

# UCSF

## UC San Francisco Previously Published Works

### Title

Chromatin accessibility reveals insights into androgen receptor activation and transcriptional specificity

### Permalink

<https://escholarship.org/uc/item/70m5496v>

### Journal

Genome Biology, 13(10)

### ISSN

1465-6906

### Authors

Tewari, Alok K  
Yardimci, Galip  
Shibata, Yoichiro  
[et al.](#)

### Publication Date

2012-10-03

### DOI

<http://dx.doi.org/10.1186/gb-2012-13-10-r88>

Peer reviewed

RESEARCH

Open Access

# Chromatin accessibility reveals insights into androgen receptor activation and transcriptional specificity

Alok K Tewari<sup>1</sup>, Galip Gürkan Yardimci<sup>1</sup>, Yoichiro Shibata<sup>1</sup>, Nathan C Sheffield<sup>1</sup>, Lingyun Song<sup>1</sup>, Barry S Taylor<sup>3</sup>, Stoyan G Georgiev<sup>1</sup>, Gerhard A Coetzee<sup>4,5</sup>, Uwe Ohler<sup>1,6,7</sup>, Terrence S Furey<sup>8</sup>, Gregory E Crawford<sup>1,9†</sup> and Phillip G Febbo<sup>2,10,11\*†</sup>

## Abstract

**Background:** Epigenetic mechanisms such as chromatin accessibility impact transcription factor binding to DNA and transcriptional specificity. The androgen receptor (AR), a master regulator of the male phenotype and prostate cancer pathogenesis, acts primarily through ligand-activated transcription of target genes. Although several determinants of AR transcriptional specificity have been elucidated, our understanding of the interplay between chromatin accessibility and AR function remains incomplete.

**Results:** We used deep sequencing to assess chromatin structure via DNase I hypersensitivity and mRNA abundance, and paired these datasets with three independent AR ChIP-seq datasets. Our analysis revealed qualitative and quantitative differences in chromatin accessibility that corresponded to both AR binding and an enrichment of motifs for potential collaborating factors, one of which was identified as SP1. These quantitative differences were significantly associated with AR-regulated mRNA transcription across the genome. Base-pair resolution of the DNase I cleavage profile revealed three distinct footprinting patterns associated with the AR-DNA interaction, suggesting multiple modes of AR interaction with the genome.

**Conclusions:** In contrast with other DNA-binding factors, AR binding to the genome does not only target regions that are accessible to DNase I cleavage prior to hormone induction. AR binding is invariably associated with an increase in chromatin accessibility and, consequently, changes in gene expression. Furthermore, we present the first *in vivo* evidence that a significant fraction of AR binds only to half of the full AR DNA motif. These findings indicate a dynamic quantitative relationship between chromatin structure and AR-DNA binding that impacts AR transcriptional specificity.

## Background

The androgen receptor (AR), a ligand-activated member of the nuclear receptor superfamily, plays a critical role in the male phenotype and prostate cancer biology. AR expression results in context-specific transformation of prostate epithelial cells [1-5], and persistent AR signaling is implicated in the progression to castration-resistant prostate cancer [6-8]. However, AR activity can be alternatively associated with promotion or inhibition of

growth. For example, AR activation by androgen induction limits proliferation in some immortalized prostate epithelial cells expressing AR [2,9], whereas AR activation most often increases proliferation in human-derived prostate cancer cell lines with endogenous AR expression (for example, LNCaP [10], LAPC-4 [11] and VCaP [12]). As the AR acts primarily through transcriptional activation of target genes, it is critical to understand the determinants of the AR-mediated transcriptional program.

AR-mediated transcriptional specificity is highly regulated, and the AR associates with proteins that possess co-activator or co-repressor function [13]. AR binding to chromatin, similar to many transcription factors, is thought to occur in competition with nucleosome histone

\* Correspondence: phillip.febbo@ucsf.edu

† Contributed equally

<sup>2</sup>Helen Diller Family Comprehensive Cancer Center, University of California at San Francisco, San Francisco, CA 94115, USA

Full list of author information is available at the end of the article

proteins, the core organizational component of chromatin [14]. Several identified AR co-factors either possess an intrinsic chromatin remodeling capability or are able to bind and recruit other chromatin modifying enzymes and facilitate AR binding. Indeed, the binding of AR to DNA across the genome (the AR cistrome) is modulated by the primary DNA sequence, chromatin structure around the AR and/or co-factor binding sites and other factors such as FOXA1, a member of the forkhead box (FOX) and hepatocyte nuclear factor transcription factor families [15].

Recent reports examining nucleosome positioning in relation to AR binding have found that local nucleosome depletion and increased chromatin accessibility accompanies the AR binding to DNA [16,17]. However, while one study observed a clear decrease in occupancy of histone 3 (H3) dimethyl lysine 4 (H3K4me2)-marked nucleosomes over AR binding sites and a concomitant increase in occupancy at flanking nucleosome positions [16], another found that the nucleosome depletion size was not increased by AR occupancy but rather nucleosome dynamics were affected by the receptor binding [17]. Interestingly, nucleosome depletion at the three enhancers studied was evident both before and after hormone treatment. Thus, chromatin structure is likely to impact the interaction between the AR and DNA, and ligand activation of the AR may result in altered chromatin structure. Our complete understanding of this process remains quite limited and, consequently, a comprehensive genome-wide analysis of AR function is needed.

The mapping of DNase I hypersensitive (DHS) sites is an accurate method to identify different types of active gene regulatory elements within accessible chromatin [18-20]. More recent high-throughput identification of all DHS sites within a single cell type using DNase-seq show high correlations with active histone modifications, regions of nucleosome depletion detected by Formaldehyde Assisted Isolation of Regulatory Elements (FAIRE) and transcription factor binding sites [21-24]. Changes in DNase I cleavage patterns have been observed at specific loci bound by nuclear receptors, supporting the finding that at least some nuclear receptors can disrupt chromatin structure [25]. The single-base-pair resolution digestion patterns from DNase-seq can identify footprints of local DNA protection that accurately predict transcription factor-DNA binding [26-28]. Thus, in a single experiment, DNase-seq can identify both larger nucleosome-depleted regions and finer resolution transcription factor binding sites within nucleosome-depleted regions.

To determine the relationship between AR-dependent chromatin accessibility changes and AR-mediated transcription, we performed DNase-seq and mRNA-seq on the well-established androgen-sensitive prostate cancer cell line LNCaP, before and after hormone induction. AR

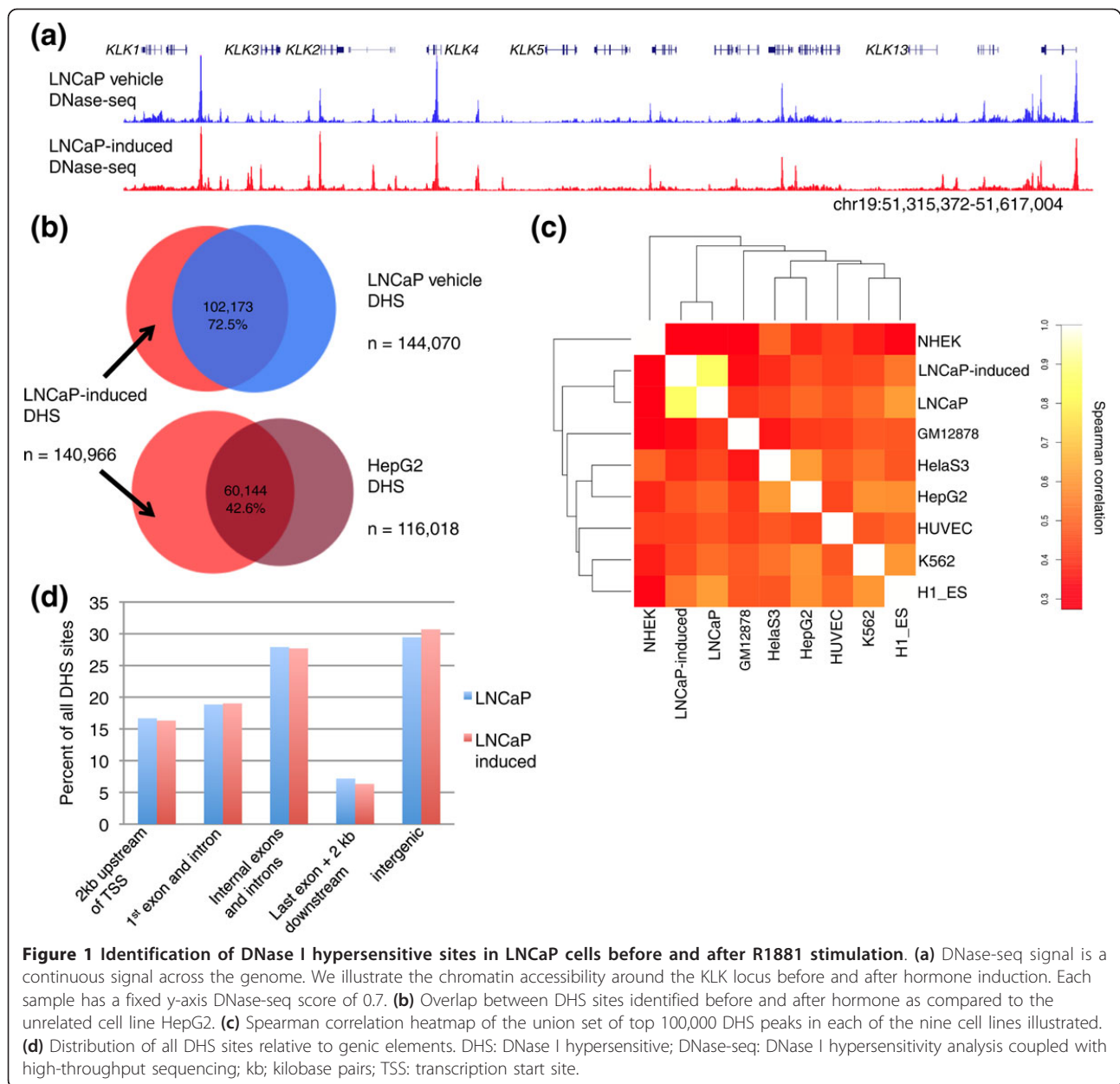
binding sites were obtained from three published studies describing AR ChIP-seq experiments on LNCaP cell lines. Another member of the nuclear receptor superfamily, the glucocorticoid receptor (GR), was recently found to bind predominantly in DHS sites that exist prior to GR ligand hormone treatment [29]. In agreement with another recently published study [30], we find that a substantial amount of AR binding occurs in accessible chromatin after hormone induction. In contrast to the GR, approximately half of these AR sites bind in DHS sites that exist prior to AR activation, with the remaining sites becoming accessible after AR activation. AR binding also significantly increases chromatin accessibility. Quantitative changes in chromatin structure correlate with AR-dependent differential gene expression and are enriched for transcription factor-DNA binding motifs that offer insight into the mechanism of AR-induced chromatin remodeling. Intriguingly, fine resolution DNase-seq profiles surrounding AR DNA binding motifs provides *in vivo* evidence of AR binding to both half and full AR DNA recognition motifs. Together, our work reveals that active chromatin remodeling occurs during androgen nuclear receptor activation.

## Results

### DNase-seq identifies changes in chromatin accessibility with androgen receptor activation

To assess the relationship between accessible chromatin and AR activation, we performed DNase-seq on independent growths of LNCaP cells that were cultured with (LNCaP-induced) or without (LNCaP) the synthetic androgen R1881 (12 hours). Using previously published methodologies and a standard analysis pipeline [31], we identified the full spectrum of DNase-seq signal across the genome (Figure 1a). We approached interpretation of DNase-seq data in two ways: calling discrete peaks, referred to as DHS sites, and comparing regions qualitatively as binary conditions (DHS site or not); and identifying regions of statistically different DNase-seq signal before and after hormone treatment, referred to as  $\Delta$ DNase regions.

From approximately 130 million post-filter sequence reads per growth condition, we identified 144,070 DHS sites in LNCaP and 140,966 DHS in LNCaP-induced cells using a *P*-value cutoff of 0.05. The DHS sites in each cell condition cover approximately 3% of the human genome (Table S1 in Additional file 1). A comparison of the DHS sites identified in LNCaP-induced and LNCaP reveals that 102,173 (72.5%) of sites overlap. To put the degree of overlap in context, we used the same criteria to identify DHS sites in seven unrelated cell lines for which high quality DNase-seq data is available (NHEK, GM12678, HeLaS3, HepG2, HUVEC, K562 and H1-ES) [24]. The average overlap between distinct



cell lines is  $50.4\% \pm 7.04\%$ , which is substantially less than the overlap between LNCaP and LNCaP-induced (Figure 1b,c). We also investigated the overall distribution of DHS sites relative to genic elements and found that AR activation does not shift this distribution (Figure 1d). These data suggest that although AR activation induces a modest amount of chromatin changes, the degree of these changes is substantially less than those detected between cell lines from unrelated tissues.

To quantitatively identify those loci with the most substantive increase or decrease in DNase-seq signal with AR activation, we used the edgeR statistical package [32]. Increases represent regions that become more accessible

after hormone treatment, and decreases become less accessible. To capture a broad spectrum of significant changes in signal, we used two statistical thresholds (strict = a false discovery rate (FDR) threshold of 5%, and loose = unadjusted *P*-value threshold of 0.05) to identify the degree of accessibility changes, which we refer to as  $\Delta$ DNase regions. At the strict threshold, we identified 2,586 regions with strict  $\Delta$ DNase increase after androgen induction and no regions of signal decrease. The loose threshold identified 18,692 regions with loose  $\Delta$ DNase increase and 1,467 regions with loose  $\Delta$ DNase decrease (Table 1). These regions suggest that AR activation results primarily in regions with increased rather than

**Table 1 Number of differential regions of DNase-seq with androgen receptor activation ( $\Delta$ DNase).**

Strict threshold	Number of regions
Strict $\Delta$ DNase increase	2,586
Strict $\Delta$ DNase decrease	0
Loose threshold	
Loose $\Delta$ DNase increase	18,692
Loose $\Delta$ DNase decrease	1,467

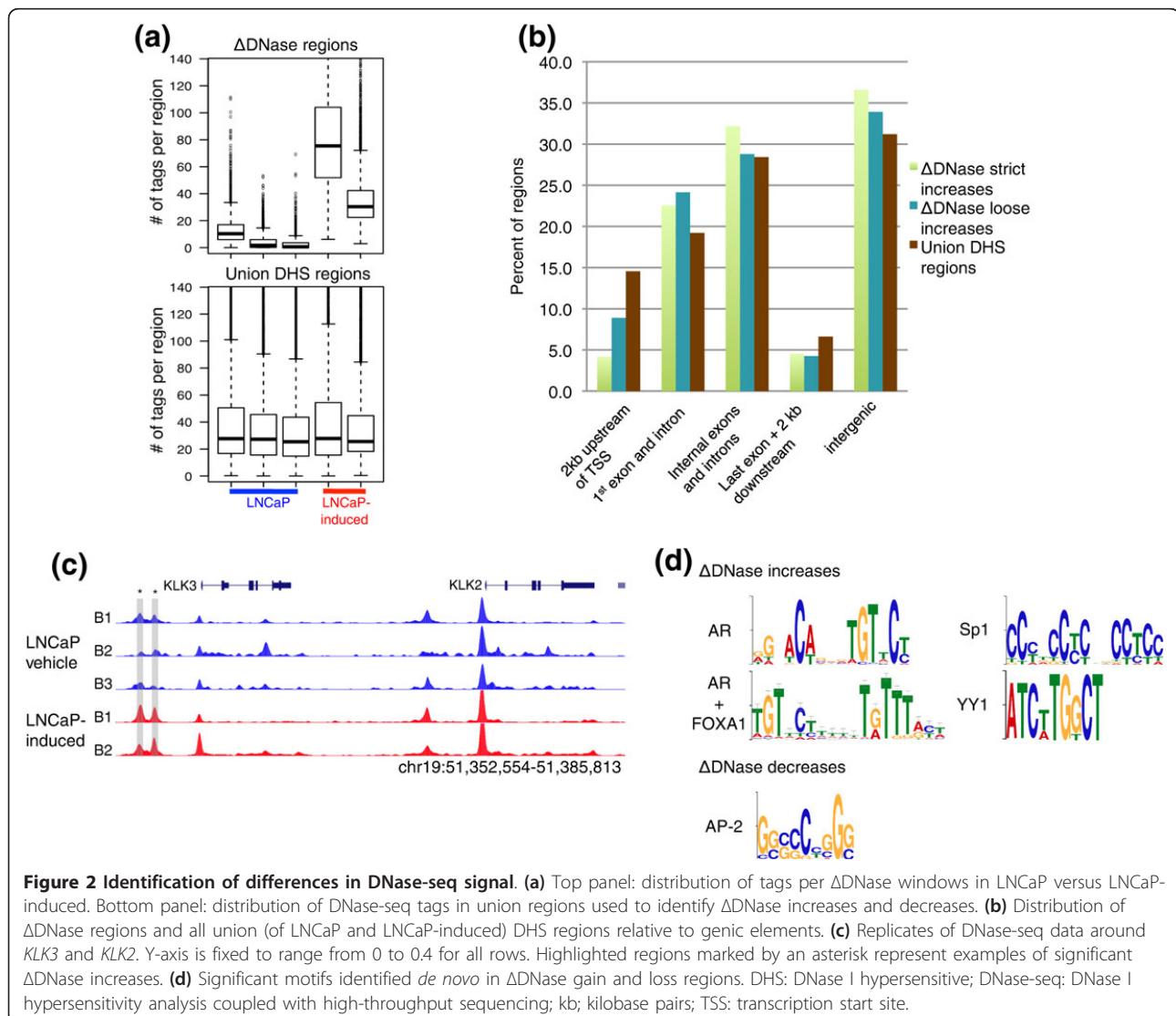
$\Delta$ DNase regions were identified using edgeR (Materials and methods). Strict  $\Delta$ DNase increases are a complete subset of the loose  $\Delta$ DNase increase regions.

decreased chromatin accessibility (Figure 2a, Figure S1A in Additional file 1).

To ensure that the observed trend towards higher levels of open chromatin is not a bias related to the edgeR algorithm, we also calculated a normalized differential DNase-seq tag count for each region in the union

set of LNCaP-induced and LNCaP DHS sites (Materials and methods). This differential count also indicated that more regions display an increase in DNase-seq signal with androgen treatment, supporting the edgeR results (Figure S1B,C in Additional file 1).

Mapping all regions of significantly changed DNase-seq signal to genic elements revealed a depletion of promoter regions and enrichment for both inter- and intragenic locations compared with all DHS sites (Figure 2b, Figure S1D in Additional file 1). For example, approximately 8% of loose  $\Delta$ DNase increases map to promoters (defined as 2 kb region upstream of the transcriptional start site) whereas close to 15% of all DHS sites fell within promoters. The opposite trend was seen for sites that overlapped the first exon and/or intron and sites contained within intergenic regions. Thus, our data show that AR activation primarily results in increased chromatin accessibility in distal regulatory elements that may be associated with



enhancer rather than proximal promoter function, exemplified by a well-characterized AR enhancer [33] 4 kb upstream of the kallikrein 3 (*KLK3*) promoter (Figure 2c).

We hypothesized that  $\Delta$ DNase regions represented locations where AR activation altered transcription factor binding. As expected, we found a strong AR motif match in regions of increased open chromatin (Materials and methods). In addition, several other significantly enriched motifs were detected in both  $\Delta$ DNase increase and decrease regions (Figure 2d, Table S2 in Additional file 1) that correspond to transcription factors such as specificity protein 1 (SP1). We also detected enrichment of an SP1 DNA recognition motif within DHS sites using a self-organizing map (SOM) (Materials and methods) that identifies highly specific LNCaP-only DHS regions that were not accessible in 113 additional cell lines (Figure S1E in Additional file 1, top panel). The SOM analysis also identified an enriched motif corresponding to E2A/TCF3 as well as FOXA1 (Figure S1E, middle panel and bottom panels). SP1 can bind directly with multiple known AR co-factors as well as the AR [34]. TCF3 is involved in the Wnt/ $\beta$ -catenin signaling pathway, which crosstalks with AR signaling in prostate cancer [35].  $\Delta$ DNase increases were also enriched for a yin and yang 1 (YY1) motif, which is a transcription factor with a known role in AR-mediated transcription [36]. To compare how often these motifs are found in  $\Delta$ DNase regions relative to other DHS sites, we calculated a relative enrichment score reflecting the relative frequency that a motif match is found in either set of regions. The score for the AR (4.82) and AR+FOXA1 (2.36) motifs suggests they are more commonly found in  $\Delta$ DNase regions, whereas the score for SP1 (0.45) suggests that this motif is more commonly found in regions in which chromatin accessibility does not change with AR activation. The score for YY1 (1.05) indicates that the motif is found with almost equal frequency in  $\Delta$ DNase regions and other DHS sites.  $\Delta$ DNase regions that underwent a decrease in chromatin accessibility with AR activation did not exhibit an enrichment of the AR motif, but we uncovered a motif consistent with activator protein 2, which has been implicated in estrogen receptor binding and function [37], and its DNA motif is found in the promoter regions of several AR-regulated genes in prostate cancer [38]. Thus, AR activation changes chromatin accessibility in regions with AR and AR co-factor binding motifs, likely due to changes in transcription factor loading at these genomic regions.

#### The androgen receptor binds both poised and remodeled chromatin accessible to DNase I cleavage

Based on our motif analysis of  $\Delta$ DNase regions and recent reports of AR binding to nucleosome-depleted regions marked by acetylated H3 [17] and H3K4me2

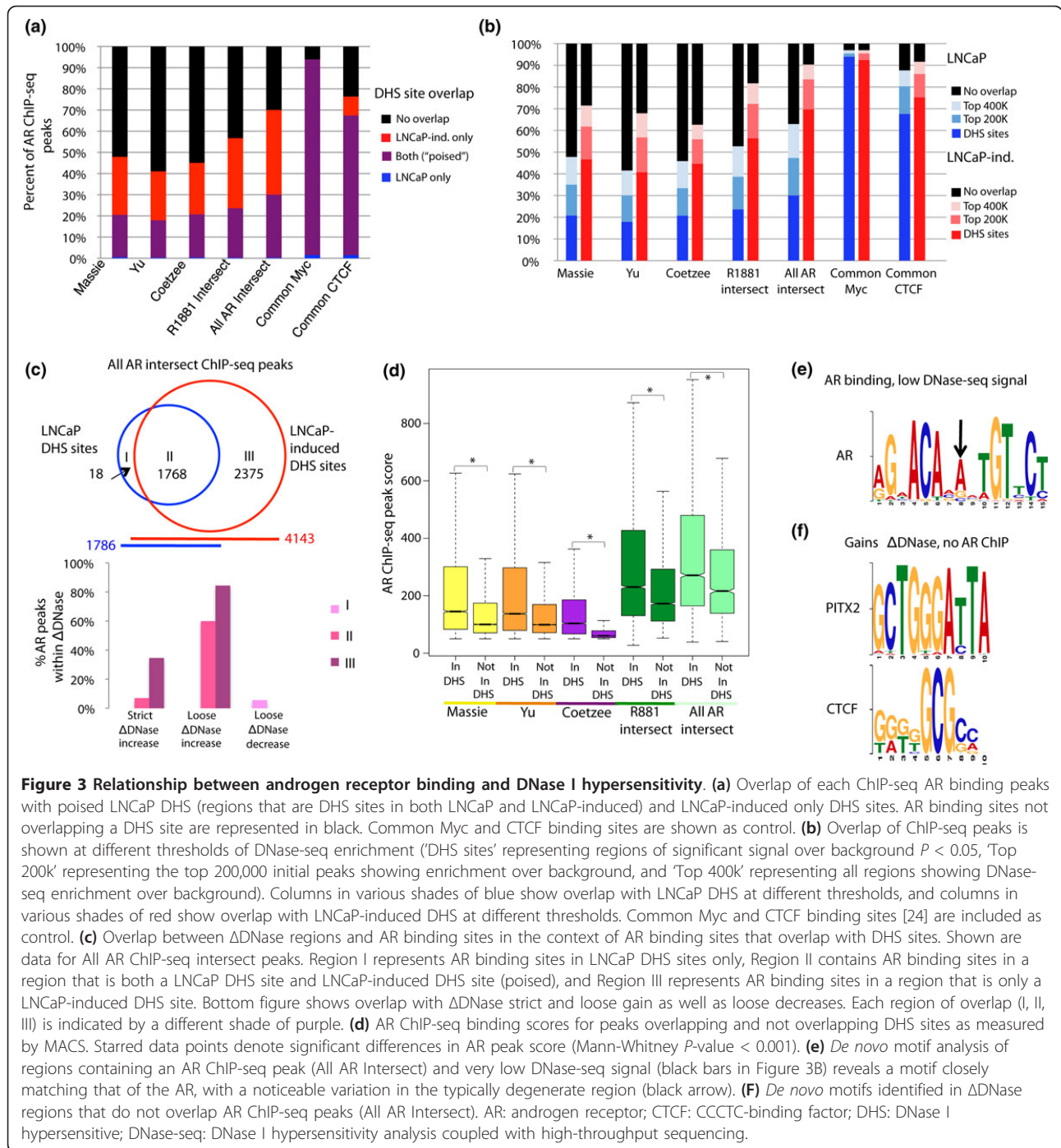
[16], we hypothesized that the AR binds primarily in DHS sites. We therefore used three sets of AR ChIP-seq data from LNCaP cells (Table 2) that we refer to as Yu [39], Massie [40] and Coetzee [17,41]. To minimize the impact of technical variation within each individual experiment, we created two high confidence sets of AR binding sites from these three sources: an 'R1881 intersect' set consisting of Yu and Massie peaks that overlap each other, as these experiments used the same AR hormone ligand as our DNase-seq experiments (R1881); and an 'All AR Intersect' dataset containing the intersection of peaks from all three data sets including the Coetzee experiment that used an alternative AR ligand, dihydrotestosterone. Intersecting AR ChIP datasets did not change the distribution of identified binding sites relative to genic elements (Figure S2A in Additional file 1), despite substantially reducing the number of total AR binding regions by including only those common to two or more experiments (Table 2).

Each of the three individual AR ChIP studies displayed consistent overlap patterns with DHS sites. In each individual experiment approximately 20% of all AR binding sites occurred within DHS sites that are present both before and after hormone treatment (poised DHS sites). An additional 20% to 30% of AR binding sites overlapped DHS sites following androgen induction. Thus, results from each dataset suggest that slightly less than half of all AR binding sites in DHS regions are poised (Figure 3a,b) and the remainder change in response to androgen treatment. The high confidence AR (R1881 intersect and All AR Intersect) binding sites displayed a similar trend. Of note, only 1% to 2% of AR binding sites map within a DHS site present in LNCaP but not LNCaP-induced cells. The amount of AR binding to both poised and LNCaP-induced DHS sites (Figure S2B in Additional file 1) is in stark contrast to Myc and CCCTC-binding factor (CTCF) binding sites [24] that almost exclusively bind within poised DHS sites (Figure 3a). Thus, of the AR binding events occurring within a DHS site, approximately half occurred in poised regions,

**Table 2 Characteristics of androgen receptor ChIP-seq datasets.**

Dataset	Number of androgen receptor binding sites
Massie	19,505
Yu	37,676
Coetzee	12,929
R1881 intersect (Massie/Yu intersect)	13,258
All AR Intersect	5,940

Number of peaks called for each AR ChIP-seq dataset. R1881 intersect represents the intersection of the Massie and Yu datasets. All AR Intersect represents high confidence androgen receptor binding sites that are found in all three datasets.



**Figure 3 Relationship between androgen receptor binding and DNase I hypersensitivity.** (a) Overlap of each CHIP-seq AR binding peaks with poised LNCaP DHS (regions that are DHS sites in both LNCaP and LNCaP-induced) and LNCaP-induced only DHS sites. AR binding sites not overlapping a DHS site are represented in black. Common Myc and CTCF binding sites are shown as control. (b) Overlap of ChIP-seq peaks is shown at different thresholds of DNase-seq enrichment ("DHS sites" representing regions of significant signal over background  $P < 0.05$ , "Top 200K" representing the top 200,000 initial peaks showing enrichment over background, and "Top 400K" representing all regions showing DNase-seq enrichment over background). Columns in various shades of blue show overlap with LNCaP DHS at different thresholds, and columns in various shades of red show overlap with LNCaP-induced DHS at different thresholds. Common Myc and CTCF binding sites [24] are included as control. (c) Overlap between DNase regions and AR binding sites in the context of AR binding sites that overlap with DHS sites. Shown are data for All AR ChIP-seq intersect peaks. Region I represents AR binding sites in LNCaP DHS sites only, Region II contains AR binding sites in a region that is both a LNCaP DHS site and LNCaP-induced DHS site (poised), and Region III represents AR binding sites in a region that is only a LNCaP-induced DHS site. Bottom figure shows overlap with  $\Delta$ DNase strict and loose gain as well as loose decreases. Each region of overlap (I, II, III) is indicated by a different shade of purple. (d) AR ChIP-seq binding scores for peaks overlapping and not overlapping DHS sites as measured by MACS. Starred data points denote significant differences in AR peak score (Mann-Whitney  $P$ -value  $< 0.001$ ). (e) *De novo* motif analysis of regions containing an AR ChIP-seq peak (All AR Intersect) and very low DNase-seq signal (black bars in Figure 3B) reveals a motif closely matching that of the AR, with a noticeable variation in the typically degenerate region (black arrow). (f) *De novo* motifs identified in  $\Delta$ DNase regions that do not overlap AR ChIP-seq peaks (All AR Intersect). AR: androgen receptor; CTCF: CCCTC-binding factor; DHS: DNase I hypersensitivity; DNase-seq: DNase I hypersensitivity analysis coupled with high-throughput sequencing.

with the majority binding to regions that displayed qualitative AR-induced chromatin remodeling.

Given the observation that a substantial number of AR binding sites occur within LNCaP-induced only DHS sites, we examined the association between AR binding events and quantitative chromatin remodeling. To test this, we evaluated AR sites that overlapped regions with increased DNase-seq signal (strict and loose  $\Delta$ DNase increases). As

expected, AR ChIP-seq peaks identified only within LNCaP-induced DHS sites (Region III, Figure 3c) show significant overlap with  $\Delta$ DNase increase regions. Interestingly, AR binding sites in peaks found in both LNCaP and LNCaP-induced cells (Region II, Figure 3c) were also enriched for  $\Delta$ DNase increases, although not to the same extent as those sites that mapped only within LNCaP-induced DHS sites. The proportions of AR binding regions

that mapped to poised, LNCaP-induced DHS sites only and to  $\Delta$ DNase regions were consistent across each AR binding data set (Figure S2C in Additional file 1). Analogously, we found that 36.5% of strict  $\Delta$ DNase increases and 16.7% of loose  $\Delta$ DNase increases overlapped the high confidence AR binding set (All AR Intersect) (Figure S2D in Additional file 1). These observations indicate that although AR binding occurred within DNA in a poised open chromatin state, a substantial increase in chromatin accessibility occurred in many of these regions after AR activation. This highlights the utility of identifying regions of  $\Delta$ DNase signal in addition to regions that simply cross the binary threshold of becoming a DHS site with androgen induction. These findings support similar observations at three previously identified poised AR enhancers [17] and suggest that AR binding more globally stabilizes DHS, allowing for more DNase I cleavage after hormone treatment.

A large percentage of AR binding sites detected by each of the individual AR ChIP-seq datasets (approximately 50%) did not overlap DHS sites. To determine if this is due to a peak-calling threshold, we decreased the stringency threshold for identifying DHS sites to either the top 200,000 or top 400,000 DNase I sensitive regions. Overlap with AR ChIP-seq indicates that the proportion of AR binding sites binding in a poised versus qualitatively remodeled region was consistent regardless of the threshold, and that a substantial proportion of AR binding occurs in non-DNase I sensitive regions of the genome even after relaxing the DHS peak thresholds (Figure 3b). Even after increasing the sequencing depth two-fold, which increased the overall overlap of DHS sites with high confidence AR binding sites, 40% of these binding sites remained only within an induced DHS site (data not shown). In addition, the AR binding signal was stronger in regions overlapping DHS sites than non-DHS regions (Figure 3d), and was the strongest for AR sites common to two or three experiments. Thus, it appears that AR binding occurs at sites with a range of DNase I sensitivity and DNase I sensitivity correlates with AR binding strength.

Finally, we examined several different combinations of regions for evidence of differential co-factor requirements using *de novo* motif analysis. First, we searched for motifs enriched in AR binding sites defined by ChIP-seq peaks that did not map within DHS sites. Only one motif was enriched within these regions by our *de novo* analysis, which resembled both the canonical AR motif and a motif derived *de novo* from AR ChIP-seq sequences that fall within poised DHS sites (Figure 4a), but displayed an increased invariant nucleotide within the degenerate 3 bp region between half sites (Figure 3e). Scanning these same regions against annotated motifs revealed enrichment, albeit with lower match scores, of motifs commonly associated with AR binding (Table S3 in Additional file 1).

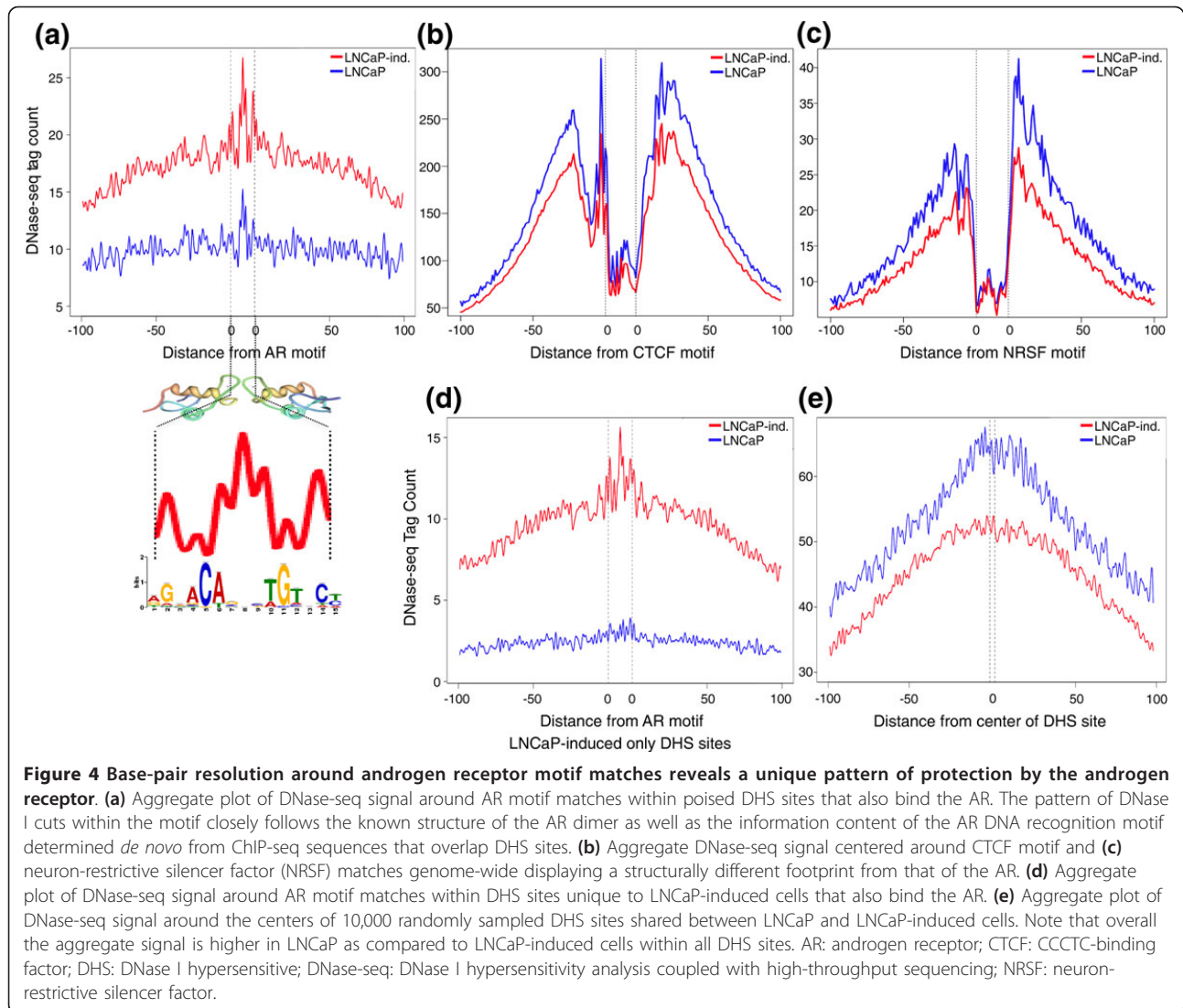
These results suggest that AR binding in regions of very low DNase-seq signal may be less associated with AR co-factors. We separated strict and loose  $\Delta$ DNase increase regions into regions with and without an AR ChIP-seq peak (from All AR Intersect set) and searched for enriched motifs *de novo*.  $\Delta$ DNase regions overlapping AR binding were enriched for motifs matching the AR and FOX family members, as expected.  $\Delta$ DNase regions without AR binding were enriched for several high information content motifs including those for paired-like homeodomain transcription factor 2 (PITX2) and CTCF (Figure 3f, Table S3 in Additional file 1).

### Changes in chromatin accessibility correlate with the androgen receptor transcriptional program

To compare  $\Delta$ DNase regions to the AR-mediated transcriptional program, we generated mRNA-seq data under conditions matched to our DNase-seq experiments and identified genes differentially regulated by androgen induction. Three replicates were generated and their expression values clustered according to hormone treatment status (Figure S3A in Additional file 1). Using edgeR [32], we identified 339 genes differentially expressed upon AR induction (FDR < 0.05), 202 of which were upregulated and 137 of which were downregulated (Figure 5a, Additional file 2). Of these, 46% were identified as AR target genes in at least one other study (Additional file 3).

We hypothesized that AR-mediated changes in chromatin accessibility contribute to the AR-mediated gene expression program. By mapping  $\Delta$ DNase regions to the closest transcriptional start site (Figure 5b), we found that strict  $\Delta$ DNase increase regions were significantly enriched near upregulated genes ( $P < 0.001$ ). Loose  $\Delta$ DNase increases were significantly enriched near both up- and downregulated genes ( $P < 0.001$ ). We noticed that both strict  $\Delta$ DNase increases and loose  $\Delta$ DNase decreases were enriched near downregulated genes with borderline significance. The reverse comparison, in which we associated differentially regulated genes to  $\Delta$ DNase regions within 20 kb of the transcriptional start site (Figure S3C in Additional file 1), confirmed the strongly significant trends mentioned. By contrast, the borderline significant associations disappeared in this reverse comparison, and also when we limited our analysis from Figure 5b to a distance cutoff of 25 kb. We performed an identical analysis using  $\Delta$ DNase regions and microarray expression data from Massie *et al.* [40], and observed similar associations (Figure S3B,D in Additional file 1). We also examined the association between AR binding events with very low DNase-seq signal and AR-regulated genes, and found these regions were not significantly enriched around either up- or downregulated genes (data not shown). Overall, our data support the hypothesis that AR activation preferentially causes distal increases in chromatin accessibility



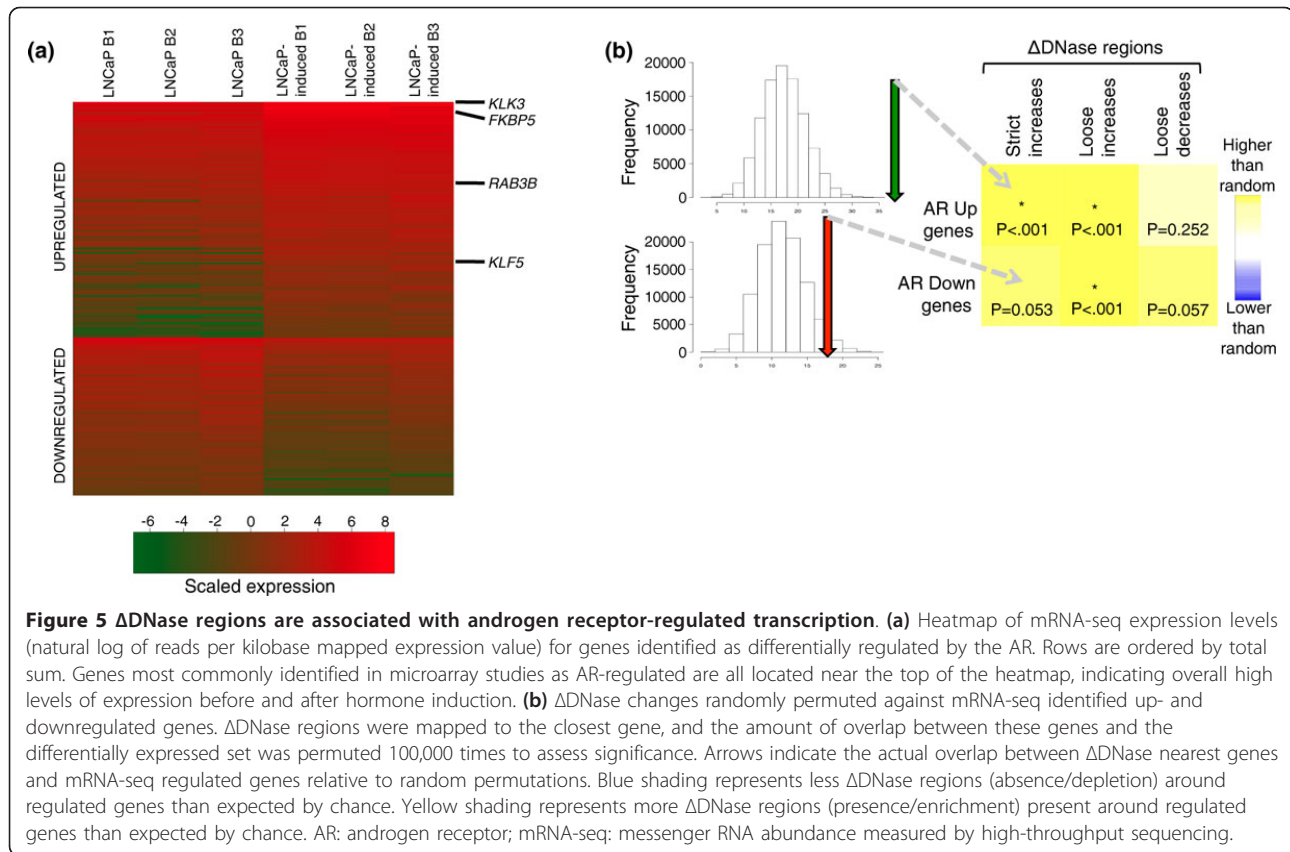


that significantly correlate with nearby gene expression changes.

#### Base-pair resolution analysis of DNase-seq reveals multiple signal profiles

Our group and others have shown that DNase-seq can detect individual transcription factor binding events via the identification of DNase I footprints and that DNase I footprints correspond to local protection of DNA from nuclease cleavage by bound transcription factors [26-28]. An overall increase in DNase signal was observed around AR motifs (Figure 4a) compared with other transcription factor motifs such as CTCF and neuron-restrictive silencer factor (NRSF) (Figures 4b,c). A symmetrical depletion of DNase-seq signal was detected around AR motifs in DHS sites that closely matches the information content of the AR binding motif dimer (Figure 4a, red line) [42]. In poised AR binding sites, we observed a similar pattern of

protection despite lower overall DNase-seq signal intensity (Figure 4a, blue line). Binding sites that became available only after androgen induction only exhibited the footprint after androgen treatment (Figure 4d, blue line). Importantly, the overall enrichment of DNase signal in LNCaP-induced cells is specific to DHS regions that bind the AR and have an AR motif, as opposed to all DHS sites (Figure 4e). The observed evidence of AR motif protection prior to androgen induction (Figure 4a) may represent binding of an alternate factor that is displaced upon AR activation, such as has been reported for specific loci by GATA binding protein 2 [17]. From the compendium of cell lines that have been processed for DNase-seq through the ENCODE project, we identified H1 embryonic stem cells and D721 medulloblastoma cells as having relatively low expression levels of the AR. DNase-seq signal around AR motifs within DHS sites in these two cell lines resemble that of LNCaP cells prior to hormone treatment

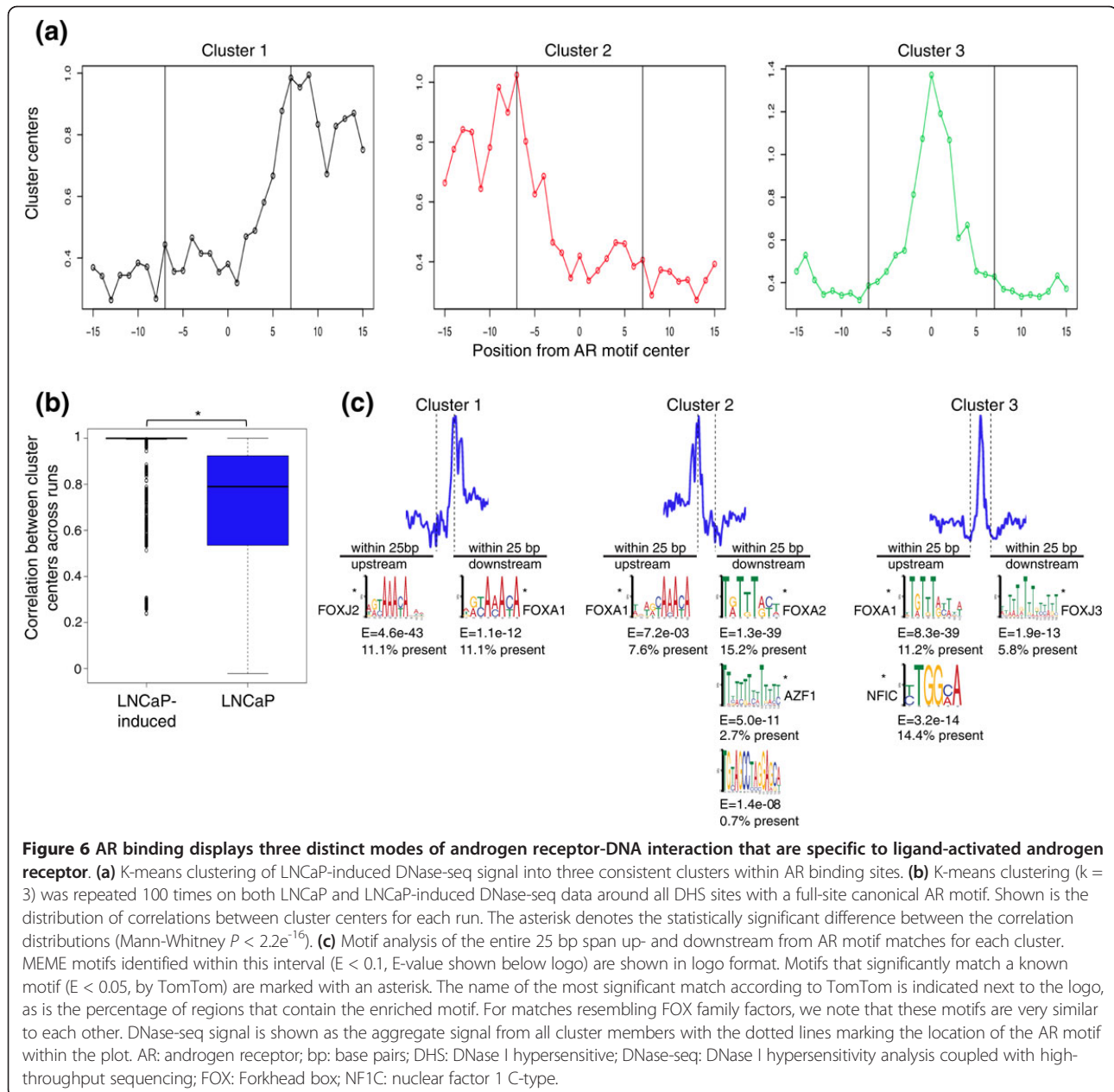


(Figure S4A in Additional file 1), suggesting that such a protection pattern in non-AR activated cell lines could result from alternative transcription factor binding to DNA at these regions.

To further investigate the AR footprint we performed k-means clustering to search for discrete DNase-seq patterns around AR motif matches (Materials and methods). DNase-seq signal was represented by a vector of DNase I cuts spanning 15 bp around the center of the AR motif. We identified three reproducible clusters, each of which represented part of the observed composite footprint (Figure 6a). These clusters were much less frequently detected across repeated iterations of clustering in untreated LNCaP cells. To quantify the degree to which these three patterns were present in LNCaP-induced cells compared with untreated cells, we examined the correlation between cluster centers obtained by performing k-means clustering 100 times for induced and uninduced LNCaP DNase-seq data. Specifically, the correlation of each cluster center to the cluster centers from all previous iterations was computed. Correlations tightly distributed around 1.0 represent highly reproducible clusters across different runs, suggesting that the three patterns are robust and consistently observed at AR motifs. Correlations loosely distributed about values less than 1.0 indicate that the three DNase-seq patterns at AR motif matches

are less reproducible. We found this correlation distribution to be significantly higher (Mann-Whitney  $P < 2.2e^{-16}$ ) for LNCaP-induced cells (Figure 6b), with the most robust clustering associated with AR binding (Figure S4B in Additional file 1) (Mann-Whitney  $P < 0.001$  between each column of correlations). Increasing the value of  $k$  consistently identified the same three general patterns in LNCaP-induced DNase-seq data within the AR motif, with multiple clusters aggregating to each general pattern (Figure S4E in Additional file 1). Using correlation analysis to analyze clusters from different values of  $k$  revealed that  $k = 3$  is the most appropriate value (Figure S4F in Additional file 1), supporting that three distinct patterns of DNase I cleavage exist within AR motifs. Overall, the three distinct patterns of DNase I protection appeared to be a robust phenomenon more often detected in LNCaP-induced DNase-seq data, suggesting that AR activation stabilizes specific chromatin structure around AR motifs.

AR binding has been associated with enrichment of palindromic full-site AR motifs (such as depicted in Figure 4a) as well as half-site motifs [43,44]. The directional footprinting in clusters 1 and 2 is indicative of only half of the full canonical AR motif being protected from DNase I cleavage, whereas cluster 3 is consistent with full-site protection. Our ability to detect this indicates that specific half-site usage is consistent across the



entire population of cells, and does not fluctuate randomly. The spike in the center of cluster 3 corresponds to the degenerate bases in the middle of the AR motif, indicating reduced DNA protection between AR proteins within the dimer. A recent report examining the dynamics of AR dimerization showed, in an exogenous system, that the AR binding enhancer element of *TMPRSS2* requires an AR dimer. Consistently, we observed a DNase-seq digestion pattern similar to that shown in cluster 3 within this enhancer element (Figure S4D in Additional file 1).

While we posited that full-site protection might reflect a stronger AR-DNA association, AR ChIP-seq peak scores

were evenly distributed between the three clusters, suggesting similar binding strength (Figure S4C in Additional file 1). We next explored if each cluster exhibited different co-factor motif enrichment by *de novo* motif analysis of the 25 base pairs upstream and downstream of the motif clusters (Figure 6c). Within these intervals, we detected more significant enrichment of FOX family motifs in the highly protected portions (dips) of clusters 1 and 2. A motif consistent with NF1C, which was also detected in an analysis of  $\Delta$ DNase regions (Table S3 in Additional file 1), was enriched only upstream of cluster 3. These analyses suggest that the two well-defined dips observed around the composite footprint (Figure 4a) correspond to

FOX factor-mediated DNA protection, which is seen to a more noticeable degree in LNCaP-induced DNase-seq data. Overall, our footprinting analysis revealed three different stable modes of DNase-seq protection with AR binding that represent two phenomena: full- or half-site protection at full-site DNA motifs.

## Discussion

The AR is a transcription factor and a primary driver of prostate cancer. Understanding the key determinants of its transcriptional specificity remains a critical issue. By integrating analysis of DNase-seq data with AR ChIP-seq and mRNA-seq, we showed that AR activation induced genome-wide changes in chromatin structure that were associated with AR binding and transcriptional response. We also uncovered multiple modes of AR utilization of its DNA recognition motif. Although a subset of AR binding occurs in qualitatively poised chromatin exhibiting nucleosome depletion prior to hormone treatment, we demonstrated that AR binding is consistently associated with a quantitatively significant increase in DNase-seq signal, suggesting stabilization of nucleosome depletion and chromatin remodeling.

Several prior reports also support AR-induced chromatin remodeling [16,17], including a very recently published study utilizing DNase-seq by He *et al.* [30]. Our data combined with these prior reports suggest a different model for nuclear receptor interaction with the genome than that proposed by John *et al.* for the GR [29], where almost all GR binding occurred in poised DHS sites. The AR and GR, though possessing similar DNA-response elements, seem to display fundamentally different interactions with chromatin and DNA. Our data represent a significant additional resource for understanding the association between chromatin accessibility and nuclear receptor function for several reasons. First, our DNase-seq experiments were sequenced very deeply (approximately 130 million reads), which is similar to the depth of sequencing with which John *et al.* observed GR binding to poised chromatin. Second, we utilized a different AR ligand (R1881) and time point of 12 hours as compared with 4 hours by He *et al.* and 1 hour by John *et al.* Similar to He *et al.*, who also utilized a quantitative measure of change in DNase-seq signal, we observed that less than half of AR binding targets poised chromatin and we were also able to associate AR-induced chromatin remodeling with AR-induced transcriptional changes, suggesting that the mechanism of chromatin remodeling and its phenotypically relevant association with differential transcription requires longer periods of receptor activation. Importantly, we used a different statistical measurement of quantitative change in DNase-seq signal to reach the same result and conclusion. In our study and those by He *et al.* and John *et al.*, we note that the degree of nuclear receptor binding

within regions of poised chromatin decreases with increased hormone treatment time (37% in He *et al.*, 88% in John *et al.* and 20% to 30% in our study). Although this observation is confounded by differences in receptor, receptor ligand, sequencing depth and DNase-seq protocol among the mentioned studies, these data suggest that more extensive comparative analyses over a full time course of ligand stimulation of both AR and GR are needed to fully understand the similarities and differences of different hormone receptors with respect to their interaction with chromatin.

While the majority of high confidence AR binding occurred in regions sensitive to DNase I cleavage, a substantial proportion of AR binding events occurred in regions of low DNase-seq signal. It is possible that inconsistent and/or intermittent nucleosome depletion at these genomic regions decreases DNA accessibility and limits detection by our assay; this attribute of nucleosome depletion appears to be associated with a slightly different AR motif. Consistently, we also found that AR binding (as measured by AR ChIP-seq signal intensity) is significantly lower in non-DHS regions than in DHS regions. Thus, it is plausible that regions that are identified with weaker AR binding and lower DNase-seq signal may experience a dynamic equilibrium of nucleosome and nuclear receptor binding, as has been previously proposed [14]. Loci with reduced DNase I cleavage and AR binding could reflect low levels of AR binding at linker regions of non-displaced nucleosomes or residual nucleosome occupancy, limiting accessibility to DNase I cleavage in the cell population.

AR footprinting analysis further revealed the complexity of the AR-DNA interaction. The aggregate DNase-seq signal around AR motifs demonstrated a relatively weak but consistent pattern of protection that corresponds to the expected binding pattern, consistent with other DNase I footprinting studies [26]. In addition, we found three distinct patterns of DNase I protection significantly associated with LNCaP cells treated with androgen. The footprint patterns suggest that either AR binds to the full AR consensus motif as a dimer (cluster 3) or only binds to half of the motif (clusters 1 and 2). We also cannot exclude the possibility that clusters 1 and 2 represent AR dimers with only one AR molecule binding to half of the consensus motif. AR binding to either half site did not appear to be random, as evidenced by reproducible detection of distinct clusters. In other words, random binding to either half site in a population of cells would not show consistent half-site protection. Intriguingly, clusters 1 and 2 may provide the first *in vivo* and endogenous evidence of functional AR monomers that have been suggested to exist as a stable subpopulation of AR molecules [45]. Only the AR binding sites that displayed a full-site dimer protection pattern (cluster 3) were enriched for the NF1C motif, which is a known co-factor of AR. Therefore, there

appears to be multiple modes that AR binds to canonical DNA motifs *in vivo*, and these modes are associated with different co-factors. These observations are consistent with a recently proposed model of a transient interaction between nuclear receptors such as the AR and DNA rather than a stronger and more stable AR-DNA interaction [46]. Our analysis also provides the first evidence of substructure within a nuclear receptor footprint

The dynamics of AR-DNA binding are likely impacted by additional co-factors that may facilitate AR binding directly or indirectly. Distal regulatory elements identified by DNase-seq displayed an enrichment of SP1 and E2A/TCF3 motifs within DHS specifically accessible in LNCaP cells compared with 113 independent cell lines. TCF3, a basic helix loop helix factor involved in Wnt/ $\beta$ -catenin signaling [47,48], represents a new putative co-factor for the AR that warrants further investigation to understand its role in AR-mediated chromatin dynamics as well as the crosstalk between AR and  $\beta$ -catenin signaling. SP1 is especially interesting both because its motif was enriched in  $\Delta$ DNase regions and also in light of a recent report that identified SP1 as necessary for the expression of a variety of chromatin modifying enzymes, such as the histone deacetylases 1 to 4 in LNCaP cells [49]. Additionally, small molecule inhibitors of histone deacetylases have been shown to decrease the growth rate of AR-positive prostate cancer cell lines [50,51] and disrupt AR-induced expression of its target genes [52]. Our relative enrichment score of less than one for the SP1 motif and an observation that SP1 motifs often co-localize with AR binding suggest complexity in the interplay between SP1 and the AR.

## Conclusions

Overall, these lines of evidence combined with our results warrant further investigation of SP1 in the context of AR binding and AR-induced chromatin remodeling. Our analyses show that qualitative and quantitative assessment of chromatin accessibility by DNase-seq is an important and useful tool for elucidating AR biology in prostate cancer cell line models.

## Materials and methods

### Cell culture

LNCaP cells were obtained from ATCC and maintained according to manufacturer instructions. Prior to cell treatment with either 1 nM R1881 (methyltrienolone) or vehicle (ethanol), cells were grown in Roswell Park Memorial Institute-1640 media containing 10% charcoal-dextran stripped serum for 60 hours.

### DNase-seq library generation and analysis

DNase-seq was performed as previously described [24,53]. Briefly,  $10 \times 10^6$  cells were harvested for each condition ( $\pm$

androgen). Nuclei were extracted and digested with optimal concentrations of DNase I enzyme. After confirmation of adequate digestion, DNase I-digested ends were blunt ended, and a biotinylated linker was ligated to these ends. Fragments with linker attached were isolated, digested with MmeI, and captured using streptavidin-conjugated magnetic beads. A second linker was ligated to the MmeI-digested end, and then the fragments were amplified and subsequently purified via gel electrophoresis. These sequencing libraries were sequenced on the Illumina GAIIx sequencing platform (Illumina, San Diego, CA, USA). Three biological replicates were processed for each cell growth condition. Sequencing results were aligned to the human reference genome (NCBI Build 37) using the Burrows-Wheeler aligner (BWA)[54]. Alignments were filtered to remove problematic repetitive regions such as alpha satellites and PCR artifacts characterized by many sequences mapped to small genomic locations. Biological replicates were compared for reproducibility, and then combined. In our cross-replicate analysis, we determined one of three biological replicates of LNCaP-induced DNase-seq to be discordant from the other two biological replicates and thus removed that replicate from the combined DNase-seq sequence set. The final base-pair resolution signal to reflect chromatin accessibility was generated using F-seq [31]. Discrete peaks were called by fitting DNase-seq signal data to a gamma distribution and then determining the signal value that corresponded to  $P < 0.05$ . Gene-relative categories were defined as previously described [24].

### Identification of increases and decreases in DNase-seq signal

To determine regions of significant change in DNase-seq signal with androgen induction, we used the edgeR bio-conductor package [32,55]. The edgeR package is designed to detect differences in count data among groups of samples containing biological and technical replicates. Prior to running the algorithm, we defined windows in which to compare DNase-seq signal across replicates by first taking the union set of all identified DHS sites in both LNCaP and LNCaP-induced cells.

This approach allows for inclusion of regions that contain an increase or decrease in DNase-seq signal such that they cross the threshold defining a DHS site. The defined union set was then divided into overlapping windows of 300 bp. DHS regions smaller than the window size were expanded to the window size. Regions larger than the window size were tiled with overlapping windows, where the overlap varies depending on the size of the hypersensitive region to tile. We start by finding the number of windows that would fit completely inside the defined DHS site using the default overlap. If these windows discard fewer than 10% of the bases on each edge

of the DHS site, we tile the site using these windows. If using the default overlap would cause us to lose more than this edge threshold, we add another window and adjust the overlap so that the windows exactly cover the entire DHS region. We find that these windows cover almost all of the DHS bases in the original, while minimizing the number of non-DHS bases considered for the downstream analysis. Our approach created approximately 550,000 windows for differential analysis among five replicates (three LNCaP, two LNCaP-induced). The number of tags mapping to each window in each replicate were extracted, and regions with a sum total of less than five reads were eliminated. We then used edgeR to call windows with significantly different counts in each pairwise comparison at two thresholds: strict (FDR < 0.05) and loose (unadjusted  $P < 0.05$ ). Finally, neighboring windows that were identified as having a significantly higher DNase-seq signal in a condition were merged. To generate a normalized differential tag count for regions, the number of DNase-seq tags within each LNCaP and LNCaP-induced DHS region was determined and normalized to the average number of tags in either all LNCaP or LNCaP-induced DHS site. For each of the 175,796 union DHS regions, the normalized number of tags in LNCaP in the region was subtracted from the normalized number of tags in LNCaP-induced to give the differential tag score for each region.

#### RNA expression analysis

RNA expression in response to androgen induction in LNCaP was analyzed using both exon microarrays and mRNA-seq. Total RNA was extracted using TRIzol (Sigma-Aldrich, St. Louis, MO, USA) from the same cell growth as used for DNase-seq and hybridized to Affymetrix Exon 1.0 ST arrays (Affymetrix, Santa Clara, CA, USA) using a standard protocol. Resulting .CEL files were summarized into expression measures at the gene-core level using Affymetrix Power Tools and Robust Multi-array Average (RMA) normalization [56]. Differential expression upon androgen induction was determined using the Statistical Analysis of Microarrays bioconductor package [57]. Two biological replicates were processed for exon array analysis.

RNA for use in mRNA-seq was isolated from three independent growths of LNCaP with or without androgen using the Ambion miRVANA miRNA isolation kit (Invitrogen, Grand Island, NY, USA). Induction of canonical AR target gene expression was confirmed by qPCR, and RNA quality was verified using an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). All RNA used for subsequent library preparation had an RNA integrity number greater than 9.0. mRNA-seq libraries were created using the Illumina mRNA-seq protocol and

kit then sequenced on the Illumina GAIIX platform. Resulting sequence data was aligned to the human reference genome (NCBI Build 37) first using BWA [54]. Reads unaligned by BWA were independently aligned with TopHat [58] to only known and annotated splice junctions. The results combined and filtered to remove non-unique reads. Technical replicates were merged such that three biological replicates ( $\pm$  androgen) were available for subsequent analysis. The reads per kilobase mapped expression measure was computed for each RefSeq gene model in each replicate, leaving out tags mapping to the 3' untranslated region of genes [59]. To identify RefSeq genes differentially expressed between LNCaP and LNCaP-induced cells, we first counted the number of mRNA-seq tags that fell within exons of RefSeq gene models in each biological replicate, resulting in a tag count value for each gene. We then used the edgeR bioconductor package to detect genes whose expression differed with AR activation, FDR < 0.05.

#### Correlation of $\Delta$ DNase increases and decreases with expression increases and decreases

To establish the relationship between differential chromatin and differential expression, we tested for significance in overlap in both directions, that is, we tested if  $\Delta$ DNase regions tend to be located near differentially expressed genes, and then tested if differentially expressed genes tend to have a  $\Delta$ DNase region nearby. We first assigned each DHS site to its nearest gene and intersected these nearest genes with each AR-regulated gene set (AR mRNA-seq upregulated, AR mRNA-seq downregulated, and four sets from Massie *et al.* [40]: Massie early up, Massie early down, Massie late up, Massie late down). We calculated the significance of the  $\Delta$ DNase association to differentially regulated genes by permuting the set of all RefSeq genes 100,000 times, randomly selecting the number of genes for each comparison, and intersecting those random sets with the genes related to AR-induced expression changes. This established a null distribution of overlaps in random intersects. We also conducted the same analysis in the opposite direction to relate expression change  $\Delta$ DNase regions. Using the University of California Santa Cruz (UCSC) Known Genes table, we merged all isoform coordinates for each gene and found all  $\Delta$ DNase increase or decrease sites within a surrounding 20 kb window. We calculated pairwise overlaps of  $\Delta$ DNase sites between each  $\Delta$ DNase increase or decrease list and these lists of all nearby  $\Delta$ DNase sites. If a gene contained a  $\Delta$ DNase site within 20 kb, it was counted as a match. We then permuted genes located all nearby DHS sites 1,000 times, and tested for overlap to create a null distribution of overlap count.

### Androgen receptor ChIP-seq

AR ChIP-seq data was obtained from accession numbers [GSE14097] and [GSE28126] through the NCBI Gene Expression Omnibus portal. [GSE28126] was recently published [40] as part of a study identifying AR ChIP-seq peaks in LNCaP cells after 4 hours of 1 nM R1881 stimulation. [GSE14097] [39] contains AR ChIP-seq data from LNCaP cells treated with either 10 nM R1881 for 16 hours or vehicle for the same length of time. Coetzee AR ChIP-seq was conducted after 4 hours of either 10 nM dihydrotestosterone or ethanol treatment of LNCaP cells [17,41]. Raw sequence files were processed through the same pipeline as our DNase-seq data [31] to obtain aligned sequences. Model-based analysis for ChIP-seq (MACS) [60] was used with default parameters to identify regions of significant AR ChIP enrichment in LNCaP-induced relative to LNCaP. To generate the common AR peaks list, we intersected the peak calls from the two data sets, considering peaks to be overlapping if they intersect by at least one base pair.

### Self-organizing maps

To identify DHS sites specific to our two cell types, we used a SOM built from DNase-seq data generated by our group from 113 lines [61]. SOMs are artificial neural networks that learn patterns in data by iteratively assigning data points to cluster centers. The SOM eventually assigns each DHS site to a cluster with the most similar hypersensitivity profile. We are using SOMs to characterize DNase I hypersensitivity profiles across over 100 cell lines (NS, in preparation). For this analysis, we were interested in clusters specific to LNCaP cell lines. We first built a data matrix by counting the number of reads mapping in each peak region in each cell type. We quantile-normalized the scores by cell type and then capped them at the 99th quantile (by setting the top 1% of scores to a maximum value), and then row-scaled the scores to a decimal between 0 and 1. After normalization, capping and scaling, we built an SOM using the Kohonen package in R. This SOM used a hexagonal  $50 \times 50$  grid (for 2,500 total nodes). We then took each node and selected the 50 regions closest in distance to the node center, and submitted these to MEME for *de novo* motif analysis [62]. We then matched these motifs to publicly available DNA binding motifs in TRANSFAC 7.0 [63] and JASPAR 2010 [64] using STAMP [65].

### Motif analysis

To determine motif enrichment in regions of interest we utilized used three algorithms: MEME, cERMIT [66], and CentDist [67]. MEME and cERMIT report identified DNA motifs not matched to known motifs. If MEME was run on regions falling within DHS sites (all but analysis in Figure 3e), we used a first order background model

common to DHS sites. cERMIT was run using  $\Delta$ DNase *P*-value as evidence for directing motif analysis. CentDist identifies motifs enriched within a region and ranks them relative to their distribution within each region, reporting publicly available motifs that are found in regions. Motifs identified in Figures 2 and 3 were identified in at least two of these algorithms. If a motif was reported as enriched by MEME and cERMIT, it was included in our results if its match to publicly known motifs, determined by STAMP or TomTom (using JASPAR and Uniprobe databases) [68], was significant ( $E < 0.05$ ). Results from CentDist are shown in Table S3 in Additional file 1.

### Androgen receptor footprint analysis

To generate an aggregate plot of DNase-seq signal around AR motifs, we scanned DHS regions containing AR binding sites using a first order log likelihood scanner with a slightly modified AR position weight matrix (PWM) from the JASPAR database. The MA0007.1 matrix was trimmed by discarding low information base pair positions surrounding positions 4 to 18, resulting in a 15 bp PWM. PWM motif scores that scored lower than the 90<sup>th</sup> percentile of all match scores were discarded, and the strand with the stronger PWM match was chosen if both strands at a location matched the PWM within the 90<sup>th</sup> percentile or higher. DNase-seq reads mapping to each base at the motif site and surrounding 100 bp were collected and the sum of each position was calculated.

For cluster correlation analysis, the k-means algorithm [69] was run 100 times to yield  $k \times 100$  cluster centers, where each cluster center is a vector of values of length 31 (clustering was performed on DNase-seq signal mapping to 15 bp on both sides of center of AR motif). Cluster centers from one run *i* to all other runs (*i*, ..., *i-1*, *i+1*, ..., 100) were compared. Each cluster center from a single run was matched to another cluster center in another run in a pairwise manner that identifies maximum correlation; this procedure was performed across all pairs of runs to assess the similarity and reproducibility of results over multiple runs of the algorithm.

For aggregate visualization of clusters, we tabulated DNase-seq tag counts 100 bp around AR PWM matches classified into each cluster within 'R1881 intersect' ChIP peaks that were DHS in both LNCaP-induced and LNCaP. MEME was used to search for *de novo* motifs 25 bp up- and downstream from PWMs classified into each cluster. TomTom was used to match significant motif matches to publicly available motifs ( $E < 0.05$ ).

### Data access

DNase-seq data from this study can be visualized using the UCSC Genome Browser [70]. Specifically, click on the 'Genome Browser' option then click on the 'configure tracks and display' button. Under the section entitled

'Regulation', find the subsection 'ENC DNase/FAIRE' for ENCODE Open Chromatin by DNase I HS and FAIRE, and then click on the 'Duke DNaseI HS' link. In the menu of cell lines that will appear, click on the two boxes in the row labeled LNCaP (one for 'no treatment' and one for 'methyltrienolone (androgen)'), and this will allow for data visualization. Processed and raw DNase-seq data has also been deposited through the NCBI GEO website under accession number [GSE32970]. Within that accession number, data for LNCaP is available under [GSM816637] (Duke\_DnaseSeq\_LNCaP), and data for LNCaP-induced is available under [GSM816634] (Duke\_DnaseSeq\_LNCaP\_androgen). Exon array expression data is publicly available through the NCBI GEO website under accession number GSE15805. Within this accession number, the two LNCaP replicates are under GSM443919 and GSM443920, and the two LNCaP-induced replicates are under GSM443921 and GSM443922. mRNA-seq data from this study is available through the NCBI GEO website under accession number GSE34780.

## Additional material

**Additional file 1: Supplemental information containing four supplemental figures and three supplemental tables.**

**Additional file 2: Table S4 detailing RefSeq genes identified as differentially expressed in response to androgen receptor activation by mRNA-seq.** mRNA from three biological replicates of LNCaP and three biological replicates of LNCaP-induced was extracted and sequenced. Genes differentially regulated by AR activation were determined using edgeR at a threshold of FDR < 0.05.

**Additional file 3: Table S5 showing a comparison of androgen receptor-regulated genes as determined by mRNA-seq to previously published androgen receptor-regulated gene lists.** Genes identified as differentially regulated by AR activation are shown along with all other RefSeq genes, ranked in order by the number of published studies that identify the gene as AR-regulated.

## Abbreviations

AR: androgen receptor; bp: base pairs; BWA: Burrows-Wheeler Aligner; ChIP-seq: chromatin immunoprecipitation coupled with high-throughput sequencing; CTCF: CCCTC-binding factor; DHS: DNase I hypersensitive; DNase-seq: DNase I hypersensitivity analysis coupled with high-throughput sequencing; FDR: false discovery rate; FOX: Forkhead box; FAIRE: formaldehyde-assisted isolation of regulatory elements; GR: glucocorticoid receptor; kb: kilobase pairs; mRNA-seq: messenger RNA abundance measured by high-throughput sequencing; NF1C: nuclear factor 1 C-type; PCR: polymerase chain reaction; PWM: position weight matrix; qPCR: quantitative polymerase chain reaction; RMA: Robust Multi-array Average; SOM: self-organizing map.

## Acknowledgements

We thank Darin London, Mark DeLong and Alan Cowles from the Duke Institute for Genome Sciences and Policy for their invaluable technical support; Lisa Bukovnik and Fangfei Ye from the Duke Sequencing Core facility for their assistance with next generation sequencing; Boris Lenhard for his assistance and insight into self-organizing maps and motif analysis; and Bala Balakumaran as well as the other past and current members of the Febbo and Crawford laboratories for their advice and encouragement. This work was supported by Department of Defense grants W81XWH-11-1-0555

to AKT and W81XWH-07-1-0033 and W81XWH-10-1-0821 to PGF, a National Science Foundation Student Fellowship and a grant from the Research Council of Norway to NCS, NIH grants R01CA109147 and R01CA136924 to GAC, and NIH/NHGRI ENCODE Consortium U54 HG004563 to TSF and GEC.

## Author details

<sup>1</sup>Institute for Genome Sciences & Policy, Duke University, Durham, NC 27708, USA. <sup>2</sup>Helen Diller Family Comprehensive Cancer Center, University of California at San Francisco, San Francisco, CA 94115, USA. <sup>3</sup>Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA. <sup>4</sup>Department of Preventive Medicine, Norris Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA. <sup>5</sup>Department of Urology, Norris Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA. <sup>6</sup>Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27708, USA. <sup>7</sup>Department of Computer Science, Duke University, Durham, NC 27708, USA. <sup>8</sup>Departments of Biology and Genetics, Carolina Center for Genome Sciences and Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. <sup>9</sup>Department of Pediatrics, Division of Medical Genetics, Duke University, Durham, NC 27708, USA. <sup>10</sup>Department of Medicine, University of California at San Francisco School of Medicine, San Francisco, CA 94115, USA. <sup>11</sup>Department of Urology, University of California at San Francisco School of Medicine, San Francisco, CA 94115, USA.

## Authors' contributions

AKT, TSF, GEC and PGF designed the experiments. AKT and LS performed the experiments. GAC contributed AR ChIP-seq data and revised the manuscript. AKT, GGY, YS, NCS, BST, SGG and UO analyzed the data. AKT, TSF, GEC and PGF wrote the manuscript. All authors revised the manuscript and approved the final version.

## Competing interests

The authors declare that they have no competing interests.

Received: 1 May 2012 Revised: 14 August 2012

Accepted: 3 October 2012 Published: 3 October 2012

## References

1. Stanbrough M, Leav I, Kwan PW, Bublej GJ, Balk SP: **Prostatic intraepithelial neoplasia in mice expressing an androgen receptor transgene in prostate epithelium.** *Proc Natl Acad Sci USA* 2001, **98**:10823-10828.
2. Berger R, Febbo PG, Majumder PK, Zhao JJ, Mukherjee S, Signoretti S, Campbell KT, Sellers WR, Roberts TM, Loda M, Golub TR, Hahn WC: **Androgen-induced differentiation and tumorigenicity of human prostate epithelial cells.** *Cancer Res* 2004, **64**:8867-8875.
3. Zong Y, Xin L, Goldstein AS, Lawson DA, Teitell MA, Witte ON: **ETS family transcription factors collaborate with alternative signaling pathways to induce carcinoma from adult murine prostate cells.** *Proc Natl Acad Sci USA* 2009, **106**:12465-12470.
4. Carver BS, Tran J, Gopalan A, Chen Z, Shaikh S, Carracedo A, Alimonti A, Nardella C, Varmeh S, Scardino PT, Cordon-Cardo C, Gerald W, Pandolfi PP: **Aberrant ERG expression cooperates with loss of PTEN to promote cancer progression in the prostate.** *Nat Genet* 2009, **41**:619-624.
5. Goldstein AS, Huang J, Guo C, Garraway IP, Witte ON: **Identification of a cell of origin for human prostate cancer.** *Science* 2010, **329**:568-571.
6. Chen CD, Welsbie DS, Tran C, Baek SH, Chen R, Vessella R, Rosenfeld MG, Sawyers CL: **Molecular determinants of resistance to antiandrogen therapy.** *Nat Med* 2004, **10**:33-39.
7. Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, Shah RB, Chinnaiyan AM: **Integrative molecular concept modeling of prostate cancer progression.** *Nat Genet* 2007, **39**:41-51.
8. Wang Q, Li W, Zhang Y, Yuan X, Xu K, Yu J, Chen Z, Beroukhir M, Wang H, Lupien M, Wu T, Regan MM, Meyer CA, Carroll JS, Manraj AK, Janne OA, Balk SP, Mehra R, Han B, Chinnaiyan AM, Rubin MA, True L, Fiorentino M, Fiore C, Loda M, Kantoff PW, Liu XS, Brown M: **Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer.** *Cell* 2009, **138**:245-256.



9. Ling MT, Chan KW, Choo CK: **Androgen induces differentiation of a human papillomavirus 16 E6/E7 immortalized prostate epithelial cell line.** *J Endocrinol* 2001, **170**:287-296.
10. Horoszewicz JS, Leong SS, Kawinski E, Karr JP, Rosenthal H, Chu TM, Mirand EA, Murphy GP: **LNCAp model of human prostatic carcinoma.** *Cancer Res* 1983, **43**:1809-1818.
11. Craft N, Shostak Y, Carey M, Sawyers CL: **A mechanism for hormone-independent prostate cancer through modulation of androgen receptor signaling by the HER-2/neu tyrosine kinase.** *Nat Med* 1999, **5**:280-285.
12. Korenchuk S, Lehr JE, L MC, Lee YG, Whitney S, Vessella R, Lin DL, Pienta KJ: **VCaP, a cell-based model system of human prostate cancer.** *In Vivo* 2001, **15**:163-168.
13. Heemers HV, Tindall DJ: **Androgen receptor (AR) coregulators: a diversity of functions converging on and regulating the AR transcriptional complex.** *Endocr Rev* 2007, **28**:778-808.
14. Segal E, Widom J: **What controls nucleosome positions?** *Trends Genet* 2009, **25**:335-343.
15. Lupien M, Eeckhoutte J, Meyer CA, Wang Q, Zhang Y, Li W, Carroll JS, Liu XS, Brown M: **FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription.** *Cell* 2008, **132**:958-970.
16. He HH, Meyer CA, Shin H, Bailey ST, Wei G, Wang Q, Zhang Y, Xu K, Ni M, Lupien M, Mieczkowski P, Lieb JD, Zhao K, Brown M, Liu XS: **Nucleosome dynamics define transcriptional enhancers.** *Nat Genet* 2010, **42**:343-347.
17. Andreu-Vieyra C, Lai J, Berman BP, Frenkel B, Jia L, Jones PA, Coetzee GA: **Dynamic nucleosome-depleted regions at androgen receptor enhancers in the absence of ligand in prostate cancer cells.** *Mol Cell Biol* 2011, **31**:4648-4662.
18. Wu C: **The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I.** *Nature* 1980, **286**:854-860.
19. Gross DS, Garrard WT: **Nuclease hypersensitive sites in chromatin.** *Annu Rev Biochem* 1988, **57**:159-197.
20. Cockerill PN: **Structure and function of active chromatin and DNase I hypersensitive sites.** *FEBS J* 2011, **278**:2182-2210.
21. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD: **FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin.** *Genome Res* 2007, **17**:877-885.
22. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE: **High-resolution mapping and characterization of open chromatin across the genome.** *Cell* 2008, **132**:311-322.
23. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B: **Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.** *Nat Genet* 2007, **39**:311-318.
24. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D, Liu Z, London D, McDaniell RM, Shibata Y, Showers KA, Simon JM, Vales T, Wang T, Winter D, Clarke ND, Birney E, Iyer VR, Crawford GE, Lieb JD, Furey TS: **Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity.** *Genome Res* 2011, **21**:1757-1767.
25. Bhattacharyya N, Dey A, Minucci S, Zimmer A, John S, Hager G, Ozato K: **Retinoid-induced chromatin structure alterations in the retinoic acid receptor beta2 promoter.** *Mol Cell Biol* 1997, **17**:6481-6490.
26. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, Fields S, Stamatoyannopoulos JA: **Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.** *Nat Methods* 2009, **6**:283-289.
27. Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS: **High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells.** *Genome Res* 2011, **21**:456-464.
28. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK: **Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data.** *Genome Res* 2011, **21**:447-455.
29. John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA: **Chromatin accessibility pre-determines glucocorticoid receptor binding patterns.** *Nat Genet* 2011, **43**:264-268.
30. He HH, Meyer CA, Chen MW, Jordan VC, Brown M, Liu XS: **Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics.** *Genome Res* 2012.
31. Boyle AP, Guinney J, Crawford GE, Furey TS: **F-Seq: a feature density estimator for high-throughput sequence tags.** *Bioinformatics* 2008, **24**:2537-2538.
32. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26**:139-140.
33. Cleutjens KB, van der Korput HA, van Eekelen CC, van Rooij HC, Faber PW, Trapman J: **An androgen response element in a far upstream enhancer region is essential for high, androgen-regulated activity of the prostate-specific antigen promoter.** *Mol Endocrinol* 1997, **11**:148-161.
34. Safe S, Abdelrahim M: **Sp transcription factor family and its role in cancer.** *Eur J Cancer* 2005, **41**:2438-2448.
35. Chesire DR, Isaacs WB: **Ligand-dependent inhibition of beta-catenin/TCF signaling by androgen receptor.** *Oncogene* 2002, **21**:8453-8469.
36. Deng Z, Wan M, Cao P, Rao A, Cramer SD, Sui G: **Yin Yang 1 regulates the transcriptional activity of androgen receptor.** *Oncogene* 2009, **28**:3746-3757.
37. Tan SK, Lin ZH, Chang CW, Varang V, Chng KR, Pan YF, Yong EL, Sung WK, Cheung E: **AP-2gamma regulates oestrogen receptor-mediated long-range chromatin interaction and gene transcription.** *EMBO J* 2011, **30**:2569-2581.
38. Eisermann K, Tandon S, Bazarov A, Brett A, Fraizer G, Piontkivska H: **Evolutionary conservation of zinc finger transcription factor binding sites in promoters of genes co-expressed with WT1 in prostate cancer.** *BMC Genomics* 2008, **9**:337.
39. Yu J, Mani RS, Cao Q, Brenner CJ, Cao X, Wang X, Wu L, Li J, Hu M, Gong Y, Cheng H, Laxman B, Vellaichamy A, Shankar S, Li Y, Dhanasekaran SM, Morey R, Barrette T, Lonigro RJ, Tomlins SA, Varambally S, Qin ZS, Chinnaiyan AM: **An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression.** *Cancer Cell* 2010, **17**:443-454.
40. Massie CE, Lynch A, Ramos-Montoya A, Boren J, Stark R, Fazli L, Warren A, Scott H, Madhu B, Sharma N, Bon H, Zecchini V, Smith DM, Denicola GM, Mathews N, Osborne M, Hadfield J, Macarthur S, Adryan B, Lyons SK, Brindle KM, Griffiths J, Gleave ME, Rennie PS, Neal DE, Mills IG: **The androgen receptor fuels prostate cancer by regulating central metabolism and biosynthesis.** *EMBO J* 2011, **30**:2719-2733.
41. Berman BP, Frenkel B, Coetzee GA, Jia L: **Androgen receptor responsive enhancers are flanked by consistently-positioned H3-acetylated nucleosomes.** *Cell Cycle* 2010, **9**:2249-2250.
42. Shaffer PL, Jivan A, Dollins DE, Claessens F, Gewirth DT: **Structural basis of androgen receptor binding to selective androgen response elements.** *Proc Natl Acad Sci USA* 2004, **101**:4758-4763.
43. Tang Q, Chen Y, Meyer C, Geistlinger T, Lupien M, Wang Q, Liu T, Zhang Y, Brown M, Liu XS: **A comprehensive view of nuclear receptor cancer cistromes.** *Cancer Res* 2011, **71**:6940-6947.
44. Wang Q, Li W, Liu XS, Carroll JS, Janne OA, Keeton EK, Chinnaiyan AM, Pienta KJ, Brown M: **A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth.** *Mol Cell* 2007, **27**:380-392.
45. van Royen ME, van Cappellen WA, de Vos C, Houtsmuller AB, Trapman J: **Stepwise androgen receptor dimerization.** *J Cell Sci* 2012, **125**:1970-1979.
46. Voss TC, Schiltz RL, Sung MH, Yen PM, Stamatoyannopoulos JA, Biddie SC, Johnson TA, Miranda TB, John S, Hager GL: **Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism.** *Cell* 2011, **146**:544-554.
47. Merrill BJ, Pasolli HA, Polak L, Rendl M, Garcia-Garcia MJ, Anderson KV, Fuchs E: **Tcf3: a transcriptional regulator of axis induction in the early embryo.** *Development* 2004, **131**:263-274.
48. Yi F, Pereira L, Hoffman JA, Shy BR, Yuen CM, Liu DR, Merrill BJ: **Opposing effects of Tcf3 and Tcf1 control Wnt stimulation of embryonic stem cell self-renewal.** *Nat Cell Biol* 2011, **13**:762-770.
49. Chen CH, Huang PH, Chu PC, Chen MC, Chou CC, Wang D, Kulp SK, Teng CM, Wang Q, Chen CS: **Energy restriction-mimetic agents induce apoptosis in prostate cancer cells in part through epigenetic activation of KLF6 tumor suppressor gene expression.** *J Biol Chem* 2011, **286**:9968-9976.
50. Butler LM, Agus DB, Scher HI, Higgins B, Rose A, Cordon-Cardo C, Thaler HT, Rifkin RA, Marks PA, Richon VM: **Suberoylanilide hydroxamic acid, an inhibitor of histone deacetylase, suppresses the growth of prostate cancer cells in vitro and in vivo.** *Cancer Res* 2000, **60**:5165-5170.
51. Rokhlin OW, Glover RB, Guseva NV, Taghiyev AF, Kohlgraf KG, Cohen MB: **Mechanisms of cell death induced by histone deacetylase inhibitors in**

- androgen receptor-positive prostate cancer cells. *Mol Cancer Res* 2006, **4**:113-123.
52. Welsbie DS, Xu J, Chen Y, Borsu L, Scher HI, Rosen N, Sawyers CL: **Histone deacetylases are required for androgen receptor function in hormone-sensitive and castrate-resistant prostate cancer.** *Cancer Res* 2009, **69**:958-966.
53. Song L, Crawford GE: **DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells.** *Cold Spring Harb Protoc* 2010, **2010**, pdb prot5384.
54. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754-1760.
55. Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, Tewari AK, London D, Song L, Lee BK, Iyer VR, Parker SC, Marguiles EH, Wray GA, Furey TS, Crawford GE: **Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection.** *PLoS Genet* 2012, **8**:e1002789.
56. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4**:249-264.
57. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98**:5116-5121.
58. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25**:1105-1111.
59. Ramskold D, Wang ET, Burge CB, Sandberg R: **An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data.** *PLoS Comput Biol* 2009, **5**:e1000598.
60. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol* 2008, **9**:R137.
61. Kohonen T: **Cortical maps.** *Nature* 1990, **346**:24.
62. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching.** *Nucleic Acids Res* 2009, **37**:W202-208.
63. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34**:D108-110.
64. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A: **JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles.** *Nucleic Acids Res* 2010, **38**:D105-110.
65. Mahony S, Benos PV: **STAMP: a web tool for exploring DNA-binding motif similarities.** *Nucleic Acids Res* 2007, **35**:W253-258.
66. Georgiev S, Boyle AP, Jayasurya K, Ding X, Mukherjee S, Ohler U: **Evidence-ranked motif identification.** *Genome Biol* 2010, **11**:R19.
67. Zhang Z, Chang CW, Goh WL, Sung WK, Cheung E: **CENTDIST: discovery of co-associated factors by motif distribution.** *Nucleic Acids Res* 2011, **39**:W391-399.
68. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS: **Quantifying similarity between motifs.** *Genome Biol* 2007, **8**:R24.
69. Hartigan JA, Wong MA: **Algorithm AS 136: a K-means clustering algorithm.** *Journal of the Royal Statistical Society Series C (Applied Statistics)* 1979, **28**:100-108.
70. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12**:996-1006.

doi:10.1186/gb-2012-13-10-R88

**Cite this article as:** Tewari et al.: Chromatin accessibility reveals insights into androgen receptor activation and transcriptional specificity. *Genome Biology* 2012 **13**:R88.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

