CCCTC-Binding Factor Confines the Distal Action of Estrogen Receptor

Chang S. Chan and Jun S. Song

The Simons Center for Systems Biology, Institute for Advanced Study, Princeton, New Jersey

Abstract

Distal enhancers have recently emerged as a common mode of gene regulation for several transcription factors, including estrogen and androgen receptors, the two key regulators of breast and prostate cancer major subtypes. Despite the rapid success in genome-wide annotation of estrogen receptor-α (ERα) binding sites in cell lines, the precise mechanism governing the gene-to-enhancer association is still unknown and no quantitative model that can predict the estrogen responsiveness of genes has been hitherto proposed. This article presents an integrative genomics approach to construct a predictive model that can explain more than 70% of estrogen-induced expression profiles. The proposed method combines a recent map of the insulator protein CCCTC-binding factor (CTCF) with previous ER location studies and expression profiling in the presence of the translation inhibitor cycloheximide, providing evidence that CTCF partitions the human genome into distinct ER-regulatory blocks. It is shown that estrogen-responsive genes with a decreased transcription level (down-regulated genes) have a markedly different relative distribution of ER binding sites compared with those with an increased transcription level (up-regulated genes). Finally, Bayesian belief networks are constructed to quantify the effects of ER-binding distance from genes as well as the insulating effects of CTCF on the estrogen responsiveness of genes. This work thus represents a stride toward understanding and predicting the distal activities of steroid hormone nuclear receptors. [Cancer Res 2008;68(21):9041–9]

Introduction

Several genome-wide studies have mapped the global binding locations of estrogen receptor-α (ER-α) in the breast cancer cell line MCF7 (1–3), shifting the traditional focus of study from proximal promoters to distal enhancers, the preponderance of which has now been shown also for other transcription factors such as androgen receptor and p63 (4, 5). Despite the success in constructing a genomic map of ER-binding sites, the mechanism by which ER regulates a certain subset of genes while not affecting others remains a profound mystery. In fact, the distal nature of ER activities renders the binding site information alone quite insufficient for predicting estrogen-responsive genes. For instance, our analysis of these data shows that only 24% of genes with ER binding in [−1 kb,1 kb] proximal promoters and only 15% of genes with ER in exons or introns are actually regulated by ER.

In addition to using chromatin immunoprecipitation (ChIP), a common approach for identifying genes regulated by a transcription factor involves time course expression profiling before and after overexpression or overactivation of the transcription factor, e.g., via estrogen induction in the case of ER (2, 6). This method, however, is complicated by the differential expression of numerous secondary targets that are not directly regulated by the transcription factor under investigation but rather by other transcriptionally induced transcription factors. Ideally, a combination of high-throughput expression and ChIP experiments can provide a powerful tool to discover functional transcription factor binding sites and a confident set of primary target genes by focusing on differentially expressed genes with transcription factor binding sites near the genes. Unfortunately, estrogen and androgen receptors pose formidable exceptions to this ideal scenario because they can bind quite far (>100 kb) from the genes that they regulate while having no effect on intermediate genes. Along this line, a crucial role has been ascribed to the so-called "pioneering factor" FoxA1, a forkhead protein that selectively localizes to H3K4me2-modified enhancer regions and remodels chromatin for subsequent ER and androgen receptor binding activities (1, 7). Around 50% of ER binding sites have nearby FoxA1, and FoxA1 may thus play an important role in establishing cell type–specific gene regulation by ER and androgen receptor (1, 2, 7). However, only a small subset (<40%) of estrogen-responsive genes can be explained by any given study to date. Understanding how the estrogen receptor selectively chooses genes that it regulates thus remains a daunting challenge.

A recent study attempted to circumvent these difficulties by using the translation inhibitor cycloheximide in time course expression profiling (6). In addition to eliminating secondary estrogen-target genes, cycloheximide can prevent potential repression of estrogen-induced primary transcription by secondary negative feedback loops and can also stabilize mRNAs by weakening the activities of RNases (8), allowing low-level transcripts to accumulate for detection. Despite these apparent advantages of using cycloheximide to identify the genes directly targeted by ER, the cytotoxicity of cycloheximide and the accompanying genetic perturbation introduced into cells may raise serious doubts against the validity of this approach. Our article revisits the use of cycloheximide in studying ER regulatory networks and provide strong support for the use of cycloheximide in this particular setting. We accomplish this task by integrating the data available from high-throughput ER ChIP and expression studies with those from another genome-wide assay of the so-called CCCTC-binding factor (CTCF; ref. 9).

CTCF is a nuclear protein that is ubiquitously expressed across cell types. It binds to diverse DNA sequences by combinatorial use of its 11-zinc finger DNA-binding domains. CTCF is essential and highly conserved from fruit fly to human, with its binding sites

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Requests for reprints: Jun S. Song, The Simons Center for Systems Biology, Institute for Advanced Study, Einstein Drive, Princeton, NJ 08540. Phone: 609-734-8038; Fax: 609-951-4438; E-mail: jssong@ias.edu.

doi:10.1158/0008-5472.CAN-08-2632

being highly conserved across vertebrates (10). It is the only major protein identified in vertebrates that is involved in the establishment of insulators that can block enhancer activities. CTCF has been shown to be involved in the regulation of gene imprinting, monoallelic gene expression, X chromosome inactivation, and escape from X-linked inactivation (11). The mechanism of insulator function, however, still remains unknown. The distribution of CTCF binding sites in the genome correlates with gene density, with ~46% of sites lying in intergenic regions, 20% near transcriptional start sites, 22% in introns, and 12% in exons (9). Regions that are depleted of CTCF binding sites often include clusters of related gene families and genes that are transcriptionally coregulated; on the other hand, regions that are enriched in CTCF binding sites tend to contain genes with tissue-specific multiple alternative promoters (9). The distribution of CTCF binding sites is thus consistent with its role as an insulator, and it remains to be shown in a genome-wide fashion whether CTCF can block the activity of distally located transcription factors such as estrogen receptor.

By combining the aforementioned data sets, we here report for the first time integrated computational models that can predict estrogen-responsive genes with 60% to 70% sensitivity and 70% to 80% specificity, demonstrating the insulator function of CTCF in a global scale. In particular, we show that CTCF may block transcriptional activation of genes by ER and that most estrogen-responsive genes do not have CTCF between them and nearby ER binding sites. We also construct Bayesian belief networks to quantify the distance effect of ER and FoxA1-binding sites as well as the insulating effect of CTCF on gene regulation.

Materials and Methods

ChIP-ChIP data analysis. ChIP-enriched sites from the microarray data set generated by Carroll and colleagues (2) were obtained at a cutoff of FDR 1% as described in ref. 7. We obtained 1,226 ER sites from ref. 3 by mapping the 1,234 sites reported in ref. (3) to the latest human genome National Center for Biotechnology Information (NCBI) Build 36, using University of California Santa Cruz Batch Coordinate Conversion tool with 95% identity. Overlapping regions from the two lists were combined to obtain a set of 6,285 distinct ER binding sites in MCF7.

Expression data normalization and analysis. The expression data from refs. 2 and 6 were normalized together using RMA (12) and using the latest probe mapping to the human genome NCBI Build 36 (13). The cutoffs for deciding down-regulated and up-regulated genes in the presence of cycloheximide were obtained by constructing a three-component mixture model of log fold-change distribution, as described in Results. Gaussian mixture models of log fold-change distribution were estimated using an expectation maximization algorithm.[1] The 5th and 95th quantiles of the middle distribution consisting of noise and weakly responsive genes were chosen as cutoffs. These cutoffs corresponded to the 2nd and 97th quantiles in the overall fold-change distribution of Bourdeau and colleagues. Equivalent cutoffs for selecting down-regulated and up-regulated genes of Carroll and colleagues were taken to be the 2nd and 97th quantiles of the distribution of the maximum statistic max(fold change at 3 h, fold change at 6 h).

Correlation of expression of estrogen-inducible genes within and across CTCF blocks. UGT1A1, UGT1A2, . . ., UGT1A10 comprise a gene family with alternative transcription start sites (TSS) and were excluded from "Within CTCF" comparisons to remove any potential bias. The data set of Wang and colleagues was obtained from Gene Expression Omnibus GSE2034.

[1] http://algorithmics.molgen.mpg.de/Software/PyMix/

Bayesian network analysis. For each gene, the distance D_{ER} from TSS to the nearest ER binding site in the genome was computed and discretized into 30 bins, and likewise for the distance D_{FoxA1} to the nearest FoxA1. Discretization was performed using the Data PreProcessor (14) with the option "Equal Frequency." The boundaries of D_{ER} bins were located at −1169719, −756064, −541760, −410268, −320005, −251895, −198610, −159728, −125557, −97298, −73124, −51441, −30947, −15326, −684, 11100, 27113, 44859, 67912, 91809, 119665, 152739, 196654, 247969, 312264, 405354, 542869, 737070, 1139259. Similarly, the boundaries of D_{FoxA1} bins were at −510165, −317673, −235483, −180471, −144864, −117117, −95399, −77092, −61167, −46743, −33846, −22280, −12887, −3749, 557, 6754, 15363, 26293, 36663, 49181, 62833, 79941, 99614, 123768, 151500, 188368, 240405, 333340, 516120. The FoxA1 sites at FDR 1% were obtained from ref. 7. BN PowerPredictor was used to select the four significant Bayesian belief networks shown in Fig. 5*A*, maximizing the area under the receiver operating characteristic curve (14) for training data sets. BN PowerPredictor estimates the conditional probabilities graphically encoded in Fig. 5*A* by simply counting the frequency of events in training data sets. We used an unbiased success rate of 0.5 as a prior for the Bernoulli variable CLASS. Posterior probability of CLASS is then computed for each test gene from these conditional probabilities and prior distribution.

For cross-validation, we formed a subset excluding the 509 up-regulated genes and trimmed the subset by discarding the top and bottom 5% of genes ranked by fold change in the presence of cycloheximide. The resulting trimmed set thus consisted of non–estrogen-responsive genes for testing the specificity of the networks. The trimmed set and the up-regulated genes were partitioned into five distinct equal-sized subsets with equal fold-change distributions for cross-validation.

Results

The effect of cycloheximide and identification of estrogen-responsive genes. Two sets of time course expression profiles in MCF7 cells are currently available: (*a*) the study of Carroll and colleagues at 0, 3, 6, and 12 hours after estrogen treatment of hormone-starved cells (2); (*b*) the study of Bourdeau and colleagues at 0 and 24 hours after estrogen treatment of hormone-starved cells with or without cycloheximide (6). Carroll and colleagues used a Welch *t* test *P* value cutoff of 0.001 to identify early (3 hours) and late (6 and 12 hours) estrogen-induced genes, in which the early up-regulated genes were interpreted as immediate targets of ER and the late up-regulated genes a mixture of primary targets that accumulate slowly and secondary targets that require other transcription factors induced by estrogen (2). It has been observed, however, that the *t* test can be problematic for analyzing differential expression with few replicates, because variance estimates can be unreliable (15). In contrast, ranking genes by fold change has been shown to correlate well with results from quantitative PCR validations and to yield a robust framework for comparing different experimental studies (15). Bourdeau and colleagues thus used a fold-change (24 hours versus 0 hour) method with a cutoff of 1.4 for up-regulated genes; we here take a similar approach and use a statistical model to provide further justification for their choice of fold-change cutoff.

To examine the effect of cycloheximide on estrogen-regulated genes, we normalized together the data from Carroll and colleagues and Bourdeau and colleagues (see Materials and Methods). Density and quantile-quantile plots of fold changes from the two data sets, shown in Fig. 1*A* and *B*, respectively, indicated that the data of Carroll and colleagues had roughly three times more differentially expressed genes than the data of Bourdeau and colleagues at a given fold-change cutoff in the same cell line MCF7, consistent with the fact that cycloheximide eliminates many secondary targets of ER from being differentially expressed after

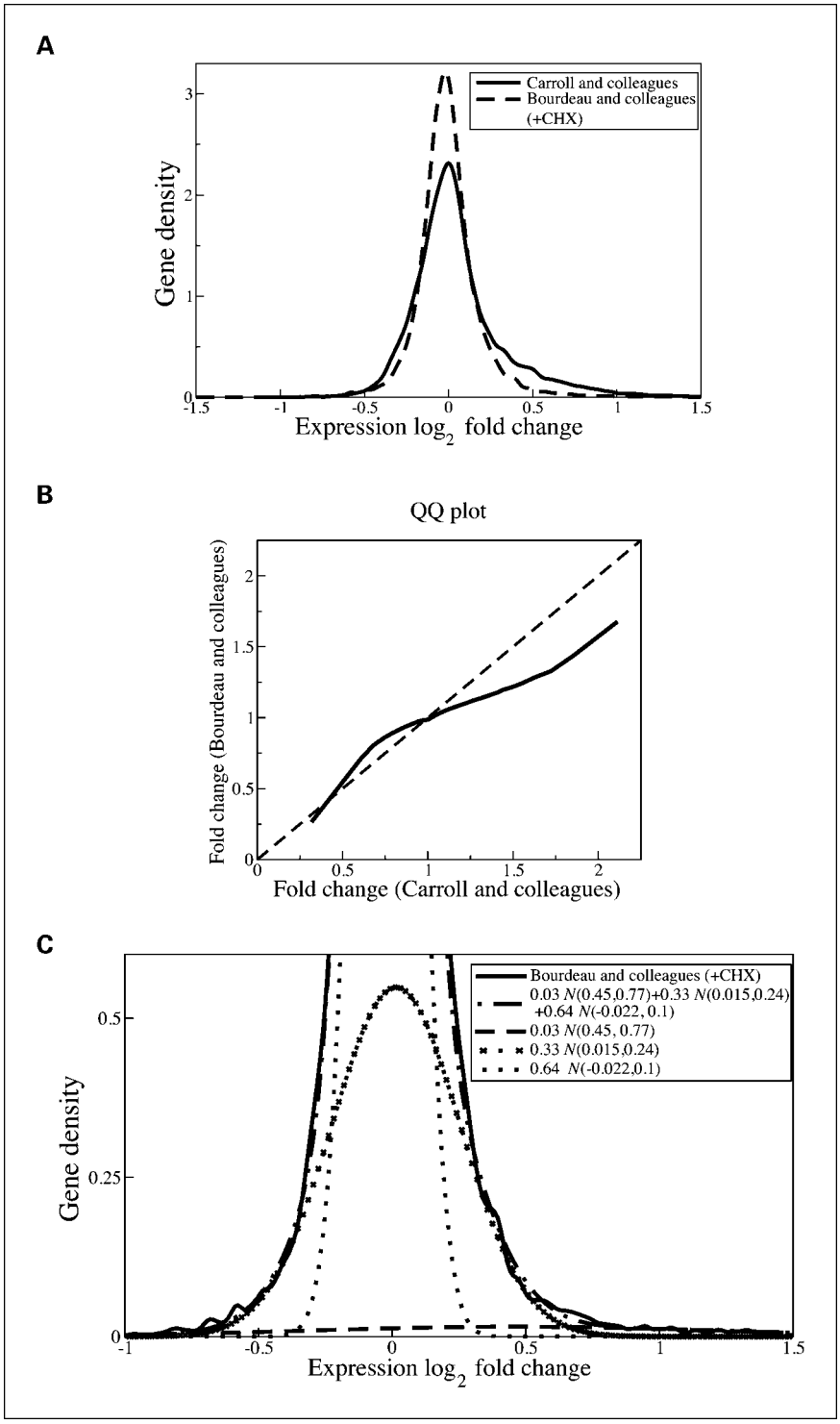

Figure 1. Distribution of estrogen-induced expression changes. *A,* maximum of 3 and 6 h fold changes from Carroll and colleagues and fold changes under cycloheximide treatment from Bourdeau and colleagues. Without cycloheximide (*CHX*), which filters out many secondary targets of ER, one overestimates the number of estrogen-responsive genes. *B,* quantile-quantile plot of the two distributions in *A*. Cycloheximide is also seen to dampen the strength of response to estrogen. *C,* a three-component mixture model can accurately describe the behavior of gene expression changes under estrogen induction.

estrogen treatment. The fold-change distribution of Bourdeau and colleagues has heavy tails and is not easily decomposed into estrogen-responsive and nonresponsive genes because of tight mixing between signal and noise. To find a statistical way of demarcating the subdistribution of up-regulated genes directly targeted by ER, we constructed a Gaussian mixture model of fold changes consisting of three components. We fitted the log fold-change distribution P using an expectation-maximization algorithm[1] as $P = 0.64\ N(-0.022,0.1) + 0.33\ N(0.015,0.24) + 0.03\ N(0.45, 0.77)$, where $N(\mu,\sigma)$ is normal distribution with mean μ and SD σ. Figure 1*C* shows that the model fits the original distribution well (Kolmogorov-Smirnov test $P > 0.13$). A simpler two-component model did not yield a good fit (Kolmogorov-Smirnov test $P = 2.7 \times 10^{-14}$; see Supplementary Fig. S1). We interpreted the component $N(-0.022,0.1)$ as describing pure noise, $N(0.015,0.24)$ as a mixture of noise and weakly responsive genes, and $N(0.45, 0.77)$

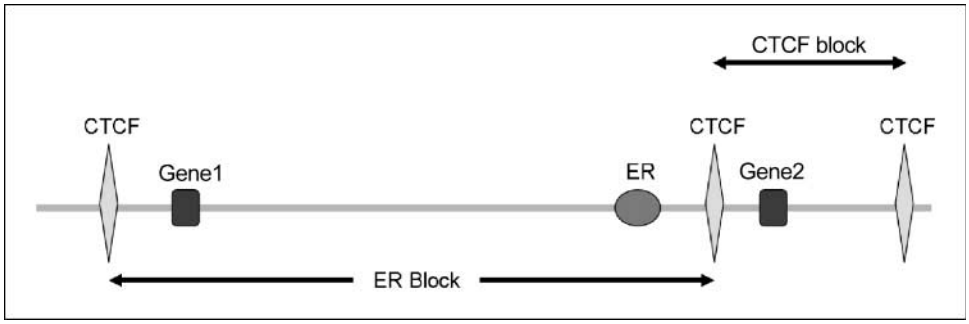

Figure 2. CTCF partitions the genome into distinct blocks. We call the genomic intervals delimited by CTCF binding sites CTCF blocks. A CTCF block may or may not contain an ER binding site, and when it does, we often refer to such a block as an ER block.

as strongly responsive genes. By taking the 5th quantile and 95th quantile of the distribution $N(0.015,0.24)$ for weakly responsive genes as cutoffs, we identified 509 up-regulated and 326 down-regulated genes, consistent with the estimates of 544 up-regulated and 235 down-regulated direct targets found by Bourdeau and colleagues using their 1.4 fold-change cutoff. We shall use the genes identified in this analysis as estrogen-responsive genes throughout this article. It should be noted that the fold-change method agreed well with significance analysis of microarrays (SAM; ref. 16), with 90% of our top 500 genes appearing in the list of top 500 genes found by SAM; fold changes and SAM q values also had a very high Pearson correlation of -0.93.

Partitioning the genome into CTCF blocks. A recent study has mapped the locations of 13,801 CTCF binding sites in IMR90 human fibroblasts (9). These sites partitioned the genome into distinct blocks of average size 213 kb and SD 601 kb, as exemplified in Fig. 2. Supplementary Fig. S2 summarizes the distribution of block sizes. RefSeq genes were grouped into 7,476 CTCF blocks of size 317 kb (SD 699 kb), whereas ER-containing blocks had a larger size of 555 kb (SD 1,061 kb), and the average number of RefSeq genes in a CTCF block was 3 ± 3. There was also a significant difference in size between RefSeq containing CTCF blocks with and without estrogen-responsive genes, 546 kb (SD 1,040 kb) and 290 kb (SD 645 kb), respectively ($P < 6.00 \times 10^{-11}$ using t test), consistent with the previous findings that ER regulation usually requires a long-range interaction between distal enhancers and promoters of regulated genes (1, 2). To incorporate the information about ER binding sites into our study, we combined the 5,782 and 1,226 ER binding sites from ChIP-Chip (2, 7) and ChIP-PET sequencing (3) experiments, respectively, to obtain 6,285 distinct ER binding sites in MCF7 (see Materials and Methods). There were in total 3,086 CTCF blocks containing ER binding sites; 25% of them had no genes and 61% had genes but no estrogen-responsive genes. Thus, only 14% of ER-containing CTCF blocks (ER blocks) had differentially expressed genes, posing a difficult computational problem of predicting estrogen-responsive genes from these ER and CTCF binding data sets. This problem will be addressed in this article.

The ER blocks showed no significant spatial clustering behavior, and 80% of the ER blocks did not have ER binding in immediately adjacent CTCF blocks (see Supplementary Table S1), indicating that CTCF groups ER binding sites into isolated units. When several ER blocks were clustered together in a row, up-regulated genes were evenly distributed across the ER blocks, with the exception that in a cluster containing precisely two adjacent ER blocks, up-regulated genes had a significant tendency to lie within only one of the two ER blocks ($P = 0.019$, binomial test; see Supplemental Table S2).

Up-regulated estrogen-responsive genes have ER binding sites within CTCF blocks. Three hundred forty-eight (68%) up-regulated genes have ER binding within the same CTCF blocks as the genes. To assess the significance of this phenomenon, we computed the percentage of genes with ER as a function of fold-change cutoff (see Fig. 3*A*). We simulated 6,285 random ER binding sites and computed the fraction of up-regulated estrogen-responsive genes (+cycloheximide) in CTCF blocks with ER, repeating the procedure 10,000 times. Figure 3*A* shows that a significant number of ER binding sites lie in the same CTCF blocks as up-regulated genes (P value $<10^{-4}$). It can also be seen that much less significant fraction of the early up-regulated genes identified by Carroll and colleagues have ER binding in CTCF blocks, thus supporting the use of cycloheximide treatment in filtering out genes not directly targeted by ER. Interestingly, because highly estrogen-responsive genes tend to have ER within 20 kb, whereas CTCF blocks are much larger, simulating random CTCF sites did not significantly change the fraction of genes having ER within CTCF blocks, indicating that Fig. 3*A* may be capturing the distance effect of ER rather than CTCF itself (see Supplementary Fig. S3*A*). Supplementary Fig. S3*B* shows the corresponding analysis for ER and FoxA1 overlapping sites. Similar graphs are obtained when one uses fold-change ranks instead of fold changes (data not shown).

In contrast, down-regulated genes have a markedly different behavior and the distribution of ER was not biased toward down-regulated genes compared with random distributions (see Supplementary Fig. S4), suggesting that the mode of ER-mediated down-regulation differs from that of up-regulation.

CTCF can act as an insulator of ER regulation. It has been shown that CTCF can act as boundary elements that group together coregulated genes into regulatory blocks (9, 10). To see whether CTCF also plays a similar role in ER-mediated gene regulation, we examined the correlation of expression of estrogen-inducible genes across cell lines in various expression profiling studies. Figure 3*B* shows the boxplots of pair-wise Pearson correlation across NCI60 cell lines (17) between up-regulated genes within same CTCF blocks and between up-regulated genes located within 100 kb but in different CTCF blocks. The correlation of genes within the same CTCF blocks was significantly higher than that of genes located within 100 kb but in different CTCF blocks ($P = 6.87 \times 10^{-9}$ for the data of Bourdeau and colleagues in the presence of cycloheximide, $P = 9.261 \times 10^{-11}$ for the data of Carroll and colleagues), consistent with the idea that the genes within CTCF blocks were coregulated by ER. Although it may be true that nearby genes can be coregulated, the observed high correlation was not because the genes were closer to each other within blocks than between blocks; in fact, Supplementary Fig. S5 shows that the

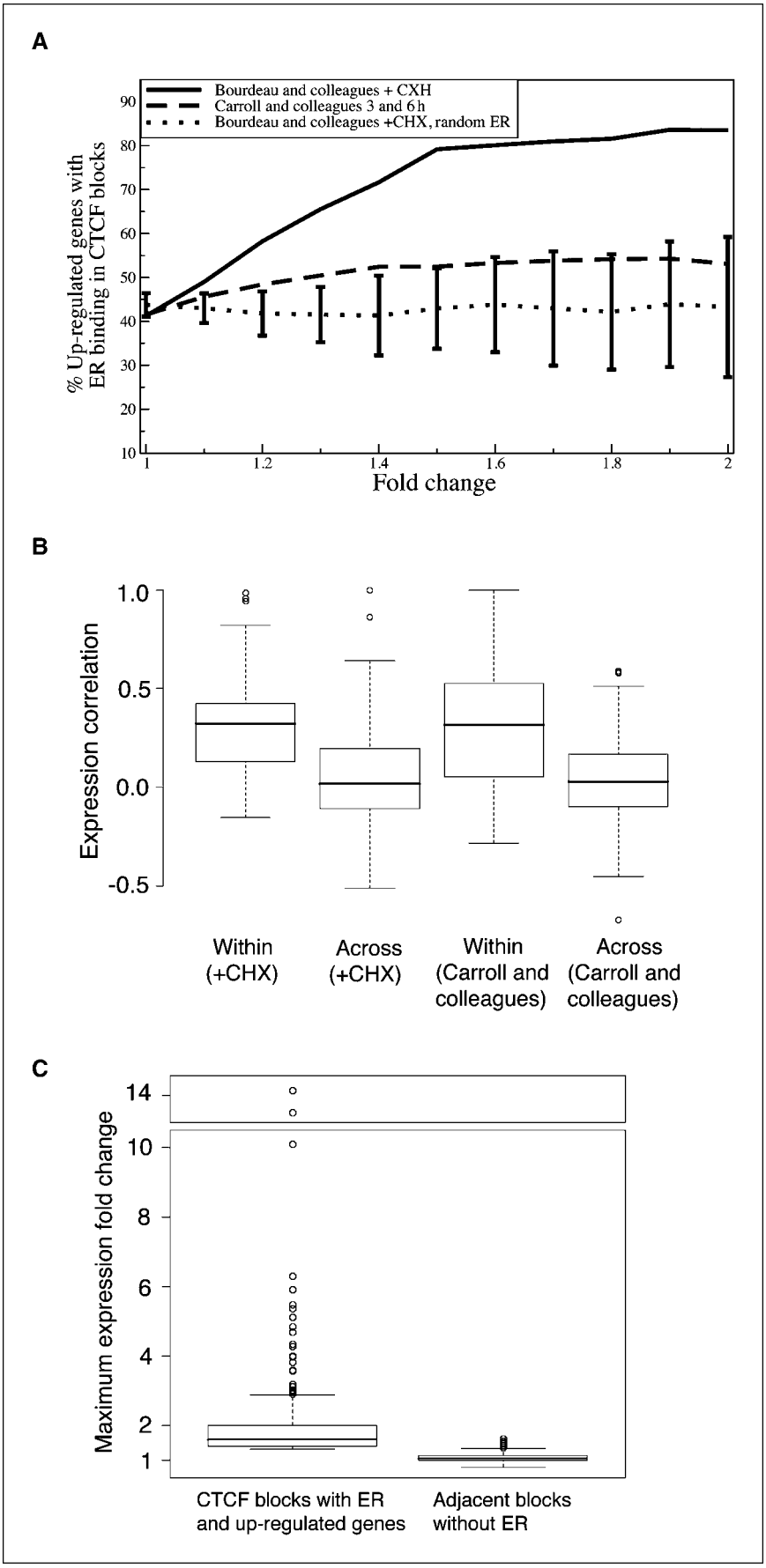

Figure 3. CTCF acts as an insulator. *A,* fractions of up-regulated genes found by Bourdeau and colleagues and Carroll and colleagues having ER binding within the same CTCF blocks as the genes. A significant proportion of up-regulated genes in the presence of cycloheximide has ER binding sites. *Dotted line,* mean proportion of the up-regulated genes of Bourdeau and colleagues from simulating random ER binding sites 10,000 times; *error bars,* maximum deviation from the mean. *B,* box plots of pair-wise Pearson correlation of estrogen-responsive genes across NCI60 cell lines, using pairs of estrogen-responsive genes within same CTCF blocks and those located within 100 kb but in different CTCF blocks. Estrogen-inducible genes within a given CTCF block are highly correlated and are thus likely to be coregulated by ER. *C,* for each CTCF block containing both an up-regulated gene and ER, we considered the immediately adjacent blocks without ER. The box plot of maximum fold changes in the blocks examined strongly indicates that CTCF can insulate adjacent blocks from the ER activation of genes.

opposite was true for the gene pairs in our analysis, as we had restricted the maximum distance between genes across CTCF blocks to be 100 kb. A significant difference in correlation was also observed across 286 primary ER+ and ER− breast cancer samples (ref. 18; see Supplementary Fig. S6*A*). To find further support for the insulator effect of CTCF on ER regulation, we randomized the CTCF boundaries by taking the midpoints of original blocks as new boundary locations, unless a block contained more than one differentially expressed gene, in which case the midpoint between two randomly chosen differentially expressed gene TSSs was chosen. The expression levels of estrogen-regulated genes within the random CTCF blocks were not significantly more correlated across NCI60 cell lines than those in different blocks ($P = 0.04$; see Supplementary Fig. S6*B*), indicating that the high correlation of estrogen-responsive genes observed in Fig. 3*B* depended on the structure of CTCF boundaries.

To further verify that CTCF can indeed act as an insulator of ER, we checked that the genes in CTCF blocks without ER (non–ER blocks) but immediately adjacent to ER blocks with up-regulated genes have significantly low fold changes ($P = 1.3 \times 10^{-16}$, using t test; see Fig. 3*C*). Supplementary Fig. S8 shows that randomizing CTCF sites yields more "leaky" ER activities across CTCF boundaries. In addition, of the 110 genes that lie in non–ER blocks but have a nearest ER binding site within 20 kb in an immediately adjacent CTCF block, only 3 were up-regulated ($P = 5.6 \times 10^{-4}$). ER activation of genes, therefore, is mostly confined to ER blocks and does not cross the boundaries into neighboring blocks. This observation, together with the fact that a significant number of up-regulated genes have ER binding sites within their CTCF blocks, supports that CTCF can indeed delimit the range of ER activities.

Distance effect of ER binding sites. Distance distribution of nearest ER binding sites from up-regulated genes in ER blocks had mean 85 kb (SD 242 kb), 10% trimmed mean 47 kb (SD 63 kb; see Materials and Methods). For down-regulated genes, the mean was 216 kb (SD 391 kb), 10% trimmed mean 151 kb (SD 118 kb), comparable with nonresponsive genes (mean 245 kb and SD 590 kb, 10% trimmed mean 148 kb and SD 180 kb). As can be seen in Fig. 4*A*, ER was located significantly closer to the TSSs of up-regulated genes than those of down-regulated or nonresponsive genes ($P < 10^{-14}$, Wilcoxon rank sum test), whereas there was no significant difference between the nearest ER locations for down-regulated and nonresponsive genes ($P > 0.7$, Wilcoxon rank sum

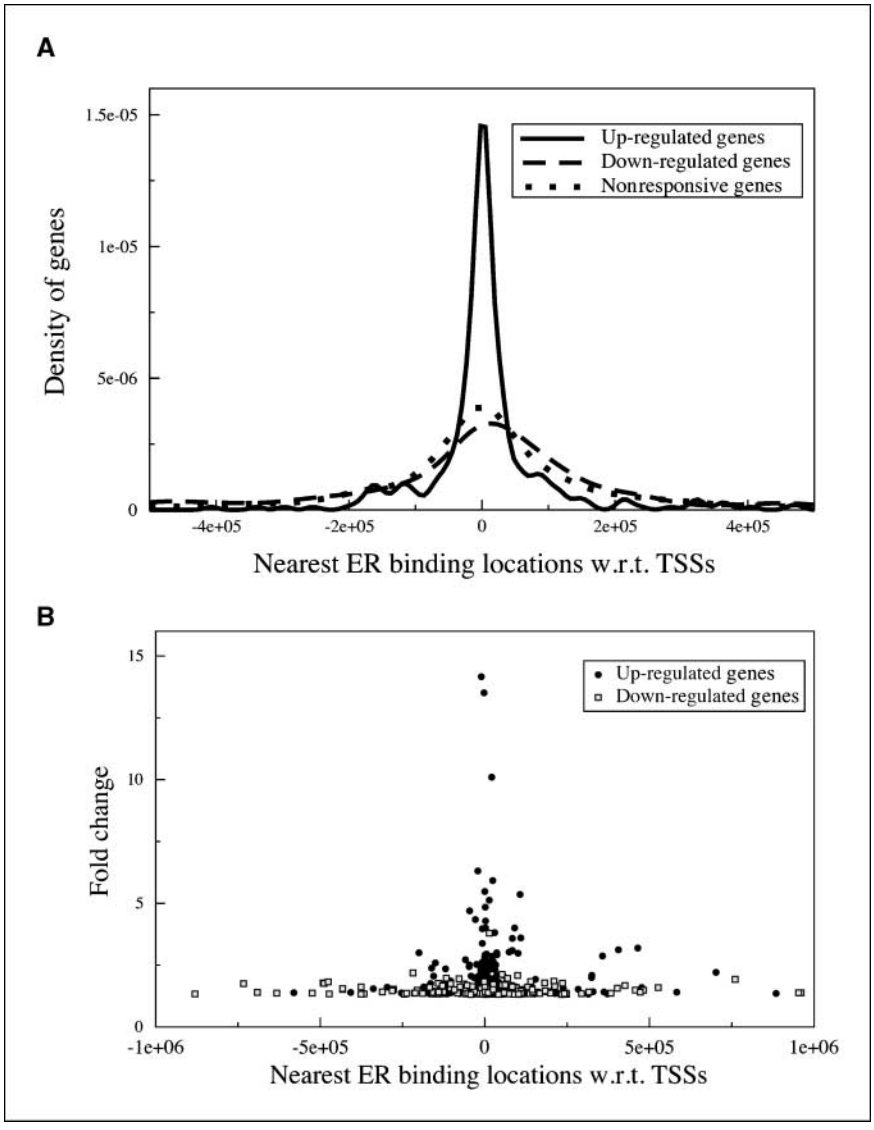

Figure 4. The effects of the nearest ER binding locations on gene expression in the presence of cycloheximide. *A,* a significant number of up-regulated genes have ER binding sites within 100 kb from TSSs, whereas down-regulated genes show no such preference. *B,* up-regulated genes with high fold changes tend to have ER binding sites within 20 kb from TSSs, but, at the same time, many genes with close ER binding sites have only modest response to estrogen and the overall correlation between the distance to nearest ER and expression change is weak. In comparison, down-regulated genes have a broader distribution of ER locations and a much weaker response to estrogen.

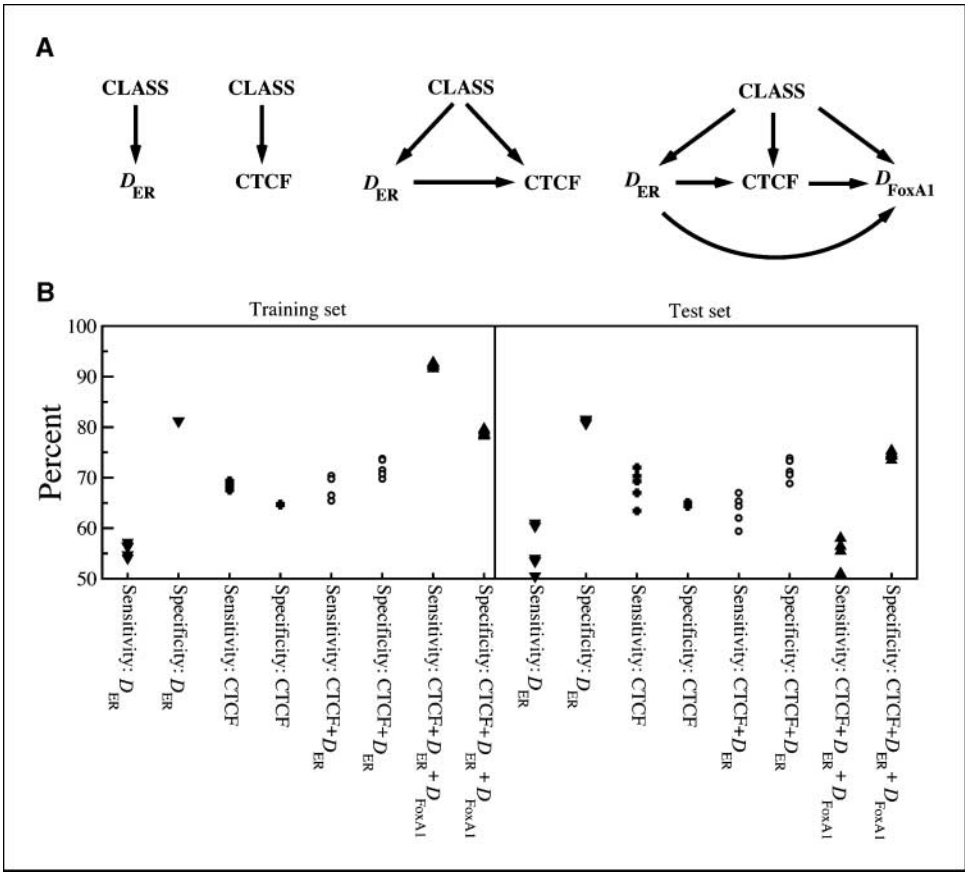

Figure 5. Bayesian network model of predicting estrogen-responsive genes. *A,* four significant Bayesian network models after feature selection. For each gene, we have CLASS = responsive/nonresponsive, D_{ER} = distance to nearest ER anywhere in the genome, CTCF = YES if there is ER within the same CTCF block as the gene, and D_{FoxA1} = distance to nearest FoxA1 anywhere in the genome. *B,* 5-fold cross-validation of the models. Including only D_{ER} has high specificity but low sensitivity because it classifies estrogen-responsive genes with distal ER enhancers as nonresponsive. The third model that also incorporates the information about having ER within the CTCF block of a gene seems to be a good compromise between sensitivity and specificity.

test). Although ER binding sites lie close to up-regulated genes within CTCF blocks and most highly up-regulated genes have ER binding within 20 kb, there was no significant global correlation between the distance of ER binding sites and fold change among up-regulated genes (see Fig. 4*B*). Pearson correlation between fold change and $\log(|D|)$ was -0.09 with a P value of 0.097.

Down-regulated genes have a broad distribution of ER binding locations with respect to TSS. They also have low fold changes, possibly because of the low levels of the corepressors AP1 and NRIP1 (2, 19), which are transcriptionally regulated by estrogen but cannot be translated in the presence of cycloheximide. Indeed, the fold change of down-regulated genes under cycloheximide treatment was ~30% less than that of down-regulated genes found by Carroll and colleagues.

Bayesian networks for predicting estrogen-responsive genes. Although 44% of the genes directly targeted by ER lie within 20 kb of ER binding sites, the precise effect of the distance between ER/FoxA1 binding sites and TSSs has not been hitherto quantified. We here used Bayesian networks to model the distance effect of ER/FoxA1 binding sites on differential expression of genes. A Bayesian network is a probabilistic graphical model that allows computation of joint probabilities in terms of a limited number of conditional probabilities, with arrows encoding the conditional dependence of the head node on the tail node. Bayesian networks have been previously applied to successfully predict 73% to 79% of the gene expression patterns in *Saccharomyces cerevisiae* (20). After feature selection, four models shown in Fig. 5*A* were analyzed using BN PowerPredictor (14) and the performance of each model was tested using 5-fold cross-validation (see Materials and Methods).

The first network in Fig. 5*A* involving only the distance D_{ER} to the nearest ER in the genome yielded 56% sensitivity and 81% specificity for predicting estrogen-responsive genes. Roughly, this model predicted a gene to be estrogen regulated if it had an ER binding site within 50 kb of its TSS. The second model that included only the information about whether a gene had an ER binding site within its CTCF block increased the sensitivity to 68% but lowered the specificity to 65%. The third model combining the information about ER in CTCF block and distance to the nearest ER had 66% sensitivity and 72% specificity (68% sensitivity on training sets). Incorporating the distance to the nearest FoxA1 site in the genome (the fourth model in Fig. 5*A*) yielded 92% sensitivity and 79% specificity on training sets, but only 55% sensitivity and 74% specificity on test sets, indicating that this model was overfitted. We have also analyzed the model containing only the information about nearest FoxA1 sites; however, it had only 50% sensitivity and 66% specificity on test sets. Overall, the third model gave the best compromise between sensitivity and specificity and also had the greatest area under the receiver operating characteristic curve (Supplementary Fig. S9).

We have examined other factors that could potentially correlate with estrogen responsiveness, but no significant influence was found: There was a slight correlation between the number of ER binding sites within CTCF blocks and expression fold change (Supplementary Fig. S10*A*). Although the ER blocks with an up-regulated gene had significantly more ER binding sites than those without an up-regulated gene (Supplementary Fig. S10*B*), the distance-normalized number of ER binding sites was similar between CTCF blocks with and without differentially expressed

genes. There was also no correlation between the number of ER motifs in nearest ER binding sites and fold change (correlation = 0.01 and $P = 8.517 \times 10^{-1}$). We also checked that there was no orientation bias in differentially expressed genes within CTCF blocks.

Discussion

Cycloheximide eliminates spurious differentially expressed genes. Cycloheximide is a potent cytotoxic agent that can introduce significant stress into cells, including DNA damage and cell cycle arrest, and can therefore cause profound changes in the transcriptional program of affected cells in response to the stress. To separate cycloheximide-induced expression changes from estrogen-induced ones, it is thus important to subject both estrogen-treated and control cells to cycloheximide and compare their differential expression profiles, as was done in ref. (6). Furthermore, the response of cells to estrogen under the perturbations of cycloheximide may not accurately reflect the normal response of estrogen-inducible genes. Despite these potential complications, this article supports the utility of cycloheximide in facilitating the discovery of direct target genes of ER. We have shown that estrogen-responsive genes in the presence of cycloheximide are correlated within their CTCF blocks and that up-regulated genes also have preferential ER binding activities. Without cycloheximide, secondary targets of ER make the distance effect of ER binding difficult to interpret and the distribution of ER becomes no more specific to estrogen-responsive genes than a random distribution (Fig. 3*A*). We have also provided further justification for the choices of fold-change cutoffs used in ref. (6) and have shown that the majority of differentially expressed genes according to these cutoffs can be correctly classified by our Bayesian network models.

Unlike the up-regulated genes, we observed that down-regulated genes had a broad distribution of ER binding sites and had low fold changes. The weak response of down-regulated genes is consistent with the fact that cycloheximide prevents the translation of AP1 and NRIP1, which are transcriptionally activated by ER and which have been shown to play critical roles in ER-mediated repression of genes (2, 19). Our study thus supports the hypothesis that ER repression of genes involves secondary cofactors that are themselves induced by estrogen.

CTCF in MCF7 versus IMR90. To date, a genome-wide map of CTCF binding sites in MCF7 is not available. It has been shown, however, that the CTCF binding sites are mostly invariant across cell types. For example, a comparison of 44 genomic regions representing 1% of the human genome (ENCODE regions) in the hematopoietic progenitor cell line U937 and primary human fibroblasts IMR90 yielded an agreement of 67% at a strict confidence level of $P < 10^{-6}$; the overlap increased at lower cutoffs (9). CTCF sites have also been mapped in $CD4^+$ T cells using ChIP-seq (21) and a genome-wide comparison between the sites and those in IMR90 had a 71% overlap. We thus believe that the majority of the CTCF binding sites found in IMR90 will also be present in MCF7 and, at the same time, that accounting for the cell type–specific CTCF sites will only increase the power of our analysis.

CTCF blocks form regulatory units for understanding and predicting ER activities. For the past 3 years, high-throughput sequencing and ChIP-ChIP studies have revealed that ER can bind in distal enhancers to regulate its targeted genes (1–3). Although it was observed that highly regulated genes tended to have ER binding within 20 kb (2, 7), various distance cutoffs were chosen somewhat arbitrarily for analyses and no quantitative model has yet been formulated to capture the precise distance effect of ER binding locations. Here, we trained a Bayesian belief network to discover that just based on the information about nearest ER locations, our model classified genes that have nearest ER within 50 kb as estrogen responsive. As expected, this naïve approach capturing only the ER distance effect yielded a high level of 44% false negatives.

We have presented in this article several lines of evidence that CTCF can demarcate the range of ER activities, grouping coregulated estrogen-responsive genes into distinct blocks. As can be seen in Fig. 5*B*, incorporating the information about whether genes reside in ER-containing CTCF blocks increased the performance of our Bayesian network. Considering conserved motifs in promoters further improved the sensitivity and specificity by ~3% (data not shown), supporting the hypothesis that the specificity of estrogen-induced genes is in part determined by the presence of particular cofactors in promoters. Interestingly, Bourdeau and colleagues could not find c-Myc as enriched in up-regulated genes, although c-Myc has been shown to localize to estrogen-responsive promoters and interact with ER (22). c-Myc is an important oncoprotein with a diverse transcriptional program in cell proliferation, growth, and apoptosis, interacting not only with ER but with other cofactors (23). The broad distribution of c-Myc and its non–ER-specific activities throughout the genome thus may make it difficult to find c-Myc as an enriched motif in estrogen-responsive genes. Our analysis shows that CTCF also helps in this regard; moreover, by considering the CTCF blocks with and without ER separately, we were able to detect c-Myc as a potential cofactor of ER in up-regulated promoters.

Many aspects of ER regulation of genes still remain unclear. For instance, about 30% of up-regulated genes did not have ER binding sites in their CTCF blocks. It could be that ER was actually present, but the specific epitopes recognized by antibodies were masked by interacting proteins such as FoxA1. Furthermore, there are more than 6,000 ER binding sites in the genome but only ~500 direct target genes, and more than 6,000 non–estrogen-responsive genes (35% of RefSeq genes studied here) had ER binding within their CTCF blocks. Therefore, we do not yet understand the functions, if any, of the majority of ER binding regions. It is possible that the specificity of ER activation of genes can be influenced by local chromatin structure or other epigenetic states, such as the recently found estrogen-induced dimethylation of histone 3 arginine 17 in the E2F1 promoter by CARM1 (24). Future studies in these directions will greatly facilitate our understanding of how ER functions in normal development and cancer.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

Received 7/9/2008; revised 8/6/2008; accepted 8/22/2008.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank H. Girgis, A.J Levine, and I. Tagkopoulos for helpful comments on the manuscript; M. Brown, M. Lupien, and C. Meyer for discussions; and Y.S. Song for bringing the work of Bourdeau and colleagues to our attention.

References

1. Carroll JS, Liu XS, Brodsky AS, et al. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell 2005;122:33–43.
2. Carroll JS, Meyer CA, Song J, et al. Genome-wide analysis of estrogen receptor binding sites. Nat Genet 2006;38:1289–97.
3. Lin C-Y, Vega VB, Thomsen JS, et al. Whole-genome cartography of estrogen receptor α binding sites. PLoS Genet 2007;3:e87.
4. Wang Q, Li W, Liu XS, et al. A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth. Mol Cell 2007;27:380–92.
5. Yang A, Zhu Z, Kapranov P, et al. Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. Mol Cell 2006;24:593–602.
6. Bourdeau V, Deschênes J, Laperrière D, Aid M, White JH, Mader S. Mechanisms of primary and secondary estrogen target gene regulation in breast cancer cells. Nucleic Acids Res 2008;36:76–93.
7. Lupien M, Eeckhoute J, Meyer CA, et al. FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell 2008;132:958–70.
8. Lopez PJ, Marchand I, Yarchuk O, Dreyfus M. Translation inhibitors stabilize *Escherichia coli* mRNAs independently of ribosome protection. Proc Natl Acad Sci U S A 1998;95:6067–72.
9. Kim TH, Abdullaev ZK, Smith AD, et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 2007;128:1231–45.
10. Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proc Natl Acad Sci U S A 2007;104:7145–50.
11. Brasset E, Vaury C. Insulators are fundamental components of the eukaryotic genomes. Heredity 2005;94:571–6.
12. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249–64.
13. Dai M, Wang P, Boyd AD, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res 2005;33:e175.
14. Cheng J, Greiner R. Learning Bayesian belief network classifiers: algorithms and system. Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence: Springer-Verlag; 2001. p. 141-51.
15. Shi L, Reid LH, Jones WD, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 2006;24:1151–61.
16. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98:5116–21.
17. Shankavaram UT, Reinhold WC, Nishizuka S, et al. Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. Mol Cancer Ther 2007;6:820–32.
18. Wang Y, Klijn JGM, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005;365:671–9.
19. Lee CH, Chinpaisal C, Wei LN. Cloning and characterization of mouse RIP140, a corepressor for nuclear orphan receptor TR2. Mol Cell Biol 1998;18:6745–55.
20. Beer MA, Tavazoie S. Predicting gene expression from sequence. Cell 2004;117:185–98.
21. Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. Cell 2007;129:823–37.
22. Cheng ASL, Jin VX, Fan M, et al. Combinatorial analysis of transcription factor partners reveals recruitment of c-MYC to estrogen receptor-α responsive promoters. Mol Cell 2006;21:393–404.
23. Sakamuro D, Prendergast GC. New Myc-interacting proteins: a second Myc network emerges. Oncogene 1999;18:2942–54.
24. Frietze S, Lupien M, Silver PA, Brown M. CARM1 regulates estrogen-stimulated breast cancer growth through up-regulation of E2F1. Cancer Res 2008;68:301–6.