Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis

Bum-Kyu Lee, Akshay A. Bhinge and Vishwanath R. Iyer*

Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, Section of Molecular Genetics and Microbiology, University of Texas at Austin, Austin, TX 78712, USA

ABSTRACT

The E2F family of transcription factors has important roles in cell cycle progression. E2F4 is an E2F family member that has been proposed to be primarily a repressor of transcription, but the scope of its binding activity and functions in transcriptional regulation is not fully known. We used ChIP sequencing (ChIP-seq) to identify around 16 000 E2F4 binding sites which potentially regulate 7346 downstream target genes with wide-ranging functions in DNA repair, cell cycle regulation, apoptosis, and other processes. While half of all E2F4 binding sites (56%) occurred near transcription start sites (TSSs), ~20% of sites occurred more than 20 kb away from any annotated TSS. These distal sites showed histone modifications suggesting that E2F4 may function as a long-range regulator, which we confirmed by functional experimental assays on a subset. Overexpression of E2F4 and its transcriptional cofactors of the retinoblastoma (Rb) family and its binding partner DP-1 revealed that E2F4 acts as an activator as well as a repressor. E2F4 binding sites also occurred near regulatory elements for miRNAs such as let-7a and mir-17, suggestive of regulation of miRNAs by E2F4. Taken together, our genome-wide analysis provided evidence of versatile roles of E2F4 and insights into its functions.

INTRODUCTION

The E2F transcription factor (TF) family is composed of eight different proteins including E2F1, E2F2, E2F3a, E2F3b (an isoform of E2F3) and E2F4–E2F8, which play pivotal roles in cell cycle progression and differentiation by activating or suppressing certain classes of E2F responsive genes (1,2). E2Fs have been classified as either activators (E2F1–E2F3) or repressors (E2F4–E2F8) (3). Interestingly, it was recently discovered that E2F1–E2F3 can switch from being activators to repressors in differentiating cells (4). The expression of E2F1–E2F3 is highly regulated during cell cycle progression whereas E2F4 and E2F5 are constitutively expressed (3).

The activity of E2F4 is regulated by several mechanisms such as sub-cellular localization, interactions with retinoblastoma (RB) proteins, post-translational modification such as phosphorylation, and decreased translation mediated by antisense transcripts (5,6). Unlike E2F1, E2F4 does not have a nuclear localization signal (NLS) and primarily exists in the cytoplasm during cell cycle progression. Upon cell cycle arrest, due to the depletion of mitogenic stimuli, cells start to enter into G_0, and cytoplasmic E2F4 forms heterodimers with a DP protein, which facilitates the localization of E2F4 complexes into the nucleus in a CRM1 mediated manner (7,8). Nuclear localized E2F4 complexes occupy target promoters and regulate diverse classes of genes which are involved in cell cycle, DNA repair and apoptosis (9). Three pocket proteins, pRB, p107/RBL1, and p130/RBL2, are critical cofactors for the regulation of E2Fs, and the expression levels of these pocket proteins change depending upon the phase of the cell cycle (10). Because of these results, it is believed that E2F4 is not involved in promoting cell cycle progression; instead, along with RBL2, it has crucial roles in mediating cell cycle arrest in G_0. The binding of the E2F4–RBL2 complex to E2F4 responsive promoters triggers the recruitment of HDAC complexes or other co-repressors, resulting in the repression of target gene expression (3,11).

Other observations however are not consistent with the view that E2F4 is exclusively a repressor of cell proliferation. Abnormal expression or mutation of E2F4 causes

*To whom correspondence should be addressed. Tel: +1 512 232 7833; Fax: +1 512 232 3472; Email: vishy@mail.utexas.edu
Present address:
Akshay A. Bhinge, Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore

the malfunction of cell cycle controls and results in malignant tumors. Transfection of E2F4 into non-transformed cells induces the oncogenic activity of E2F4 (12). In addition, E2F4 over-expressing transgenic mice develop tumors, providing evidence for the oncogenic activity of E2F4 (13). Mutated E2F4 has been reported in various tumors such as colorectal carcinomas, endometrial cancers, gastric adenocarcinomas, prostatic carcinomas, and ulcerative colitis-associated neoplasms, further emphasizing the important role E2F4 plays in tumorigenesis (12,14). The recent finding that E2F1–E2F3 can switch roles from activators to repressors suggests that the function of other members of this family of regulators may be also be more malleable that previously thought (4). In order to better understand the physiological roles of E2F4 and reconstruct its regulatory network, it is essential to identify genome-wide E2F4 targets and establish how target promoters respond to it.

Several chromatin immunoprecipitation (ChIP) followed by chip (ChIP-chip) experiments have been performed on core-promoter arrays with quiescent and continuously growing cells. These studies have led to the identification of several hundred E2F4 targets that are involved in diverse functions such as cell cycle regulation, DNA damage repair, apoptosis, mRNA processing, ubiquitination, etc. (9,15). A recent ChIP-chip study of E2F4 using tiled ENCODE arrays identified 187 E2F4 binding sites in 1% of the human genome in lymphoblastoid cells (16), suggesting the possibility that E2F4 may have more than 10 000 binding sites across the entire human genome. In addition, even though E2F4 showed a strong binding preference to promoters, some E2F4 binding sites were discovered in non-promoter regions. Without a comprehensive and unbiased genome-wide target analysis of E2F4, it is difficult to evaluate its promoter binding preferences or gain a complete understanding of its functions as a transcription factor.

The recently developed ChIP followed by sequencing (ChIP-seq) technique makes it possible to conduct a genome-wide unbiased search for binding sites of TFs (17–20). We used ChIP-seq to catalog E2F4 binding sites across the genome in the human B-lymphoblastoid cell line, GM06990. We discovered 16 246 putative E2F4 binding sites distributed across promoters to coding and non-coding regions, providing evidence to support diverse roles of E2F4, which were not reported in previous studies. Furthermore, gene expression profiling in response to overexpression of E2F4 in the presence of its cofactors showed that it can function as an activator as well as a repressor of transcription.

MATERIALS AND METHODS

Cell culture

The lymphoblastoid (GM06990) cell line was purchased from Coriell and cultured in RPMI medium containing 15% fetal bovine serum (FBS) and 1% antibiotics (penicillin/streptomycin). In order to perform time course ChIP experiments, cells were harvested every 24 h for 5 days. Serum starvation was achieved by washing cells cultured for 72 h three times with RPMI medium without FBS, then adding low-serum RPMI medium containing 0.1% FBS and then cultivating them for 2 days. For serum activation, low serum medium was replaced with RPMI containing 15% FBS. Cells were then cultivated and harvested at 3, 9 and 18 h.

ChIP sequencing

ChIP assays were performed as described previously (21). Briefly, GM06990 cells cultured for 72 h were cross-linked with 1% formaldehyde and incubated for 7 min at room temperature. Formaldehyde was deactivated by the addition of glycine (125 mM final concentration). Sonicated cell lysate containing an average size of 500 bp DNA fragments was used for immunoprecipitation to enrich E2F4-DNA complexes using an anti-E2F4 antibody (SC-1082X, Santa Cruz Biotech). Immunoprecipitated DNA was sequenced using Illumina sequencing technology (single end sequencing). Data from this study is available at the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/projects/geo/, GSE21488 and GSE21439).

Quantitative-PCR validation

Primer pairs for 42 targets and a negative control region (Supplementary Table S1) for normalization were designed using Primer3 (http://frodo.wi.mit.edu/primer3/input.htm). Quantitative-PCR (qPCR) was performed using the SYBR green PCR kit from Applied Biosystems with 1 ng of ChIP and input DNA. Fold enrichment of targets in ChIP DNA relative to input was calculated from an average of three replicate qPCR reactions.

E2F4 overexpression and expression microarrays

Full length E2F4, DP-1 and RBL2 clones were purchased from Open Biosystems and subcloned into the pcDNA 3.1 vector (Invitrogen). Either full-length expression constructs or empty vectors as a control were transfected into HeLa cells using Lipofectamine 2000 from Invitrogen. Total RNA was extracted from cells transfected with combinations of the three factors (E2F4, DP-1 and RBL2) or vector transfected cells using Trizol. Microarray experiments were performed using spotted HEEBO oligonucleotide human arrays (22), which has 44 308 probes, using the protocol described previously (23). Briefly, total RNA was converted into cDNA and labeled with Cy dyes (Cy3 for control and Cy5 for TF overexpression). Dye-coupled cDNA was combined and hybridized onto the oligo arrays for 14 hr. Cy5/Cy3 ratios were calculated from scanned intensity data from each channel. Data were normalized and analyzed by the error model described previously (24).

TaqMan assay for miRNA expression

Total RNA was extracted from relevant samples by Trizol. All primer sets for specific miRNAs and PCR reagents for TaqMan miRNA assay were purchased

from Applied Biosystems and real time PCR was performed using a 7900HT real time PCR machine from Applied Biosystems. RNU66 was used as an internal control for normalization. miRNA gene expression levels relative to the control was calculated from an average of four replicate qPCR reactions.

Luciferase reporter gene assay

Around 700 bp of PCR-amplified insert from each of 10 distal binding sites was cloned into the upstream position of a SV40 promoter in a pGL3 plasmid (Promega cat. # E1761) between the KpnI and XhoI restriction sites. All primers used for cloning are listed in Supplementary Table S1. For the luciferase reporter gene assay, approximately 3×10^5 HEK 293 cells were co-transfected with 200 ng of the pGL3 vector or reporter construct containing the Firefly reporter gene as well as 10 ng of pRL-PK vector (Promega cat. # E2241) containing a Renilla reporter gene, which served as an internal control reporter, using 1 μl of lipofectamine 2000 (Invitrogen) and Opti-MEM medium (Invitrogen). After transfection and incubation for 24 h, cells were washed with PBS once, lysed, and assayed to measure luciferase activity using the Dual Luciferase assay Kit (Promega cat. # E1910) according to manufacturer's instructions. Firefly luciferase activity from the pGL3 construct was then normalized to Renilla luciferase activity. To calculate relative expression fold change, firefly activity of the pGL3 vector containing a distal E2F4 binding site were further normalized with that of an empty pGL3 vector. *P*-value was calculated from three independent transfections using a *t*-test.

Identification of ChIP-seq binding peaks and sites

Illumina sequencing generated 23–32 base pair short reads from the ends of ChIP-enriched DNA fragments. These short reads were mapped back to the genome using the ELAND algorithm. We obtained 6 508 011 uniquely aligning reads from the E2F4 chip library and 8 474 489 uniquely aligning reads from the input library. To identify E2F4 binding sites from high-throughput Illumina sequencing data, we used a Parzen window based algorithm as described previously with minor modifications (25). Each read was assigned a score that was essentially the frequency of observing that read in the sequencing library. The plus and the minus strand reads were analyzed separately to find peaks on the plus and minus strands, respectively. The algorithm begins by assigning the score of each read to its neighboring nucleotides as a function of the read's distance from that nucleotide. The function used to assign scores was a Gaussian kernel with a defined band-width. Local maxima on the plus and minus strands were defined as peaks. High scoring plus peaks that are upstream and within 500 bp of minus strand peaks were considered to be paired and the distance between the paired plus and the minus peak was calculated as the fragment length. A second iteration of peak finding was then carried out, where all aligning reads were extended in the 3′ direction by half of the

previously estimated fragment length, to effectively represent the center of the ChIP fragment. The peak-finding algorithm described above was used again on these positions to find local maxima across the genome thereby defining binding sites. The score associated with the nucleotide corresponding to the maxima was assigned to the binding site.

Input correction

In order to correct high ChIP-seq scores arising from repetitive sequences or copy number repeats rather than true ChIP enrichment, we normalized the E2F4 scores by the parallel input sequencing scores. Scores for E2F4 peaks that were within 500 bp from any input peak were divided by the corresponding input peak score. If a given E2F4 peak overlapped with more than one input peak, the higher scoring input peak was used for the correction. E2F4 peaks mapping to within 10 000 bp from any transcription start site (TSS) of a gene were not input corrected, since peaks near promoters in sonicated crosslinked chromatin can arise even in input DNA due to transcription factor binding (26). This restriction applied to only a small fraction (1.3%) of all E2F4 sites reported here.

False discovery rate

We ran the peak-finding algorithm on a set of randomly simulated read coordinates equal in number to the ChIP-seq data. These simulations were repeated 20 times. At each of a series of different score thresholds, the number of E2F4 peaks found after input correction was compared to those found in the random simulations to give the false discovery rate (FDR).

Saturation

We used a capture–recapture analysis to estimate saturation of binding sites in our E2F4 data. Capture-recapture analysis has been used to estimate population sizes of animals in a given area. The reads from the E2F4 chip library were obtained in two sets or 'lanes', the first set having 2 305 280 reads and the second set having 4 202 731 reads. Each set was treated as an independent capture. The entire genome was binned into 500 bp bins and reads mapping to each bin were counted for the two sets separately. Each bin was now assigned a *P*-value value dependent on the number of reads observed within that bin according to a random Poisson model. At different *P*-value thresholds, we calculated the following:

$N1$: number of bins in set 1;
$N2$: number of bins in set 2;
K: number of bins common to set 1 and set 2;
$E1$: expected number of bins in set 1 according to a random Poisson model;
$E2$: expected number of bins in set 2 according to a random Poisson model;
FDR1: $(E1/N1) \times 100$;
FDR2: $(E2/N2) \times 100$;

*E*1 and *E*2 represent the expected number of false positives at each enrichment cut-off.

The estimated number of E2F4 bins at each *P*-value cut-off was calculated as:

$$P = (N1 - E1)(N2 - E2)/\min[k(1 - E1/N1), k(1 - E2/N2)].$$

Whereas the observed number of bins was calculated as:

$$O = (N1 - E1) + (N2 - E2) - \min[k(1 - E1/N1), \\ k(1 - E2/N2)].$$

The percentage saturation was calculated as:

$$S = (O/P) \times 100.$$

For each *P*-value cut-off, we calculated the average FDR as the geometric mean of FDR1 and FDR2. The percentage saturation (*S*) was now plotted as a function of the average FDR.

Mapping binding sites to gene features

To detect E2F4 target genes, E2F4 sites were mapped to within 2 kb from the TSS of all genes annotated in the RefSeq database. In order to estimate the number of sites mapping to different gene features, it was necessary to assign one site to one and only one gene feature. Since E2F4 has been known to preferentially bind near the TSSs of genes, we used the following hierarchy to assign sites to features: core > upstream > intron > exon > intergenic. Core was defined as 2000 bp upstream and downstream from the TSS, upstream was defined as >2000 bp upstream to a maximum of 20 000 bp upstream from the TSS. Binding sites that could not be mapped to within 20 000 bp upstream of any TSS and were not assigned to any intron or exon were termed intergenic. Genes that had E2F4 binding sites within the core were defined as targets.

Mapping binding sites to miRNAs

In order to identify miRNA targets of E2F4, we excluded binding sites that mapped to core promoters, as it was not possible to unequivocally assign such sites to the annotated gene or the miRNA using binding data alone. We included miR-22 as a special case for further characterization because we have identified a role for miR-22 in the cell cycle (A.A. Bhinge and V.R. Iyer, unpublished data). Sites mapping to intergenic regions, introns and exons were mapped to within 10 000 bp of the annotated starts of mature miRNAs. The data for mature miRNA start/stop coordinates was downloaded from miRBase (www.mirbase.org).

Generating TSS/TTS profiles

A region of 20 kb around the TSS (10 kb upstream and 10 kb downstream), was binned in 50 bp size bins and E2F4 sites were mapped to each bin. Each bin was assigned the score of the peak that mapped to it. Corresponding bin scores were averaged across all genes to generate an average peak score profile across the TSS. This average profile was smoothened by a moving window of three bins. The same procedure was used to generate transcription termination site (TTS) profiles.

Motif analysis

Since attempting to run motif discovery algorithms on all binding sites would have been computationally expensive, we divided the binding sites into strong (score $\geq$24.93), moderate (scores 8.01–9.01) and weak (scores 5.6–5.75) categories and considered the top 500 sites from each category for motif discovery. A 200 bp region centered on each site was extracted from the human genome assembly hg18. Motif discovery was performed using the software DRIM (27) on each category separately at a *P*-value threshold of 1×10^{-5}. A random background was generated by sampling 200 000 sequences of 200 bp from the genome. E2F4 sites (55.9%) occurred within 2 kb upstream and downstream of TSSs of genes and this ratio was maintained in the random sample. In addition, we calculated the enrichment of each motif with respect to the random background as a function of the peak score. To analyze the relationship between each motif and different gene features like core, upstream, intron, exon and intergenic (as defined above), we divided sites into different features such that each site was assigned to one and only one feature (as described above) and then extracted sequences associated with these feature-specific sites (200 bp centered on the site). Then, for each feature, we counted the number of sites that had a given motif and divided this number by the total number of sites that had that specific motif. This was repeated for each of the five analyzed features for a given motif and the data was displayed as a heat-map.

Motif co-enrichment

Conserved TF binding site data for the human genome assembly hg18 was obtained from http://genome.ucsc.edu. This data contains TF bind sites (TFBS) for 398 TFs from the TRANSFAC database that are conserved between human, mouse and rat (http://genome.ucsc.edu/cgi/bin/hgTrackUi?hgsid = 118849564&c = chr13&g = tfbsConsSites). A TFBS was considered associated with an E2F4 site if it was found within 250 bp of that site. The frequencies of TFBS associated with binding sites were calculated for E2F4 peaks as well as for the randomly generated peaks. The analysis was performed on the strong, moderate, and weak categories separately and *P*-values were calculated according to a binomial model. We excluded TFBS that were not enriched at a *P*-value of $<1 \times 10^{-6}$ and were associated with less than 4% of the E2F4 sites under consideration.

Motif co-occurrence

We counted the number of different motifs that were associated with each binding site peak (i.e. within 100 bp on either side of the peak), and compared the distribution of these counts between randomly generated peaks and E2F4 binding site peaks.

RESULTS

Optimal cell culture conditions for E2F4 ChIP in lymphoblastoid cells

Since E2F4 is thought to occupy its target promoters in quiescent cells which could arise after long periods in culture, we first examined E2F4 ChIP efficiency in lymphoblastoid cells that were grown for 24, 48, 72 and 96 h after replating, and hybridizing the ChIP-enriched DNA to a core-promoter array that covered from −750 to +200 bp from the TSS of ∼9000 genes. We also investigated E2F4 binding to its target promoters in serum-starved and serum-fed cells using core promoter arrays because it is known that upon serum starvation, E2F4 accumulates in the nucleus and is recruited to its target promoters, resulting in cell cycle arrest (28). Strong occupancy signals for E2F4 were detected at 72 h and were maintained at 96 h. We found that the E2F4 binding profile was not significantly affected by serum-starvation or stimulation in lymphoblastoid cells (Supplementary Figure S1). Based on the above results, we generated ChIP-seq data for E2F4 in lymphoblastoid cells that were maintained in culture for 72 h.

Identification and verification of E2F4 binding sites from ChIP-seq data

Illumina sequencing of E2F4 ChIP generated around 6.5-million sequence reads that uniquely mapped to the reference human genome sequence. In order to identify E2F4 binding sites from this data, we developed a peak detection program using a Parzen windows density estimation algorithm we have previously used to map nucleosome positions (25). We also sequenced a parallel sample of input DNA from the same cells as a control to ensure that enriched sites were not an artifact of the processing and sequencing. Our algorithm identifies discrete regions of ∼150 bp around each ChIP-seq peak which we refer to as sites. Known E2F4 targets were easily detected in our ChIP-seq data (Figure 1A). We observed that some strong peaks in the ChIP dataset corresponded to equally strong peaks in the input sequencing data (Figure 1B). Such peaks may arise due to the presence of repeat regions in the genome and are thus false positives. To minimize such false positive targets, we first normalized the ChIP peaks by dividing the ChIP peak scores with their corresponding input peak scores. Next, we calculated a FDR based on random simulations to decide an appropriate significance threshold (Figure 1C). At 1% FDR, which corresponded to an input-corrected peak score of 4.4, we identified 16 246 putative E2F4 binding sites across the entire genome (Supplementary Table S2). To verify the quality of the putative E2F4 binding sites, we investigated overlaps between ChIP-seq targets with those from our core promoter arrays. Around 84% of core promoter targets were also found in the ChIP-seq data (Supplementary Figure S2). As further verification, we assayed 42 randomly selected targets by quantitative-PCR (qPCR), including 30 targets with a score between 4.4 and 8, as well as 12 targets with a score <4.4, which was below our 1% FDR threshold. Overall, binding sites

with stronger ChIP-seq scores showed higher fold enrichment by qPCR. Specifically, more than 90% of the targets above the 1% FDR threshold showed an enrichment of at least 1.5-fold by qPCR. On the other hand, only 41% of sites below the 1% FDR threshold showed enrichment by qPCR (Figure 1D). We also estimated the extent to which we identified all E2F4 sites in the genome, using a tagging-recapture saturation analysis (see Materials and Methods section). At the FDR threshold of 1% and our given sequencing depth, we estimated that we identified more than 80% of all E2F4 binding sites in the human genome in lymphoblastoid cells (Figure 1E).

Distribution of E2F4 binding sites in relation to gene annotations

We first examined the relationship of E2F4 binding with the distribution of genes. E2F4 binding sites are correlated well with gene density across the genome ($r^2 = 0.75$) (Figure 2A). Next, we investigated the distribution of E2F4 binding sites across five different categories of genomic elements including promoter, exon, intron, intergenic and upstream regions by mapping E2F4 sites relative to RefSeq annotated genes. Approximately 56% of the sites occurred within promoters (Figure 2B). In addition, the binding profile of E2F4 around TSSs also provided evidence that E2F4 had a preference for binding promoters, especially near the TSS, whereas no significant binding preference was observed near the TTSs (Figure 2C). This result is consistent with previous studies using selective arrays, showing that the binding sites of several TFs, including E2F1, were mainly distributed near the TSSs (16,29,30) Since E2F4 showed preferential binding at promoters, we examined whether the binding strength as measured by the peak score of E2F4 sites at promoters was stronger than sites at other genomic regions. E2F4 promoter sites showed significantly higher peak scores compared to those from other genomic regions (Figure 2D). Taken together, these results show that not only does E2F4 bind preferentially to promoters, but it also binds with higher occupancy to promoters as compared to other genomic regions.

E2F4 and bidirectional promoters

It has been reported that consensus E2F4 motifs are significantly overrepresented in bidirectional promoters (31). Approximately 11% of all genes have been reported to have bidirectional promoters in mammalian genomes (32,33). For our analysis, we first defined bidirectional promoters as the region of DNA between the TSSs of two genes that were divergently transcribed from opposite strands and separated by <2 kb. Based on this criterion, we identified 918 bidirectional promoters corresponding to 1836 genes among all the 18 693 human genes annotated in RefSeq. We then investigated whether E2F4 binding sites showed a bias toward binding to bidirectional promoters. Of the 4599 E2F4 target genes where its binding site was within the same specified distance (2 kb) from the TSS, 572 were bidirectionally transcribed, which was a significant enrichment over background
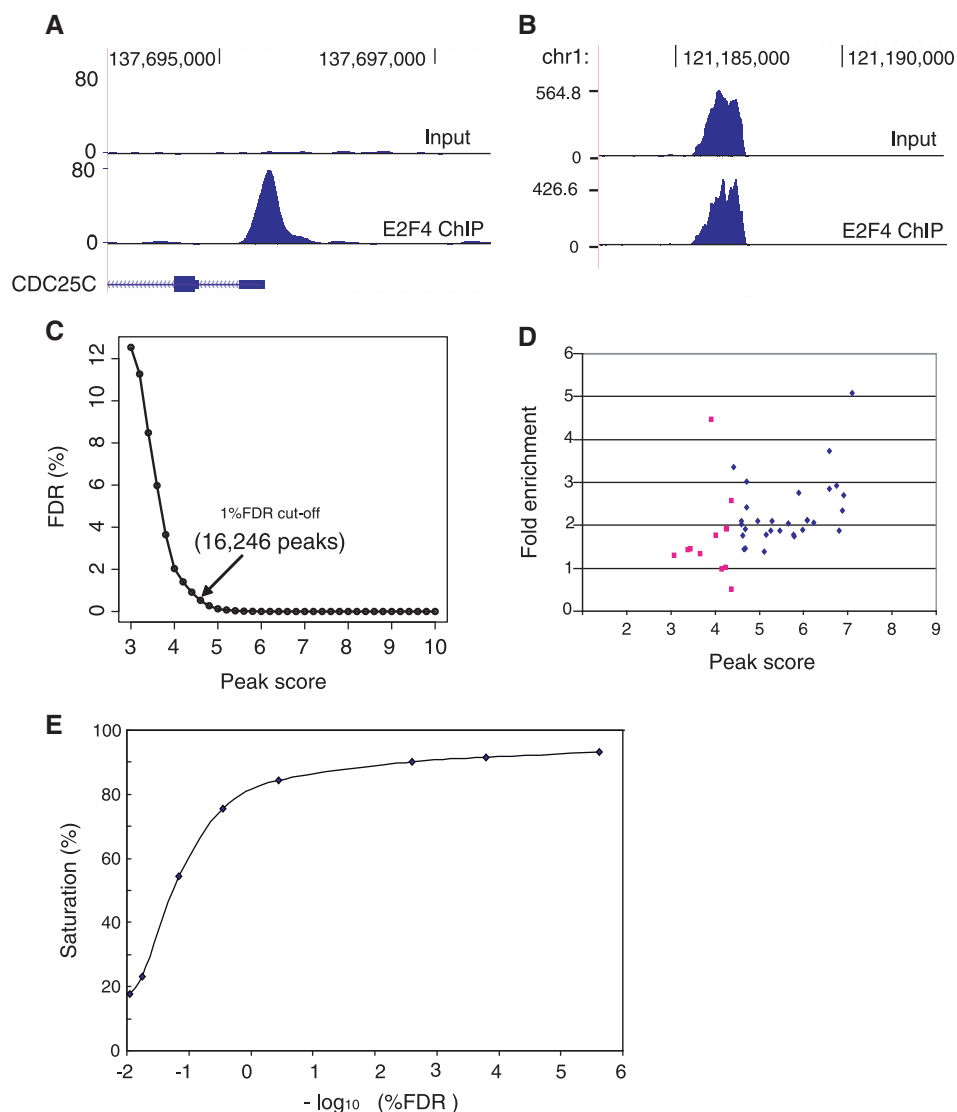
Figure 1. E2F4 ChIP-seq reveals genome-wide E2F4 binding sites. (**A**) An example of a known E2F4 binding site that was identified in our ChIP-seq data. Chromosome coordinates are indicated on top. The plot in the middle shows the density of ChIP-seq reads, with the peak score indicated on the Y axis. The bottom track shows the CDC25C gene with coding regions, exons and introns indicated by thick or thin boxes and line, respectively. The direction of transcription is indicated by the arrows from right to left. (**B**) An example of strong peaks discovered in both input and ChIP likely due to copy number differences between the cell genome and the reference sequence. Such sites were removed by input correction ('Materials and Methods' section). (**C**) FDR calculation based on random simulations. The 1% FDR threshold was used for further analysis. (**D**) qPCR verification of 42 randomly selected targets identified by ChIP-seq. Blue diamonds represent targets which passed the 1% FDR ChIP-seq threshold, and red squares represent targets below this threshold. (**E**) Capture-recapture analysis to estimate saturation for E2F4 targets (see Methods). *x*-axis represents $-\log_{10}$ FDR and the *y*-axis shows the saturation as a percentage of expected sites that were discovered at each FDR.

(hypergeometric $P < 1.1 \times 10^{-11}$) (Supplementary Table S3), indicating that many divergently transcribed genes in the genome might be co-regulated by E2F4. Divergently transcribed E2F4 target genes were highly overrepresented in the categories of RNA processing, DNA repair, protein folding and cell cycle.

Distal E2F4 sites could be enhancers or other regulatory elements

In addition to strong promoter occupancy and bidirectional promoter enrichment of E2F4 binding sites, a proportion of E2F4 sites were also detected in introns (7.8%),

upstream regions (13.2%) and intergenic regions (21.7%) (Figure 2B). Previous approaches have not identified this latter class of E2F4 sites that are not at the core promoter. To exclude the possibility that the limited number of TSS annotated in Refseq (~18 000) was resulting in an overestimate of the number of intergenic binding sites, we mapped all E2F4 sites to an expanded data set of ~60 000 TSSs derived by combining RefFlat annotated genes with additional gene annotations obtained from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables) and further filtered to remove redundant TSS coordinates. Even with the much larger number of TSSs used in this analysis, the number of intergenic E2F4
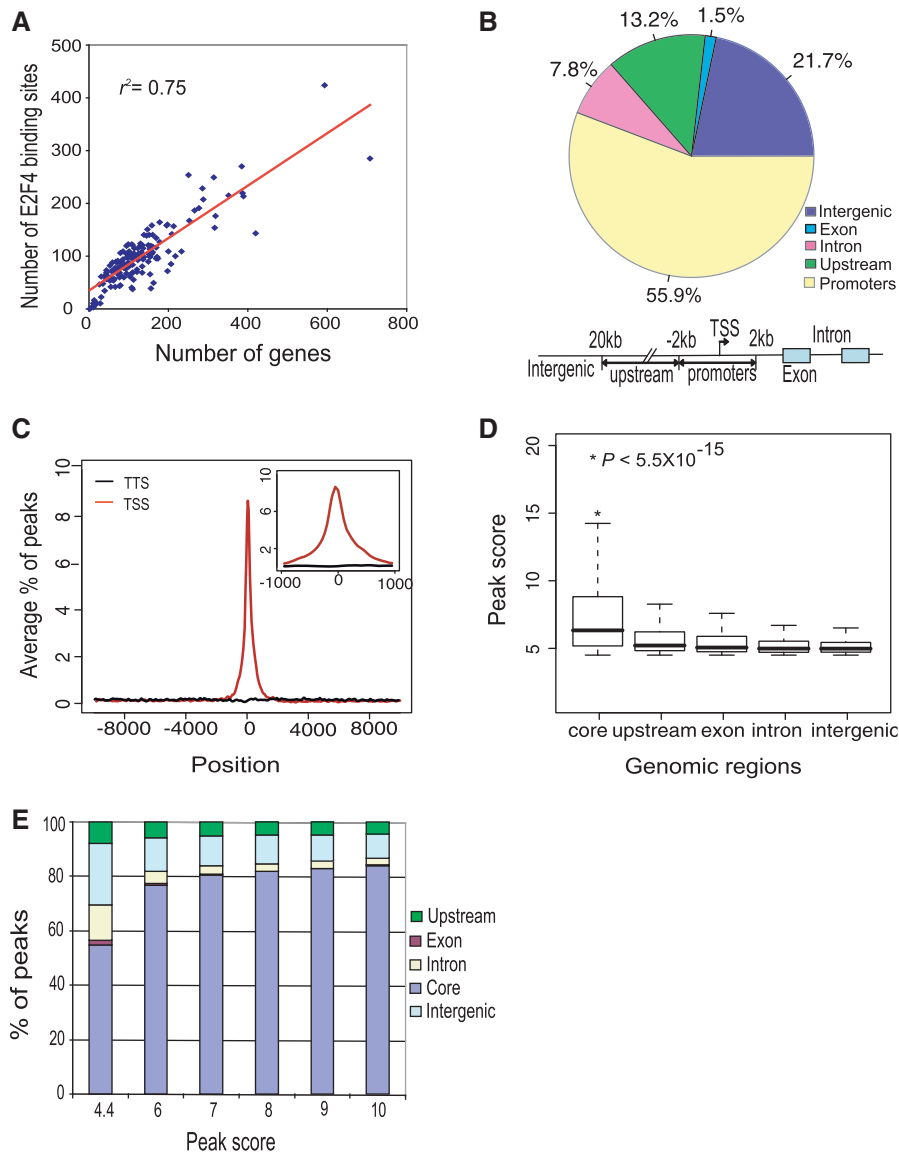
Figure 2. The genome-wide distribution pattern of E2F4 binding sites. (**A**) The correlation between E2F4 binding sites and gene density. Each point on the plot represents a 20 Mb bin. (**B**) A pie chart representation of the distribution of E2F4 binding sites in five different genomic regions. The definition of each genomic region is described below. Core promoters are within ±2 kb from the TSS, upstream is from 2 to 20 kb upstream from the TSS, and intergenic is a region not included as a promoter, upstream region, intron or exon. (**C**) Distribution of E2F4 binding sites within ±10kb. Inset shows a close up of a 1 kb region centered on the TSS. (**D**) A box-plot shows the ChIP-seq peak score distribution across five different genomic regions. Peak scores in core promoters were significantly higher compared to those from other genomic regions ($P < 5.5 \times 10^{-15}$, Wilcoxon test with Bonferroni correction). (**E**) Distribution of E2F4 binding sites depending on peak scores. Even though the number of intergenic sites decreased with increasing score, a substantial proportion of intergenic sites (10%) still remained at a score of 10.

sites showed only a modest decrease from 22.5 to 17.5%, indicating that a significant proportion of E2F4 sites are truly intergenic. We then analyzed the distribution of E2F4 sites at different peak score cut-offs to investigate whether the percentage of intergenic sites was dependent on the score. Although the proportion of intergenic sites decreased with an increase in the score threshold, it remained fairly constant above a score cut-off of 10, where ~10% of E2F4 sites were deemed intergenic (Figure 2E). Overall, these results indicate the existence of a significant number of strong E2F4 sites that were found <20 kb away from any annotated TSS.

It is well-established that TFs are able to regulate the expression of target genes by binding to promoters or long-range regulatory elements such as enhancers and insulators. To investigate the possibility that some of the distal (upstream and intergenic) E2F4 binding sites represent enhancers, we examined these distal E2F4 binding sites (based on the expanded TSS annotations above) for the presence of characteristic enhancer signature marks based on published histone modification data (17,34,35). Genome-wide histone signature analyses have revealed that the three forms of H3K4 methylation (H3K4me1, H3K4me2 and H3K4me3), H3K9me1, H3K18ac and the
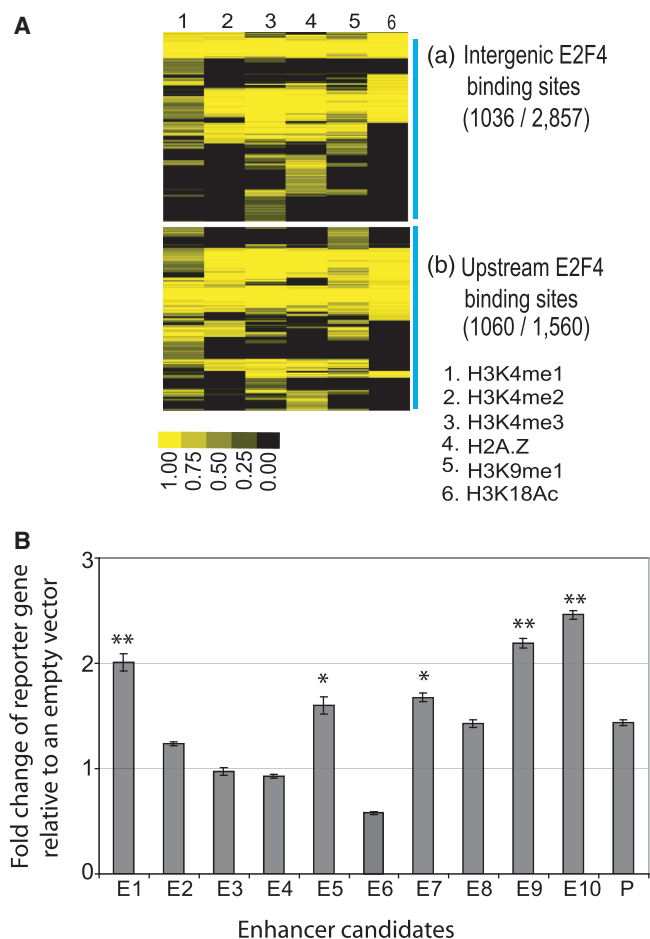
A

B

Figure 3. Some E2F4-bound distal sites function as enhancers. (**A**) Relationship of histone enhancer marks with the 2857 intergenic E2F4 sites (a) or 1560 upstream E2F4 sites (b). Data for histone modifications indicative of enhancers was obtained from previous studies (17,35), assigned to E2F4 binding sites identified here and hierarchically clustered for display. The relative strength of the histone modification signal is indicated in the heat-map according to the color table. Sixty-eight percent of upstream and 36% of intergenic E2F4 sites contained at least one enhancer mark. (**B**) Luciferase reporter gene assays for randomly chosen 10 distal E2F4 binding sites. The *y*-axis represents the expression fold change of a luciferase reporter gene normalized to an empty-vector control. *P*-values were calculated using *t*-test from three independent transfections. E1 through E10 represent 10 enhancer candidates randomly selected from among distal E2F4 binding sites. 'P' represents a positive control enhancer selected based on a previously published study (64). Single and double asterisks indicates $P < 0.005$ and 0.001, respectively.

variant H2A.Z are highly enriched in enhancer sites. We found that 36% of the intergenic E2F4 binding sites and 68% of the upstream sites had at least one histone enhancer marker, with most of these E2F4 sites showing multiple enhancer makers (Figure 3A). To verify the possibility that distal E2F4 binding sites could function as enhancers, we first tested whether they showed binding by p300, a known marker of enhancers in mammalian cells (36). We tested p300 binding at 10 randomly selected distal E2F4 enhancer candidate sites using ChIP followed by real-time PCR. We found that 9 out of 10 sites showed significant binding by p300 (Supplementary Figure S3), which supported the possibility that some of

the distal E2F4 binding sites we identified by ChIP-seq could function as enhancers. To functionally test this possibility further, we cloned these 10 candidates into pGL3 promoter-containing enhancer reporter vectors and performed luciferase reporter gene assays. Five of the 10 sites conferred significant increase in expression of their reporter genes ($P < 0.005$) at levels comparable to or greater than a positive control enhancer, confirming that a subset of distal E2F4 binding sites could indeed function as enhancers (Figure 3B). Based on the reporter assays, we can estimate that ~1048 out of 4147 distal E2F4 binding sites may function as enhancers. Distal E2F4 sites that did not show any enhancer marks may potentially regulate non-coding genes such as miRNAs whose promoters are not well defined, or these distal sites could be transcriptionally neutral.

Putative E2F4 target genes are involved in a broad range of biological processes

In order to investigate the functions of E2F4 suggested by its genome-wide binding profile, we first considered all E2F4 sites that occurred within ± 2 kb from the TSS of all genes annotated in the RefSeq database. We found 7346 genes that had E2F4 binding sites in their promoters, which cover ~30% of all annotated human genes (Supplementary Table S4). Next, we analyzed functional categories among these putative E2F4 target genes using the database for annotation, visualization and integrated discovery (DAVID) (37). In agreement with previously reported results, E2F4 target genes were highly enriched for cell cycle, DNA repair, RNA processing, stress response, apoptosis and ubiquitination (Table 1). We also found significant enrichment among E2F4 targets for additional functions that have not previously been associated with E2F4, such as protein transport and targeting, protein folding and I-kappaB kinase/NF-kappaB cascade. KEGG pathway analysis also showed strong enrichment of E2F4 in the categories of cell cycle, ubiquitin-mediated proteolysis, p53 signaling pathway and chronic myeloid leukemia (Supplementary Table S5). We also found that E2F4 binds to the promoters of 780 TFs out of the approximately 2000 known TFs in the human genome (Supplementary Table S6), which suggests that E2F4 regulates broad classes of genes indirectly. Taken together, these results suggest that E2F4 could regulate more diverse biological processes than previously suspected.

E2F4 potentially regulates other E2F family members and its cofactors

Previous ChIP-chip studies have revealed that members of the E2F family transcriptionally regulate each other (9,29). For instance, E2F4 occupies the promoters of activator E2Fs (E2F1, E2F2 and E2F3) and represses their expression to cause cell-cycle arrest. We found that E2F4 occupied the promoters of all E2F family genes including E2F7 and E2F8, which are RB independent repressors (2). The notable exceptions were E2F4 itself, and E2F6 (Supplementary Table S7), which interestingly, functions redundantly with E2F4 as a repressor. E2F4 also occupied

Table 1. Functional categories of E2F4 target genes

Biological functions	Count (Percent)	Fold enrichment	*P*-value	FDR
Biopolymer metabolic process	2259 (34.81)	1.36	6.32*E*-103	1.21*E*-99
Nucleotide and nucleic acid metabolic Process	1736 (26.75)	1.38	1.86*E*-77	3.57*E*-74
Cell cycle	490 (7.55)	1.77	8.05*E*-52	1.54*E*-48
Gene expression	1499 (23.10)	1.31	2.09*E*-46	4.00*E*-43
RNA processing	277 (4.27)	1.96	1.55*E*-39	2.97*E*-36
Organelle organization and biogenesis	579 (8.92)	1.57	3.36*E*-39	6.42*E*-36
Response to DNA damage stimulus	208 (3.21)	2.07	7.92*E*-35	1.51*E*-31
Biopolymer modification	811 (12.50)	1.4	1.06*E*-32	2.03*E*-29
mRNA processing	172 (2.65)	2.14	3.14*E*-31	6.01*E*-28
RNA splicing	156 (2.40)	2.22	3.23*E*-31	6.18*E*-28
Protein modification process	772 (11.90)	1.38	1.98*E*-29	3.78*E*-26
DNA repair	171 (2.64)	2.07	1.48*E*-28	2.83*E*-25
Ubiquitin cycle	278 (4.28)	1.74	7.46*E*-28	1.43*E*-24
Response to endogenous stimulus	230 (3.54)	1.84	1.38*E*-27	2.64*E*-24
Macromolecule localization	401 (6.18)	1.54	2.09*E*-25	4.00*E*-22
Protein transport	345 (5.32)	1.6	2.64*E*-25	5.05*E*-22
Chromosome organization and biogenesis	221 (3.41)	1.81	3.61*E*-25	6.91*E*-22
Post-translational protein modification	653 (10.06)	1.39	4.68*E*-25	8.96*E*-22
Transcription	1061 (16.35)	1.27	1.67*E*-24	3.19*E*-21
Protein localization	373 (5.75)	1.53	1.34*E*-22	2.56*E*-19
DNA replication	145 (2.23)	1.87	1.42*E*-18	2.73*E*-15
Apoptosis	356 (5.49)	1.47	3.15*E*-18	6.02*E*-15
Chromatin modification	118 (1.82)	1.9	1.11*E*-15	2.12*E*-12
Establishment and maintenance of Chromatin	165 (2.54)	1.69	4.09*E*-15	7.86*E*-12
DNA packaging	166 (2.56)	1.67	1.48*E*-14	2.85*E*-11
Chromosome segregation	50 (0.77)	2.56	2.65*E*-14	5.05*E*-11
Ribonucleoprotein complex biogenesis and assembly	116 (1.79)	1.85	2.84*E*-14	5.44*E*-11
Protein targeting	122 (1.88)	1.8	7.12*E*-14	1.36*E*-10
Cell development	489 (7.54)	1.27	1.16*E*-10	2.21*E*-07
RNA localization	58 (0.89)	2.1	1.36*E*-10	2.61*E*-07
Protein modification by small protein conjugation	55 (0.85)	2.14	1.73*E*-10	3.30*E*-07
Sister chromatid segregation	28 (0.43)	2.82	6.29*E*-10	1.20*E*-06
Protein ubiquitination	51 (0.79)	2.11	1.76*E*-09	3.37*E*-06
Ubiquitin-dependent protein catabolic process	98 (1.51)	1.68	3.04*E*-09	5.81*E*-06
Ribosome biogenesis and assembly	55 (0.85)	1.95	2.16*E*-08	4.14*E*-05
Protein kinase cascade	174 (2.68)	1.43	2.96*E*-08	5.67*E*-05
Response to stress	416 (6.41)	1.24	6.16*E*-08	1.18*E*-04
Spindle organization and biogenesis	19 (0.29)	3.06	6.67*E*-08	1.28*E*-04
Phosphate metabolic process	399 (6.15)	1.25	1.07*E*-07	2.04*E*-04
Phosphorus metabolic process	399 (6.15)	1.25	1.07*E*-07	2.04*E*-04
Protein folding	127 (1.96)	1.49	2.01*E*-07	3.85*E*-04
DNA damage response, signal transduction	33 (0.51)	2.22	3.98*E*-07	7.62*E*-04
Protein–RNA complex assembly	62 (0.96)	1.73	1.07*E*-06	0.002
Regulation of gene expression, epigenetic	34 (0.52)	2.11	1.43*E*-06	0.002
Chromatin assembly or disassembly	73 (1.12)	1.63	2.06*E*-06	0.003
Lipid biosynthetic process	124 (1.91)	1.44	2.09*E*-06	0.004
Microtubule organization and biogenesis	17 (0.26)	2.89	2.53*E*-06	0.004
Centrosome organization and biogenesis	17 (0.26)	2.89	2.53*E*-06	0.004
I-kappaB kinase/NF-kappaB cascade	69 (1.06)	1.63	3.98*E*-06	0.007

Count represents the number of genes in the biological function category. Percent shows the proportion of E2F4 targets among the count. FDR is the false discovery rate. Functional categories were as defined by the online database DAVID (37). *P*-values and FDR were also as calculated using their online tool, based on the list of E2F4 target genes identified in this study.

the promoters of the three RB family proteins (pRB, p107/RBL1 and p130/RBL2), and two binding partners (DP1 and DP2). In particular, E2F4 showed strong binding at the promoters of E2F1–E2F3, which are genes involved in cell cycle progression. To identify targets that were common to E2F1 and E2F4, we compared our E2F4 targets with previously published E2F1 targets obtained from ChIP-chip data in the same cell line (16). Among the top 2000 known E2F1 targets, 1416 targets (∼70%) overlapped with our E2F4 targets identified by ChIP-seq (Supplementary Table S8). Functional analysis revealed that cell cycle and DNA repair functions were highly enriched among these genes that were occupied by both E2F1 and E2F4.

The RB protein family has important roles in the regulation of E2F activity. Several studies have showed that E2F4 can form a complex with one of three RB proteins, and that the abundance of the E2F4-RB complex varies depending on the cell cycle state (8,15). We compared the overlap of our E2F4 targets with previously known RBL1 and RBL2 targets identified by ChIP-chip using a core-promoter array of 14 000 genes (9). We found that ∼87% of previously identified RBL1 and RBL2 targets were also bound by E2F4 in our ChIP-seq data.

Additionally, we found that 75% of genes previously reported as being bound RBL1 alone were in fact occupied by E2F4 in our dataset, and the majority of the remaining 25% of 'RBL1-alone' targets contained E2F4 sites in their promoters that were just below our 1% FDR threshold. This suggests that almost all RBL1/p107 and RBL2/p130 targets are in fact also occupied by E2F4 (Supplementary Table S9).

Motif analysis of E2F4 binding sites

E2F family proteins bind to DNA as a heterodimeric E2F-DP complex to the motif TTTc/gGCGCc/g (38). We examined the presence of the consensus E2F motif (TTTSSCGC) over all E2F4 binding sites identified by ChIP-seq. Interestingly, we found that only 5% of E2F4

sites contained the consensus motif, suggesting that E2F4 might be recruited to its sites either through a novel motif or via interaction with other proteins. In order to discover alternative E2F4 motifs, we performed a *de novo* motif search using the discovering rank imbalanced motifs (DRIM) algorithm (27). We first classified E2F4 sites into three different groups, namely strong, moderate and weak, based on binding strength. Next, we extracted the top 500 sites from each group and then executed DRIM on each set separately. We found a total of five different motifs that were significantly enriched over background ($P < 1 \times 10^{-5}$) (Figure 4A). Of these, Motifs 2 and 3 were similar to motifs recently identified using microarray based *in vitro* binding experiments for mouse E2F2 and E2F3 (39). To investigate motif occurrence and enrichment, and their dependence on binding strength, we
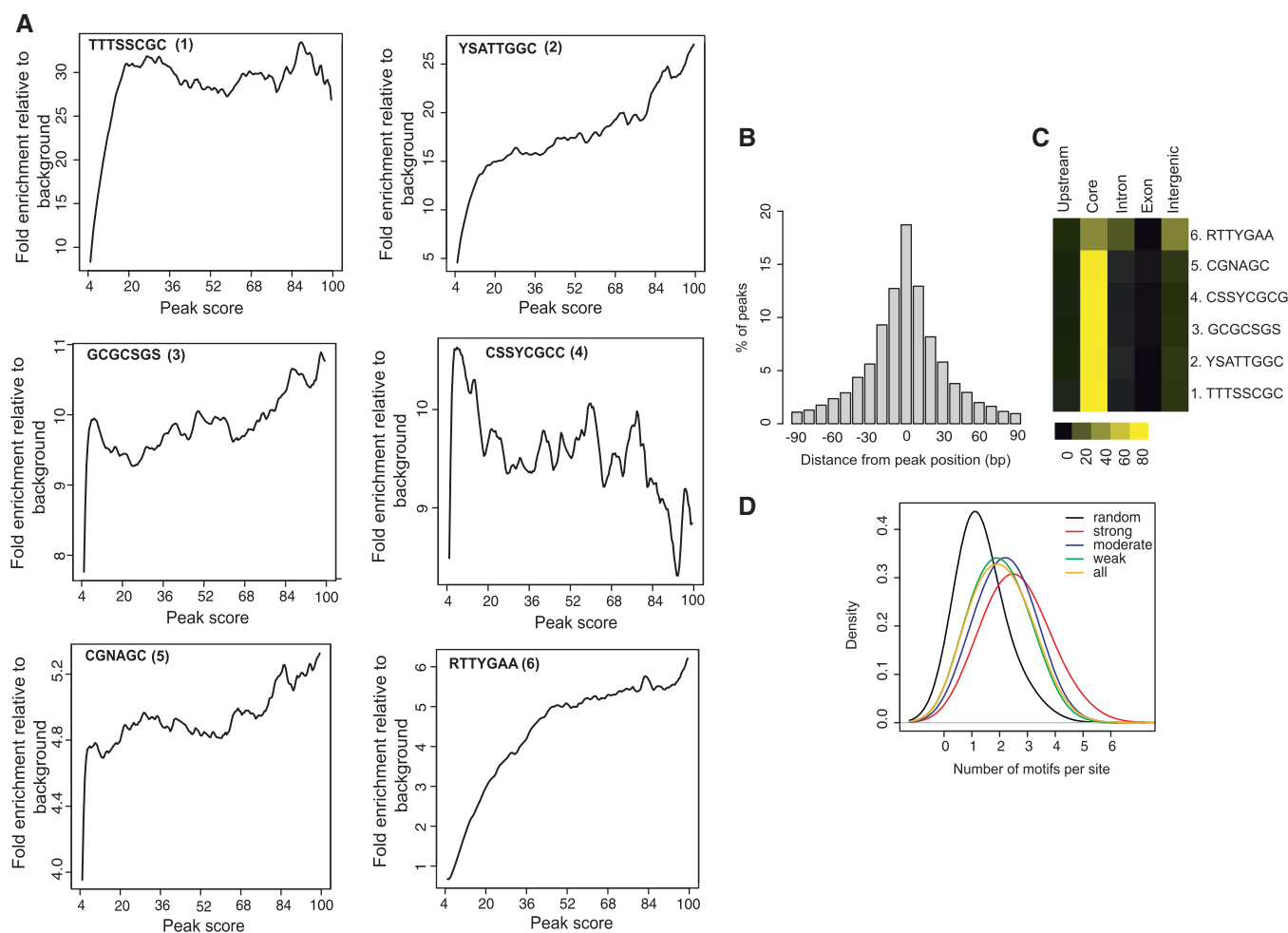

Figure 4. E2F4 motif analysis. (**A**) Enrichment of indicated motifs over background is plotted on the Y axis, as a function of ChIP-seq peak score plotted on the *x*-axis. (**B**) Distribution of motifs around E2F4 binding sites identified by ChIP-seq. E2F4 motifs were mapped to E2F4 binding sites and the distance of the identified motif from the maxima of the binding site was plotted as a histogram. The *y*-axis shows the percentage of peaks that had an E2F4 motif within the specified distance shown on the *x*-axis. The figure indicates that the majority of E2F4 peaks had an E2F4 motif within 20 bp of the indicated nucleotide that was designated as the binding site. (**C**) Frequency of motif occurrence in five different genomic regions. The heat-map shows the percentage distribution of E2F4 binding sites found in each genomic region for each of the six different E2F4 motifs used in this study. E2F4 motifs 1–5 were found predominantly in sites that mapped to the core promoter except motif 6. Motif 6 was found at almost equal frequency in sites that mapped to the core and intergenic regions. Color bar indicates percentage of a motif in a given genomic region. For a given motif, the sum of the percentages across all five different genomic regions is 100%. (**D**) Number of motifs discovered within E2F4 sites segregated by their ChIP-seq score. The density plot shows the relative frequency of sites on the *y*-axis containing each indicated number of motifs on the *x*-axis. Sites with stronger ChIP-seq scores had more motifs and overall, E2F4 sites had approximately two motifs per site on average.

mapped all six motifs (one canonical and five newly discovered) back to all E2F4 sites. Figure 4A represents the enrichment of each motif over background as a function of ChIP-seq peak score. The canonical motif (motif 1) and motif 2 showed the strongest enrichment over background indicating that these two motifs correspond to high occupancy E2F4 binding sites. Motifs 3 and 4 showed moderate enrichment, and motifs 5 and 6 showed weak enrichment over background. All motifs except motif 4 showed a gradual increase in the enrichment as well as in the percentage prevalence amongst binding sites as a function of peak score. Motif 4 hit the highest enrichment around score 10 and showed an apparent decrease of enrichment over background at higher ChIP-seq peak scores. This was mainly because only a small number of high-scoring E2F4 sites had this motif. Additionally, we found that for most binding sites, at least one of the six different E2F4 motifs was found <20 bp from our estimated peak position (Figure 4B).

To investigate whether different motifs are used to recruit E2F4 to different genomic regions, we examined the percentage occurrence of all six motifs at five different genomic regions (promoters, upstream, intergenic, introns and exons) to investigate any regional binding bias of each motif. Most motifs, with the exception of motif 6, were highly overrepresented in promoters, consistent with the occurrence and scores of peaks in these five genomic regions (Figure 2B and D). Motif 6 was distinct in that it showed comparable enrichment in promoters and intergenic regions as well as in introns, but not in upstream and exons (Figure 4C), suggesting that motif 6 may have distinct regulatory roles in intergenic and intronic regions. We also examined the number of motifs within each binding site. On average, each E2F4 site contained two motifs, while sites with higher scores contained more than two motifs (Figure 4D), suggesting the possibility of either multiple E2F4 DNA interactions per regulatory region, or usage of distinct motifs under different physiological conditions. In order to investigate whether specific binding motifs were associated with specific functions, we grouped genes based on the presence of specific motifs in their promoters and performed KEGG pathway analysis for each group (Table 2). All E2F4 motifs were used to regulate the cell cycle pathway and most motifs were used in several different pathways. However, we found that some pathways were significantly enriched

Table 2. Motif usage of E2F4 within different biological pathways

KEGG pathway terms	Motifs
Cell cycle	1, 2, 3, 4, 5, 6
Ubiquitin mediated proteolysis	3, 4, 5
Pyrimidine metabolism	1, 3, 5
DNA polymerase	1, 3, 5
p53 signaling pathway	3, 5
Chronic myeloid leukemia	4
N-Glycan biosynthesis	3
Biosynthesis of steroids	2

Each number indicates one of six E2F4 motifs, assigned to a KEGG pathway category.

with only one motif. For instance, motif 2 was overrepresented in the biosynthesis of steroids pathway while motif 3 was enriched in the *N*-glycan biosynthesis pathway, and motif 4 in the chronic myeloid leukemia pathway. Additionally, we found that motif 6 was associated exclusively with cell cycle genes. These results suggest that E2F4 may use distinct motifs to perform specific physiological functions. We found several other TF motifs to be significantly co-enriched within 500 bp of E2F4 binding sites (Supplementary Table S10). Among them, 10 TFs (EGR1-3, ELK1, PAX5, RFX1, SP1, STAT1, RFX1 and YY1) were also targets of E2F4 in our ChIP-seq data. Many cell cycle progression-related TFs such as AP1, MAZ, ELK1 were also highly enriched in the neighborhood of E2F4 binding sites and such TFs may regulate genes along with E2F4 in a combinatorial or competitive manner.

Overexpression of E2F4 and its cofactors reveal that E2F4 functions as an activator and a repressor

It has been reported that siRNA-mediated E2F4 knock-down leads to drug- or irradiation-induced apoptosis (3) while E2F4 knock-down in T98G cells does not affect gene expression due to its functional redundancy with E2F5 and E2F6 (9). In order to identify genes whose expression levels are affected by E2F4 and address whether E2F4 functions as an activator or a repressor, we perturbed E2F4 expression levels by transient overexpression and analyzed gene expression using microarrays. We initially tried overexpressing E2F4 in lymphoblastoid cells; however, the transfection efficiency was too low to discriminate overexpression effects given the background of untransfected cells. We therefore used HeLa cells for gene expression profiling. Before performing overexpression experiments, we compared our E2F4 binding targets from lymphoblastoid cells with targets from HeLa cells obtained from previously published data (16) to confirm that E2F4 binding profiles were comparable between the two cell lines. About 81% of the E2F4 targets from HeLa cells were also found in lymphoblastoid cells (Supplementary Figure S4), justifying the use of HeLa cells to assay the effects of E2F4 overexpression. Transient overexpression of E2F4 alone did not trigger dramatic expression changes, even though some E2F4 responsive genes were up- or down-regulated. This result is likely due to low levels of its cofactors, which are required for E2F4 localization and binding. We therefore performed expression profiling after co-transfecting E2F4 with its cofactors (DP-1 and RBL2). Co-transfection resulted in increased levels of mRNA and protein for E2F4 and its cofactors (Figure 5A and B). We performed at least two biological replicates of the expression arrays and used an error-model to identify statistically significant genes whose expression was altered in response to overexpression of the regulators (24). Compared to overexpression of E2F4 alone, co-transfection of E2F4 and its cofactors increased the number of targets that showed significant expression changes (Table 3 and Supplementary Table S11). Overall, combinatorial
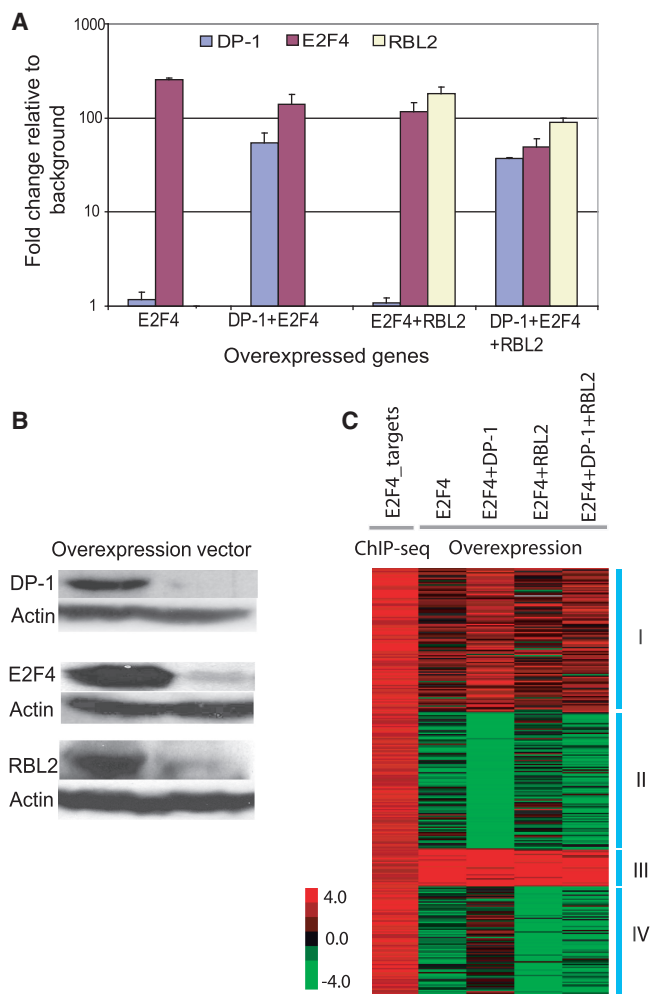
Figure 5. Overexpression of E2F4 and its cofactors (DP-1 and RBL2). (**A**) qPCR verification of increase in mRNA of E2F4 and its cofactors. GAPDH was used as an internal control and the log-scaled *y*-axis shows the fold increase of the indicated mRNA relative to the empty vector control. (**B**) Western blotting confirming overexpression of E2F4 and its cofactors at the protein level. Empty vector was used as a control. (**C**) K-means clustering of E2F4 targets identified by ChIP-seq along with gene expression data obtained in four different overexpression conditions. The data plotted is the expression value relative to that of a vehicle transfection control. The significance value (*X*) obtained from error model analysis was used for the clustering. A significance value of 3.3 corresponds to a *P*-value of 0.001. ChIP score was transformed to natural log.

overexpression of E2F4 and RBL2 or E2F4 and DP-1 caused the down-regulation of more E2F4 target genes than overexpression of E2F4 alone. However, co-expression of E2F4 and DP-1 also activated several genes. K-means clustering revealed four distinct clusters of genes whose expression was significantly altered: genes activated by overexpression of E2F4+DP-1 (cluster I), genes repressed by overexpression of E2F4+DP-1 (cluster II), genes activated by E2F4 alone (cluster III) and genes repressed by E2F4+RBL2 (cluster IV) (Figure 5C). Cell cycle genes were highly enriched in both clusters I and IV. Cluster I contained DNA replication and repair genes, while response to endogenous stimulus and programmed cell death were categories

Table 3. Number of up- or down-regulated genes after overexpression of E2F4 and its cofactors

(*P* < 0.001)	Number of expression-changed genes	
Overexpression	Up-regulated	Down-regulated
E2F4	167	128
E2F4 + DP-1	314	341
E2F4 + RBL	105	171
E2F4 + DP-1 + RBL2	228	281

P-value was calculated using an error model (24)

enriched in cluster III. These results indicate that E2F4 can function as either an activator or a repressor of transcription, and is involved in diverse physiological processes. More importantly, several genes whose expression is positively associated with cell cycle progression were activated by E2F4 and its cofactors overexpression (CDC6, CDCA5, CEP55, MYBL2, RPA1, SGOL2 and SMC3). Only a subset of E2F4 binding targets showed significant expression changes, which suggests that the regulation of E2F4 target genes may be more complex than currently perceived. Interestingly, even the low-scoring ChIP-seq binding targets were just as likely to be differentially expressed as the high-scoring ChIP-seq targets, and are therefore likely to be just as biologically meaningful (Supplementary Figure S5).

E2F4 can regulate microRNAs

miRNAs have been implicated in fine tuning gene expression by cleaving target mRNAs or inhibiting their translation, and some of them cooperate to regulate specific cellular events (40–42). The expression of miRNAs is modulated by TFs and reciprocally, TFs are also regulated by miRNAs (43–45). To address whether E2F4 potentially regulates miRNAs, we first compared E2F4 binding sites with predicted human miRNA promoters that were identified based on histone modification signatures (46) and found 41 putative miRNA targets of E2F4 (Supplementary Table S12). Since miRNA promoters are not well-defined, we also examined E2F4 binding sites located within 10 kb upstream of mature miRNA coding sequences. For this latter analysis, we excluded miRNAs present within exonic or intronic regions as it was not possible to assign E2F4 binding sites unambiguously to the miRNA or its parent gene. We thus identified an additional 161 miRNAs that showed E2F4 binding within 10 kb upstream of their coding regions (Supplementary Table S13). E2F4 showed strong binding to the putative promoters of the mir-17-92 cluster and let-7a, which are highly conserved miRNAs, as well as miR-22, an exonic miRNA (Figure 6A and Supplementary Figure S6). Quantitative ChIP-PCR confirmed that E2F4 was indeed recruited to these three miRNA promoters in lymphoblastoid cells (Figure 6B). To investigate whether E2F4, either by itself or in combination with its cofactors, could regulate
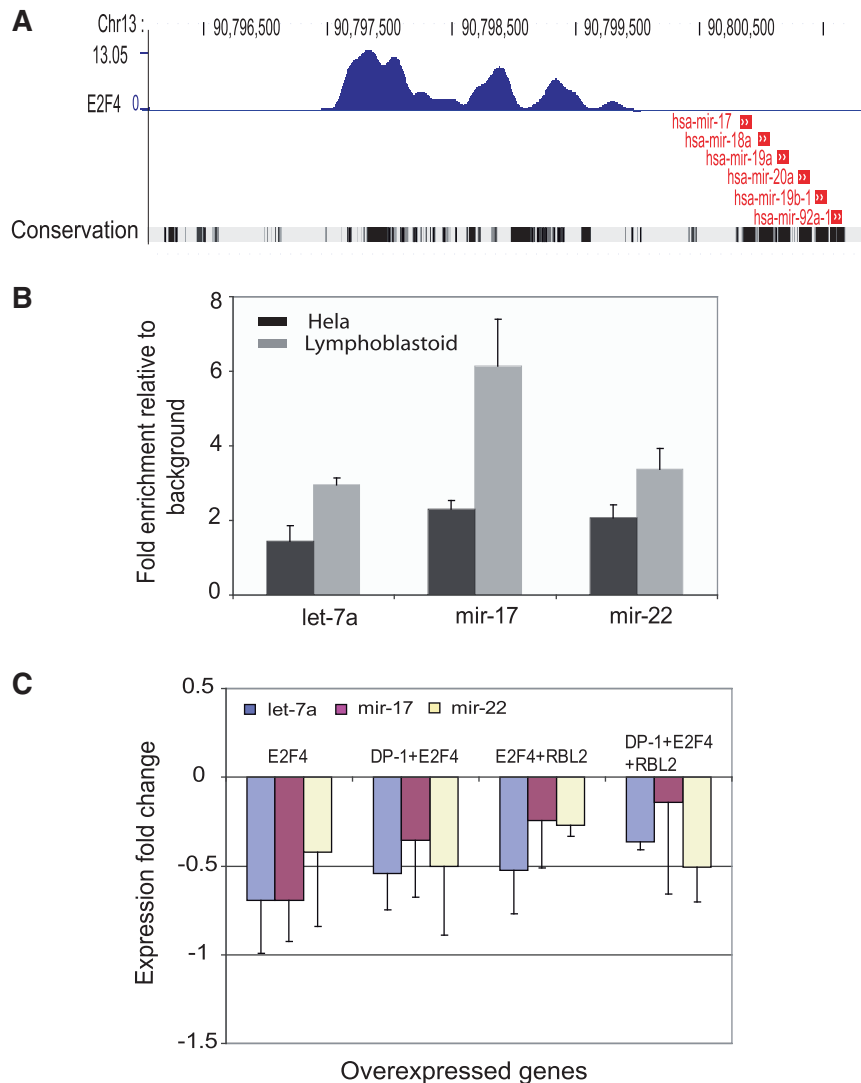
Figure 6. E2F4 can regulate miRNAs. (**A**) ChIP-seq data showing E2F4 binding within 10 kb upstream of the mir-17–92 cluster. The positions of the miRNAs are shown in red. The bottom track shows phylogenetic conservation across vertebrates species (Vertebrate Multiz Alignment & PhastCons Conservation: http://genome.ucsc.edu) with darker vertical bars indicating greater conservation. (**B**) qPCR verification of E2F4 binding sites upstream of indicated miRNAs in lymphoblastoid and HeLa cells. (**C**) TaqMan qPCR data for miRNA expression upon overexpression of E2F4 and its cofactors. Different combinations of E2F4 overexpression with its cofactors caused a modest decrease in the expression of all three miRNAs. The data plotted is the $\log_2$ of the expression relative to RNU66 which served as the internal control.

these miRNAs, we first established that E2F4 did bind to the promoters of these miRNAs in HeLa cells also (Figure 6B). We then used a quantitative TaqMan qPCR assay to measure expression changes of those three miRNAs in response to overexpression of E2F4 and its cofactors. All three miRNAs showed modest down-regulation upon overexpression of E2F4 alone or in combination with its cofactors (Figure 6C).

DISCUSSION

The 16 246 E2F4 binding sites in the human genome that we identified by ChIP-seq are consistent with the number estimated by extrapolation from the sites previously identified using tiling microarrays covering 1% of the

human genome (187 sites) (16). While this manuscript was in preparation, another study reported identification of ~15 000 E2F4 binding sites in K 562 erythroleukemia cells (47). Despite the different cell lines used, we noted that 54% of the binding sites we identified overlapped sites identified by this independent study.

About 56% of E2F4 sites in our study were found at promoters, and the average binding profile of E2F4 relative to a gene showed a preference of E2F4 to bind near the TSS. This finding also agrees with previously published E2F4 ChIP-chip data (16). Overall, our E2F4 binding site analysis suggests that E2F4 mainly regulates the expression of target genes by being recruited to their core promoters. However, our unbiased ChIP-seq approach revealed a significant proportion of distal E2F4 sites that have not been noted before. This implies

that in addition to regulating target genes by binding to the core promoter, E2F4 may be involved in additional modes of gene regulation. Many TF binding sites in eukaryotes occur far away from TSSs and these distal regulatory regions are believed to have important physiological roles (48,49). For instance, a TF can play diverse roles by interacting with different *cis*-regulatory elements such as enhancers, insulators, or silencers. It is reasonable to speculate that some distal E2F4 binding sites function as enhancers or silencers to modulate target gene expression. Half of the distal E2F4 sites showed histone modification signatures characteristic of enhancers, suggesting that E2F4 may act like an enhancer at specific loci. Based on luciferase reporter gene assays we confirmed that some of our distal E2F4 sites can function as enhancers. However, 5 out of 10 distal E2F4 binding sites did not show enhancer activity in the luciferase reporter assays even though those sites were highly enriched with enhancer marks of histone and p300 binding. This discrepancy may be in part because enhancers are cell-type specific, and the histone modification data were generated in a different cell type from the ChIP-seq data, and in part because histone modifications are not sufficient to fully specify enhancers. Nonetheless, our study suggests that in addition to a role at core promoters, E2F4 may act as a long-range regulator.

We found that E2F4 binding sites were highly enriched in bidirectional promoters, which is consistent with previously published data (31). Bidirectional promoters may be an efficient way to modulate gene expression where the same DNA element regulates two different downstream genes at the same time. Genome-wide studies of bidirectional promoters in several mammalian genomes have suggested that they are evolutionarily conserved and functionally related in certain categories like DNA repair (32,33). We also found that E2F4 can modulate most E2F family members, as well as its own cofactors. Comparing our E2F4 target genes with all known RBL1 and RBL2 targets revealed that almost all cofactor targets overlapped with E2F4 targets. A number of studies have shown that E2F promoter specificity is determined by its cofactors. For instance, E2F4-p130/RBL2 is a major complex in quiescent cells, whereas an E2F4-pRB or –p170/RBL1 complex is important in the G_1 phase of the cell cycle, which suggests distinct roles of cofactors in different cell cycle stages (50,51). The facts that E2F4 binds and may directly regulate its cofactors and family members suggest the possibility of feedback loops where the activity of E2F4 can be potentiated. Motif analysis revealed that the canonical E2F4 motif was present in only 5% of the 16 246 E2F4 binding sites. This result implies the possibility that E2F4 uses other unknown motifs or is recruited to target promoters by the aid of other cofactors. *De novo* motif analysis using DRIM discovered five putative novel motifs. All motifs were found to be positioned near the peak of the ChIP-seq signal. As a corollary, this suggests that the peak position identified by our algorithm from ChIP-seq read data denotes the actual binding site of the protein. This level of resolution has not been achieved

before in previous studies of E2F4 binding since they all used lower resolution tiling array approaches to identify binding sites.

Pathway analysis suggested that all E2F4 motifs were likely used to regulate the cell cycle pathway. Specifically, motif 1 (RTTYGAA) which was similar to a cell cycle repressor element, CHR (cell-cycle homology region; TT GAA) where E2F4/RB complexes were recruited (52,53), was highly enriched only among cell cycle pathway genes. Co-enrichment analysis identified several TFs that may co-regulate genes with E2F4. For example, many E2F4 targets such as E2F1, b-MYB and HSORC1 contain SP1 motifs near E2F binding sites (54); MYB and YY1 are known to be transcriptional partners of several E2F proteins (55–57); the constitutively expressed factor, NF-Y, binds to several cell cycle related E2F target promoters, and helps other regulatory proteins (PCAF and p300) gain access to target promoters to activate downstream genes (58).

E2F4 was classified as a repressor of cell proliferation because it binds to its target promoters involved in cell cycle progression in G_0/G_1 and represses them. Even though E2F4 was previously known as a repressor, some studies introduced the possibility that E2F4 may function as an activator by showing that it was able to trigger cell proliferation. In addition, overexpression of E2F4 in transgenic mice induced cell propagation in the basal layer of the epidermis (59,60). Our genome-wide identification of E2F4 binding targets and transient perturbation of E2F4 followed by gene expression profiling indicate that E2F4 can indeed function as either an activator or a repressor of transcription. In particular, cluster I, consisting of genes activated by E2F4 and DP-1, contains genes implicated in positive regulation of the cell cycle, suggesting that E2F4 may function as a cell cycle activator. Our data further revealed that E2F4 is capable of repressing the expression of several miRNAs such as the mir-17–92 cluster, mir-22, and let-7a, albeit by modest amounts. The mir-17–92 cluster, encoding six miRNAs, is known to be regulated by MYC, E2F1 and E2F3, and this regulation promotes cell proliferation (44,61,62). E2F4 may mediate its anti-proliferative role partly by repressing the mir-17-92 cluster. E2F4 is not only capable of repressing the expression of E2F1–E2F3, thus indirectly down-regulating the mir-17–92 cluster, but also binds to the mir-17–92 cluster promoter and directly regulates it, suggesting that it mediates a feedback loop for the regulation of the miR-17–92 cluster. Another miRNA target of E2F4, let-7a, is able to down-regulate the expression of MYC as well as trigger cell cycle arrest (45). Thus, E2F4 can not only regulate the expression of MYC directly (63), but also indirectly via let-7a, suggestive of another regulatory feedback loop. In summary, our genome-wide E2F4 target analysis reveals diverse functions of E2F4 and provides support for E2F4 functioning both as a long-range transcriptional regulators of mRNAs as well as a miRNA regulator, which allowed us to gain insights into understanding the versatile roles of this member of the E2F family.

REFERENCES

1. Attwooll,C., Lazzerini Denchi,E. and Helin,K. (2004) The E2F family: specific functions and overlapping interests. *EMBO J.*, **23**, 4709–4716.
2. Rowland,B.D. and Bernards,R. (2006) Re-evaluating cell-cycle regulation by E2Fs. *Cell*, **127**, 871–874.
3. Crosby,M.E. and Almasan,A. (2004) Opposing roles of E2Fs in cell proliferation and death. *Cancer Biol. Ther.*, **3**, 1208–1211.
4. Chong,J.L., Wenzel,P.L., Saenz-Robles,M.T., Nair,V., Ferrey,A., Hagan,J.P., Gomez,Y.M., Sharma,N., Chen,H.Z., Ouseph,M. *et al.* (2009) E2f1-3 switch from activators in progenitor cells to repressors in differentiating cells. *Nature*, **462**, 930–934.
5. Lindeman,G.J., Gaubatz,S., Livingston,D.M. and Ginsberg,D. (1997) The subcellular localization of E2F-4 is cell-cycle dependent. *Proc. Natl Acad. Sci. USA*, **94**, 5095–5100.
6. Yochum,G.S., Cleland,R., McWeeney,S. and Goodman,R.H. (2007) An antisense transcript induced by Wnt/beta-catenin signaling decreases E2F4. *J. Biol. Chem.*, **282**, 871–878.
7. Gaubatz,S., Lees,J.A., Lindeman,G.J. and Livingston,D.M. (2001) E2F4 is exported from the nucleus in a CRM1-dependent manner. *Mol. Cell. Biol.*, **21**, 1384–1392.
8. Moberg,K., Starz,M.A. and Lees,J.A. (1996) E2F-4 switches from p130 to p107 and pRB in response to cell cycle reentry. *Mol. Cell. Biol.*, **16**, 1436–1449.
9. Balciunaite,E., Spektor,A., Lents,N.H., Cam,H., Te Riele,H., Scime,A., Rudnicki,M.A., Young,R. and Dynlacht,B.D. (2005) Pocket protein complexes are recruited to distinct targets in quiescent and proliferating cells. *Mol. Cell. Biol.*, **25**, 8166–8178.
10. Roman,A. (2006) The human papillomavirus E7 protein shines a spotlight on the pRB family member, p130. *Cell Cycle*, **5**, 567–568.
11. Meloni,A.R., Smith,E.J. and Nevins,J.R. (1999) A mechanism for Rb/p130-mediated transcription repression involving recruitment of the CtBP corepressor. *Proc. Natl Acad. Sci. USA*, **96**, 9574–9579.
12. Souza,R.F., Yin,J., Smolinski,K.N., Zou,T.T., Wang,S., Shi,Y.Q., Rhyu,M.G., Cottrell,J., Abraham,J.M., Biden,K. *et al.* (1997) Frequent mutation of the E2F-4 cell cycle gene in primary human gastrointestinal tumors. *Cancer Res.*, **57**, 2350–2353.
13. Wang,D., Russell,J.L. and Johnson,D.G. (2000) E2F4 and E2F1 have similar proliferative properties but different apoptotic and oncogenic properties in vivo. *Mol. Cell. Biol.*, **20**, 3417–3424.
14. Schwemmle,S. and Pfeifer,G.P. (2000) Genomic structure and mutation screening of the E2F4 gene in human tumors. *Int. J. Cancer*, **86**, 672–677.
15. Cam,H., Balciunaite,E., Blais,A., Spektor,A., Scarpulla,R.C., Young,R., Kluger,Y. and Dynlacht,B.D. (2004) A common set of gene regulatory networks links metabolism and growth inhibition. *Mol. Cell*, **16**, 399–411.
16. Xu,X., Bieda,M., Jin,V.X., Rabinovich,A., Oberley,M.J., Green,R. and Farnham,P.J. (2007) A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. *Genome Res.*, **17**, 1550–1561.
17. 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.
18. Ji,H., Jiang,H., Ma,W., Johnson,D.S., Myers,R.M. and Wong,W.H. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.
19. 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.
20. Valouev,A., Johnson,D.S., Sundquist,A., Medina,C., Anton,E., Batzoglou,S., Myers,R.M. and Sidow,A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.
21. Kim,J., Lee,J.H. and Iyer,V.R. (2008) Global identification of Myc target genes reveals its direct role in mitochondrial biogenesis and its E-box usage in vivo. *PLoS ONE*, **3**, e1798.
22. Kim,Y.H. and Pollack,J.R. (2009) Comparative genomic hybridization on spotted oligonucleotide microarrays. *Methods Mol. Biol.*, **556**, 21–32.
23. Gu,J. and Iyer,V.R. (2006) PI3K signaling and miRNA expression during the response of quiescent human fibroblasts to distinct proliferative stimuli. *Genome Biol.*, **7**, R42.
24. Hu,Z., Killion,P.J. and Iyer,V.R. (2007) Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.*, **39**, 683–687.
25. Shivaswamy,S., Bhinge,A., Zhao,Y., Jones,S., Hirst,M. and Iyer,V.R. (2008) Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.*, **6**, e65.
26. Auerbach,R.K., Euskirchen,G., Rozowsky,J., Lamarre-Vincent,N., Moqtaderi,Z., Lefrancois,P., Struhl,K., Gerstein,M. and Snyder,M. (2009) Mapping accessible chromatin regions using Sono-Seq. *Proc. Natl Acad. Sci. USA*, **106**, 14926–14931.
27. Eden,E., Lipson,D., Yogev,S. and Yakhini,Z. (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.*, **3**, e39.
28. Deschenes,C., Alvarez,L., Lizotte,M.E., Vezina,A. and Rivard,N. (2004) The nucleocytoplasmic shuttling of E2F4 is involved in the regulation of human intestinal epithelial cell proliferation and differentiation. *J. Cell. Physiol.*, **199**, 262–273.
29. Ren,B., Cam,H., Takahashi,Y., Volkert,T., Terragni,J., Young,R.A. and Dynlacht,B.D. (2002) E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev.*, **16**, 245–256.
30. Tabach,Y., Brosh,R., Buganim,Y., Reiner,A., Zuk,O., Yitzhaky,A., Koudritsky,M., Rotter,V. and Domany,E. (2007) Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site. *PLoS ONE*, **2**, e807.
31. Lin,J.M., Collins,P.J., Trinklein,N.D., Fu,Y., Xi,H., Myers,R.M. and Weng,Z. (2007) Transcription factor binding and modified histones in human bidirectional promoters. *Genome Res.*, **17**, 818–827.
32. Adachi,N. and Lieber,M.R. (2002) Bidirectional gene organization: a common architectural feature of the human genome. *Cell*, **109**, 807–809.
33. Trinklein,N.D., Aldred,S.F., Hartman,S.J., Schroeder,D.I., Otillar,R.P. and Myers,R.M. (2004) An abundance of bidirectional promoters in the human genome. *Genome Res.*, **14**, 62–66.
34. 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.
35. Wang,Z., Zang,C., Rosenfeld,J.A., Schones,D.E., Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Peng,W., Zhang,M.Q. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.

36. Visel,A., Blow,M.J., Li,Z., Zhang,T., Akiyama,J.A., Holt,A., Plajzer-Frick,I., Shoukry,M., Wright,C., Chen,F. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.

37. Dennis,G. Jr, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.

38. Zheng,N., Fraenkel,E., Pabo,C.O. and Pavletich,N.P. (1999) Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. *Genes Dev.*, **13**, 666–674.

39. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.

40. Croce,C.M. (2009) Causes and consequences of microRNA dysregulation in cancer. *Nature Rev. Genet.*, **10**, 704–714.

41. Stark,A., Brennecke,J., Bushati,N., Russell,R.B. and Cohen,S.M. (2005) Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell*, **123**, 1133–1146.

42. Sun,M., Hurst,L.D., Carmichael,G.G. and Chen,J. (2005) Evidence for a preferential targeting of 3′-UTRs by cis-encoded natural antisense transcripts. *Nucleic Acids Res.*, **33**, 5533–5543.

43. Lal,A., Navarro,F., Maher,C.A., Maliszewski,L.E., Yan,N., O'Day,E., Chowdhury,D., Dykxhoorn,D.M., Tsai,P., Hofmann,O. *et al.* (2009) miR-24 Inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle genes via binding to 'seedless' 3′UTR microRNA recognition elements. *Mol. Cell*, **35**, 610–625.

44. 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.

45. Sampson,V.B., Rong,N.H., Han,J., Yang,Q., Aris,V., Soteropoulos,P., Petrelli,N.J., Dunn,S.P. and Krueger,L.J. (2007) MicroRNA let-7a down-regulates MYC and reverts MYC-induced growth in Burkitt lymphoma cells. *Cancer Res.*, **67**, 9762–9770.

46. 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.

47. Blahnik,K.R., Dou,L., O'Geen,H., McPhillips,T., Xu,X., Cao,A.R., Iyengar,S., Nicolet,C.M., Ludascher,B., Korf,I. *et al.* (2010) Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. *Nucleic Acids Res.*, **38**, e13.

48. Kim,T.H., Abdullaev,Z.K., Smith,A.D., Ching,K.A., Loukinov,D.I., Green,R.D., Zhang,M.Q., Lobanenkov,V.V. and Ren,B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, **128**, 1231–1245.

49. Kimura-Yoshida,C., Kitajima,K., Oda-Ishii,I., Tian,E., Suzuki,M., Yamamoto,M., Suzuki,T., Kobayashi,M., Aizawa,S. and Matsuo,I. (2004) Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development*, **131**, 57–71.

50. Ikeda,M.A., Jakoi,L. and Nevins,J.R. (1996) A unique role for the Rb protein in controlling E2F accumulation during cell growth and differentiation. *Proc. Natl Acad. Sci. USA*, **93**, 3215–3220.

51. van der Sman,J., Thomas,N.S. and Lam,E.W. (1999) Modulation of E2F complexes during G0 to S phase transition in human primary B-lymphocytes. *J. Biol. Chem.*, **274**, 12009–12016.

52. Yang,J., Song,K., Krebs,T.L., Jackson,M.W. and Danielpour,D. (2008) Rb/E2F4 and Smad2/3 link survivin to TGF-beta-induced apoptosis and tumor progression. *Oncogene*, **27**, 5326–5338.

53. Zwicker,J., Lucibello,F.C., Wolfraim,L.A., Gross,C., Truss,M., Engeland,K. and Muller,R. (1995) Cell cycle regulation of the cyclin A, cdc25C and cdc2 genes is based on a common mechanism of transcriptional repression. *EMBO J.*, **14**, 4514–4522.

54. Li,J.M., Hu,P.P., Shen,X., Yu,Y. and Wang,X.F. (1997) E2F4-RB and E2F4-p107 complexes suppress gene expression by transforming growth factor beta through E2F binding sites. *Proc. Natl Acad. Sci. USA*, **94**, 4948–4953.

55. Giangrande,P.H., Zhu,W., Rempel,R.E., Laakso,N. and Nevins,J.R. (2004) Combinatorial gene control involving E2F and E Box family members. *EMBO J.*, **23**, 1336–1347.

56. Schlisio,S., Halperin,T., Vidal,M. and Nevins,J.R. (2002) Interaction of YY1 with E2Fs, mediated by RYBP, provides a mechanism for specificity of E2F function. *EMBO J.*, **21**, 5775–5786.

57. Zhu,W., Giangrande,P.H. and Nevins,J.R. (2004) E2Fs link the control of G1/S and G2/M transcription. *EMBO J.*, **23**, 4615–4626.

58. Caretti,G., Salsi,V., Vecchi,C., Imbriano,C. and Mantovani,R. (2003) Dynamic recruitment of NF-Y and histone acetyltransferases on cell-cycle promoters. *J. Biol. Chem.*, **278**, 30435–30440.

59. Lukas,J., Petersen,B.O., Holm,K., Bartek,J. and Helin,K. (1996) Deregulated expression of E2F family members induces S-phase entry and overcomes p16INK4A-mediated growth suppression. *Mol. Cell. Biol.*, **16**, 1047–1057.

60. Pierce,A.M., Gimenez-Conti,I.B., Schneider-Broussard,R., Martinez,L.A., Conti,C.J. and Johnson,D.G. (1998) Increased E2F1 activity induces skin tumors in mice heterozygous and nullizygous for p53. *Proc. Natl Acad. Sci. USA*, **95**, 8858–8863.

61. Pickering,M.T., Stadler,B.M. and Kowalik,T.F. (2009) miR-17 and miR-20a temper an E2F1-induced G1 checkpoint to regulate cell cycle progression. *Oncogene*, **28**, 140–145.

62. Woods,K., Thomson,J.M. and Hammond,S.M. (2007) Direct regulation of an oncogenic micro-RNA cluster by E2F transcription factors. *J. Biol. Chem.*, **282**, 2130–2134.

63. Chen,C.R., Kang,Y., Siegel,P.M. and Massague,J. (2002) E2F4/5 and p107 as Smad cofactors linking the TGFbeta receptor to c-myc repression. *Cell*, **110**, 19–32.

64. Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.