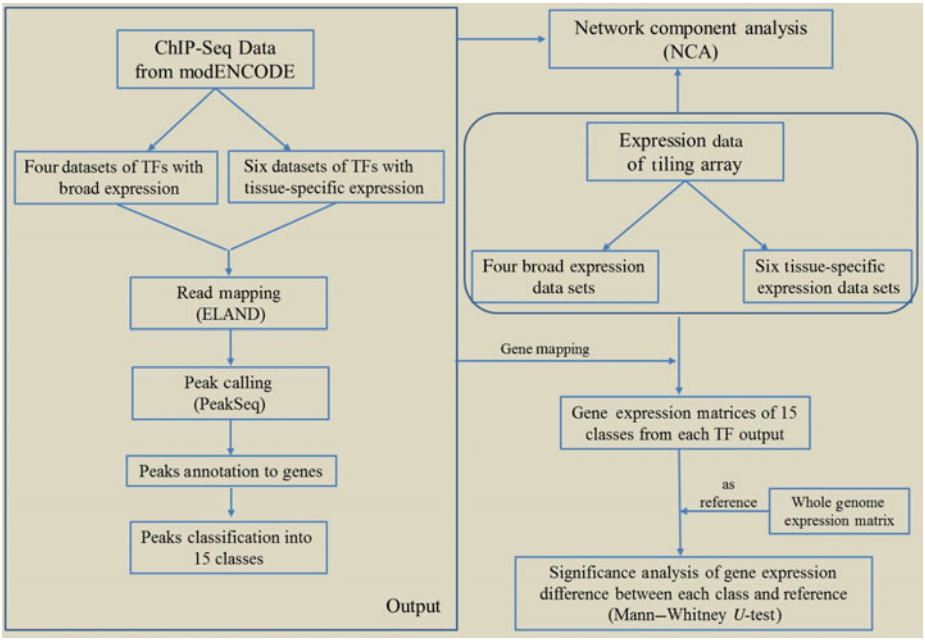

Fig. 1. The workflow of gene expression and ChIP-seq data collection and integration analysis.

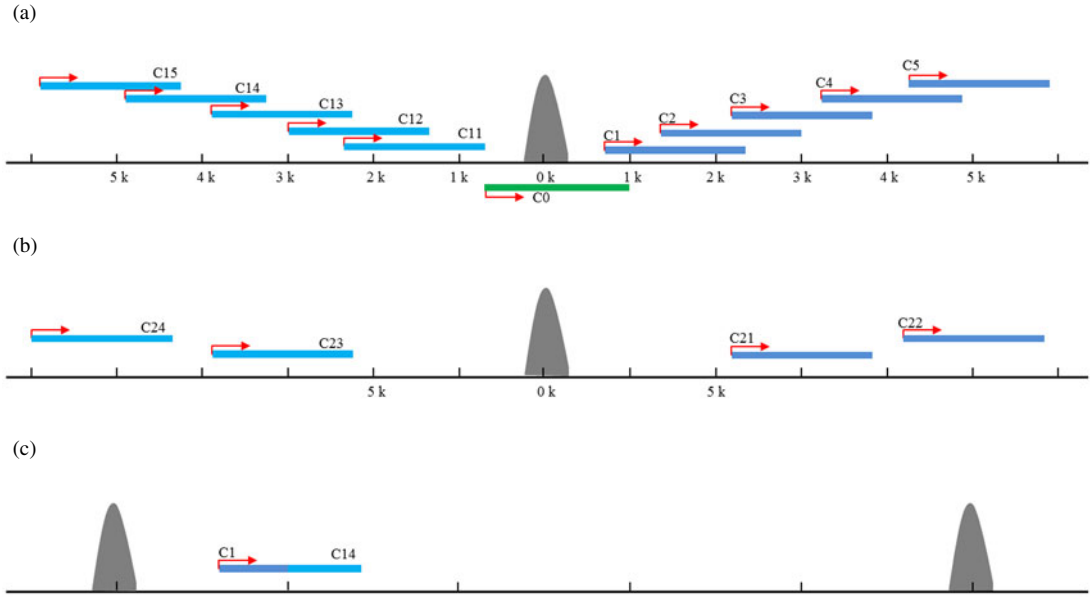

Fig. 2. The target genes assignment and classification. The targets of each ChIP-seq study were assigned to unique gene and the target genes were classified into 15 classes based on their positions relative to the peak summits. Those with peak summit within the gene region were defined as the class of C0; those with peak summit located 0–5 kb upstream the TSS were defined as C1–5 for each 1 kb interval, respectively, whereas 0–5 kb downstream the TES as C11–15, respectively (*a*). For those peaks with no candidate genes in the 5-kb flanking region, the two nearest genes were found. The nearest gene upstream the peak was defined as C21, and the second nearest was C22. Similarly, the nearest gene downstream the peak was defined as C23, and the second nearest was C24 (*b*). If the same gene could be assigned to several classes due to the different peaks nearby, it was assigned into a class with the smallest number (*c*).

(iii) *Association of binding site locations with tissue-specific or broad expression*

For association of the target genes of each TF with those enriched in tissue-specific cells or broad expression of whole animal based on microarray expression assays, expression values of all genes with corresponding patterns were derived from tiling array data (Spencer *et al.*, 2011). The summary of involved ChIP-seq and gene expression data were shown in Table 1. For each gene, the normalized log2 intensities from three replicates were averaged

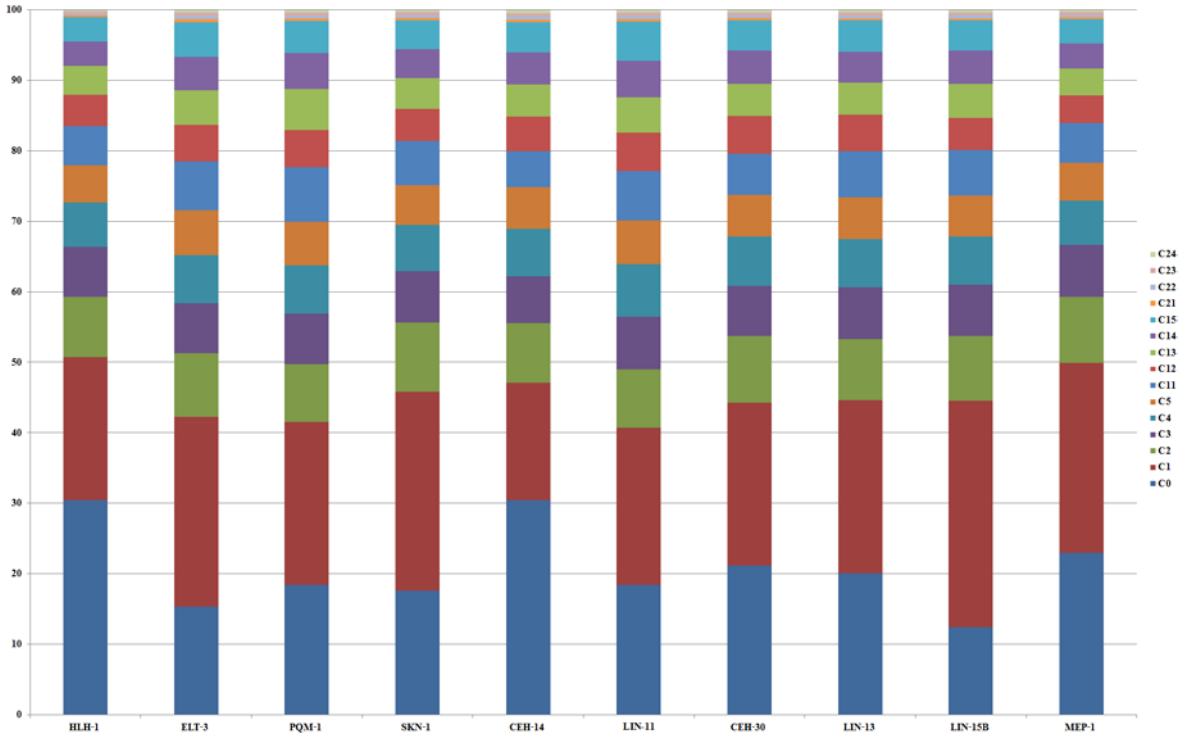

Fig. 3. The proportion of target genes in each class of ten ChIP-seq studies. The bar charts show the proportion of target genes in each class of ten ChIP-seq studies, including the TFs of HLH-1, ELT-3, PQM-1, SKN-1, CEH-14 and LIN-11 with tissue-specific expression patterns as well as CEH-30, LIN-13, LIN-15B and MEP-1 with broad expression patterns. *X*-axis represents the different ChIP-seq studies; *Y*-axis represents the percentage (%) of target genes in each class (C0–C24).

to give rise to its expression value. The expression values of the target genes were extracted for each gene. To plot the tissue-specific or broad expression of different classes of the target genes, those that all of whole genome genes were used as a reference and the plotting was performed with 'vioplot' package in *R* (Hintze JaN, 1998). Mann–Whitney *U*-test was performed to calculate the significance levels (*P*-value) of gene expression difference between reference and the target genes of each class with cutoff of *P*-value as 0·001.

(iv) *Calculation of TF activity*

With both expression values and initial binding relationship provided, we can estimate transcription factor activity using proper mathematical model. Here, we applied network component analysis (NCA) method to calculate the transcription factor activity in 15 different classes for each TF (Liao *et al.*, 2003). Generally, the values for target genes were set to be 1 in initial transcription factor activity matrix, whereas all others were set to be 0. The final transcription factor activity was estimated by matrix decomposition, and the value will vary from 0 to some number bigger than 1.

3. Results and discussion

(i) *Most of binding sites are located within 4 kb upstream of TSS*

Based on the assignment and classification of ChIP-seq targets for all ten TFs, the target genes were classified into 15 classes, from C0 to C24 (Fig. 2, see in the Materials and methods). According to the proportion of target genes in each class of ten ChIP-seq studies, more than 60% of targets were classified from C0 to C4 (Fig. 3). It suggested that most of binding sites are located within 4 kb upstream of TSS, not only for the regulation of TFs with tissue-specific expression pattern but also for the regulation of TFs with broad expression pattern.

(ii) *Integration analysis of binding site locations with tissue-specific expression data*

Actually, there are several major transcriptional regulators of tissue-specific differentiation in *C. elegans*, which are expressed in specific tissues where they are known to bind and regulate diverse well-studied tissue-specific genes. Among the 22 TFs involved in the study of ChIP-seq as part of the modENCODE Consortium, there are some of tissue-specific TFs, including HLH-1 for body wall muscle (BWM), ELT-3

Table 2. *The number of target genes in each class for ten TFs*

TF (number)	C0	C1	C2	C3	C4	C5	C11	C12	C13	C14	C15	C21	C22	C23	C24	Total
HLH-1	4079	2716	1134	955	844	708	736	599	542	466	465	34	32	42	32	13 384
ELT-3	1224	2151	724	564	544	509	557	414	389	380	390	38	34	35	36	7989
PQM-1	1537	1925	684	601	569	512	643	444	483	426	382	29	30	37	35	8337
SKN-1	1860	2975	1038	765	692	593	665	478	466	432	431	40	40	46	32	10 553
CEH-14	2359	1295	654	520	517	460	396	382	352	356	333	24	40	34	37	7759
LIN-11	802	975	358	325	326	272	306	235	220	226	242	16	20	18	19	4360
CEH-30	2000	2188	899	675	661	561	552	508	429	445	409	38	33	34	37	9469
LIN-13	1642	2013	707	599	562	483	539	421	377	353	365	22	34	32	36	8185
LIN-15B	925	2391	683	543	509	433	481	338	363	347	322	22	30	30	28	7445
MEP-1	2761	3233	1123	894	750	644	682	468	455	432	412	33	37	47	42	12 013

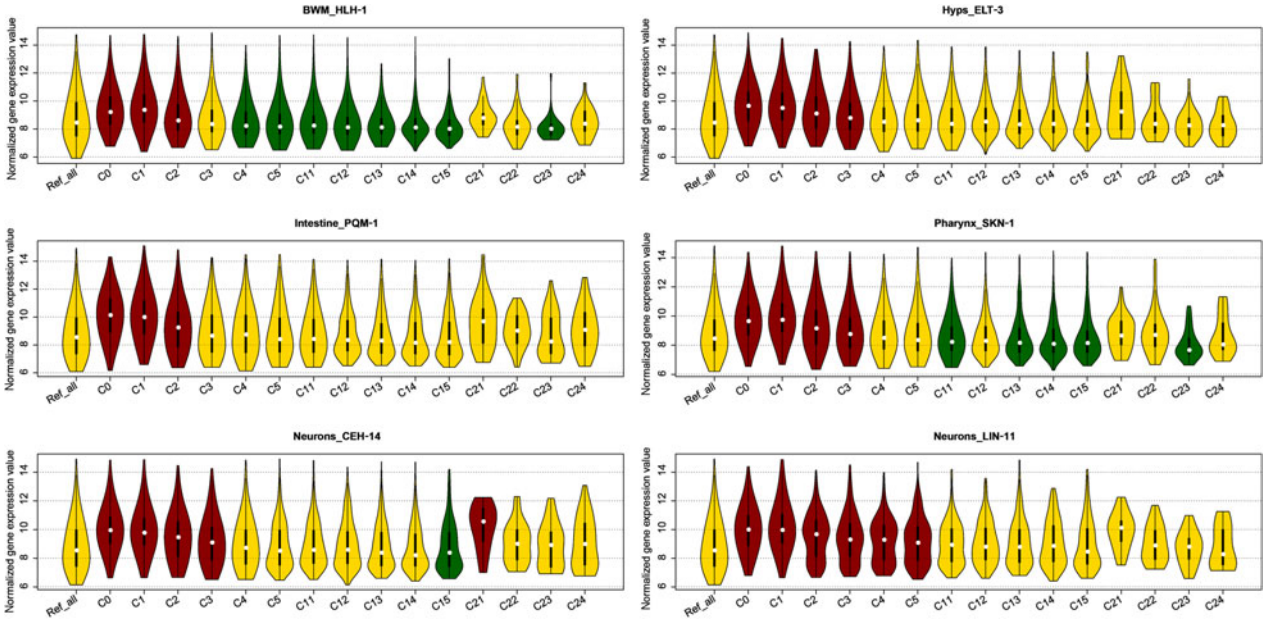

Fig. 4. The expression patterns of tissue specific TFs target genes. The Violin plots of normalized expression levels in specific tissues for different categories of corresponding TF target genes were performed for six TFs with tissue-specific expression pattern, including HLH-1 for body wall muscle (BWM), ELT-3 for hypodermis (Hyps), PQM-1 for intestine, SKN-1 for pharynx as well as CEH-14 and LIN-11 both for neurons, respectively. *X*-axis represents 15 different classes of target genes and reference of genome gene expression; *Y*-axis represents the normalized gene expression value. Significant levels by Mann–Whitney *U*-test ($P < 10^{-3}$) are indicated in the plots, dark-red for significantly up-regulated and dark-green for significantly down-regulated.

for hypodermis (Hyps), PQM-1 for intestine, SKN-1 for pharynx as well as CEH-14 and LIN-11 both for neurons (Table 1) (Niu *et al.*, 2011). Through read mapping with ELAND and peak calling with PeakSeq as well as the peak gene annotation, 13 384 target genes for HLH-1, 7989 target genes for ELT-3, 8337target genes for PQM-1, 10 553 target genes for SKN-1, 7759 target genes for CEH-14 and 4360 target genes for LIN-11 were identified, respectively. According to the binding site locations, the target genes were classified into 15 different categories (C0–C24) for all the six TFs studies (Table 2, see in Section 2).

Numerous target genes have been identified by the ChIP-seq assays, but we are not certain whether the bindings are functional or not *in vivo* as well as whether the corresponding expression behaviours of each tissue-specific TF target genes are similar or not. We addressed the question by the integration analysis of the tissue-specific expression data derived from tiling arrays (Spencer *et al.*, 2011). The results of normalized expression levels in specific tissues for different categories of corresponding TF target genes, respectively, can be seen in Fig. 4. Moreover, the expression patterns of the mean expression distribution of each class in six tissue-specific studies were

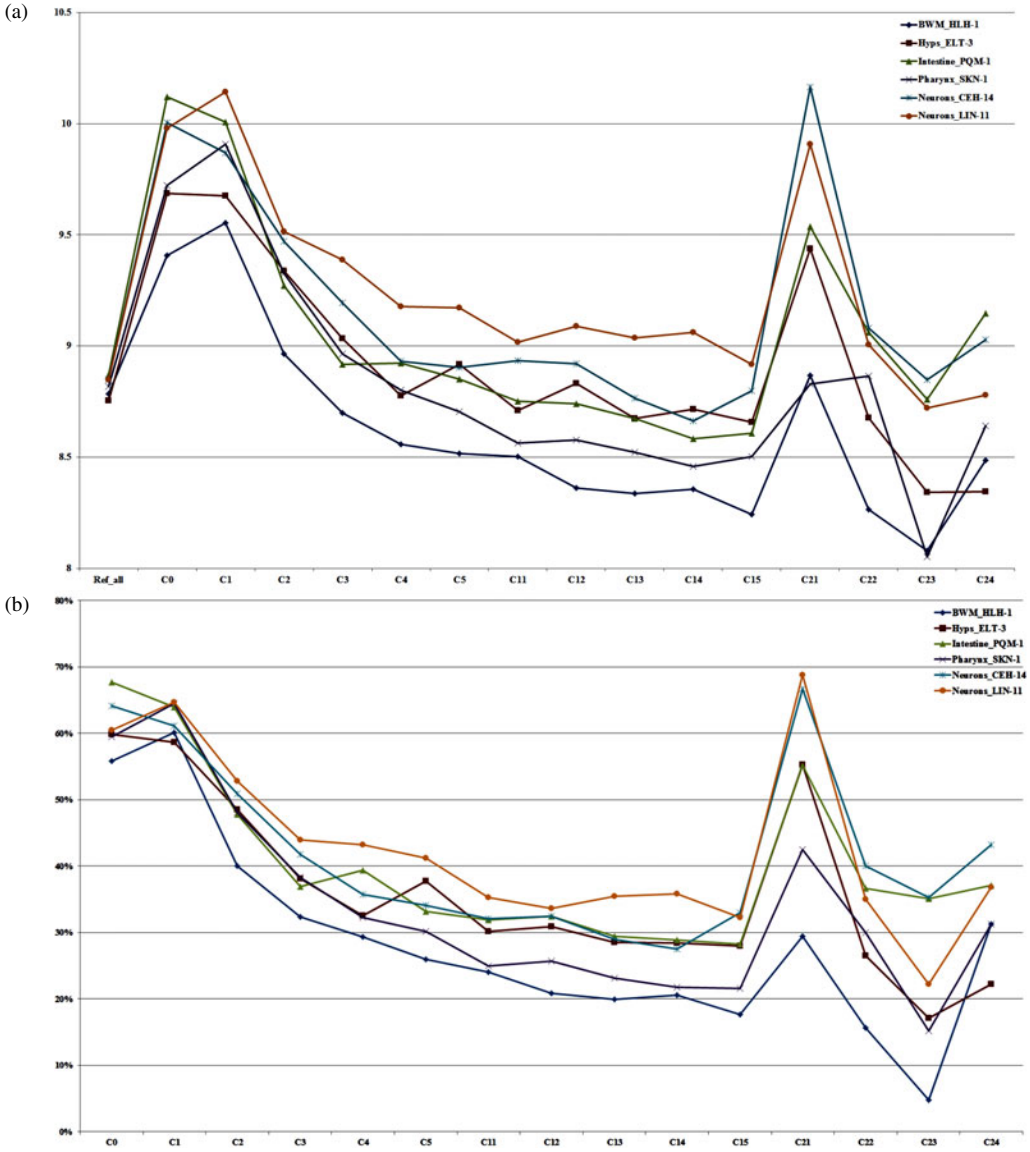

Fig. 5. The mean expression distribution and NCA results of each class in six tissue-specific studies. (*a*) It showed the mean expression distribution of each class in six tissue-specific studies, which is based on the integration analysis of ChIP-seq and gene expression data, including HLH-1 for body wall muscle (BWM), ELT-3 for hypodermis (Hyps), PQM-1 for intestine, SKN-1 for pharynx as well as CEH-14 and LIN-11 both for neurons, respectively. *X*-axis represents 15 different classes of target genes and reference of genome gene expression; *Y*-axis represents the mean level of normalized gene expression value. (*b*) It showed the transcription factor activity in 15 different classes for each of six TFs with tissue-specific expression pattern based on NCA method, including HLH-1 for BWM, ELT-3 for Hyps, PQM-1 for intestine, SKN-1 for pharynx as well as CEH-14 and LIN-11 both for neurons, respectively. *X*-axis represents 15 different classes of target genes and reference of genome gene expression; *Y*-axis represents the percentage of target genes with the corresponding *A* scores larger than 1 compared with the raw number of targets in ChIP-seq data.

shown in Fig. 5(*a*). In common, all first three classes (C0–C2) of the target genes showed significantly elevated level of expression as opposed to that of reference whole genome genes for those six different TFs (Mann–Whitney *U*-test, $P < 10^{-3}$), supporting that the majority of binding events are likely to be functional mostly through up-regulating its target genes within 2 kb upstream from TSS. Another class C3 was also reported as significantly elevated one in some of studies, including for ELT-3, SKN-1, CEH-14 and LIN-11. It indicated that bindings of those four TFs within 2–3 kb upstream from a gene's TSS may be also considered as one candidate group that more likely control its function by up-regulating its activities. Intriguingly, expression of all first six classes of the LIN-11 target genes representing genes with a peak inside (C0) or 0–5 kb upstream (C1–C5) from their TSS showed significantly higher than reference. In contrast, most classes of the 0–5 kb downstream (C11–C15) target genes for

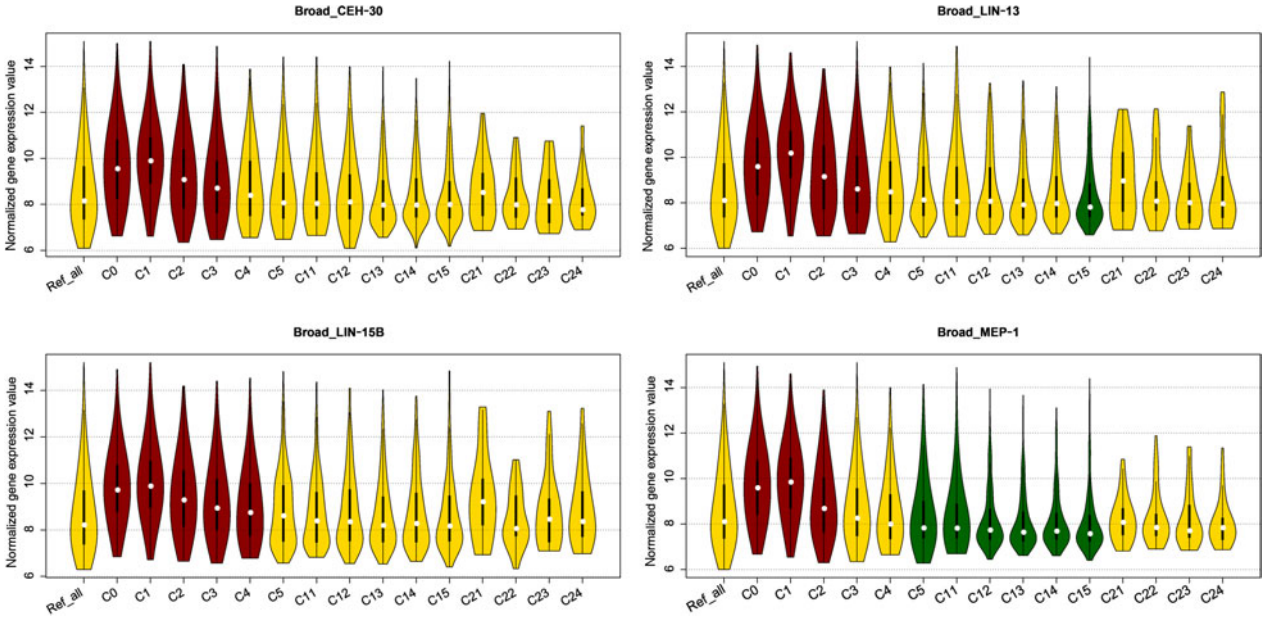

Fig. 6. The expression patterns of broad expressed TFs target genes. The Violin plots of normalized expression levels in whole animal for different categories of corresponding TF target genes were performed for four TFs with broad expression pattern, including CEH-30, LIN-13, LIN-15B and MEP-1, respectively. *X*-axis represents 15 different classes of target genes and reference of genome gene expression; *Y*-axis represents the normalized gene expression value. Significant levels by Mann–Whitney *U*-test ($P < 10^{-3}$) are indicated in the plots, dark-red for significantly up-regulated and dark-green for significantly down-regulated.

those six TFs showed decreased level of expression, suggesting a diverse expression behaviour of binding targets downstream from TSS that likely inhibit the TF activities.

Obviously, the mean expression level of target genes in C0 and C1 classes reached the summit as opposed to that of the other classes for all six TFs (Fig. 5(*a*)), which also suggested their most important roles in regulatory function. What is more interesting is that all of the class of C21 representing the nearest target genes out of 5 kb upstream from TSS showed sharply increased level of expression. However, the gene expression difference test compared with the reference is not significant, which is mainly due to the small sample size of genes count in C21.

In order to confirm our results, we have also performed NCA on the ChIP-seq and gene expression data. Compared to the traditional statistical methods of principal component analysis and independent component analysis, NCA is more comparable to determine low-dimensional representations of high-dimensional datasets for reconstruction of regulatory signals in biological systems (Liao *et al.*, 2003). Here, we applied NCA method to calculate the transcription factor activity in 15 different classes for each TF. According to the percentage of target genes with the corresponding *A* scores larger than 1 compared with the raw number of targets in ChIP-seq data, there appears to be maximum plot ratio in the class of C0, C1 and C21 (Fig. 5(*b*)). It suggests that they are three of most essential classes of target genes in the TF transcriptional regulation issues, which is consistent with our above integration analysis.

(iii) *Integration analysis of binding site locations with broad expression data*

Similarly, there are also several involved TFs with the broad expression pattern (expressed in the whole animal) in TF binding sites ChIP-seq database of modENCODE project, including CEH-30, LIN-13, LIN-15B and MEP-1. There were 9469 target genes for CEH-30, 8185 target genes for LIN-13, 7445 target genes for LIN-15B and 12013 target genes for MEP-1, respectively, identified in those four broad expression TFs ChIP-seq. We then integrated all of the ChIP-seq targets information with the gene expression data in different whole animal developmental stages derived from the same tiling arrays (Spencer *et al.*, 2011). As expected, the bindings within 2 kb upstream of or within its target gene (C0–C2) showed significantly elevated level of expression compared to that of the reference for those four broad expression TFs (Mann–Whitney *U*-test, $P < 10^{-3}$) (Fig. 6), which is the same as above results of six TFs with

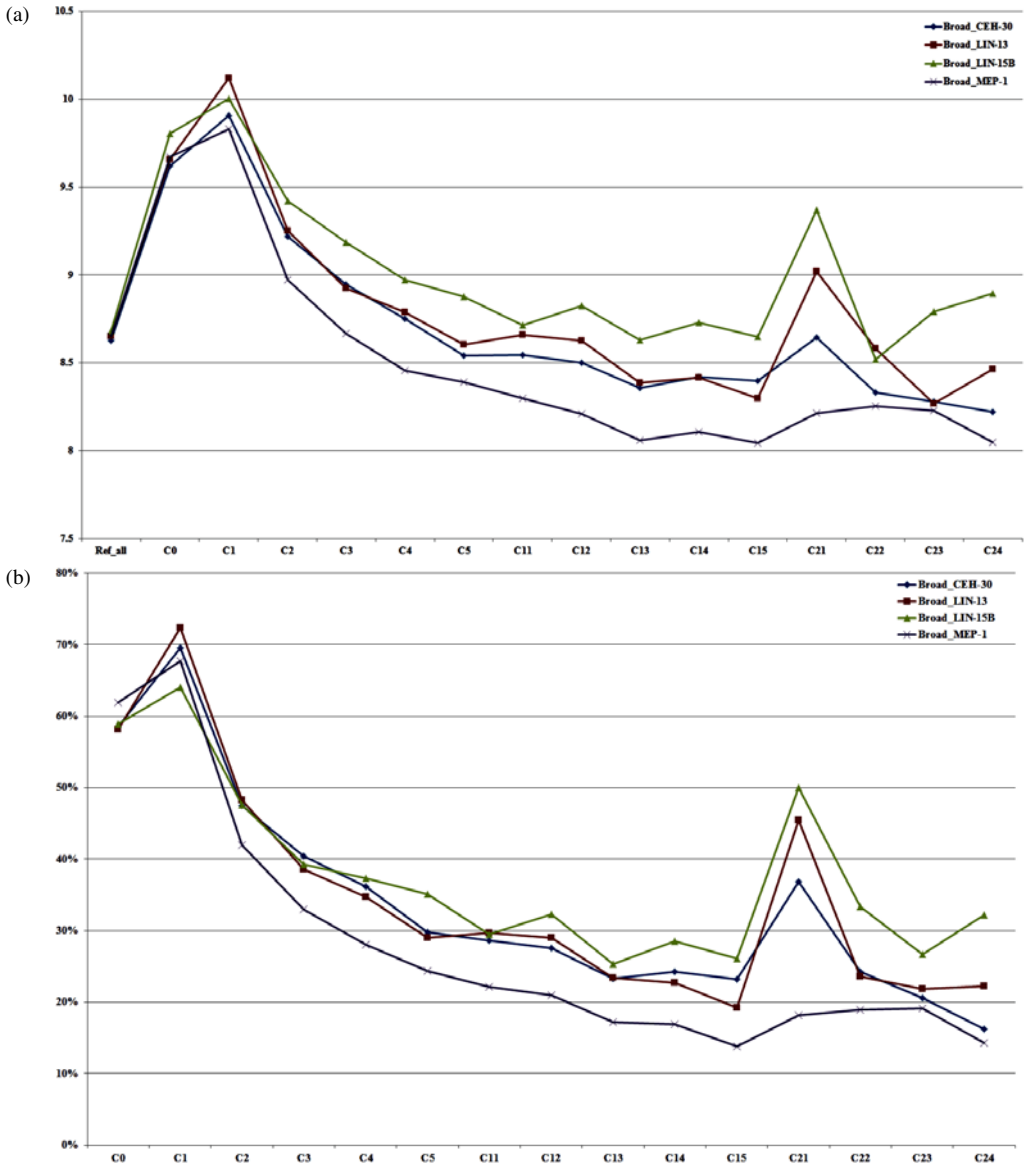

Fig. 7. The mean expression distribution and NCA results of each class in four broad expression TFs studies. (*a*) It showed the mean expression distribution of each class in four broad expression TFs studies, which is based on the integration analysis of ChIP-seq and gene expression data, including CEH-30, LIN-13, LIN-15B and MEP-1, respectively. *X*-axis represents 15 different classes of target genes and reference of genome gene expression; *Y*-axis represents the mean level of normalized gene expression value. (*b*) It showed the transcription factor activity in 15 different classes for each of four TFs with broad expression pattern based on NCA method, including CEH-30, LIN-13, LIN-15B and MEP-1, respectively. *X*-axis represents 15 different classes of target genes and reference of genome gene expression; *Y*-axis represents the percentage of target genes with the corresponding *A* scores larger than 1 compared with the raw number of targets in ChIP-seq data.

tissue-specific expression pattern. Moreover, similar results were observed for the bindings within 2–3 kb upstream (C3) of those TFs target genes except for MEP-1 as well as for the bindings within 3–4 kb upstream (C4) of LIN-15B. Generally, the bindings within 0–5 kb downstream (C11–C15) of those TFs showed constant or decreased expression values as opposed to reference, the significance appeared in C15 for LIN-13 and C5 to C15 for MEP-1. According to the mean expression distribution of each class in those four TF studies, the mean expression level of nearest bindings out of 5 kb upstream of target genes (C21) was also found to have a summit like C0 and C1 (Fig. 7(*a*)), which suggested its equally essential roles in regulatory function of TFs with broad expression pattern. The results were validated by the following NCA approach for those four TFs (Fig. 7(*b*)).

(iv) *Similar transcriptional regulation mechanisms of tissue-specific or broad expression transcription factors in* C. elegans

Although diverse transcription factor binding features revealed by genome-wide ChIP-seq in *C. elegans* (Niu *et al.*, 2011), some of similar issues in regulatory mechanisms may be found between tissue-specific and broad expression TFs. Firstly, the majority of binding events in *C. elegans* are likely to be functional mostly through up-regulating its target genes, particularly three regions of within 2 kb upstream or within its target genes rather than those located elsewhere from its targets. Most of transcriptional regulation issues were previously reported to occur by a TF binding in a promoter region, which was defined as 2 kb from a TSS, particularly in simple organisms such as bacteria or yeast (Capaldi *et al.*, 2008). However, the regulation within the target needs further validation. Secondly, some of regulation issues in *C. elegans* were also found to occur through long-range regions like enhancers, including 2–3 kb (C3) or 3–4 kb (C4) upstream from TSS. It looks similar as more complex species, in which more regulation occurs through long-range enhancers often spanning many tens of thousands of base pairs and the enhancer interactions become the principal form of gene regulation (Arnosti & Kulkarni, 2005; Lee *et al.*, 2005).

As we all know, transcription regulatory networks play a pivotal role in the development, function and pathology of metazoan organisms, which comprises direct or indirect protein–protein or protein–DNA interactions between TFs and their target genes. For our ten TFs, there existed synergistic effects or antagonistic effects between the TFs by co-binding with the same DNA, which have been reported in the previous studies (Lum *et al.*, 2000; Vermeirssen *et al.*, 2007; Cheng *et al.*, 2011).

This work was funded by the Collaborative Research Fund (CRF) of Research Grants Council (RGC) in Hong Kong and Faculty Research Grants (FRG) from Hong Kong Baptist University as well as the Scientific Research Foundation and Academic & Technology Leaders Introduction Project, and 211 Project of Anhui University.

Declaration of interest

The authors declare that they have no competing interests.

Authors' contributions

KH designed the study, collected the datasets from databases and analysed the data, then prepared the original draft the manuscript. DL, JS and ZZ helped to design the study and reviewed the manuscript. All authors read and approved the final manuscript.

References

Arnosti, D. N. & Kulkarni, M. M. (2005). Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *Journal of Cellular Biochemistry* **94**, 890–898.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., Gifford, D. K., Melton, D. A., Jaenisch, R. & Young, R. A. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–56.

Capaldi, A. P., Kaplan, T., Liu, Y., Habib, N., Regev, A., Friedman, N. & O'Shea, E. K. (2008). Structure and function of a transcriptional network activated by the MAPK Hog1. *Nat Genet* **40**, 1300–6.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y. H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W. K., Clarke, N. D., Wei, C. L. & Ng, H. H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–17.

Cheng, C., Yan, K. K., Hwang, W., Qian, J., Bhardwaj, N., Rozowsky, J., Lu, Z. J., Niu, W., Alves, P., Kato, M., Snyder, M. & Gerstein, M. (2011). Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput Biol* **7**, e1002190.

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., Alves, P., Chateigner, A., Perry, M., Morris, M., Auerbach, R. K., Feng, X., Leng, J., Vielle, A., Niu, W., Rhrissorrakrai, K., Agarwal, A., Alexander, R. P., Barber, G., Brdlik, C. M., Brennan, J., Brouillet, J. J., Carr, A., Cheung, M. S., Clawson, H., Contrino, S., Dannenberg, L. O., Dernburg, A. F., Desai, A., Dick, L., Dose, A. C., Du, J., Egelhofer, T., Ercan, S., Euskirchen, G., Ewing, B., Feingold, E. A., Gassmann, R., Good, P. J., Green, P., Gullier, F., Gutwein, M., Guyer, M. S., Habegger, L., Han, T., Henikoff, J. G., Henz, S. R., Hinrichs, A., Holster, H., Hyman, T., Iniguez, A. L., Janette, J., Jensen, M., Kato, M., Kent, W. J., Kephart, E., Khivansara, V., Khurana, E., Kim, J. K., Kolasinska-Zwierz, P., Lai, E. C., Latorre, I., Leahey, A., Lewis, S., Lloyd, P., Lochovsky, L., Lowdon, R. F., Lubling, Y., Lyne, R., MacCoss, M., Mackowiak, S. D., Mangone, M., McKay, S., Mecenas, D., Merrihew, G., Miller, D. M. 3rd, Muroyama, A., Murray, J. I., Ooi, S. L., Pham, H., Phippen, T., Preston, E. A., Rajewsky, N., Ratsch, G., Rosenbaum, H., Rozowsky, J., Rutherford, K., Ruzanov, P., Sarov, M., Sasidharan, R., Sboner, A., Scheid, P., Segal, E., Shin, H., Shou, C., Slack, F. J., Slightam, C., Smith, R., Spencer, W. C., Stinson, E. O., Taing, S., Takasaki, T., Vafeados, D., Voronina, K., Wang, G., Washington, N. L., Whittle, C. M., Wu, B., Yan, K. K., Zeller, G., Zha, Z., Zhong, M., Zhou, X., Ahringer, J., Strome, S., Gunsalus, K. C., Micklem, G., Liu, X. S.,

Reinke, V., Kim, S. K., Hillier, L. W., Henikoff, S., Piano, F., Snyder, M., Stein, L., Lieb, J. D. & Waterston, R. H. (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**, 1775–87.

Guan, D., Shao, J., Deng, Y., Wang, P., Zhao, Z., Liang, Y., Wang, J. & Yan, B. (2014). CMGRN: a web server for constructing multilevel gene regulatory networks using ChIP-seq and gene expression data. *Bioinformatics*.

Hintze JaN, R. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician* **52**, 181–184.

Kim, J., Chu, J., Shen, X., Wang, J. & Orkin, S. H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132**, 1049–1061.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J. & Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**, 1813–31.

Lee, G. R., Spilianakis, C. G. & Flavell, R. A. (2005). Hypersensitive site 7 of the TH2 locus control region is essential for expressing TH2 cytokine genes and for long-range intrachromosomal interactions. *Nature Immunology* **6**, 42–48.

Liao, J. C., Boscolo, R., Yang, Y. L., Tran, L. M., Sabatti, C. & Roychowdhury, V. P. (2003). Network component analysis: reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 15522–15527.

Lum, D. H., Kuwabara, P. E., Zarkower, D. & Spence, A. M. (2000). Direct protein-protein interaction between the intracellular domain of TRA-2 and the transcription factor TRA-1A modulates feminizing activity in *C. elegans*. *Genes and Development* **14**, 3153–3165.

Niu, W., Lu, Z. J., Zhong, M., Sarov, M., Murray, J. I., Brdlik, C. M., Janette, J., Chen, C., Alves, P., Preston, E., Slightham, C., Jiang, L., Hyman, A. A., Kim, S. K., Waterston, R. H., Gerstein, M., Snyder, M. & Reinke, V. (2011). Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. *elegans*. Genome Res **21**, 245–54.

Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M. & Gerstein, M. B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* **27**, 66–75.

Spencer, W. C., Zeller, G., Watson, J. D., Henz, S. R., Watkins, K. L., McWhirter, R. D., Petersen, S., Sreedharan, V. T., Widmer, C., Jo, J., Reinke, V., Petrella, L., Strome, S., Von Stetina, S. E., Katz, M., Shaham, S., Ratsch, G. & Miller, D. M. 3rd (2011). A spatial and temporal map of C. elegans gene expression. *Genome Res* **21**, 325–41.

Vermeirssen, V., Barrasa, M. I., Hidalgo, C. A., Babon, J. A., Sequerra, R., Doucette-Stamm, L., Barabasi, A. L. & Walhout, A. J. (2007). Transcription factor modularity in a gene-centered C. elegans core neuronal protein-DNA interaction network. *Genome Res* **17**, 1061–71.

Yan, B., Li, H., Yang, X., Shao, J., Jang, M., Guan, D., Zou, S., Van Waes, C., Chen, Z. & Zhan, M. (2013). Unraveling regulatory programs for NF-kappaB, p53 and microRNAs in head and neck squamous cell carcinoma. *PLoS One* **8**, e73656.

Zhong, M., Niu, W., Lu, Z. J., Sarov, M., Murray, J. I., Janette, J., Raha, D., Sheaffer, K. L., Lam, H. Y., Preston, E., Slightham, C., Hillier, L. W., Brock, T., Agarwal, A., Auerbach, R., Hyman, A. A., Gerstein, M., Mango, S. E., Kim, S. K., Waterston, R. H., Reinke, V. & Snyder, M. (2010). Genome-wide identification of binding sites defines distinct functions for Caenorhabditis elegans PHA-4/FOXA in development and environmental response. *PLoS Genet* **6**, e1000848.