Nucleic Acids Research Advance Access published February 16, 2012

Nucleic Acids Research, 2012, 1–11
doi:10.1093/nar/gks139

Integrated analysis identifies a class of androgen-responsive genes regulated by short combinatorial long-range mechanism facilitated by CTCF

Cenny Taslim[1,2], Zhong Chen[3], Kun Huang[4], Tim Hui-Ming Huang[2], Qianben Wang[3,*] and Shili Lin[1,*]

[1]Department of Statistics, [2]Department of Molecular Virology, Immunology & Medical Genetics, [3]Department of Molecular and Cellular Biochemistry and the Comprehensive Cancer Center and [4]Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA

Received July 6, 2011; Revised January 18, 2012; Accepted January 21, 2012

ABSTRACT

Recently, much attention has been given to elucidate how long-range gene regulation comes into play and how histone modifications and distal transcription factor binding contribute toward this mechanism. Androgen receptor (AR), a key regulator of prostate cancer, has been shown to regulate its target genes via distal enhancers, leading to the hypothesis of global long-range gene regulation. However, despite numerous flows of newly generated data, the precise mechanism with respect to AR-mediated long-range gene regulation is still largely unknown. In this study, we carried out an integrated analysis combining several types of high-throughput data, including genome-wide distribution data of H3K4 di-methylation (H3K4me2), CCCTC binding factor (CTCF), AR and FoxA1 cistrome data as well as androgen-regulated gene expression data. We found that a subset of androgen-responsive genes was significantly enriched near AR/H3K4me2 overlapping regions and FoxA1 binding sites within the same CTCF block. Importantly, genes in this class were enriched in cancer-related pathways and were downregulated in clinical metastatic versus localized prostate cancer. Our results suggest a relatively short combinatorial long-range regulation mechanism facilitated by CTCF blocking. Under such a mechanism, H3K4me2, AR and FoxA1 within the same CTCF block combinatorially regulate a subset of distally located androgen-responsive genes involved in prostate carcinogenesis.

INTRODUCTION

Androgens, functioning through the androgen receptor (AR), are not only essential in sexual development of males but also drive the onset and subsequent progression of prostate cancer. As a result, androgen ablation has been used as an effective first-line therapy for treatment of advanced androgen-dependent prostate cancer (ADPC). However, ADPC will ultimately progress to an AR-dependent, castration-resistant stage (CRPC) (1–3). Despite the importance of AR in both ADPC and CRPC, the precise mechanism by which AR regulates AR-dependent genes remains unknown.

Recent developments of chromatin immunoprecipitation (ChIP) techniques followed by high-throughput screening (e.g. ChIP-chip and ChIP-seq) have enabled studies to map AR cistrome in prostate cancer. An important finding from these genome-wide studies is that the majority of AR binding sites is located within non-promoter regions such as enhancers. For instance, in both ADPC (LNCaP and VCaP) and CRPC (LNCaP-abl and C4-2B) cell models, over 85% of the AR-binding sites are distal from AR-dependent genes (4–6). These studies reinforce the concept that steroid receptors [e.g. estrogen receptor (ER) and AR] are distal binding factors (7,8). Consistent with the hypothesis that chromatin looping is a mechanism for interactions between distal enhancers and proximal promoters (9), recent studies using chromatin conformation capture (3C) assay have demonstrated that AR directly upregulates two canonical androgen-responsive genes *PSA* and *TMPRSS2* (5,8,10,11). However, the efficiency and the underlying mechanisms for AR-mediated global long-range regulation are not fully understood.

*To whom correspondence should be addressed. Tel: +1 614 292 7404; Fax: +1 614 292 2096; Email: shili@stat.ohio-state.edu
Correspondence may also be addressed to Qianben Wang. Tel: +1 614 247 1609; Fax: +1 614 688 4181; Email: Qianben.Wang@osumc.edu

In addition to distal binding and long-range regulation, another important finding from analysis of AR cistrome and gene expression data is that AR and its cooperating transcription factor FoxA1 contribute to a subset of AR target gene expression in a combinatorial fashion (5,12–15). Furthermore, comparing AR and FoxA1 cistromes with genome-wide enriched regions of histone H3 lysine 4 di-methylation (H3K4me2), an active histone mark present on both enhancer and promoter regions (16,17), revealed a significant overlapping between H3K4me2 enriched regions and AR/FoxA1 binding regions, suggesting that AR/FoxA1 binding regions are associated with a permissive chromatin structure (18). Despite these interesting findings, it is unknown how H3K4me2, AR and FoxA1 combinatorially contribute to AR-mediated global long-range regulation.

CCCTC-binding factor (CTCF) is a highly conserved transcription factor that plays important roles in transcriptional activation or repression, enhancer blocking (EB), genetic imprinting and X chromosome inactivation (19,20). Consistent with this notion, Chan and Song (21) found that CTCF acts as an EB insulator for ER, preventing a remote ER enhancer from activating a promoter when it is located between these regulatory elements. Recently, CTCF is proposed to function as an organizer of global chromatin structure, facilitating its own regulation of essential biological processes such as transcription and insulation (19,22). Therefore, CTCF may also be involved in AR-driven long-range gene regulation.

Here, we performed an integrated analysis using four sets of genome-wide protein-DNA binding data (i.e. AR, FoxA1 and CTCF cistrome data as well as H3K4me2 ChIP-seq data) to elucidate the regulatory mechanism of androgen responsive genes in prostate cancer. We selected to use non-promoter H3K4me2 instead of H3K4me1 as a representative of enhancers in our analysis due to the fact that there is a higher fraction of overlapping between me2 with FoxA1 and AR (18). Among the androgen-responsive genes is a class of genes that are more highly expressed compared to the rest. Functional data analysis shows that genes in this class are significantly enriched in cancer related pathways. Further, these genes are located in CTCF blocks in which there are also AR/FoxA1 binding sites that overlap with those of H3K4me2. This suggests that AR-mediated target gene expression requires CTCF insulated, combinatorial regulation of H3K4me2, AR and FoxA1. We hypothesize that CTCF blocking leads to a relatively short but efficient long-range combinatorial regulation mechanism. Our study provides evidence on the importance of interplay between chromatin mark, chromatin organizer and specific transcription factors, which will likely help in unraveling the complex regulatory mechanism of AR.

MATERIALS AND METHODS

Data sources

ChIP-seq data of CTCF-bound genomic sites from HeLa, Jurkat and $CD4^+$ T cells are obtained from (23,24). Sites shared at least one base pair (bp) among the different cells are considered to be common. ChIP-chip data of CTCF bound genomic sites from LNCaP cells are obtained from ref. (25). Genes expression values in LNCaP cells after 4 h of dihydrotestosterone (DHT) (dihydrotestosterone) treatment are obtained from ref. (5). ChIP-chip experiment on AR and FoxA1 performed on LNCaP cell line are obtained from (5,12,26). High-throughput profiling of active histone modification marker in prostate cancer cells (H3K4me2 ChIP-seq data from LNCaP after 4 h of DHT) are obtained from (18). Large-scale gene expression profiles in 79 human tissues are obtained from (27). Expression of genes in LCM-capture epithelial cell populations representing prostate cancer progression from benign/normal epithelium to metastatic stage are obtained from (28).

Binding sites and responsive genes detection

Site Identification from Short Sequence Reads (SISSRs) is used to identify CTCF binding sites (29). The longest overlapping binding sites in all three cells (i.e. $CD4^+$ T, Jurkat and HeLa) are defined as the CTCF-bound region. Model-based Analysis of Tiling array (MAT) are used to obtain the FoxA1 and AR binding sites with FDR cut-off = 5% (30). Androgen-responsive genes are defined as genes that display differential expression after 4 h of DHT treatment with q-values (FDR) less than 0.05. All gene expression is normalized with RMA (31) and false discovery rate was calculated using Significance Analysis of Microarrays (SAM) algorithm (32). Note since gene expression values are obtained from microarray experiment (5), not all genes in the RefSeq database have gene expression values. Active histone marker (H3K4me2) enrichment sites were studied in prostate cancer cells treated with and without DHT treatment for 4 h. Peaks of H3K4me2 enrichment are determined using Model-based Analysis of ChIP-Seq (MACS) with P-value cut-off of 10^{-5} (33). Overlap between histone markers and transcription factors are defined as any overlap of at least 1 bp.

Statistical analysis

To study distinct mechanisms of AR regulation of androgen-responsive genes due to different interplay among transcriptional factors and active histone modification (H3K4me2), we performed supervised classification on the androgen responsive genes based on their expression levels and their relations to the factors including AR, CTCF, FoxA1 and H3K4me2. Specifically, these factors are used to build a logistic regression model: the effects of distance from genes to transcription factor binding sites [i.e. distances from transcription start sites (TSS) of androgen-responsive genes to AR and FoxA1 binding sites], the role of CTCF as a facilitator (whether AR and FoxA1 are in the same block) and their overlay with active histone mark (i.e. whether AR or FoxA1 binding sites overlap with H3K4me2 region of enrichment. Operationally, for each gene, the transformed distance AR_{Dist} from TSS to the nearest AR binding site and the transformed distance $FoxA1_{Dist}$ to the nearest FoxA1 binding site are computed to alleviate severe skewness. Transformed distance is obtained by subjecting distance

to the power of 1/5. Models with up to 6 factors including AR_{Dist}, $FoxA1_{Dist}$, *AR* (whether AR exist inside the block), *AR-H3K4me2* (whether AR overlap with H3K4me2), *FoxA1* (whether FoxA1 binding site present in the block) and *FoxA1-H3K4me2* (whether FoxA1 overlap with H3K4me2) are considered initially. We then selected the best model with the best specificity and sensitivity in a 5-fold cross validation. Bayes Factor (34) like criterion was used to classify androgen-responsive genes based on the final logic model.

To compare the sizes of different blocks bounded by CTCF, we used the Wilcoxon rank sum (i.e. the Mann–Whitney) test. We chose to use this non-parametric test instead of the T-test to minimize the effect of outlying observations since the distributions of the block sizes are extremely skewed. Likewise, for comparing expression of genes in different tissues, we used the paired Wilcoxon rank sum test.

RNA interference

siRNA targeting CTCF (siCTCF) or a control siRNA (siControl) were purchased from Dharmacon (ON TARGET plus siRNA), and were transfected into LNCaP cells using Lipofectamine 2000 (Invitrogen).

Real-time RT–PCR

Real-time RT–PCR was performed as before (10). Briefly, total RNA was isolated using an RNeasy kit (Qiagen, Valencia, CA, USA). cDNA was reverse transcribed from total RNA (1 mg) using a High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, Foster City, CA, USA). Real-time PCR was performed on the Applied Biosystem® TaqMan® Array Custom Plate containing 36 responsive genes in CTCF blocks with AR, 24 responsive genes in CTCF block without AR and 4 candidate endogenous control genes. TaqMan® Fast Universal PCR Master Mix was used on the StepOnePlus™ Real-Time PCR System (Applied Biosystems) following the manufacturer's instructions. All genes examined are listed in Supplementary Table S1.

ChIP

CTCF ChIP assays were performed as previously described (10,26). The anti-CTCF antibodies (07–729) were purchased from Millipore. Primer sequences are listed in Supplementary Table S2.

RESULTS

Common CTCF binding sites

Recent studies have mapped the locations of CTCF binding sites in CD4, Jurkat and HeLa cell lines (23,24), with 11 553 common sites between the three cells. We found that these shared CTCF binding sites are mostly located in intergenic regions (52%) but over 27% are in intronic regions and many are occupying promoter and exonic regions (Supplementary Figure S1). These CTCF sites are also evolutionarily conserved (Supplementary Method and Supplementary Figure S2), which is shown by the figure that conservation scores around the CTCF binding sites (center of plot) are twice as large as the scores denoting the genomic background (both ends of plot). Also, around 59% of these sites overlap with DNase I hypersensitive zones (see Supplementary Method), which might indicate that CTCF can partition the genome into physically distinct domains of gene expression since hypersensitive sites are found in relaxed chromatin in which the associated gene is active.

Genomic CTCF blocks

In order to study the long-range regulation of androgen-responsive genes involving AR, together with other factors such as CTCF, FoxA1 and H3K4me2, we first identified androgen-responsive genes using expression levels in LNCaP cell line after 4 h of DHT treatment [a more potent agonist for AR activation than testosterone (T)] compared to their basal expression (5). We identified a total of 388 androgen-responsive genes (315 upregulated and 73 downregulated) using q-value (FDR) cut-off of 0.05. Since we focus on androgen-responsive genes throughout this article, these genes are referred to as responsive genes (up/down-regulated) hereafter. We identified AR and FoxA1 binding sites from ChIP-chip experiments on LNCaP cell line (5,12,26). We found a total of 8663 AR binding sites and 20 826 FoxA1 binding sites. We also identified H3K4me2 enrichment region using ChIP-seq experiment on LNCaP cells (18) and found a total of 90 053 peaks.

Next, we divided the genome into distinct CTCF blocks to study the combinatorial relationship of AR-mediated gene expression with other factors within and across different blocks. Since these 11 553 common CTCF binding sites are evolutionary conserved [see Supplementary Figure S2 and Reference (20)], they are used to partition the human genome. These sites demarcate the genome into 11 576 blocks of varying sizes with a median of 81 kb and an IQR (interquartile range) of 198 kb. Here, a block is defined as a genomic region with two consecutive CTCF-bound sites as borders and include the beginning (ending) of each chromosome to (from) the first (last) binding sites on the chromosome (Figure 1A).

Of the 11 576 blocks, 4164 (36%) contain no RefSeq genes. The remaining 7412 blocks (64%) consist of at least one RefSeq genes, but 256 blocks have RefSeq genes that were not included in our gene expression arrays. This leaves us with 7156 blocks (62%) that contain at least one RefSeq gene that are on our microarray. Of these, a small number of blocks (365 or 5%) include at least one responsive gene (referred to as responsive blocks), while 6791 (95%) have only non-responsive genes (referred to as non-responsive blocks). Supplementary Table S3 provides a summary of these four types of blocks.

There are a total of 3669 blocks (32%) containing 8663 AR binding sites (referred to as AR blocks) indicating many blocks with multiple AR binding sites. For brevity, we refer to AR binding sites as AR sites. The remaining 7907 blocks (68%) contains no AR sites. Among the AR blocks, 296 (8% relative to AR blocks)

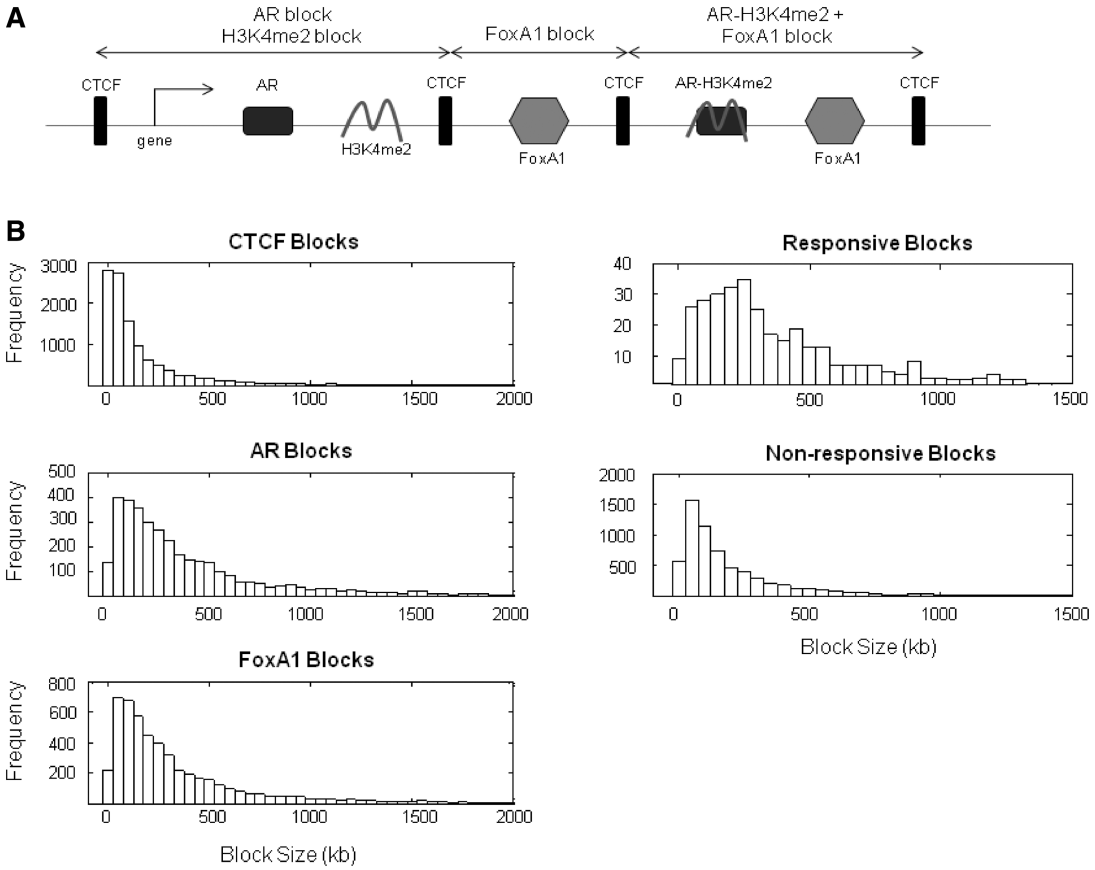

Figure 1. (**A**) Cartoon illustration of various blocks. Bar: CTCF binding sites; rectangle: AR binding site; hexagon: FoxA1 binding site; line: H3K4me2 enrichment site. A genomic region with CTCF binding sites as borders is defined as a block. The first block from the left containing a RefSeq gene, AR binding site and H3K4me2 enrichment is both an AR block and a H3K4me2 block. The second block is referred to as FoxA1 block since it contains only FoxA1 binding site. The last block is both an AR-H3K4me2 block (AR binding site overlap with H3K4me2 enrichment) and a FoxA1 block. (**B**) Distribution of length of blocks in kb. In addition to the overall distribution, the figure also shows the distribution of blocks containing AR binding sites (AR blocks), blocks containing FoxA1 (FoxA1 blocks), blocks containing at least one responsive genes (responsive blocks), and blocks consist of only non-responsive genes (non-responsive genes) in LNCaP cell line 4 h after being induced by androgen.

contain at least one responsive gene. These blocks are referred to as AR block with AR-responsive gene (AR–ARG), which consist a total of 316 responsive genes representing 81% of all responsive genes. The majority of these AR blocks have only one responsive gene but a few have multiple ones. About 2729 AR blocks (74%) have RefSeq genes that are all non-responsive and the remaining 644 AR blocks contain no RefSeq genes that were studied. Supplementary Table S4 provides a summary of these AR blocks.

In contrast, there are 5231 blocks (45%) containing 20 826 FoxA1 binding sites (referred to as FoxA1 blocks). For brevity, we refer to FoxA1 binding sites as FoxA1 sites. The remaining 6345 blocks (55%) contains no FoxA1 sites. Among the FoxA1 blocks, 334 (6%) contain at least one responsive gene with a total of 354 responsive genes representing 91% of all responsive genes. About 3804 FoxA1 blocks have RefSeq genes that are all non-responsive and the remaining 1093 FoxA1 blocks contain no RefSeq gene that were studied. Supplementary Table S5 provides a summary of these FoxA1 blocks.

Considering AR, FoxA1 and RefSeq genes jointly, there are 2620 blocks that contain both AR, and FoxA1 sites and at least one RefSeq genes which were on our microarray. Of these, 289 blocks (11%) consist of at least one responsive gene. The remaining 2331 blocks (89%) have genes that are all non-responsive. Supplementary Table S6 provides a summary on these and other types of combined blocks. As can be seen from Supplementary Tables S2 and S4, out of the 296 AR blocks with at least one responsive gene, 98% (289) of these also contain FoxA1 (See supplementary Figure S3 for intersection between AR, FoxA1, responsive genes and RefSeq genes blocks). This perhaps suggests that almost all of the responsive genes in AR blocks are co-regulated by FoxA1.

It is interesting to note that AR blocks with responsive genes (median = 396 kb and IQR = 553 kb) are significantly longer than AR blocks with only non-responsive genes (median = 302 kb and IQR = 480 kb) (Mann–Whitney test, P-value = 1.50×10^{-5}), which appears to support the theory that AR regulates genes via long-range interaction (8,35). As such, AR blocks with responsive genes are longer than AR blocks with genes that are not being regulated by AR. Figure 1B shows the distribution of different types of blocks.

CTCF's role as facilitator of AR-regulation of androgen-responsive genes

CTCF has been touted as an enhancer insulator in the literature. For example, Chan and Song (21) provided some evidence that CTCF can block distal action of ER. We set out to find whether CTCF has a similar role in relation to AR by trying to obtain answers to the following questions: Does AR potentially regulate genes in the same block? Can AR also regulate genes in nearby blocks? To this end, we compared the log fold change of the expression levels of all genes in 'AR-ARG blocks' to the log fold change of the expression of all genes in 'nearby-no-AR blocks' (i.e. blocks nearest to AR-ARG blocks but not containing AR themselves). The log fold change is computed from the expression levels of genes after 4-h of DHT stimulation versus vehicle control. To exclude genes that are too far to be regulated by AR, we filter out genes that are more than 100 kb away from any AR binding site. We have 1064 total genes in 296 'AR-ARG blocks' and 71 total genes in 58 'nearby-no-AR blocks'. Figure 2A provides an illustration of the genes in the two types of blocks being compared (i.e. all genes in 'AR-ARG blocks' versus all genes in 'nearby-no-AR blocks'). Figure 2B shows that the change of expression levels of all genes in 'AR-ARG blocks' is significantly higher than genes located in 'nearby-no-AR blocks' (Mann–Whitney test, P-value $= 1.42 \times 10^{-5}$). Figure 2C shows the individual gene expression fold change in 'upAR-ARG blocks' (red bar, 'AR blocks with up-regulated gene'), 'downAR-ARG blocks' (green bar, 'AR blocks with down-regulated gene') and in 'nearby-no-AR blocks' (black bar, 'blocks without AR nearest to AR-ARG blocks'). Although there are some responsive genes, in general genes in 'nearby-no-AR blocks' do not show differential expression (with black lines spread out through the entire range), whereas genes in 'upAR-ARG blocks' and in 'downAR-ARG blocks' mostly show high and low expression level, respectively. This indicates that distal AR may regulate genes within the same 'CTCF block'. Consistent with this notion, we have experimentally demonstrated that 4 distal AR binding sites located at −12 kb, −14 kb, −20 kb and −73 kb away from transcription start site (TSS) of *TMPRSS2* gene form chromatin loops with the TMPRSS2 promoter within the same CTCF block (10).

We have also observed responsive genes that do not have AR binding sites in the same block. However, we found the expression fold changes of the 316 responsive genes in blocks with AR are significantly higher than the 72 responsive genes in blocks without AR (Figure 2D, Mann–Whitney test, P-value $= 4.44 \times 10^{-4}$). To experimentally validate these findings and investigate whether CTCF binding affects the expression of responsive genes in blocks with AR and without AR, we have now examined expression of a subset of the responsive genes (60 genes) after CTCF silencing in LNCaP cells by using real-time RT–PCR. As expected, silencing of CTCF significantly decreased CTCF protein expression (Figure 2E) and CTCF binding at selected CTCF blocks with or without AR (Figure 2F). Consistent with gene expression microarray results (Figure 2D), expression fold changes of responsive genes in blocks with AR were significantly higher than responsive genes in blocks without AR in control silenced cells (Figure 2G, Mann–Whitney test, P-value $= 3.24 \times 10^{-9}$). Importantly, while silencing of CTCF does not lead to appreciable fold changes of the responsive genes in blocks without AR (Figure 2G, P-value $= 0.20$), knocking down of this transcription factor significantly decreased expression fold changes of responsive genes in blocks with AR (Figure 2G, P-value $= 1.61 \times 10^{-4}$), suggesting that CTCF facilitates AR regulation of target genes within the same block. Since AR binding sites and the TSSs of most responsive genes within the same CTCF blocks can be separated by more than 4 kb, this lead to the hypothesis of a relatively short but efficient long-range AR regulation mechanism facilitated by CTCF blocking.

Supervised classification of androgen-responsive genes further supports the hypothesis of short combinatorial long-range regulation

As shown in the previous section, not all responsive genes have AR binding sites in the same CTCF block. This leads to the hypothesis of two distinct mechanisms of AR regulation of responsive genes due to different interplay among transcriptional factors and active histone modification (H3K4me2). Indeed, analysis based on the logistic regression model selection and classification strategy as described in methods section reveals two distinct classes of responsive genes. Our results show that these classifications are determined by their distances to the AR and FoxA1 binding sites, whether the gene is in an 'AR block', and whether there is an overlap between AR and H3K4me2 enriched regions. Class 1 contains mainly responsive genes that are close to AR and FoxA1 binding sites (median distance from TSS to AR and FoxA1 binding sites are 12 and 7.6 kb, respectively; Table 1) and with most genes (94%) having AR/H3K4me2 overlapping region within the same CTCF block. On the other hand, Class 2 contains responsive genes that are farther away from AR and FoxA1 binding sites (median distance from TSS to AR and FoxA1 binding sites are 96 and 54 kb, respectively; Table 1) and only a small percentage of genes (24%) have AR/H3K4me2 overlapping region within the same block. These characterizations of the two classes are based on the logit model (left panel of Table 1). In addition, as shown in Table 1, almost all of the genes (93%) in Class 1 has an AR/H3K4me2 overlapping region and also FoxA1 binding sites in the same block compared to Class 2 that have only around 22% showing such overlap (right panel of Table 1).

The two classes indicate two distinct mechanisms of long-range regulation of androgen-responsive genes. In particular, genes in Class 1 correspond to 'short' combinatorial long-range regulation that we have hypothesized, whereas the long-range regulation mechanism of genes in Class 2 is less clear, although it is consistent with a long-range regulation mechanism that requires mediation from factor(s) that are located much farther away from the genes being regulated. To elaborate on the effect of

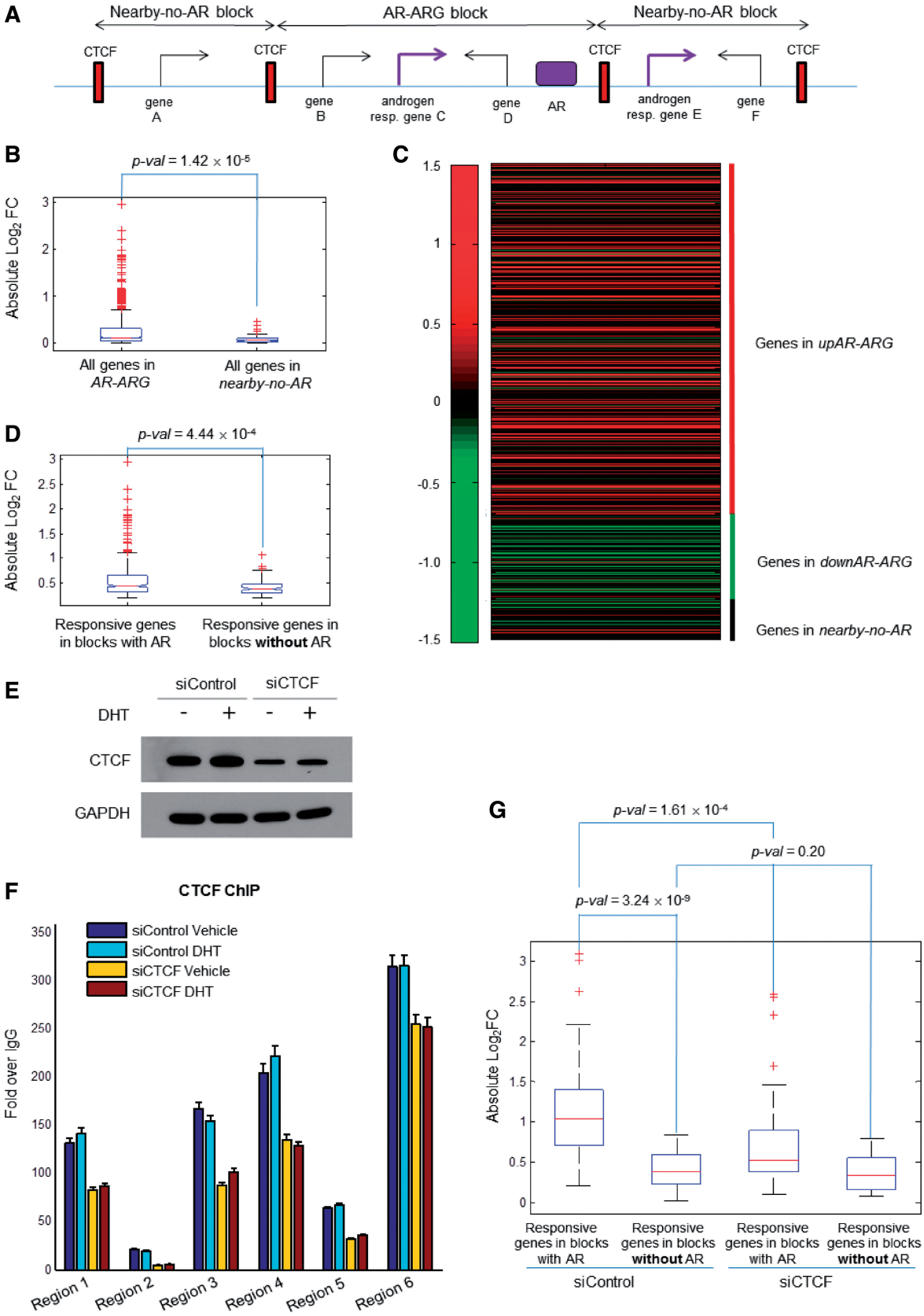

Figure 2. Genes with AR in the same block exhibit higher levels of gene expression fold change. (**A**) Illustrations of two types of blocks that are being compared in (**B**). AR-ARG block is defined as block with both AR binding site (purple box) and androgen-responsive gene (purple arrow). Nearby non-AR block is defined as block adjacent to AR-ARG blocks without AR binding site. (**B**) Expression fold-change of all genes in AR-ARG blocks (blocks with androgen-responsive gene and AR binding site) are significantly higher than genes in nearby non-AR blocks. (**C**) Heatmap

(continued)

Table 1. Characteristics of genes in the two classes

	Model characteristics				Implications	
	AR (%)	AR_{Dist} (kb)	$FoxA1_{Dist}$ (kb)	AR-H3K4me2 (%)	FoxA1 (%)	AR-H3K4me2 + FoxA1 (%)
Class 1	98	12	7.6	94	97	93
Class 2	38	96	54	24	76	22

The left panel, 'Model characteristics', shows significant factors identified from the supervised classification model. The right panel displays two factors that were not included in the model but are indirectly implied from the classification of the two classes. AR: percentage of genes which have AR in the same block, AR_{Dist}: median distance (in kb) from the TSS of genes to the nearest AR binding sites, $FoxA1_{Dist}$: median distance (in kb) from the TSS of genes to the nearest FoxA1 binding sites. AR-H3K4me2: percentage of genes which have AR binding site that overlap with H3K4me2 enrichment in the same block, FoxA1: percentage of genes which contain FoxA1 in the same block, AR-H3K4me2 + FoxA1: percentage of genes which have AR binding sites that overlap with H3K4me2 enrichment as well as FoxA1 binding site in the same block.

distances on the classification of these two classes, we plotted the density function (smoothed using spline function) of the distances, stratified according to classes and up/down regulation of genes (Figure 3). As shown in Figure 3, genes in Class 1 are closer to AR and FoxA1 binding sites compared to those in Class 2 with both up- and down-regulated genes showing similar patterns. This phenomenon is more obviously displayed in upregulated genes.

Interestingly, genes in Class 1 shows significantly higher change of expression level compared to genes in Class 2 (Mann–Whitney, P-value $= 4.34 \times 10^{-10}$), with around 93% of the genes in Class 1 having AR/H3K4me2 overlapping region as well as having FoxA1 binding sites in the same CTCF blocks (Figure 4). Our analysis thus suggests that higher expression level may be due to *short* combinatorial long-range regulation being much more efficient than other mechanisms of long-range regulation. We next provide three lines of evidence to support our observation and conclusion using existing data and resources/databases, as we detail in the following sections.

Functional analysis of classified androgen-responsive genes

Three lines of evidence are shown to substantiate the finding that 'short' combinatorial long-range regulation is much more efficient than other long-range regulation. First, we performed KEGG pathway analysis to show that Class 1 is significantly enriched in prostate cancer related genes. Second, we showed that the genes in Class 1 are preferentially expressed in prostate tissues. Finally, we assess the biological significance of the set of genes in Class 1 on cells isolated using laser-capture microdissection (LCM) representing prostate cancer progression.

Genes in Class 1 are enriched in cancer related pathway. In order to assess the biological significance of the two classes identified using supervised classification, we performed KEGG pathway analysis (Table 2). Genes in Class 1 demonstrate a significant enrichment in cancer and prostate cancer pathways with P-values = 0.0009 and 0.0015, respectively. In contrast, genes in Class 2 do not show enrichment in cancer pathway in general nor prostate cancer in particular. Instead they show enrichment in Focal adhesion pathway, among others (Table 2).

Preferential expression of Class 1 in prostate tissues. We further substantiate our finding in prostate tissues by considering the microarray data published in (27). There were a total of 79 tissue types, with one being prostate tissue and the remaining being other tissues. Each tissue type has two samples. We performed statistical analysis to investigate whether genes in either or both classes defined previously are preferentially expressed in the prostate tissue compared to the other tissues. Our results reveal that genes in Class 1 indeed exhibit much higher difference of expression level in prostate tissues. Specifically, we demonstrated that the levels of expression for genes in Class 1 are much higher in prostate cells (Figure 5; Paired Wilcoxon rank test, P-value $= 2.2 \times 10^{-16}$) compared to the other tissues types. Although genes in Class 2 also exhibit higher levels of expression in prostate cells (paired Wilcoxon rank test, P-value $= 2.15 \times 10^{-8}$) compared to the other tissues, their magnitudes differ (P-value of 10^{-16} versus P-value of 10^{-8}). Taken together, this shows that genes in both

Figure 2. Continued
showing the log2 fold-change of expression level of all genes in nearby non-AR blocks (black bar), genes in upAR-ARG blocks (blocks with upregulated androgen-responsive gene and AR, red bar) and genes in downAR-ARG blocks (blocks with downregulated androgen-responsive genes and AR, green bar). Color represents $\log_2$ fold-change of expression level of genes after 4 h DHT versus 0 hr DHT (basal level). (**D**) Expression level of androgen-responsive genes in AR blocks are significantly higher than those in blocks without AR (non-AR blocks). (**E**) Silencing of CTCF decreases CTCF protein expression. LNCaP cells were transfected with siControl or siCTCF, and treated with vehicle or DHT for 4 h. Western blots were performed using antibodies indicated. (**F**) Silencing of CTCF decreases CTCF binding at the CTCF blocks with AR (regions 1–4) and without AR (regions 5–6). LNCaP cells were transfected with siControl or siCTCF, and stimulated with vehicle or DHT. ChIP assays were performed using antibodies against CTCF. (**G**) Silencing of CTCF significantly decreases expression fold changes of responsive genes in blocks with AR. siControl or siCTCF transfected LNCaP cells were stimulated with vehicle or DHT for 4 h. Total RNA was isolated and amplified with gene-specific primers.

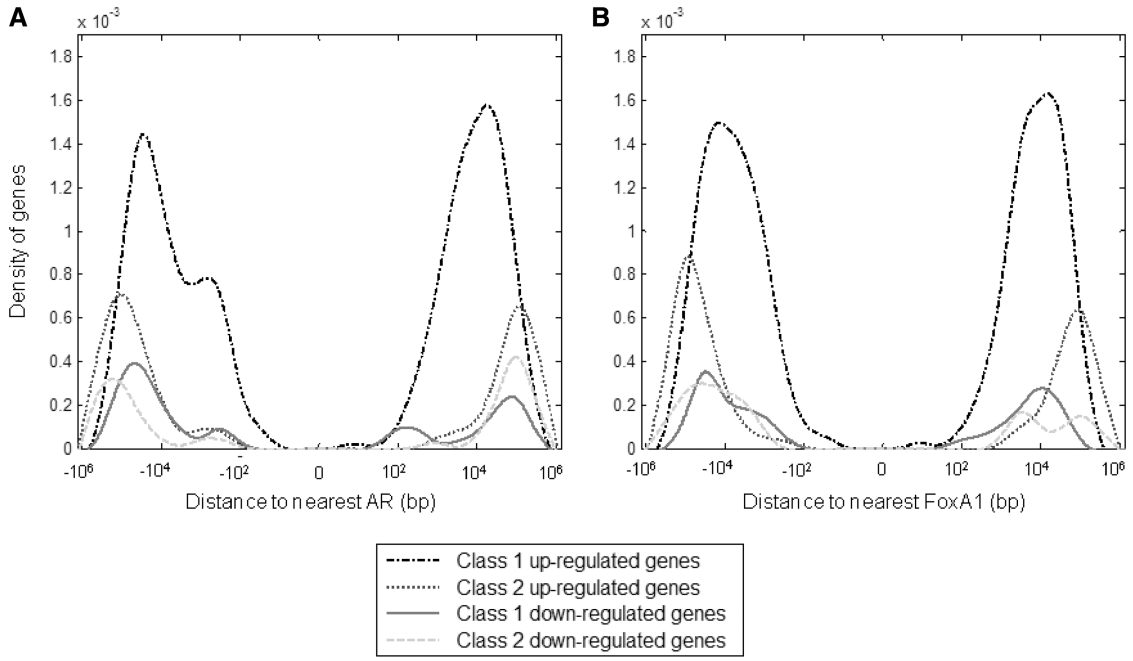

Figure 3. Genes in Class 1 are closer to a transcription factor AR and to FoxA1 binding sites compared to genes in Class 2. The same patterns are observed in both up- and down-regulated genes, although this distance preferential are more obvious for over-expressed genes.

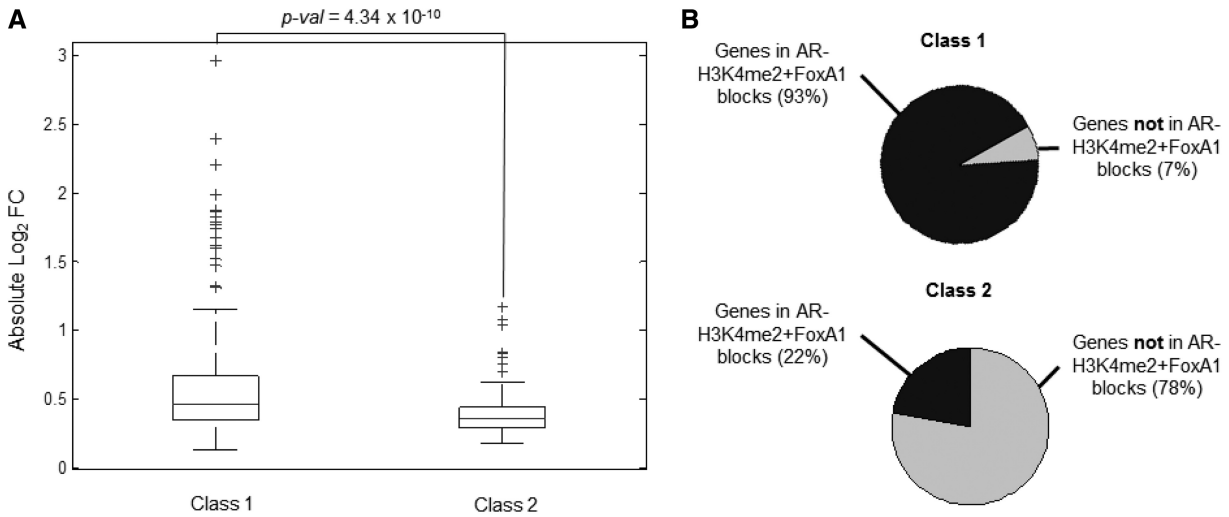

Figure 4. Comparison of genes in Class 1 and Class 2 in terms of expression level and block classification. (**A**) Genes in Class 1 exhibit higher change of expression level than those in Class 2. (**B**) 93% of genes in Class 1 have AR binding sites which overlap with H3K4me2 as well as having FoxA1 bindings in the same CTCF block. In contrast, only 22% of genes in Class 2 have AR overlap with H3K4me2 and FoxA1 in the same block.

classes are preferentially expressed in prostate tissues, but with Class 1 having higher difference in expression than Class 2 (Figure 5). The fact that genes in both classes show greater activity in prostate tissues is not surprising; as both classes contain responsive genes. However, the observation that genes in Class 1 exhibit greater changes of expression in prostate tissues than those in Class 2 suggest that Class 1 genes might play a more important role than Class 2 genes in the growth and maintenance of the prostate.

Class 1 genes are related to prostate cancer metastasis. Finally, we evaluated the biological significance of the genes (focusing on Class 1) for prostate cancer progression. We assess the expression of these genes in the analysis of LCM (Laser Capture Microdissection) epithelial cell populations representing prostate cancer progression from benign epithelium to metastatic disease (28). Interestingly, genes in Class 1 show a similar pattern in benign/normal epithelial, PIN (prostatic intra epithelial neoplasia), and localized

Table 2. Top functional terms for genes within the two classes obtained using supervised classification method

Class 1		Class 2	
Pathway	*P*-value	Pathway	*P*-value
Pathways in cancer	**0.0009**	Focal adhesion	0.0115
B-cell receptor signaling pathway	0.0009	TGF-beta signaling pathway	0.0115
Type II diabetes mellitus	0.0009	Bladder Cancer	0.0143
Steroid biosynthesis	0.0009	O-Glycan biosynthesis	0.0143
Adipocytokine signaling pathway	0.0009	Cytokine-cytokine receptor interaction	0.0143
Neuroactive ligand-receptor interaction	0.0009	Cell adhesion molecules (CAMs)	0.0143
Prostate cancer	**0.0015**	Arginine and proline metabolism	0.0173
P53 signaling pathway	0.0041	mTOR signaling pathway	0.0173
endocytosis	0.0041	VEGF signaling pathway	0.0212
Neurotrophin signaling pathway	0.0042	P53 signaling pathway	0.0212

Benjamini–Hochberg corrected *P*-values are noted. Categories that are cancer related are in bold.

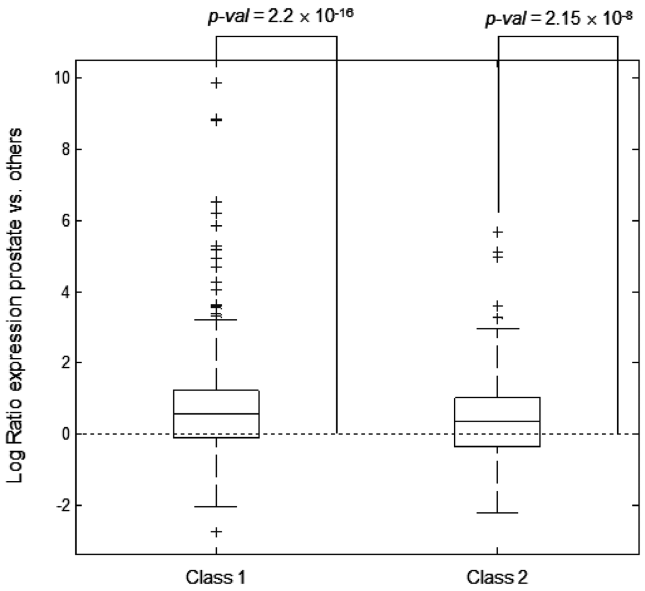

Figure 5. Genes in Class 1 are preferentially expressed in prostate cells compared to other tissues. Genes in Class 1 exhibit significantly higher expression level in prostate compared to other tissues (paired Wilcoxon test, *P*-value = 2.2×10^{-16}). Genes in Class 2 also show significantly higher expression level in prostate cells versus the remaining tissues (paired Wilcoxon test, *P*-value = 2.15×10^{-8}).

prostate cancer, but a distinctive pattern in metastatic prostate cancer (Figure 6, left panel). In contrast, genes in Class 2 show similar patterns in all categories (Figure 6, right panel). Since Class 1 is more responsive to androgen, our finding that these genes are more reduced in metastatic samples compared to Class 2 is consistent with previous findings that androgen signaling activity is decreased in metastatic prostate cancer (28). The distinctive pattern in metastatic prostate cancer for Class 1 genes indicates their promising potential as novel biomarkers in delineating metastatic prostate cancer. Specifically, genes in Class 1 that are highly expressed and are hypothesized to be involved in 'short' combinatorial long-range regulation were downregulated in metastatic versus localized prostate cancer.

DISCUSSION

The mechanisms by which transcription factors regulate distally located genes have become one of the most intriguing topics in recent years. Transcription factors such as AR, FoxA1, CTCF and H3K4me2 (an active histone mark) are all believed to play an important role in gene regulation. Such factors are likely to act in concert rather than independently, and thus it is hypothesized that understanding of the combinatorial effects of known gene regulators would lead to unraveling of the complexity of long-range regulation mechanism. To the best of our knowledge, the interacting effects of individual regulators have not been studied previously, which is the challenge that we have taken up in this article.

Our statistical analysis and experimental exploration show that CTCF can act as a facilitator where AR regulation within CTCF blocks is more effective than across blocks. This is consistent with our previous finding that inserting the chicken β-globin insulator core fragment FII containing CTCF binding sequence between the PSA gene enhancer and promoter significantly decreases the PSA reporter gene expression, suggesting that CTCF can function as an insulator to block AR target gene expression (11).

We have also performed genome-wide integrated analysis using all factors jointly followed by supervised classification, which enabled us to discover a subset of androgen-responsive genes. Almost all of the genes (93%) in this set have AR/H3K4me2 overlapping region and FoxA1 binding sites located in the same CTCF block. Furthermore, they are much closer to AR and FoxA1 binding sites compared to other androgen responsive genes (median distances to AR and FoxA1 binding sites are 12 kb and 7.6 kb versus 96 kb and 54 kb). Genes in this subset also show higher preferential expression in prostate tissues and are enriched in prostate cancer and cancer related pathways. Furthermore, these genes were downregulated in metastatic versus localized prostate cancer. Thus, our result reveals for the first time that AR target genes under CTCF-facilitated 'short' combinatorial long-range control may play more important roles in prostate development and prostate cancer progression than those genes under 'long' long-range control.

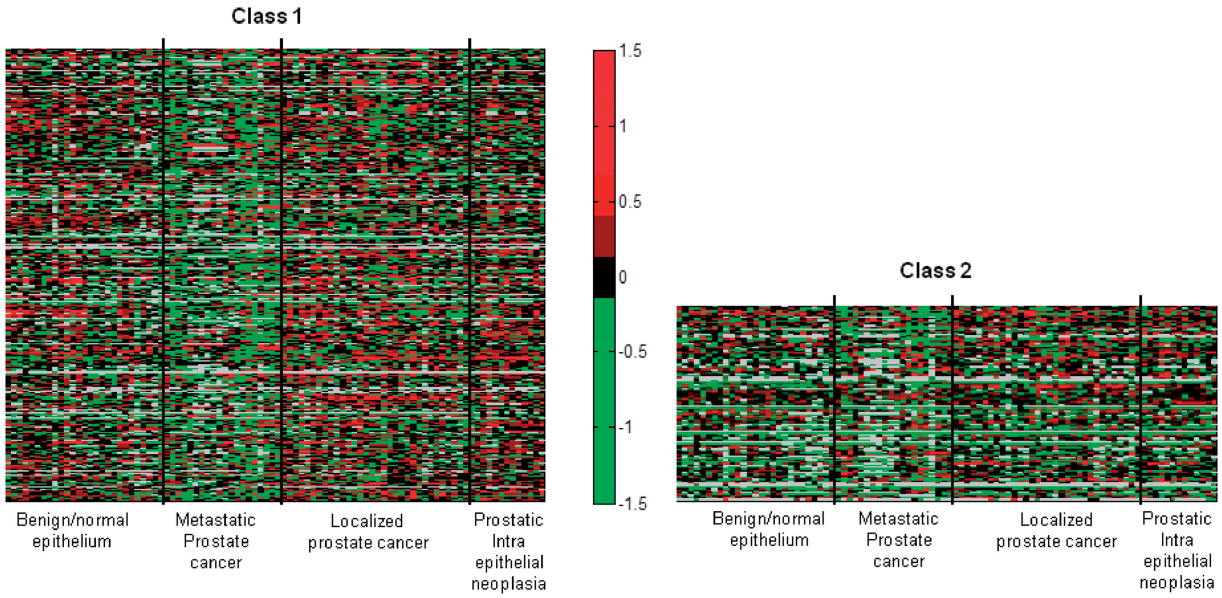

Figure 6. Genes in Class 1 shows distinctive pattern in metastatic prostate cancer while Class 2 shows less distinctive markings. Color represents $\log_2$ ratios of expression. When the expression value is not available, it is denoted by gray color.

Fullwood *et al.* (36) developed a new method called ChIA-PET (chromatin interaction analysis by paired-end tag) to detect global chromatin looping mediated by a specific protein. Their results suggest that many ER-α binding sites interact with gene promoters through long-range chromatin looping. Studying the same issue but from a different angle, our results indicates that AR-mediated regulation appear to be similar to ER in that many genes are regulated through a remote mechanism. Most interestingly, for the first time, our results shed new lights on long-range regulation mechanisms by showing that AR-mediated long-range regulation may involve short chromatin loops with a number of factors including histone modification marks, and that such short combinatorial long-range regulation may be more efficient than other long-range regulation mechanisms.

Recently, a new CTCF ChIP-chip experiment in LNCaP cells covering chromosomes 8, 11 and 12 was carried out by (25). We found CTCF binding sites discovered in their experiments have huge overlap with the CTCF ChIP-seq data that we are using. More than 80% of the CTCF in LNCaP cells ($fdr < 10\%$) overlapped with the CTCF in Jurkat, CD4+ T and HeLa cells on those three chromosomes. In this article, we decided to use the shared CTCF binding sites found in the three cells because of their large overlap with CTCF binding sites in LNCaP. Furthermore, CTCF experiment in LNCaP only has limited coverage (only done on 3 chromosomes). In the future, as more data become available, it would be useful to confirm our findings using additional genome-wide CTCF bound data in LNCaP.

ACCESSION NUMBERS

[All the data used in this study are available in NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession no. GSE12889, GSE21513, GSE7868, GSE20042, GSE1133, GSE6099 and GSE26329 http://research.dfci.harvard.edu/brownlab/datasets/]

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Methods, Supplementary Figures 1–3, Supplementary Tables 1–6, and Supplementary References [37–39].

ACKNOWLEDGEMENTS

We thank Dr. Housheng He (Dana-Farber Cancer Institute) for helpful discussions. We also thank the referees for their constructive comments.

FUNDING

Funding for open access charge: The National Cancer Institute grants U54 CA113001 (to T.H-M.H., C.T., K.H., Q.W., and S.L.), R01 CA151979 (to Q.W. and S.L.), 2011 V Foundation V Scholar Award (to Q.W.), and the National Science Foundation grant DMS-1042946 (to S.L.).

Conflict of interest statement. None declared.

REFERENCES

1. Hääg,P., Bektic,J., Bartsch,G., Klocker,H. and Eder,I.E. (2005) Androgen receptor down regulation by small interference RNA induces cell growth inhibition in androgen sensitive as well as in androgen independent prostate cancer cells. *J. Steroid Biochem. Mol. Biol.*, **96**, 251–258.

2. Knudsen,K.E. and Penning,T.M. (2010) Partners in crime: deregulation of AR activity and androgen synthesis in prostate cancer. *Trends Endocrinol. Metab.*, **21**, 315–324.
3. Scher,H.I., Liebertz,C., Kelly,W.K., Mazumdar,M., Brett,C., Schwartz,L., Kolvenbag,G., Shapiro,L. and Schwartz,M. (1997) Bicalutamide for advanced prostate cancer: the natural versus treated history of disease. *J. Clin. Oncol.*, **15**, 2928–2938.
4. Jia,L., Berman,B.P., Jariwala,U., Yan,X., Cogan,J.P., Walters,A., Chen,T., Buchanan,G., Frenkel,B. and Coetzee,G.A. (2008) Genomic androgen receptor-occupied regions with different functions, defined by histone acetylation, coregulators and transcriptional capacity. *PLoS One*, **3**, e3645.
5. Wang,Q., Li,W., Zhang,Y., Yuan,X., Xu,K., Yu,J., Chen,Z., Beroukhim,R., Wang,H., Lupien,M. *et al.* (2009) Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. *Cell*, **138**, 245–256.
6. Yu,J., Yu,J., Mani,R.S., Cao,Q., Brenner,C.J., Cao,X., Wang,X., Wu,L., Li,J., Hu,M. *et al.* (2010) An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell*, **17**, 443–454.
7. Carroll,J.S., Liu,X.S., Brodsky,A.S., Li,W., Meyer,C.A., Szary,A.J., Eeckhoute,J., Shao,W., Hestermann,E.V., Geistlinger,T.R. *et al.* (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, **122**, 33–43.
8. Wang,Q., Li,W., Liu,X.S., Carroll,J.S., Jänne,O.A., Keeton,E.K., Chinnaiyan,A.M., Pienta,K.J. and Brown,M. (2007) A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth. *Mol. Cell*, **27**, 380–392.
9. Bulger,M. and Groudine,M. (2011) Functional and mechanistic diversity of distal transcription enhancers. *Cell*, **144**, 327–339.
10. Chen,Z., Zhang,C., Wu,D., Chen,H., Rorick,A., Zhang,X. and Wang,Q. (2011) Phospho-MED1-enhanced UBE2C locus looping drives castration-resistant prostate cancer growth. *EMBO J.*, **30**, 2405–2419.
11. Wang,Q., Carroll,J.S. and Brown,M. (2005) Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Mol. Cell*, **19**, 631–642.
12. Lupien,M., Eeckhoute,J., Meyer,C.A., Wang,Q., Zhang,Y., Li,W., Carroll,J.S., Liu,X.S. and Brown,M. (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, **132**, 958–970.
13. Wang,D., Garcia-Bassets,I., Benner,C., Li,W., Su,X., Zhou,Y., Qiu,J., Liu,W., Kaikkonen,M.U., Ohgi,K.A. *et al.* (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, **474**, 390–394.
14. Sahu,B., Laakso,M., Ovaska,K., Mirtti,T., Lundin,J., Rannikko,A., Sankila,A., Turunen,J.P., Lundin,M., Konsti,J. *et al.* (2011) Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. *EMBO J*, **30**, 3962–3976.
15. Augello,M.A., Hickey,T.E. and Knudsen,K.E. (2011) FOXA1: master of steroid receptor function in cancer. *EMBO J*, **30**, 3885–3894.
16. 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.
17. 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.
18. He,H.H., Meyer,C.A., Shin,H., Bailey,S.T., Wei,G., Wang,Q., Zhang,Y., Xu,K., Ni,M., Lupien,M. *et al.* (2010) Nucleosome dynamics define transcriptional enhancers. *Nat. Genet.*, **42**, 343–347.
19. Phillips,J.E. and Corces,V.G. (2009) CTCF: master weaver of the genome. *Cell*, **137**, 1194–1211.
20. Xie,X., Mikkelsen,T.S., Gnirke,A., Lindblad-Toh,K., Kellis,M. and Lander,E.S. (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl Acad. Sci. USA*, **104**, 7145–7150.
21. Chan,C.S. and Song,J.S. (2008) CCCTC-binding factor confines the distal action of estrogen receptor. *Cancer Res.*, **68**, 9041–9049.
22. Handoko,L., Xu,H., Li,G., Ngan,C.Y., Chew,E., Schnapp,M., Lee,C.W., Ye,C., Ping,J.L., Mulawadi,F. *et al.* (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, **43**, 630–638.
23. 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.
24. Cuddapah,S., Jothi,R., Schones,D.E., Roh,T.Y., Cui,K. and Zhao,K. (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.*, **19**, 24–32.
25. Sérandour,A.A., Avner,S., Percevault,F., Demay,F., Bizot,M., Lucchetti-Miganeh,C., Barloy-Hubler,F., Brown,M., Lupien,M., Métivier,R. *et al.* (2011) Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. *Genome Res.*, **21**, 555–565.
26. Zhang,C., Wang,L., Wu,D., Chen,H., Chen,Z., Thomas-Ahner,J., Zynger,D., Eeckhoute,J., Yu,J., Luo,J. *et al.* (2011) Definition of a FoxA1 cistrome that is crucial for G1 to S-Phase cell- cycle transit in castration-resistant prostate cancer. *Cancer Res.*, **71**, 6738–6748.
27. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
28. Tomlins,S.A., Mehra,R., Rhodes,D.R., Cao,X., Wang,L., Dhanasekaran,S.M., Kalyana-Sundaram,S., Wei,J.T., Rubin,M.A., Pienta,K.J. *et al.* (2007) Integrative molecular concept modeling of prostate cancer progression. *Nat. Genet.*, **39**, 41–51.
29. Jothi,R., Cuddapah,S., Barski,A., Cui,K. and Zhao,K. (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, **36**, 5221–5231.
30. Johnson,W.E., Li,W., Meyer,C.A., Gottardo,R., Carroll,J.S., Brown,M. and Liu,X.S. (2006) Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl Acad. Sci. USA*, **103**, 12457–12462.
31. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
32. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
33. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
34. Raftery,A.E. (1996) Hypothesis testing and model selection. In: Gilks,W.R., Richardson,S. and Spiegelhalter,D.J. (eds), *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, pp. 163–188.
35. Wu,D., Zhang,C., Shen,Y., Nephew,K.P. and Wang,Q. (2011) Androgen receptor-driven chromatin looping in prostate cancer. *Trends Endocrinol. Metab.*, **22**, 474–480.
36. Fullwood,M.J., Liu,M.H., Pan,Y.F., Liu,J., Xu,H., Mohamed,Y.B., Orlov,Y.L., Velkov,S., Ho,A., Mei,P.H. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, 58–64.
37. Ji,X., Li,W., Song,J., Wei,L. and Liu,X.S. (2006) CEAS: cis-regulatory element annotation system. *Nucleic Acids Res.*, **34**, W551–W554.
38. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
39. Sabo,P.J., Hawrylycz,M., Wallace,J.C., Humbert,R., Yu,M., Shafer,A., Kawamoto,J., Hall,R., Mack,J., Dorschner,M.O. *et al.* (2004) Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc. Natl Acad. Sci. USA*, **101**, 16837–16842.