Published online 31 August 2012

Nucleic Acids Research, 2013, Vol. 41, No. 1 ***e5***
doi:10.1093/nar/gks821

Deciphering the transcriptional regulation of microRNA genes in humans with ACTLocater

Zhen-Dong Xiao, Li-Ting Diao, Jian-Hua Yang, Hui Xu, Mian-Bo Huang, Yong-Jin Deng, Hui Zhou 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, People's Republic of China

Received May 27, 2012; Revised August 4, 2012; Accepted August 6, 2012

ABSTRACT

Understanding the transcriptional regulation of microRNAs (miRNAs) is extremely important for determining the specific roles they play in signaling cascades. However, precise identification of transcription factor binding sites (TFBSs) orchestrating the expressions of miRNAs remains a challenge. By combining accessible chromatin sequences of 12 cell types released by the ENCODE Project, we found that a significant fraction (~80%) of such integrated sequences, evolutionary conserved and in regions upstream of human miRNA genes that are independently transcribed, were preserved across cell types. Accordingly, we developed a computational method, Accessible and Conserved TFBSs Locater (ACTLocater), incorporating this chromatin feature and evolutionary conservation to identify the TFBSs associated with human miRNA genes. ACTLocater achieved high positive predictive values, as revealed by the experimental validation of FOXA1 predictions and by the comparison of its predictions of some other transcription factors (TFs) to empirical ChIP-seq data. Most notably, ACTLocater was widely applicable as indicated by the successful prediction of TF→miRNA interactions in cell types whose chromatin accessibility profiles were not incorporated. By applying ACTLocater to TFs with characterized binding specificities, we compiled a novel repository of putative TF→miRNA interactions and displayed it in ACTViewer, providing a promising foundation for future investigations to elucidate the regulatory mechanisms of miRNA transcription in humans.

INTRODUCTION

Transcriptional regulation is mediated by *cis*-elements and the transcription factors (TFs) that bind to these elements (1). Precise identification of *cis*-elements or transcription factor binding sites (TFBSs) is fundamentally important to decipher the complex transcription regulatory networks. Chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing (ChIP-seq) has become the most powerful tool to profile TFBSs (2,3). *Cis*-elements are, however, often active only in specific cell types or during certain development stages; therefore, a comprehensive catalog of all *cis*-elements would require a thorough investigation of various physiological conditions. Current applications of ChIP-seq are limited by the availability of ChIP-grade antibodies, and the reported binding sites cannot be determined at nucleotide-level resolution. Computational methods have long been sought as an alternative to time-consuming experimental studies; however, owing to the short, degenerate nature of TFBSs, predictions usually contained an overwhelming number of false positives (FPs), which has been termed the 'futility theorem' (4). To improve the accuracy of predictions, existing methods often focus on promoter regions of protein-coding genes and have employed other criteria, such as conservation, co-operation and co-regulation (4,5). Recently, owing to the rapid progress in measures of chromatin accessibility by DNase-seq (6,7) and FAIRE-seq (Formaldehyde Assisted Isolation of Regulatory Elements) (8,9), as well as the elucidation of histone modification profiles based on ChIP-seq (10,11), chromatin accessibility and histone modification profiles of human cell lines have been cataloged (12,13). The accumulation of these data has provided an unprecedented opportunity to improve TFBS predictions in humans and methods incorporating such information have been successfully developed to work in a cell-specific manner (14–16). However, the human body contains more than

*To whom correspondence should be addressed. Tel: +86 20 84112399; Fax: +86 20 84036551; Email: lssqlh@mail.sysu.edu.cn

200 unique cell types, which could be under various physiological and pathological conditions. It is unpractical to thoroughly profile the chromatin accessibility and histone modification of all types of human cells under all conditions, and thus the applicability of such methods is limited to cell types whose experimental data are available. Currently, the major computational problem to solve is to enhance the applicability of predictions.

While the methods for predicting TFBSs associated with protein-coding genes are fairly comprehensive, unfortunately, transcriptional regulation genomics of non-coding RNAs (ncRNAs), such as microRNAs (miRNAs), which have been found to collaborate with proteins in essential biological processes, have been much less investigated. Accurate maps of TFBSs associated with ncRNA genes will be essential for a comprehensive understanding of protein-coding and ncRNA genes coordinate regulation network. miRNAs are a class of small ncRNAs, which are expressed in a spatio-temporal manner and play key roles in diverse biological processes through targeting and suppressing the expression of protein genes (17–19). Extensive studies have been performed on miRNA gene identification, expression analysis and function. For example, more than 1000 human miRNA genes have been documented in miRBase (20); expression patterns have been profiled from about 200 major human organs and cell types (21) and more than 3000 confirmed target genes of human, mouse or rat miRNAs are included in miRWalk (22). Although previous studies have shown that miRNAs were under tight transcriptional regulation (23–27), only few investigations have been conducted on the transcriptional regulation of human miRNA genes. Currently, only 221 human TF→miRNA (miRNA regulation by TF) interactions have been reported in TransmiR (28). This certainly represents only a small fraction of the total number of regulatory networks, and large-scale investigations are needed to decode the full set of pathways governing the expression of human miRNAs. Human miRNAs are either located in the introns of protein-coding genes (intronic miRNAs) or between protein-coding genes (intergenic miRNAs). Intronic miRNAs are likely to be transcribed along with their host genes (29,30), and existing TFBS prediction techniques for protein-coding genes can be used for these systems. However, for intergenic miRNAs, which are independently transcribed, the distances between transcription initiation sites (TSSs) and miRNA-coding regions dramatically vary, ranging from a few hundred bases (31) to 30-kb upstream (32). Also, TFs might not almost exclusively bind at proximal promoters (12); it is likely that a sufficient number of distal TFBSs would locate in regions between miRNA promoters and their coding regions providing additional regulatory information. To fully explore the transcriptional regulation of human intergenic miRNAs, it is necessary to examine large genomic regions to locate all the possible TFBSs, and thus existing methods employing only simple pattern matching (33,34) would lack a reasonable accuracy. Even filtered out with significant conservation (such as Conserved TFBS track in UCSC human genome browser and its web interface, PuTmiR) (35,36), there would be still a large amount of FPs due to the presence of slowly evolving neutral regions (37). On the other hand, recent methods based on cell-specific experimental data (38) have improved in accuracy, but the scope of their applicability is limited. Prediction methods that are suitable for human intergenic miRNAs, particularly those with high accuracy and a wide applicable range, are still lacking.

Here, we showed that the conserved and accessible chromatin sequences integrated from 12 cell types, immediately upstream of human intergenic miRNAs found also in the mouse and rat (referred as HMR intergenic miRNAs), were highly preserved across cell types. Therefore, we developed ACTLocater (Accessible and Conserved TFBSs Locater) incorporating known chromatin features to identify TFBSs associated with HMR intergenic miRNAs. Applying ACTLocater to selected TFs showed that the positive predictive values (PPVs) of predictions were greatly improved compared to conventional methods based on sequence conservation alone. Although ACTLocater was based on information from a limited number of cell types, it successfully predicted TF→miRNA interactions in cells whose chromatin accessibility information was not incorporated, suggesting that ACTLocater could be applied to a wide range of cell types. By using ACTLocater for TFs with known binding specificities, we established a comprehensive human TF→miRNA interaction database, ACTViewer. The resultant maps provided a solid foundation for understanding the regulatory pathways underlying HMR intergenic miRNA expression.

MATERIALS AND METHODS

HMR intergenic miRNA identification and TSS prediction

All human miRNAs and genome annotations were obtained from the UCSC Genome Browser (35) hg18 assembly. We considered miRNAs that reside between the protein-coding genes as intergenic miRNAs and identified human intergenic miRNAs conserved in mice and rats by the miRviewer (39). We next adopted the TSS predictions of these miRNAs from previous studies (26,40,41) and manually calibrated them according to the full-length cDNA data, SwitchGear TSS predictions and ENCODE promoter-associated histone mark (H3K4me3) on nine cell lines track in the UCSC Genome Browser. If any full-length cDNAs overlapped with the miRNAs, the 5′-terminal region of the full-length cDNA was used as the TSS for the corresponding miRNA; or if SwitchGear TSS prediction in the miRNA upstream region coming along with H3K4me3 peaks present in multiple cell lines, the SwitchGear prediction was used; otherwise, TSSs for the corresponding miRNAs were directly adopted from previous studies. Same-strand miRNAs with common predicted TSSs were considered as one transcriptional unit.

HMR accessible alignments extraction and shuffling

We extracted the genomic coordinates of the upstream 40-kb flanking regions or truncated flanking regions

of genes upstream of HMR intergenic miRNAs from human genome. For clustered miRNAs, the regions upstream of each individual sequence were merged. We collected a total of 14 DNase-seq and four FAIRE-seq datasets (as listed in Supplementary Table S2) from the Data Coordination Center (http://genome.ucsc.edu/ENCODE/) of the ENCODE Project (12,13) and merged them as chromatin accessibility reference. We then extracted Human/Mouse/Rat alignments within the chromatin accessibility reference and in the regions upstream of HMR intergenic miRNA units from 17-way genome-wide multiple alignments using Galaxy (42).

To calculate the signal-to-noise ratios of FOXA1 predictions by ACTLocater and the Conservation method, we shuffled (99 runs) HMR accessible alignments and Human/Mouse/Rat alignments to generate alignments with the same length, base composition and patterns of gaps and conservation (43), respectively. The signal was defined as the number of FOXA1-binding sites predicted with the true alignments, whereas the noise was defined as the number of sites predicted with the shuffled alignments.

TFBSs prediction by ACTLocater

To locate the TFBSs from the HMR accessibility alignments, we implemented two searching methods, CScanner and MScanner, based on the consensus model and the PSSM model (44).

We designed CScanner to search both strands of every human sequence for sequence windows, and its corresponding orthologous sequences in mouse and rat matched the consensus. We implemented MScanner to score sequences on both strands as potential TFBSs using the formula (44):

$$species\_score(i) = \sum_b f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

where i is the position within the site, p_b is the relative frequency of base b in the genome and $f_{b,i}$ is the observed relative frequency of base b at that position (from the matrix).

For each species, scores were normalized to a 100-point scale. The average score was defined as the arithmetic mean of the species scores. Predictions were based on the thresholds of the species scores and the corresponding average score.

PSSMs and/or consensus sequences used for selected TFs predictions by ACTLocater were as follows: for FOXA1, M00131 from TRANSFAC (45), MA0148 from JASPAR (46) and consensus sequence TRTTKRYTY (47); for c-Myc, M00118, M00123 and M00615 from TRANSFAC, MA0059 and MA0147 from JASPAR, and consensus sequence CACGTG (48); for the E2F family, M00024, M00050 and M00516 from TRANSFAC, MA0024 from JASPAR and consensus sequence TTTSCGC (49); for MyoD, a PSSM derived from ChIP-seq data (50) (details in 'Supplementary Methods' section); for NRSF, M00256 from TRANSFAC, MA0138 from JASPAR and two PSSMs from previous studies (51,52); for Oct4, M00135, M00138, M00161, M00195 and M00210 from TRANSFAC, MA0142 from JASPAR and consensus sequence ATGCWAAT (53). The cutoff values for species scores and average score for MScanner were set to 80 and 85, respectively, for all the PSSMs except NRSF. Because this protein recognized a long motif, its species and average score cutoffs were set to 75 and 80, respectively.

Validation of predictions by genome-wide ChIP data

We collected the binding peaks from the ENCODE Project (12,13) and previous genome-wide ChIP experiments (54–56) to evaluate the TFBS predictions. ChIP peaks of NRSF have been identified by MICSA (57) using the raw data from the ENCODE Project. For each TF, we merged all ChIP peaks. We next assessed all predictions for FOXA1, c-Myc, the E2F family and NRSF by comparing the TFBSs predicted in the human genome with the corresponding genome-wide ChIP datasets generated by using human cell lines. A prediction was considered as a true positive (TP) if the predicted site was inside a merged ChIP peak; otherwise, the prediction was considered as a FP. We calculated the PPV as TP/(TP + FP). A merged ChIP peak was successfully predicted (TP) if there was at least one predicted TFBS within it; otherwise, the peak was annotated as a false negative (FN). And we calculated the sensitivity as TP/(TP + FN).

We assessed the evidence supported MyoD-binding sites by comparing the TFBSs predicted in the mouse genome with ChIP-seq datasets obtained from mouse cells (50) (ChIP-seq peaks were identified as described in 'Supplementary Methods' section). And we assessed evidence supported Oct4-binding sites by comparing the TFBSs predicted in the human and mouse genomes with ChIP-seq datasets obtained from human and mouse embryonic stem cells (26), respectively.

Definition of accessible chromatin regions for skeletal muscle and embryonic stem cells

We defined accessible chromatin regions of skeletal muscle cells by merging the DNase-seq peak regions of HSMM, HSMMtube and SKMC skeletal muscle cells and defined those of embryonic stem cells by merging the DNase-seq or FAIRE-seq peak regions of H1-hESC, H7-hESC and H9ES embryonic stem cells. All the DNase-seq and FAIRE-seq datasets were obtained from the Data Coordination Center of the ENCODE Project, as listed in Supplementary Table S2.

ACTViewer construction

To construct the ACTViewer database, we collected PSSMs from TRANSFAC, JASPAR and UniPROBE. Then, we applied MScanner to the HMR accessible alignments for each PSSM. Thresholds for species and average scores were set to 80 and 85, respectively. ACTViewer database was developed as described previously (58).

Cell culture and stable cell line generation

MDA-MB231 cells were a gift from Prof. Yan Zhang (Sun Yat-sen University, Guangzhou). MCF-7 cells and HEK293T cells were obtained from the Shanghai Cellular Institute of Chinese Academy of Sciences (Shanghai, China). All cells were grown in Dulbecco's modified Eagle's medium (GIBCO) supplemented with 10% fetal bovine serum (GIBCO) and were cultured in a 37°C incubator with 5% CO_2.

The FOXA1 cDNA lentiviral vector and control empty vector were purchased from Fulengen Corporation (Guangzhou, China). To produce the lentivirus, we transiently co-transfected FOXA1 cDNA lentiviral vector or empty vector with the ViraPower™ Lentiviral Packaging Mix (Invitrogen) into HEK293T cells using Lipofectamine 2000 reagent (Invitrogen). MDA-MB231 cells were infected and then selected with puromycin (Sigma–Aldrich).

Chromatin immunoprecipitation assay

We performed ChIP assays in MCF-7 cells by using a chromatin immunoprecipitation kit (Millipore) according to the manufacturer's instructions. Protein–DNA complexes were precipitated with a control IgG or an anti-FOXA1 antibody (ab5089, Abcam). PCR primers are listed in Supplementary Table S5.

Western blotting analysis

Proteins were extracted with RIPA lysis buffer. FOXA1 protein was revealed with a polyclonal antibody (C-20, Santa Cruz). Signals were detected by the Super Signal Western Pico chemiluminescent substrate (Pierce). Western blotting of GAPDH on the same membrane was used as a loading control.

miRNA microarray

miRNA microarrays were performed at the Beijing CapitalBio Corporation. Total RNA content extracted by TRIzol reagent (Invitrogen) from MDA-MB231 cells stably overexpressing FOXA1 or control cells was labeled with biotin. Labeled samples were hybridized to GeneChip®miRNA Array (V2.0). Raw data were normalized and analyzed using the miRNA QC tool software (Affymetrix). We considered the HMR intergenic miRNAs with a fold change >1.5 or <0.75 in the FOXA1 overexpressing MDA-MB231 cells as FOXA1-affected HMR intergenic miRNAs.

Real-time RT–PCR

Total RNA content was reverse-transcribed to cDNA using the Primescript RT reagent kit (Takara). Real-time PCR was carried out using SYBR Premix ExTaq (Takara) according to the manufacturer's instructions. The relative expression of *FOXA1* was calculated using the comparative $2^{-\Delta\Delta Ct}$ method and was normalized to *GAPDH*. The following primer sets were used: for *FOXA1*, 5′-GAAGATGGAAGGGCATGAAA-3′ (forward) and 5′-GCCTGAGTTCATGTTGCTGA-3′ (reverse); for *GAPDH*, 5′-CCATGGGGAAGGTGAAGGTC-3′ (forward) and 5′-GAAGGGGTCATTGATGGCAAC-3′ (reverse). To detect the miRNA expression levels, Bulge-Loop™miRNA qPCR primer sets and U6 primer sets were purchased from RiBoBio Corporation (Guangzhou, China). The relative expression levels of miRNAs were normalized to U6 snRNA. We performed all real-time RT–PCR experiments in triplicate.

RESULTS

ACTLocater: integrating evolutionary conservation and chromatin structure to predict TFBSs associated with HMR intergenic miRNAs

As components of conserved regulatory systems (59), we reasoned that HMR intergenic miRNAs rely not only on the conservation of the miRNA sequences but also on the transcriptional regulation elements. Via sequence examination and homology searches, we identified 203 HMR intergenic miRNAs, which were grouped into 106 transcriptional units (Supplementary Table S1). We arbitrarily chose a 40-kb region upstream of miRNA as the TFBS search area, such space was sufficient as indicated by the fact that a significant fraction of TSSs of human intergenic miRNAs are within 10-kb upstream regions (60,61).

To explore the features of accessible chromatin sequences upstream of HMR intergenic miRNAs, we mapped chromatin accessibility information of 12 cell types (Supplementary Table S2) obtained from the ENCODE Project (12,13) to the search area. In humans, mice and rats, 63.7% of the accessible sequences were conserved, whereas only 43.0% of the inaccessible sequences were conserved, indicating that accessible regions are rich in functional elements. Conserved accessible sequences were found upstream of 101 HMR intergenic miRNAs (Figure 1A and Supplementary Dataset S1). We next employed a leave-one-out cross-validation method to evaluate the conserved accessible chromatin information across cell types. Each round, conserved accessible sequences of one cell type were left out, and the sequences of left-out cell type recovered from those of the remaining cell types were then examined. The percentages of sequences recovered ranged from 74.7% to 90.0% (Figure 1B). On average, for each examined cell type, 79.7% of the conserved accessible sequences could be recovered from the remaining cell types. Such high values were not due to the similarities across cell types (Supplementary Table S3). SK-N-SH_RA and SKMC, which did not have any similar cell types in our collection, still had high sequence recovery values (80.2% and 78.5%, respectively). These results indicated that a significant fraction of conserved accessible sequences integrated from diverse sources were preserved across cell types and also implied that such integrated sequences may provide valuable clues for other cells whose chromatin accessibility has not been investigated.

Based on these observations, we designed the ACTLocater algorithm to incorporate chromatin accessibility and evolutionary conservation profiles, as shown in Figure 1C. First, we constructed a reference for accessible

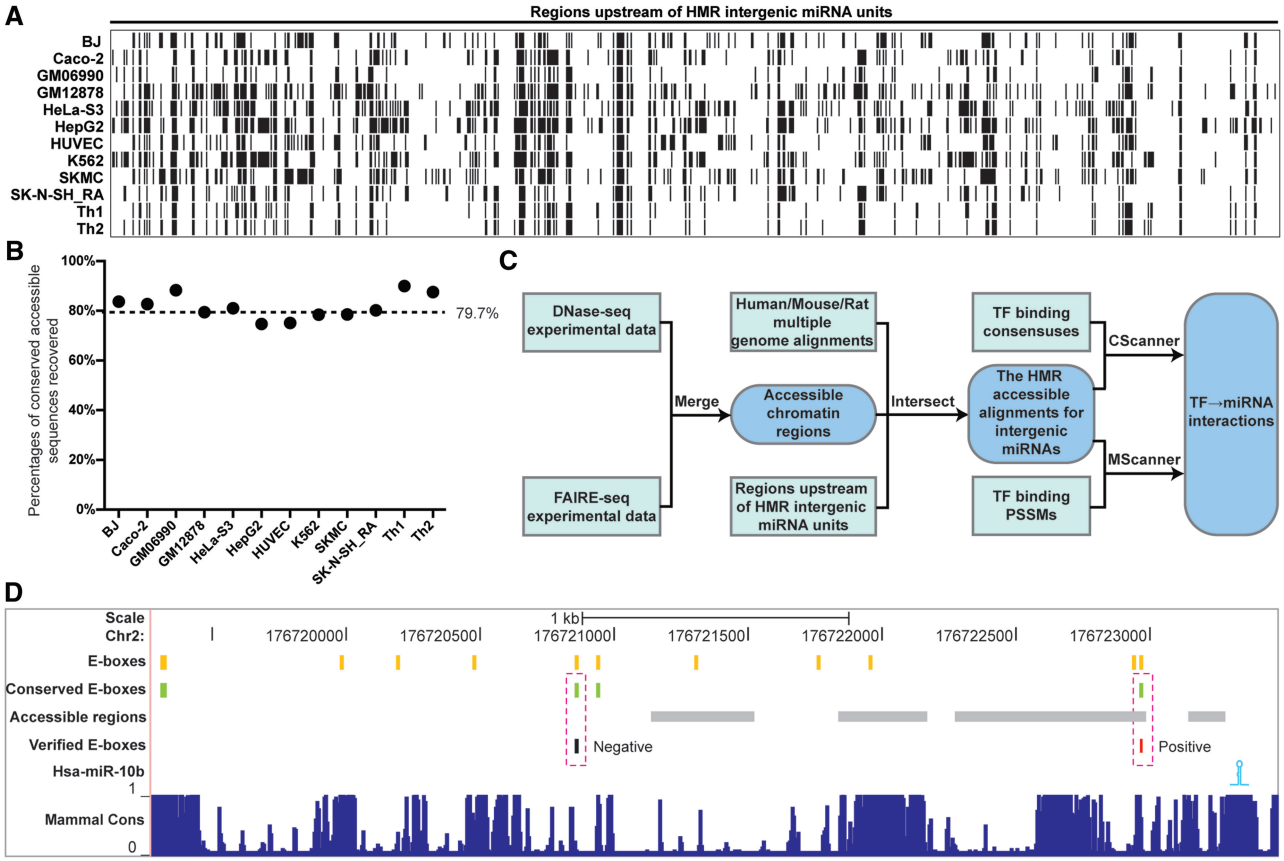

Figure 1. Development of the ACTLocater method. (**A**) Summary of conserved accessible sequences upstream of 101 HMR intergenic miRNA units. The 101 regions upstream of miRNA units were arranged side by side. Within each region, conserved accessible sequences of 12 cell types were aligned according to their genomic coordinates. (**B**) Results of the leave-one-out cross-validation performed in 12 cell types. The horizontal axis shows the cell type left out. The vertical axis shows the percentage of conserved accessible sequences recovered by the remaining cell types. The average is marked by the dashed line. (**C**) Flowchart for the ACTLocater method. PSSMs—position specific scoring matrices. (**D**) The screenshot shows genome browser tracks for E-boxes (in human genomic sequence), conserved E-boxes (found in human, mouse and rat), verified E-boxes [two E-boxes verified as negative result and positive result in Ma *et al.* (24), respectively], accessible regions (accessible chromatin regions merged from 12 cell types) and placental mammal basewise conservation profile (35) for the 4-kb flanking region upstream of Hsa-miR-10b.

chromatin in the human genome by merging the DNase-seq and FAIRE-seq peak regions of 12 cell types. Then, we extracted multiple alignments within the accessible reference regions (referred to as HMR accessible alignments) from the Human/Mouse/Rat genome alignments. Finally, we identified TFBSs in HMR accessible alignments by CScanner and/or MScanner (details in 'Materials and Methods' section).

Initially, the search area contained 9750 conserved sequence blocks (~1.45 Mb human DNA sequences). When restricted to the accessible chromatin reference regions, only 3217 blocks (~0.34 Mb human sequences) remained. As the search area was reduced by 76.6%, the prediction specificity was greatly improved. To illustrate, we made a simple prediction of the Twist→hsa-miR-10b interaction similarly to that in Ma *et al.* (24) and visualized it in the UCSC genome browser (Figure 1D). Within the 4-kb region upstream of hsa-miR-10b, we identified 12 E-boxes (CANNTG), of which 5 were conserved. By comparing the predictions against the accessible chromatin reference, only one E-box remained, which was the proven Twist-binding site (24), demonstrating the high efficiency of ACTLocater.

Prediction and experimental validation of FOXA1-targeted miRNAs

To assess the performance of its predictions, we applied ACTLocater to predict HMR intergenic miRNAs regulated by FOXA1. A total of 52 binding sites were predicted (Supplementary Table S4). In contrast, the Conservation method (applying ACTLocater without considering chromatin accessibility) predicted 223 sites, indicating that incorporating chromatin accessibility information can greatly reduce the number of predictions.

To evaluate the significance of the predicted FOXA1-binding sites, we repeated predictions with shuffled input alignments. As shown in Figure 2A, the predictions of ACTLocater had significantly higher signal-to-noise ratios than those of the Conservation method (Two-sided two-sample Student's *t*-test, $P = 2.2 \times 10^{-19}$). We then used FOXA1-binding peaks from previous

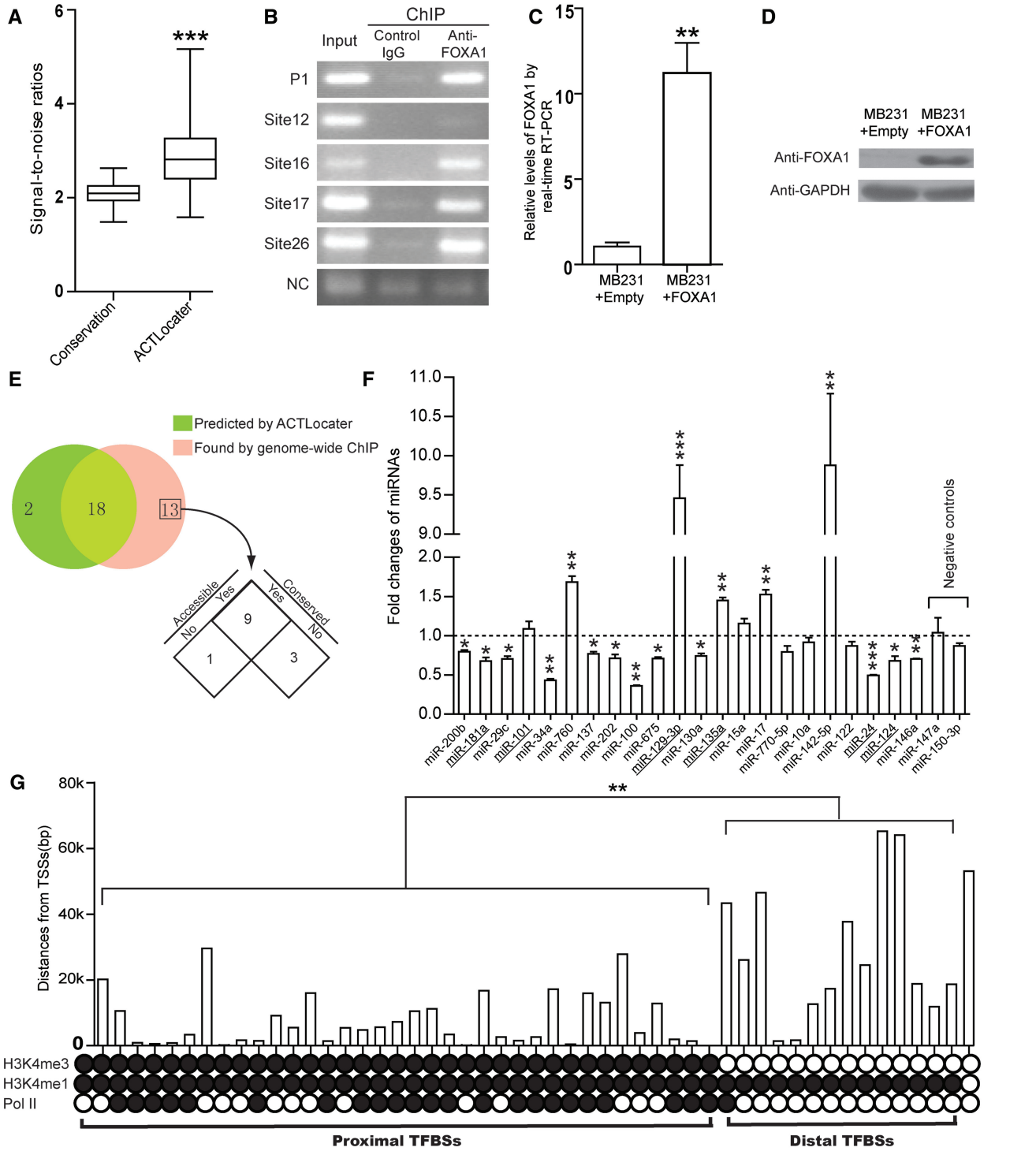

Figure 2. Prediction and validation of FOXA1-targeted HMR intergenic miRNAs. (**A**) A boxplot shows the signal-to-noise ratios of FOXA1 predictions by the Conservation method and ACTLocater method, respectively. ***$P < 0.001$. (**B**) Examples of ChIP assays in MCF-7 cells. Primers against predicted regions were used to amplify DNA immunoprecipitated with FOXA1 antibody (lane 3) or control rabbit IgG (lane 2). Genomic DNA was used as a positive control (lane 1). Site P2 validated by previous studies (54–56) and site NC with no evidence of FOXA1 binding (54) were used as a positive and a negative control for ChIP enrichment, respectively. (**C**) Real-time RT–PCR and (**D**) western blotting analysis of FOXA1 from MDA-MB231 cells stably expressing empty vector or FOXA1. Error bars indicate s.e.m.; *$P < 0.05$. (**E**) Venn diagram of FOXA1-targeted HMR intergenic miRNA units identified by ACTLocater or previous genome-wide ChIP studies. miRNA units found only in the genome-wide ChIP studies were classified according to the chromatin accessibility and conservation of corresponding ChIP peak regions. (**F**) Fold changes of randomly selected and negative control miRNAs in FOXA1 overexpressing MDA-MB231 cells relative to the empty vector controls. miRNAs with multiple locus in the human genome are marked underlines. Hsa-miR-147a and hsa-miR-150-3p, which there were no FOXA1-binding sites associated with, were used as negative controls. Error bars indicate s.e.m.; two-sided one-sample Student's *t*-test; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$. (**G**) Classification of predicted FOXA1-binding sites. Sites overlapped with ChIP peaks of H3K4me1, H3K4me3 or Pol II are marked as solid circles, otherwise are marked as unfilled circles. The plot shows the distances (absolute value) between sites and their corresponding TSSs. The TSS of the first miRNA unit is not available. **$P < 0.01$.

genome-wide ChIP studies (54–56) to evaluate the quality of these predictions. Of the binding sites predicted by the Conservation method, 13.9% (31/223) could be validated against the known set of binding sites. For ACTLocater, the number of predicted sites reduced to 52, 15 of which were known binding sites and the PPV increased to 28.8%. The sensitivities of the Conservation method and ACTLocater were 15.6% and 8.4%, respectively. ACTLocater demonstrated substantially higher specificity than the Conservation method, while reducing the sensitivity. We next conducted ChIP assays in MCF-7 cells to verify predictions, which were not validated by previous genome-wide ChIP studies. Representative examples of ChIP results are shown in Figure 2B, and full results can be found in Supplementary Figure S1. 81.1% (30/37) of these sites were confirmed in MCF-7 cells, and thus the overall PPV of ACTLocater predictions was 86.5% (45/52), emphasizing its specificity in predicting FOXA1-binding sites. Furthermore, these results also indicated that ACTLocater could discover novel sites which were missed by high-throughput experimental methods.

To examine the functional importance of the predicted sites, we employed miRNA arrays to monitor the miRNA expression changes in FOXA1 overexpressed MDA-MB231 cells (Supplementary Table S6). FOXA1 overexpression was validated by real-time RT–PCR (Figure 2C; two-sided two-sample Student's *t*-test, $P = 0.0048$) and western blotting (Figure 2D). We found that 50.0% (20/40) of the HMR intergenic miRNA units predicted by ACTLocater and 46.3% (31/67) of those found by previous genome-wide ChIP studies (54–56) showed a change in expression. 90.0% (18/20) of the FOXA1-affected HMR intergenic miRNA units predicted by ACTLocater were also identified by the genome-wide ChIP studies, indicating that the miRNA units detected by ACTLocater correlate well with those found by previous genome-wide ChIP studies. However, 13 miRNA units detected by the genome-wide ChIP studies were not predicted by ACTLocater. A detailed examination of the genomic features of the corresponding ChIP peak regions showed that these sequences lacked a conserved FOXA1-binding motif (Figure 2E), which is necessary for ACTLocater prediction. Finally, we used a more sensitive method, real-time RT–PCR, to further investigate the effect of ectopic expression of FOXA1 on miRNAs. We randomly selected 22 miRNA units and chose one member miRNA for testing from each unit. 77.3% (17/22) of the randomly selected miRNAs showed significant changes in expression levels, while two negative-control miRNAs showed no significant changes (Figure 2F). Hsa-miR-130a, of which the predicted site was not experimental supported, was likely regulated by FOXA1 indirectly. It should be noted that some of the examined miRNAs have multiple copies in the human genome, and the detection of real-time RT–PCR cannot ensure that the expression changes of multi-locus miRNAs were from the locus with predicted FOXA1-binding sites. When considering the miRNAs associated with experimental supported FOXA1-binding sites and derived from a unique locus, there are still 73.3% (11/15) miRNAs showed significant changes. These results indicated that most of the predicted FOXA1-binding sites were functional. Hsa-miR-202, hsa-miR-129-3p and hsa-miR-24, members of three miRNA units uniquely predicted by ACTLocater, all showed a significant change in expression levels, suggesting that ACTLocater could predict novel functional FOXA1-binding sites missed by genome-wide experimental assays.

Distinct histone signatures have been found for proximal promoters and distal enhancers. Previous studies have shown that strong H3K4me1 and H3K4me3 signals were associated with promoters, whereas strong H3K4me1 and weak H3K4me3 signals were associated with enhancers (10,62). To examine whether ACTLocater could predict not only proximal but also distal FOXA1-binding sites, we collected ChIP-seq peaks of histone modifications from various cells released by the ENCODE Project (12,13) (datasets are listed in Supplementary Table S2) and used them to classify the FOXA1-binding sites. Of the 51 predicted binding sites with histone modifications available, 37 were classified as proximal TFBSs and 14 were classified as distal TFBSs (Figure 2G). These classifications were supported by the fact that proximal TFBSs were usually associated with the Pol II ChIP-seq peaks, but not the case for distal TFBSs (Two-sided Fisher's exact test, $P = 0.00044$; Pol II ChIP-seq datasets were obtained from the ENCODE Project, as listed in Supplementary Table S2). Moreover, the distances of proximal TFBSs from the corresponding TSSs were significantly less than distances between distal TFBSs and their TSSs (Two-sided two-sample Student's *t*-test, $P = 0.0028$). These data indicated that ACTLocater predicted both proximal and distal FOXA1-binding sites. Six sites were found to be located in genomic regions between TSSs and miRNAs, suggesting that these regions could contain additional functional regulatory elements.

Evaluation on other selected TFs

To further assess its performance, we applied ACTLocater to identify the target miRNAs of c-Myc, the E2F family and NRSF (also known as REST). The program predicted 29 c-Myc, 25 E2F and eight NRSF-binding sites (Supplementary Table S7, S8 and S9), corresponding to 22 c-Myc→miRNA, 18 E2F→miRNA and seven NRSF→miRNA interactions, respectively. The PPVs of these predictions based on ACTLocater were all superior to that of the Conservation method, while the sensitivities of these two methods were comparable (Table 1). In previous studies (23,25), 10 HMR intergenic miRNA units have been identified as directly transcriptional targets of c-Myc in P493-6 cells. ACTLocater successfully predicted 50% (5/10) of these miRNA units. As the c-Myc binding was assayed in conserved regions without considering the presence of a binding motif in Chang *et al.* (25), we examined the binding regions (PCR regions in the previous study spanned 1 kb) of the missing units and found that c-Myc binds to these regions via non-canonical motifs, causing them to be overlooked by ACTLocater. Additionally, some c-Myc predictions not validated by ChIP-seq were supported by other studies. For example,

Table 1. Summary of FOXA1, c-Myc, the E2F family and NRSF predictions by the conservation method and ACTLocater

TF	Conservation[a]			ACTLocater		
	No. sites[b]	PPV[c] (%)	Sensitivity[d] (%)	No. sites[b]	PPV[c] (%)	Sensitivity[d] (%)
FOXA1	223	13.9	15.6	52	28.8	8.4
c-Myc	41	51.2	5.8	29	72.4	5.8
E2F	38	47.4	10.2	25	64.0	9.4
NRSF	17	52.9	33.3	8	87.5	25.9

[a]The Conservation method was the same as ACTLocater but without considering chromatin accessibility reference.
[b]Number of sites predicted by the Conservation method and ACTLocater, respectively.
[c]PPV and [d]Sensitivity of FOXA1 predictions were assessed by the genome-wide ChIP studies (54–56) and that of remaining TFs were assessed by the corresponding ChIP-seq datasets obtained from the ENCODE Project, as listed in Supplementary Table S2.

H19 RNA, the precursor of hsa-miR-675 and hsa-miR-9-3 have been found to be direct targets of c-Myc (63,64). These results showed that ACTLocater can yield consistent high-specificity predictions.

Assessment of the applicability of ACTLocater

To determine the scope of our method, we used ACTLocater to predict cell-specific MyoD→miRNA and Oct4→miRNA interactions that occur in skeletal muscle and embryonic stem cells, respectively. Meanwhile, we generated two predictions with the chromatin accessibility data of skeletal muscle and embryonic stem cells and used them as independent verifications of the ACTLocater predictions.

To test whether MyoD→miRNA interactions could be predicted by the chromatin accessibility data of non-muscle cell types, we removed the chromatin accessibility data of SKMC cell from our chromatin accessibility reference. A total of 22 MyoD-binding sites were predicted by ACTLocater (Figure 3A and Supplementary Table S10). Also, we made a similar prediction with the chromatin accessibility data of skeletal muscle cells and yielded 26 MyoD-binding sites (Figure 3A and Supplementary Table S11). To avoid the potential differences in FP predictions between two datasets, we only considered sites with evidence supported. As shown in Figure 3B, 71.4% (10/14) of evidence supported sites (50) made with the chromatin accessibility data of skeletal muscle cells could be successfully predicted with those of non-muscle cells.

We next performed a similar analysis between the two Oct4 predictions, which were made with chromatin accessibility data of non-stem cells (our chromatin accessibility reference) and embryonic stem cells, respectively. Both predictions reported 16 Oct4-binding sites (Figure 3C; Supplementary Table S12 and S13). As shown in Figure 3D, 77.8% (7/9) of evidence supported sites (26,65) made with the chromatin accessibility data of stem cells could be successfully predicted with those of non-stem cells.

Taken together, most cell specific TF→miRNA interactions were successfully predicted, indicating that ACTLocater could be applied to a wide range of cell types, such as cells whose chromatin accessibility profiles had not yet been characterized.

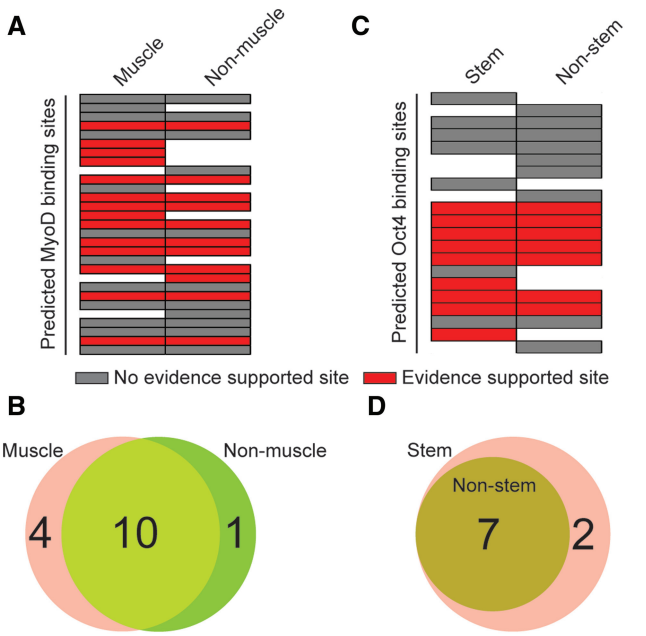

Figure 3. Applicability assessment of ACTLocater. (**A**) Summary of MyoD-binding sites predicted with chromatin accessibility data of skeletal muscle and non-muscle cells, respectively. Sites are aligned according to the genomic coordinates. (**B**) Venn diagram of evidence supported MyoD sites. (**C**) Summary of Oct4-binding sites predicted with chromatin accessibility data of embryonic stem and non-stem cells, respectively. Sites are aligned according to the genomic coordinates. (**D**) Venn diagram of evidence supported Oct4 sites.

ACTViewer: regulatory map of HMR intergenic miRNAs

Given that ACTLocater achieved high PPVs and could be widely applied, we constructed a regulatory map composed of TFBSs associated with HMR intergenic miRNAs by applying ACTLocater to the PSSMs from TRANSFAC (45), JASPAR (46) and UniPROBE (66). We found that different names in these database entries sometimes corresponded to TF isoforms or even to the same TF, which poses a serious challenge to consistent TF nomenclature. For the present study, we separated the predictions with the names and accession numbers retained. We retrieved a total of 295 PSSMs from TRANSFAC and predicted 13 688 TFBSs, corresponding to 4085 TF→miRNA interactions. Additionally, we

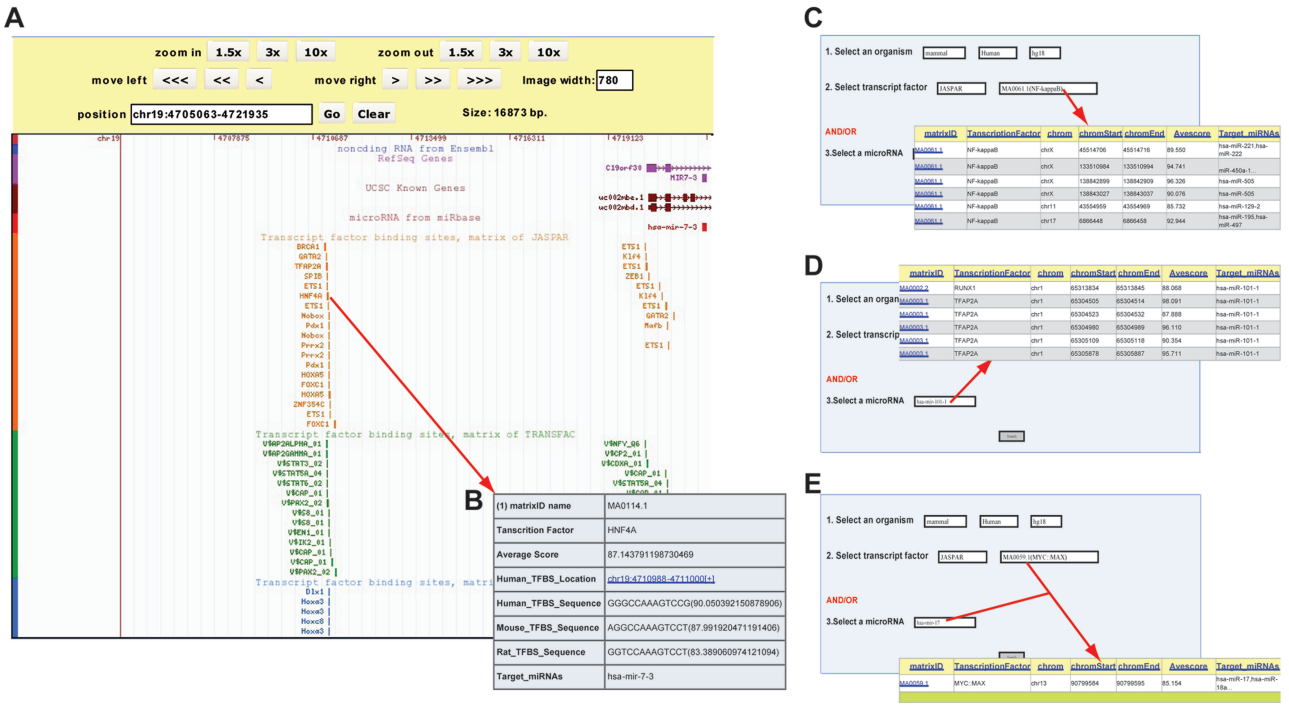

Figure 4. ACTViewer: a database documented the ACTLocater predictions on available PSSMs. (**A**) A snapshot of the ACTViewer genome browser. The controls (at the top of the browser) position the browser over a specific region in the genome. Annotations are displayed as individual tracks along the genomic regions. (**B**) A pop-up window containing detailed information about a predicted HNF4A-binding site. (**C**, **D** and **E**) Snapshots of searching ACTViewer for target miRNAs of NF-κB (C), TFBSs associated with hsa-miR-101-1 (D) and c-Myc→hsa-miR-17 interaction (E). Search results for target miRNAs of NF-κB and TFBSs associated with hsa-miR-101-1 are partially shown.

retrieved 129 PSSMs from JASPAR and predicted 13 170 TFBSs, corresponding to 3320 TF→miRNA interactions. Finally, we retrieved 389 PSSMs from UniPROBE and predicted 938 TFBSs, corresponding to 724 TF→miRNA interactions. Although false predictions could not be avoided, a large number of predictions were assumed to be true TFBSs according to the high precision of ACTLocater, indicating that a complex regulatory network governing the expression of HMR intergenic miRNAs remains to be decoded.

To provide an effective view of the TFBSs predicted by ACTLocater, we developed the ACTViewer database (ACTViewer is accessible at http://deepbase.sysu.edu.cn/ACTViewer/). We provided a genome browser and search forms in ACTViewer to facilitate data access. Predictions from ACTLocater could be viewed using the ACTViewer browser (Figure 4A). The browser displayed annotation tracks beneath genome coordinate positions, allowing rapid visual correlation of different types of genome annotations. Zooming and scrolling controls helped to narrow or broaden the displayed chromosomal range to focus on the exact region of interest. We displayed ACTLocater predictions as tracks according to the PSSMs data sources and added secondary links from entries within prediction tracks to lead to corresponding details from each entry (Figure 4B). With the ACTViewer browser, users can conveniently examine TFBSs that are clustered, overlapping or located in special genomic regions. To access the predictions made about a specific TF or miRNA, we also provided a list of search forms according to TF and/or miRNA. A query using a specific TF could retrieve the miRNAs predicted as its target genes, and using a specific miRNA could allow users to find the predicted TFBSs associated with the miRNA input (Figure 4C and D, respectively). Additionally, users could retrieve interactions from ACTViewer by specifying a TF and a miRNA simultaneously (Figure 4E). In conclusion, ACTViewer provided interfaces for integrating visualization and rapid retrieval of the ACTLocater predictions.

DISCUSSION

This work has presented a new computational method called ACTLocater that integrates evolutionary conservation and chromatin structure to predict TF→miRNA interactions with significantly improved PPVs. Most notably, ACTLocater uses chromatin accessibility data from a limited number of cell types for reference, but could also be applied to other cell types whose chromatin accessibility has not yet been characterized. Thus, the ACTLocater method and the ACTViewer resource have considerable potential to advance studies of the miRNA transcriptional regulation that underlies diverse human biological processes.

Previously, several studies have also attempted to predict the TFBSs associated with human miRNAs. For example, miRGen (33) has mapped all vertebrate TF

matrices from TRANSFAC to the regions spanning 5-kb upstream to 1-kb downstream of the miRNA TSSs. MIR@NT@N (34) has predicted potential TFBSs in the promoter regions of human miRNAs based on PSSMs from JASPAR and the 'CpG island' signal. However, these methods are based on simple pattern matching and are thus prone to false predictions. Until this point, the comparative genomics approach, such as the TFBS conserved track in the UCSC genome browser (35), was the most efficient method to predict potential TFBSs associated with miRNAs. As shown by our results for several TFs, ACTLocater yields a significant improvement in PPV compared to the comparative genomics approach using the same parameters (improvement in PPV could also be observed by comparing selected TF predictions to the UCSC conserved TFBS track, and the benefit of incorporating chromatin structure features could be noticed by comparing these predictions of the UCSC conserved TFBS track before and after filtering by our chromatin accessibility reference; see Supplementary Figure S2). Owing to the accumulation of cell-specific experimental data, computational methods integrating this information have also been developed to identify human TFBSs associated with miRNAs or on a genome-wide scale. For example, a method based on sequence features, histone modifications, and DNase I hypersensitivity has been developed and applied to human $CD4^+$ T cells (15). MITF-regulated miRNAs have also been successfully identified by combining nucleosome information and motif matching in human melanoma cells (38). Furthermore, CENTIPEDE (16) has been used in human lymphoblastoid cells by incorporating DNase I hypersensitivity. In contrast to these cell-specific methods, ACTLocater was based on a chromatin accessibility reference, which was derived from a panel of cell lines and was highly preserved across cell types, making ACTLocater widely applicable.

ChIP-seq has been considered the state-of-the-art experimental technique for profiling TFBSs including species-specific and non-canonical binding sites. Additionally, ChIP-seq has been able to avoid ambiguity for TFs that share similar binding preferences; however, ChIP-seq has to be carried out for only one TF using one set of conditions at a time. Moreover, because of the sequencing depth and performance of data analysis, some true binding sites are likely to be missed (67,68). ACTLocater, by contrast, can evaluate all TFs simultaneously, to provide nucleotide precision and to uncover novel binding sites missed by ChIP-seq assays. Because of these abilities, ACTLocater and ChIP-seq technology could be complementary tools to provide exhaustive information regarding *cis*-regulatory networks.

Although the performance of our method is encouraging, integrating more source data is likely to improve the ability of ACTLocater. Currently, chromatin accessibility datasets from only 12 cell types were used; we expect that by incorporating more chromatin accessibility datasets from other cell types released by the ENCODE Project (12,13), the coverage of all possible accessible and conserved regions will increase (improvement in the coverage of these regions by increasing cellular datasets can be seen in Supplementary Figure S3), and thereby improve the sensitivity of ACTLocater. Another limitation of ACTLocater is its reliance on binding motifs for the recognition of TFBSs. The human genome encodes about 1400 TFs with sequence-specific DNA-binding properties (69), and binding preferences were only available for a small proportion of these factors. Establishing the binding specificities of these TFs using techniques such as protein-binding arrays and high-throughput SELEX (70) will undoubtedly expand the predictive ability of ACTLocater.

Currently, ACTLocater focuses on the transcriptional regulation of intergenic miRNAs in humans, but modified algorithms based on the same principles could be applied for other genes or on a genome-wide scale as long as the required data are provided. Furthermore, the multiple genome alignments of *Drosophila* and *Caenorhabditis* species were already available (35), and the monENCODE Project continues to produce a large number of chromatin accessibility data of *D. melanogaster* (71) and *C. elegans* (72). Integrating these datasets, ACTLocater could also be expanded to these species.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–13, Supplementary Figures 1–3, Supplementary Methods, Supplementary Dataset 1, Supplementary Software 1 and Supplementary Reference [73].

ACKNOWLEDGEMENTS

The authors thank Prof. Yan Zhang at Sun Yat-sen University for providing the MDA-MB231 cells and Dr Jason Carroll at Cancer Research UK for providing the FOXA1 ChIP-seq data. The authors thank Xiao-Hong Chen, Qiao-Juan Huang and Yi-Ling Chen for technical assistance.

FUNDING

The National Basic Research Program from the Ministry of Science and Technology of China [No. 2011CB811300]; the National Natural Science Foundation of China [No. 30830066, 81070589 and 30900820]. Funding for open access charge: the National Natural Science Foundation of China [No. 30830066].

Conflict of interest statement. None declared.

REFERENCES

1. Lemon,B. and Tjian,R. (2000) Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.*, **14**, 2551–2569.
2. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
3. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using

chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.
4. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
5. Elnitski,L., Jin,V.X., Farnham,P.J. and Jones,S.J. (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res.*, **16**, 1455–1464.
6. Crawford,G.E., Holt,I.E., Whittle,J., Webb,B.D., Tai,D., Davis,S., Margulies,E.H., Chen,Y., Bernat,J.A., Ginsburg,D. *et al.* (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.*, **16**, 123–131.
7. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
8. Giresi,P.G., Kim,J., McDaniell,R.M., Iyer,V.R. and Lieb,J.D. (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.*, **17**, 877–885.
9. Giresi,P.G. and Lieb,J.D. (2009) Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). *Methods*, **48**, 233–239.
10. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
11. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
12. The ENCODE Project Consortium. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
13. The ENCODE Project Consortium. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
14. Whitington,T., Perkins,A.C. and Bailey,T.L. (2009) High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Res.*, **37**, 14–25.
15. Ernst,J., Plasterer,H.L., Simon,I. and Bar-Joseph,Z. (2010) Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.*, **20**, 526–536.
16. Pique-Regi,R., Degner,J.F., Pai,A.A., Gaffney,D.J., Gilad,Y. and Pritchard,J.K. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.
17. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
18. Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.
19. Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.
20. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
21. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
22. Dweep,H., Sticht,C., Pandey,P. and Gretz,N. (2011) miRWalk-database: prediction of possible miRNA binding sites by 'walking' the genes of three genomes. *J. Biomed. Inform.*, **44**, 839–847.
23. O'Donnell,K.A., Wentzel,E.A., Zeller,K.I., Dang,C.V. and Mendell,J.T. (2005) c-Myc-regulated microRNAs modulate E2F1 expression. *Nature*, **435**, 839–843.
24. Ma,L., Teruya-Feldstein,J. and Weinberg,R.A. (2007) Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. *Nature*, **449**, 682–688.
25. Chang,T.C., Yu,D., Lee,Y.S., Wentzel,E.A., Arking,D.E., West,K.M., Dang,C.V., Thomas-Tikhonenko,A. and Mendell,J.T. (2008) Widespread microRNA repression by Myc contributes to tumorigenesis. *Nat. Genet.*, **40**, 43–50.
26. 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.
27. Xu,H., He,J.H., Xiao,Z.D., Zhang,Q.Q., Chen,Y.Q., Zhou,H. and Qu,L.H. (2010) Liver-enriched transcription factors regulate microRNA-122 that targets CUTL1 during liver development. *Hepatology*, **52**, 1431–1442.
28. Wang,J., Lu,M., Qiu,C. and Cui,Q. (2010) TransmiR: a transcription factor-microRNA regulation database. *Nucleic Acids Res.*, **38**, D119–D122.
29. Rodriguez,A., Griffiths-Jones,S., Ashurst,J.L. and Bradley,A. (2004) Identification of mammalian microRNA host genes and transcription units. *Genome Res.*, **14**, 1902–1910.
30. Baskerville,S. and Bartel,D.P. (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, **11**, 241–247.
31. Lee,Y., Kim,M., Han,J., Yeom,K.H., Lee,S., Baek,S.H. and Kim,V.N. (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.*, **23**, 4051–4060.
32. Chang,T.C., Wentzel,E.A., Kent,O.A., Ramachandran,K., Mullendore,M., Lee,K.H., Feldmann,G., Yamakuchi,M., Ferlito,M., Lowenstein,C.J. *et al.* (2007) Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis. *Mol. Cell*, **26**, 745–752.
33. Alexiou,P., Vergoulis,T., Gleditzsch,M., Prekas,G., Dalamagas,T., Megraw,M., Grosse,I., Sellis,T. and Hatzigeorgiou,A.G. (2010) miRGen 2.0: a database of microRNA genomic information and regulation. *Nucleic Acids Res.*, **38**, D137–D141.
34. Le Béchec,A., Portales-Casamar,E., Vetter,G., Moes,M., Zindy,P.J., Saumet,A., Arenillas,D., Theillet,C., Wasserman,W.W., Lecellier,C.H. *et al.* (2011) MIR@NT@N: a framework integrating transcription factors, microRNAs and their targets to identify sub-network motifs in a meta-regulation network model. *BMC Bioinformatics*, **12**, 67.
35. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
36. Bandyopadhyay,S. and Bhattacharyya,M. (2010) PuTmiR: a database for extracting neighboring transcription factors of human microRNAs. *BMC Bioinformatics*, **11**, 190.
37. Stone,E.A., Cooper,G.M. and Sidow,A. (2005) Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu. Rev. Genomics Hum. Genet.*, **6**, 143–164.
38. Ozsolak,F., Poling,L.L., Wang,Z., Liu,H., Liu,X.S., Roeder,R.G., Zhang,X., Song,J.S. and Fisher,D.E. (2008) Chromatin structure analyses identify miRNA promoters. *Genes Dev.*, **22**, 3172–3183.
39. Kiezun,A., Artzi,S., Modai,S., Volk,N., Isakov,O. and Shomron,N. (2012) miRviewer: a multispecies microRNA homologous viewer. *BMC Res. Notes*, **5**, 92.
40. Barski,A., Jothi,R., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E. and Zhao,K. (2009) Chromatin poises miRNA- and protein-coding genes for expression. *Genome Res.*, **19**, 1742–1751.
41. Zhou,A.D., Diao,L.T., Xu,H., Xiao,Z.D., Li,J.H., Zhou,H. and Qu,L.H. (2012) beta-Catenin/LEF1 transactivates the microRNA-371-373 cluster that modulates the Wnt/beta-catenin-signaling pathway. *Oncogene*, **31**, 2968–2978.
42. Giardine,B., Riemer,C., Hardison,R.C., Burhans,R., Elnitski,L., Shah,P., Zhang,Y., Blankenberg,D., Albert,I., Taylor,J. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.
43. Washietl,S. and Hofacker,I.L. (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.*, **342**, 19–30.
44. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
45. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module

TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
46. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.
47. Overdier,D.G., Porcella,A. and Costa,R.H. (1994) The DNA-binding specificity of the hepatocyte nuclear factor 3/ forkhead domain is influenced by amino-acid residues adjacent to the recognition helix. *Mol. Cell Biol.*, **14**, 2755–2766.
48. Dang,C.V. (1999) c-Myc target genes involved in cell growth, apoptosis, and metabolism. *Mol. Cell Biol.*, **19**, 1–11.
49. Rabinovich,A., Jin,V.X., Rabinovich,R., Xu,X. and Farnham,P.J. (2008) E2F in vivo binding specificity: comparison of consensus versus nonconsensus binding sites. *Genome Res.*, **18**, 1763–1777.
50. Cao,Y., Yao,Z., Sarkar,D., Lawrence,M., Sanchez,G.J., Parker,M.H., MacQuarrie,K.L., Davison,J., Morgan,M.T., Ruzzo,W.L. *et al.* (2010) Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. *Dev. Cell*, **18**, 662–674.
51. Johnson,R., Gamblin,R.J., Ooi,L., Bruce,A.W., Donaldson,I.J., Westhead,D.R., Wood,I.C., Jackson,R.M. and Buckley,N.J. (2006) Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res.*, **34**, 3862–3877.
52. Johnson,R., Teh,C.H., Kunarso,G., Wong,K.Y., Srinivasan,G., Cooper,M.L., Volta,M., Chan,S.S., Lipovich,L., Pollard,S.M. *et al.* (2008) REST regulates distinct transcriptional networks in embryonic and neural stem cells. *PLoS Biol.*, **6**, e256.
53. Ettwiller,L., Paten,B., Ramialison,M., Birney,E. and Wittbrodt,J. (2007) Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat. Methods*, **4**, 563–565.
54. Lupien,M., Eeckhoute,J., Meyer,C.A., Wang,Q., Zhang,Y., Li,W., Carroll,J.S., Liu,X.S. and Brown,M. (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, **132**, 958–970.
55. Motallebipour,M., Ameur,A., Reddy Bysani,M.S., Patra,K., Wallerman,O., Mangion,J., Barker,M.A., McKernan,K.J., Komorowski,J. and Wadelius,C. (2009) Differential binding and co-binding pattern of FOXA1 and FOXA3 and their relation to H3K4me3 in HepG2 cells revealed by ChIP-seq. *Genome Biol.*, **10**, R129.
56. Hurtado,A., Holmes,K.A., Ross-Innes,C.S., Schmidt,D. and Carroll,J.S. (2011) FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat. Genet.*, **43**, 27–33.
57. Boeva,V., Surdez,D., Guillon,N., Tirode,F., Fejes,A.P., Delattre,O. and Barillot,E. (2010) De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. *Nucleic Acids Res.*, **38**, e126.
58. Yang,J.H., Shao,P., Zhou,H., Chen,Y.Q. and Qu,L.H. (2010) deepBase: a database for deeply annotating and mining deep sequencing data. *Nucleic Acids Res.*, **38**, D123–D130.
59. Christodoulou,F., Raible,F., Tomer,R., Simakov,O., Trachana,K., Klaus,S., Snyman,H., Hannon,G.J., Bork,P. and Arendt,D. (2010) Ancient animal microRNAs and the evolution of tissue identity. *Nature*, **463**, 1084–1088.
60. Saini,H.K., Griffiths-Jones,S. and Enright,A.J. (2007) Genomic analysis of human microRNA transcripts. *Proc. Natl Acad. Sci. USA*, **104**, 17719–17724.
61. Wang,X., Xuan,Z., Zhao,X., Li,Y. and Zhang,M.Q. (2009) High-resolution human core-promoter prediction with CoreBoost_HM. *Genome Res.*, **19**, 266–275.
62. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.
63. Barsyte-Lovejoy,D., Lau,S.K., Boutros,P.C., Khosravi,F., Jurisica,I., Andrulis,I.L., Tsao,M.S. and Penn,L.Z. (2006) The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis. *Cancer Res.*, **66**, 5330–5337.
64. Ma,L., Young,J., Prabhala,H., Pan,E., Mestdagh,P., Muth,D., Teruya-Feldstein,J., Reinhardt,F., Onder,T.T., Valastyan,S. *et al.* (2010) miR-9, a MYC/MYCN-activated microRNA, regulates E-cadherin and cancer metastasis. *Nat. Cell Biol.*, **12**, 247–256.
65. Xu,N., Papagiannakopoulos,T., Pan,G., Thomson,J.A. and Kosik,K.S. (2009) MicroRNA-145 regulates OCT4, SOX2, and KLF4 and represses pluripotency in human embryonic stem cells. *Cell*, **137**, 647–658.
66. Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**, D77–D82.
67. Kharchenko,P.V., Tolstorukov,M.Y. and Park,P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.
68. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.
69. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
70. Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, **11**, 751–760.
71. The modENCODE Consortium *et al.* (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, **330**, 1787–1797.
72. The modENCODE Consortium *et al.* (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, **330**, 1775–1787.
73. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.