# Agonist and antagonist switch DNA motifs recognized by human androgen receptor in prostate cancer

**Zhong Chen, Xun Lan, Jennifer M. Thomas-Ahner, Dayong Wu, Xiangtao Liu, Zhenqing Ye, Liguo Wang, Benjamin Sunkel, Cassandra Grenade, Junsheng Chen, Debra L.Zynger, Pearlly S. Yan, Jiaoti Huang, Kenneth P. Nephew, Tim H.-M. Huang, Shili Lin, Steven K. Clinton, Wei Li, Victor X. Jin, Qianben Wang**

## Supplementary Figures

**Figure S1.** Schematic diagram of the modified ChIP-exo approach. After ChIP with a specific antibody recognizing transcription factor A (TF-A) within a transcription complex (only transcription factor B [TF-B] is shown here), sonicated ends of DNA fragments will be labeled by ligating A2 adaptors. Lambda exonuclease (5' → 3') (exo) will then be used to digest single-stranded DNA up to the point where it is protected by the transcription complex (border). After reverse crosslinking and A2 extension to generate double-stranded DNA, A1 sequencing adapters will then be ligated to the exo-digested ends. Sequencing will be performed from both the 5' exo-digested end (R1) and 3' sonicated end (R2). This paired-end sequencing will allow a precise definition of multiple transcription factor motifs between multiple borders (blue and red shown in the figure) within a binding peak using our developed border pattern-based motif defining approach (See "ChIP-exo", "ChIP-exo peak and border analysis" and "ChIP-exo motif analysis" in Expanded Experimental Procedures).

**Figure S2.** An overview of AR ChIP-exo.

(A) Bioanalyzer (Agilent) analysis of non-size-selected ChIP-exo libraries before sequencing.

(B) The length of DNA fragments sequenced was calculated using PE information.

(C) The distribution of two end reads on two strands was shown as the distance to the center of enriched locations (top 10,000 locations were used).

(D) Correlation between biological replicates of AR ChIP-exo under vehicle treatment.

(E) Correlation between biological replicates of AR ChIP-exo under 10 nM DHT treatment.

(F) Enriched locations identified in AR ChIP-exo.

(G) Profile of location length.

(H) Profile of borders found in locations.

(I) Identified locations and borders.

(J-L) Comparison of ChIP-exo, ChIP-seq, and DNase-seq at the same sequencing depth. Tag density correlation (ln) showing the quantitative relationship between ChIP-exo and ChIP-seq (J, r=0.7252, p=0.000), ChIP-exo and DNase-seq (K, r=0.6644, p=0.000), and ChIP-seq and DNase-seq (L, r=0.9109, p=0.000).

(M) Comparison of tag distribution between ChIP-exo, ChIP-seq and DNase-seq. The window indicates $\pm$ 500 bp regions from the center of AR binding locations.

(N) *de novo* motif analysis of protected sequences showed that the most significantly enriched consensus sequence resembles the Forkhead motif. The second most significantly enriched motif is the half-site ARE. JASPAR database was used for motif comparison.

(O) Comparison of ChIP-exo signal distribution at the same locations between AR ChIP-exo and FOXA1 ChIP-exo. To confirm whether those Forkhead motifs defined in AR ChIP-exo are bound by FoxA1 *in vivo*, we performed FoxA1 ChIP-exo in LNCaP cell line with or without 10 nM DHT treatment for 4 hr. Raw tag distribution over Forkhead motif (within 80 bp) was shown for the forward (blue) and reverse (red) strands, separately.

(P and Q) The aggregated tag density for AR ChIP-exo (P) and FOXA1 ChIP-exo (Q) over Forkhead motif (within 80 bp) shown for the forward and reverse strand signal separately.

**Figure S3.** Motif analysis on agonist-liganded ARBEs and downstream sequences.

(A) *de novo* motif analysis was performed using short border-extended sequences. ARBE candidates were selected based on the similarity with JASPAR database or each other.

(B) Overview of the relative nucleotides frequencies downstream of ARBEs (i.e. $10^{th}$-$15^{th}$ nucleotides).

(C) *de novo* motif analysis was performed using 9 bp sequences downstream of ARBEs.

(D) 3 full-length ARBE motif logos obtained from our AR ChIP-seq data, a published study (Denayer et al, 2010) and JASPAR database.

(E) The percentage of canonical full-length ARBEs calculated by scanning 3 ARBE matrices in AR ChIP-exo defined ARBEs (precisely defined ARBE plus downstream extension).

(F) Comparison of canonical full-length ARBEs with full-length ARBEs comprised of two precisely defined ARBEs.

(G) Detailed information of full-length ARBEs comprised of two precisely defined ARBEs.

(H) The four types of ARBEs were used for scanning within (100 bp) or outside (2 kb) of peak regions identified in AR ChIP-seq, regardless of border signal. Same numbers of motif sequences of each set (I, II, and III) were shown.

(I) Example of ARBE2 showing raw tag distribution over ARBE2-matrix sequences defined only in regular ChIP-seq, demonstrating border-dependent ARBE discovery.

(J) Canonical full-length ARBEs were calculated by scanning 15 bp extended defined ARBE sequences within and outside of AR ChIP-seq peaks with the ARE matrix obtained from our AR ChIP-seq data. *(K-M)* Clustering of ARBEs based on the motif probability matrix within or between second-half ARBEs. ARE matrices either developed using our AR ChIP-seq or obtained from JASPAR were applied to further matrix computation. ARBE2 (K), ARBE3 (L) and ARBE4 (M) were classified into 3 categories: canonical full-length, non-canonical full-length, and half-site-like ARBEs.

**Figure S4.** Characterization of agonist-liganed ARBEs.

(A) The distribution of each type of ARBE at AR binding locations. The *x*-axis represents the distance to the center of the locations.

(B) Averaged MNase-seq tag distribution of H3K4me1, H3K4me3, H3K9me3 and H3K27me3 over precisely defined ARBEs in AR binding locations. The *x*-axis represents the distance to the motifs.

(C) Comparison of ChIP-exo signal between the forward (upstream) strand and the reverse (downstream) strand at ARBE locations. The *x*-axis represents the distance to the motifs.

(D) Relative ChIP-exo signal density was calculated within 100 bp upstream of motifs for the forward strand, and 100 bp downstream of motifs for the reverse strand, respectively.

(E) Aggregated sequencing tag distribution of H3K4me2 MNase ChIP-seq (Vehicle *vs.* DHT treatment) over 3 clusters of each type of ARBE locations.

(F) Heat map showing the supervised clustering of the 4 types of ARBE locations using different active histone marks (H3K4me1 and H3K4me3). The order of sequences in each cluster of ARBEs is the same as the order of sequences shown in Figure 4B.

(G) Heat map showing DNase-seq tag distribution around the 4 types of ARBEs within ± 1 kb. The order of sequences in each cluster of ARBEs is the same as the order of sequences shown in Figure 4B.

(H) AR ChIP-exo defines 3 types of high-confidence FOXA1 motifs. Numbers indicate motif counts. The first two columns of each motif panel show raw tags distribution over FOXA1 motifs (within 80 bp) on the forward (blue) and reverse (red) strands, separately. The aggregated tag density is shown at the top of the raw signal plots. The third column of each motif panel represents the bound motif sequences (within 30 bp) ordered as in the right two columns.

(I) The distribution of FOXA1 motifs across locations detected in AR ChIP-exo assay.

(J) Percentage of each type of FOXA1 motif located within 1 kb of TSS. Percentage of total ARBEs situated around TSS is also shown.

(K) Venn diagram showing the number of unique and overlapping locations containing ARBEs and FOXA1 motifs.

(L) To evaluate signatures of selection at particular nucleotides, average phyloP conservation score was plotted around each type of core ARBE.

(M) Null distributions of SNP frequency of simulated sequence sets for the four ARBEs. The black areas represent the null distributions, the red lines mark the observed SNP frequency for the precisely defined ARBE, and *P* values were added in red text.

(N) SNP frequency comparisons between the 4 ARBEs and simulated sequence sets within PCaR. The observation that SNP frequencies in ARBEs were higher than (for ARBE2) or equal to (for ARBE1, 3 and 4) those in PCaR controls suggested that SNPs in PCaR (already 6.2% higher than genome controls, data not shown) were enriched in ARBE2 motif and equally distributed between ARBE1, 3 and 4 and other sequences (e.g. other transcription factor motifs).

(O) SNP frequency comparisons between the 4 ARBEs and simulated sequence sets outside of the PCaR. Color bars represent the observed SNP frequencies, black bars represent the median SNP frequencies of genomic simulated regions, and error bars represent the empirical 5% and 95% quantiles. *$P$ <1e-6.

**Figure S5.** Characterization of antagonist-liganded ARBEs.

(A) Heat map showing H3K4me2 signal on the antagonist-responsive locations with or without DHT treatment. The order of sequences in each cluster of ARBEs is the same as the order of sequences shown in Figure 6A.

(B) Averaged MNase-seq tag distribution of H3K4me2 over antagonist-responsive locations.

(C) Heat map showing H3K4me1 signal on the antagonist-responsive locations without DHT. The order of sequences in each cluster of ARBEs is the same as the order of sequences shown in Figure 6A.

(D) Heat map showing H3K4me3 signal on the antagonist-responsive locations without DHT. The order of sequences in each cluster of ARBEs is the same as the order of sequences shown in Figure 6A.

(E) Heat map showing DNase-seq tag distribution around agonist- and antagonist-responsive locations within ± 1 kb. The order of sequences is the same as the order of sequences shown in Fig. 1. Right panel showing averaged DNase-seq tag distribution.

(F) Heat map showing FOXA1 ChIP-exo tag distribution around agonist- and antagonist-responsive locations within ± 2 kb. The order of sequences is the same as the order of sequences shown in Fig. 1. Right panel representing regular ChIP results. After FOXA1 or control silencing, the AR binding in agonist responsive regions (3) and antagonist responsive regions (3) were examined in the presence of 100 nM DHT (first two column) and 10 μM enzalutamide (last two column), respectively.

(G) EMSA was performed to validate AR binding to antagonist ARBEs. Anti-AR antibody was added as indicated at the top. Arrow indicates the position of the shifted and blocked specific probes.

(H) UCSC genome browser views of sequencing data at the *PKIB* and *RNABP9* loci under different conditions. The blue regions represent antagonist-responsive AR locations, while the two pink regions in each locus indicate agonist-responsive AR locations.

(I) Regular AR ChIP validation on antagonist-responsive AR locations after treatment of 1 μM enzalutamide. The data are the mean of triplicates ±SD.

(J) mRNA levels of *PKIB* and *RNABP9* were examined in cells treated with or without DHT, bicalutamide, and enzalutamide (10 μM). The data are the mean of triplicates ±SD.

(K) mRNA levels of *CPEB4*, *PKIB* and *RNABP9* were examined in cells treated with or without enzalutamide (1 μM). The data are the mean of triplicates ±SD.

(L) Silencing of enzalutamide-liganded AR target genes enhanced the inhibitory effects of enzalutamide in LNCaP cells. LNCaP cells were transfected with a control siRNA or siRNAs targeting *RANBP9* or *PKIB* (siRNA ON-TARGET pool). Cells were then treated with 1 μM enzalutamide and cell numbers were determined with a direct viable cell count assay.

**Figure S6.** Analysis of ARBEs in malignant prostate tumors and NATs.

(A) To analyze the dissimilarity of clinical tissue samples, randomly selected 2,000 AR ChIP-exo peaks from each sample were pooled together, and used in clustering analysis.

(B) Normalized tag density of AR binding locations enhanced in prostate tumors versus NATs. Each individual tissue is shown.

(C) Comparison of average signal density of enhanced AR binding locations in tumors versus NATs.

(D) Comparison of average signal density of common AR binding locations between tumors and NATs.

(E) Comparison of average signal density of enhanced AR binding locations in NATs versus tumors.

(F) A Venn diagram showing AR binding locations in LNCaP cells, tumors and NATs.

(G) Identification of agonist-liganded ARBEs and FOXA1 motif in enhanced tumor AR binding locations. The first two columns of each motif show raw tags distribution over agonist-liganded ARBEs and FOXA1 motif (within 80 bp) on the forward (blue) and reverse (red) strands, separately. The third column of each motif represents the bound motif sequences (within 30 bp) ordered as in the right two columns.

(H) Classification of ARBEs identified in enhanced tumor AR binding locations into canonical full-length, non-canonical full-length, and half-site ARBEs.

(I) Motif numbers of the 4 types of ARBEs identified in enhanced tumor AR binding locations compared with common locations.

(J) Percentage of each type of ARBE located within 100 bp of a FOXA1 motif in enhanced tumor AR and common AR binding locations.

(K) Percentage of each type of ARBE located within 1 kb of TSS in enhanced tumor AR binding and common AR binding locations.

(L) LNCaP cells were treated with DHT, R1881 or ethanol for 4, 12 and 24 h, and real-time RT-PCR was performed using gene-specific primers.

(M) mTOR enhancer, BIRC5 enhancer or BCL2L1 enhancer ARBE wild-type (WT) or Mutant (Mut) constructs were transiently transfected into LNCaP cells. Cells were stimulated with DHT, R1881, or vehicle for 24 h, and luciferase activities were measured. The results were presented as the mean±SD of the triplicated transfections.

(N) Comparison of expression of *mTOR*, *BIRC5* and *BCL2L1* genes in 30 prostate tumors with matched “normal” tissues using TCGA RNA-seq data. ***P*<0.0001.

(O) Heat map showing the tissue AR ChIP-exo signal intensity of AR binding in antagonist-responsive regions.

(P) Comparison of average signal density of antagonist-responsive locations in tumors versus NATs.

(Q) Oncomine analysis found that genes within 50kb of antagonist-liganded ARBE locations were highly expressed in prostate cancer but not in another eight cancer types. 1. Bladder Cancer (8)  2. Breast Cancer (26)  3. Colorectal Cancer (23)  4. Kidney Cancer (11)  5. Liver Cancer (7)  6. Lung Cancer (28)  7. Ovarian Cancer (27)  8. Pancreatic Cancer (6)  9. Prostate Cancer (26).

(R) Survival analysis was performed using genes in (O). The *P* values were calculated by LogRank test.

## Supplementary Tables

**Table S1.** Reads number summaries of ChIP-exo, ChIP-seq, MNase ChIP-seq and RNA-seq

**Table S1-1.** Reads number summary of **AR ChIP-exo** in LNCaP cells

| Experiment | Replicates | End1 | | | End2 | | |
|---|---|---|---|---|---|---|---|
| | | Total reads | Mapped reads | Unique mapped reads | Total reads | Mapped reads | Unique mapped reads |
| Vehicle | 1 | 157,696,023 | 116,087,452 | 95,650,131 | 157,696,023 | 113,345,179 | 94,031,620 |
| | 2 | 162,301,671 | 122,031,628 | 100,900,077 | 162,301,671 | 118,627,440 | 98,810,999 |
| DHT | 1 | 161,900,680 | 111,096,663 | 94,964,766 | 161,900,680 | 106,992,987 | 91,798,645 |
| | 2 | 172,227,011 | 131,783,301 | 112,119,538 | 172,227,011 | 128,291,772 | 109,774,433 |
| Bicalutamide | 1 | 195,090,199 | 157,650,827 | 131,280,833 | 195,090,199 | 152,546,241 | 127,777,885 |
| | 2 | 233,914,571 | 174,294,153 | 145,045,393 | 233,914,571 | 184,888,900 | 143,509,962 |
| Enzalutamide | 1 | 199,838,233 | 161,938,198 | 133,808,521 | 199,838,233 | 168,698,924 | 130,217,219 |
| | 2 | 208,232,273 | 169,311,954 | 139,933,740 | 208,232,273 | 164,357,993 | 136,801,808 |

**Table S1-2.** Reads number summary of **AR ChIP-exo** in patient samples

| Sample | Replicates | End1 | | | End2 | | |
|---|---|---|---|---|---|---|---|
| | | Total reads | Mapped reads | Unique mapped reads | Total reads | Mapped reads | Unique mapped reads |
| Tumor 1 | 1 | 122,991,333 | 105,267,308 | 87,788,488 | 122,991,333 | 101,532,131 | 85,683,058 |
| | 2 | 103,307,599 | 88,860,184 | 74,209,724 | 103,307,599 | 86,299,772 | 72,172,489 |
| Tumor 2 | 1 | 119,777,507 | 99,629,663 | 84,186,983 | 119,777,507 | 96,763,157 | 82,506,998 |
| | 2 | 135,286,702 | 116,474,273 | 98,034,397 | 135,286,702 | 113,941,087 | 96,008,562 |
| Tumor 3 | 1 | 120,565,559 | 84,479,968 | 71,238,570 | 120,565,559 | 81,891,370 | 69,490,180 |
| | 2 | 132,995,558 | 97,687,696 | 82,355,229 | 132,995,558 | 97,374,756 | 80,394,472 |
| Tumor 4 | 1 | 111,765,500 | 75,915,212 | 63,462,419 | 111,765,500 | 73,272,102 | 61,752,855 |
| | 2 | 122,624,359 | 87,352,559 | 73,002,389 | 122,624,359 | 87,522,742 | 71,374,461 |
| Normal 1 | 1 | 138,753,453 | 93,793,596 | 79,097,711 | 138,753,453 | 94,245,515 | 77,033,150 |
| | 2 | 138,538,561 | 92,953,854 | 78,349,526 | 138,538,561 | 93,449,455 | 76,153,917 |
| Normal 2 | 1 | 134,022,574 | 97,305,198 | 81,930,503 | 134,022,574 | 97,044,933 | 79,620,485 |
| | 2 | 128,768,887 | 93,856,010 | 78,947,849 | 128,768,887 | 93,471,216 | 76,745,412 |

**Table S1-3.** Reads number summary of **FOXA1 ChIP-exo** in LNCaP cells

| Experiment | Replicates | End1 | | | End2 | | |
|---|---|---|---|---|---|---|---|
| | | Total reads | Mapped reads | Unique mapped reads | Total reads | Mapped reads | Unique mapped reads |
| Vehicle | 1 | 117,660,218 | 106,942,063 | 93,393,461 | 117,660,218 | 102,096,490 | 88,533,921 |
| 10 nM DHT | 1 | 112,524,447 | 97,042,218 | 84,183,916 | 112,524,447 | 89,268,486 | 76,321,170 |

**Table S1-4.** Reads number summary of **AR ChIP-seq** in LNCaP cells

| Experiment | Replicates | Total Number of reads | Mapped reads | Unique mapped reads |
|---|---|---|---|---|
| Vehicle | 1 | 46,205,142 | 40,862,962 | 30,988,589 |
| | 2 | 47,745,017 | 42,370,685 | 32,293,624 |
| 10 nM DHT | 1 | 52,644,155 | 45,633,688 | 34,300,317 |
| | 2 | 50,161,537 | 39,860,250 | 29,526,689 |

**Table S1-5.** Reads number summary of **RNA-seq** in LNCaP cells

| Experiment | Replicates | End1 | End2 |
|---|---|---|---|
| | | Unique mapped reads | Unique mapped reads |
| Vehicle | 1 | 201,382,380 | 183,492,614 |
| | 2 | 206,493,439 | 189,262,763 |
| 10 nM DHT | 1 | 211,445,978 | 197,061,360 |
| | 2 | 207,824,631 | 193,036,498 |

**Table S1-6.** Reads number summary of **MNase ChIP-seq** in LNCaP cells

| Experiment | Replicates | Unique mapped reads |
|---|---|---|
| H3K4me1 | 1+2 | 22,259,965 |
| H3K4me2 | 1+2 | 19,082,643 |
| H3K4me3 | 1+2 | 19,120,158 |
| H3K9me3 | 1+2 | 43,184,978 |
| H3K27me3 | 1+2 | 52,436,584 |

**Table S3.** Primer, Plasmid and Probe sequences

| RT-PCR primers | |
|---|---|
| mTOR+ | GAGCCGAAGGAGATGCAGAA |
| mTOR- | TCCCGATTCATGCCCTTCTC |
| BIRC5+ | GGTTGCGCTTTCCTTTCTGT |
| BIRC5- | GCACTTTCTCCGCAGTTTCC |
| BCL2L1+ | GGTCGCATTGTGGCCTTTTTC |
| BCL2L1- | TCCAAGGCTCTAGGTGGTCA |
| Plasmid sequences | |
| chr20:56348508 (-) WT | GCCACAGGCAACATGCAAATAACGAAGGATGGCCGTGTTCCCATAAAACTTCATTTATTGACTCCGCTGTTTGAGT |
| chr20:56348508 (-) Mut | GCCACAGGCAACATGCAAATAACGAAGGATGGCCGCWTAAAACTTCATTTATTGACTCCGCTGTTTGAGT |
| chr1:108016828 (+) WT | ATTTTAAAATGAACCTTTCTGTTTACTGACATAGGACACTCTAAGACTGTCAATAACAATAAAGAGAGAGTT |
| chr1:108016828 (+) Mut | ATTTTAAAATGAACCTTTCTGTTTACTGACATCTCTAAGACTGTCAATAACAATAAAGAGAGAGTT |
| chr8:27140163 (+) WT | GTTGCTGAGTGTTTGGGCAGAAAAAGGAGACAGACTGTTCTAGTACATTTTGTACCTTTGGAATTCACAGCCATGT |
| chr8:27140163 (+) Mut | GTTGCTGAGTGTTTGGGCAGAAAAAGGAGACAGACTGTTCTTTTGGAATTCACAGCCATGT |
| chr17:8057778 (+) WT | GCCAGCCCGAGAGGGCGGCGGCCAGCGCACTTGGGAACATCATGTTCTCTTGGCTGGTGGCCCAGGACATGCACCCAGC |
| chr17:8057778 (+) Mut | GCCAGCCCGAGAGGGCGGCGGCCAGCGCACTTGCTTGGCTGGTGGCCCAGGACATGCACCCAGC |
| chr6:160974526 (-) WT | AAAGGGTTTTTTTGTTTATTTGTTTGTTTTGGCTGGACATGTCGTCCTAGGTTGATCTGTTTTTATTTCTCTTTAT |
| chr6:160974526 (-) Mut | AAAGGGTTTTTTTGTTTATTTGTTTGTTTTGGCTGGACATGTAGGTTGATCTGTTTTTATTTCTCTTTAT |
| chr21:37853917 (+) WT | TATTACACTGTGACAGTGAATGAAATGGACGTAACGTTCTGTTTTGTAGATATTTTGTGAAATCTTTATTAGTCTT |
| chr21:37853917 (+) Mut | TATTACACTGTGACAGTGAATGAAAGTTTTGTAGATATTTTGTGAAATCTTTATTAGTCTT |
| chr2:10592583 (+) WT | CCAAAAGGATCTGCTGGAAGCAGAGGACGCCCATGTGTAAGTGGAAGTCCCAGGATGTTTGTTTTATAAAGTGTTA |
| chr2:10592583 (+) Mut | CCAAAAGGATCTGCTGGAAGCAGCCCATGTGTAAGTGGAAGTCCCAGGATGTTTGTTTTATAAAGTGTTA |
| chr17:53572810 (-) WT | TTATTACACTTTGTGATCTAATCTGTTTGGATTCGTTCATGAGACACACTTATAAAAATCTCTGTATAAACAGTGCTTAT |
| chr17:53572810 (-) | TTATTACACTTTGTGATCTAATCTGTTTGGATTTGAGACACACTTATAAAAATCTCTGTATAAACAGTGCTTAT |

| | |
|---|---|
| Mut | |
| chr1:11316430 (-) WT | ACAGAGGAAGGAAAGCAAAATTGATTAAACATCTCAGTGCTCATTATCCTAGGTTGTTTACAAACACAGAATGCTT |
| chr1:11316430 (-) Mut | ACAGAGGAAGGAAAGCAAAATTGATTAAACATCTCATTATCCTAGGTTGTTTACAAACACAGAATGCTT |
| chr1:227342398 (+) WT | ATGGATTAATAGTAAACATCATGGAACTGTATTAACAGACATAAAACCTGAAACAATAAAAAGAAACACAACATTC |
| chr1:227342398 (+) Mut | ATGGATTAATAGTAAACATCATTATTAACAGACATAAAACCTGAAACAATAAAAAGAAACACAACATTC |
| chr2:157057096 (+) WT | GATCCAAAGCCAACAGAGCACTGGTGTCTGAGGAGTTAATTAGTGTAAATGCAAGAGTACACCCTGGGACTTCTCC |
| chr2:157057096 (+) Mut | GATCCAAAGCCAACAGGTGTCTGAGGAGTTAATTAGTGTAAATGCAAGAGTACACCCTGGGACTTCTCC |
| chr17:77029339 (+) WT | TCACGTGAGCACTTAAAATGTGGCTCCTTGGCCGAGAACTGGATGTTTCATTTTGTTTAACTGACATAGCTCCATG |
| chr17:77029339 (+) Mut | TCACGTGAGCACTTAAAATGTGGCTCCTTGGCCGGATGTTTCATTTTGTTTAACTGACATAGCTCCATG |
| chr20:40279351 (+) WT | ATTTAAATGCTGTATTTACATTGTTACCAGCTGTATTTGCACAGCTAGAACTTCCACCTGCTTGATAATTTACTTA |
| chr20:40279351 (+) Mut | ATTTAAATGCTGTATTTACATTGTTACCAGCTGTATTTGCACAGCTCCACCTGCTTGATAATTTACTTA |
| chr6:45992316 (-) WT | GAGCCAAGATACATAGAGAGGAAGCCCAAGTTCTGACTTGGATGTCTCAGTGCCTGAAACTTGATACCACTAGATT |
| chr6:45992316 (-) Mut | GAGCCAAGATACATAGAGAGGAAGCCCGACTTGGATGTCTCAGTGCCTGAAACTTGATACCACTAGATT |
| chr6:102932166 (+) WT | AGGCAAAAGACAATTAGTTTTCCAAATAGTGCTTCTTGGGCCATAAATACAAGCTGTCTTTTTTTGTTTGGTTTGG |
| chr6:102932166 (+) Mut | AGGCAAAAGACAATTAGTTTTCCAAATCTTGGGCCATAAATACAAGCTGTCTTTTTTTGTTTGGTTTGG |
| chr18:70570322 (-) WT | AGACTATATTTCCCCCAAGTAAATAACATGAAACACCAAGTATTTTTCTTATTTGTATAAATTTAATGGATACAAT |
| chr18:70570322 (-) Mut | AGACTATATTTCCCCCAAGTAAATAACATGAAACACCTTTCTTATTTGTATAAATTTAATGGATACAAT |
| TMPRSS2 WT | GGTGAAGTGCAGATGCTAATCAGATATGAGTACCTGCCGTACCCTTTAAAGCTTTTAAGACATGCAGCACTAGCT |
| TMPRSS2 Mut1 | GGTGAAGTGCAGATGCTAATCAGATATGGCCGTACCCTTTAAAGCTTTTAAGACATGCAGCACTAGCT |
| TMPRSS2 Mut2 | GGTGAAGTGCAGATGCTAATCAGATATGAGTACCTGCCTTTAAAGCTTTTAAGACATGCAGCACTAGCT |
| TMPRSS2 Mut3 | GGTGAAGTGCAGATGCTAATCAGATATGCTTTAAAGCTTTTAAGACATGCAGCACTAGCT |
| chr1:11316487 (-) WT | TCCTAGGTTGTTTACAAACACAGAATGCTTGTAATGTTTGAACATAAGTTCTGGAAGTGCAGATGTGTTTTATGTT |

| | |
|---|---|
| chr1:11316487 (-) Mut | TCCTAGGTTGTTTACAAACACAGAATGCTTGTAATGTTTGAACATGGAAGTGCAGATGTGTTTTATGTT |
| chr17:76207378 (+) WT | ATGCACCTGCTGATCGCAGGAATGATATGTACTTGGTACGCACTGATCGTACCTCGGGGTGGGAGAAGAGAGGGCA |
| chr17:76207378 (+) Mut | ATGCACCTGCTGATCGCAGGAATGATATGTACTTCACTGATCGTACCTCGGGGTGGGAGAAGAGAGGGCA |
| chr20:30267824 (-) WT | AAAAATCAGTGGCTCCATTTGTGCTTCTCATAGTGTCTCTCTTATTTGTAGCTATAAACAAACAAAAAGTAACAGCTA |
| chr20:30267824 (-) Mut | AAAAATCAGTGGCTCCATTTCTCATAGTGTCTCTCTTATTTGTAGCTATAAACAAACAAAAAGTAACAGCTA |
| EMSA Probes | |
| M1 (WT) | 5'- /5BioTinTEG/AC CAG GAC AGG TAA AAC CGG TCT GCC A -3' |
| M1 (Mut) | 5'- /5BioTinTEG/AC CAG GAA AGA TAA AAA CGA TCT GCC A -3' |
| M2 (WT) | 5'- /5BioTinTEG/CA GCA AAC CAG CCA AAC CAG TTT CTT C -3' |
| M2 (Mut) | 5'- /5BioTinTEG/CA GCA AAA CAA CCA AAA CAA TTT CTT C -3' |
| M3 (WT) | 5'- /5BioTinTEG/CC AAG AAC CGG TCT GAC CGG TTA CCA C -3' |
| M3 (Mut) | 5'- /5BioTinTEG/CC AAG AAA CGA TCT GAA CGA TTA CCA C -3' |
| M4 (WT) | 5'- /5BioTinTEG/CA ACA AAC TGG TTT TTC CTG CTC CTG A -3' |
| M4 (Mut) | 5'- /5BioTinTEG/CA ACA AAA TGA TTT TTA CTA CTC CTG A -3' |

**Table S5.** Detailed clinical information of the 4 individual patients

| Tissue | Number | Sex | Age | Ethnicity | Organ | Pathology diagnosis | Gleason Score | TNM | Type | Tumor % |
|---|---|---|---|---|---|---|---|---|---|---|
| M1 | 1120354A | M | 73 | White | Prostate | Adenocarcinoma | 3+4 | pT2c pN0 | Malignant | 60 |
| M2 | 1120299A | M | 60 | White | Prostate | Adenocarcinoma | 4+3 | pT3a N0 | Malignant | 95 |
| M3[a] | 1120511A | M | 62 | White | Prostate | Adenocarcinoma | 3+4 | pT2c pN0 | Malignant | 70 |
| N1[a] | 1120511B | M | 62 | White | Prostate | non-neoplastic Prostate | / | / | NAT | 0 |
| M4[b] | 1120513A | M | 59 | White | Prostate | Adenocarcinoma | 5+4 | pT2c pN0 | Malignant | 80 |
| N2[b] | 1120513B | M | 59 | White | Prostate | non-neoplastic Prostate | / | / | NAT | 0 |

**Table S6. D**ata sets used in Oncomine analysis and survival analysis

| Figure 7D | | |
|---|---|---|
| 1 | Prostate Carcinoma vs. Normal | *Arredouani Prostate, Clin Cancer Res, 2009* |
| 2 | Prostate Carcinoma vs. Normal | *Grasso Prostate, Nature, 2012* |
| 3 | Prostate Carcinoma vs. Normal | *Holzbeierlein Prostate, Am J Pathol, 2004* |
| 4 | Prostate Carcinoma vs. Normal | *Lapointe Prostate, Proc Natl Acad Sci U S A, 2004* |
| 5 | Prostate Carcinoma vs. Normal | *LaTulippe Prostate, Cancer Res, 2002* |
| 6 | Prostate Carcinoma vs. Normal | *Liu Prostate, Cancer Res, 2006* |
| 7 | Prostate Carcinoma vs. Normal | *Luo Prostate 2, Mol Carcinog, 2002* |
| 8 | Prostate Carcinoma vs. Normal | *Magee Prostate, Cancer Res, 2001* |
| 9 | Prostate Carcinoma vs. Normal | *Singh Prostate, Cancer Cell, 2002* |
| 10 | Prostate Carcinoma vs. Normal | *Taylor Prostate 3, Cancer Cell, 2010* |
| 11 | Prostate Carcinoma Epithelia vs. Normal | *Tomlins Prostate, Nat Genet, 2007* |
| 12 | Prostate Adenocarcinoma vs. Normal | *Vanaja Prostate, Cancer Res, 2003* |
| 13 | Prostate Carcinoma vs. Normal | *Varambally Prostate, Cancer Cell, 2005* |
| 14 | Prostate Adenocarcinoma vs. Normal | *Wallace Prostate, Cancer Res, 2008* |
| 15 | Prostate Carcinoma vs. Normal | *Welsh Prostate, Cancer Res, 2001* |
| 16 | Prostate Carcinoma vs. Normal | *Yu Prostate, J Clin Oncol, 2004* |
| Figure 7E | | |
| 1 | *(Nakagawa et al, 2008)* | |
| 2 | *(Taylor et al, 2010)* | |
| 3 | *(Glinsky et al, 2004)* | |
| 4 | *(Barwick et al, 2010)* | |

**Table S7.** Genes used for Oncomine, recurrence-free survival analysis and pathway analysis

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Figure 7E | | | | | | | | |
| Nakagawa *et al.* | | | | | | | | |
| BIRC5 | RELA | FRZB | SKI | FYN | PLAUR | REL | ERBB4 | MMP7 |
| MBD2 | ARAF | BCL2L1 | CDK6 | CRK | IGF2R | RARB | MTOR | IGFBP2 |
| ATAD2 | TPD52 | | | | | | | |
| Glinsky *et al.* | | | | | | | | |
| SMARCA2 | POU2F1 | PIAS1 | PPIG | IGF2R | CRK | HNRNPA1 | NFIB | MBD2 |
| DDC | SCAP | ACACA | GNAQ | PRKAB1 | ATAD2 | PDZD2 | NEDD4L | DUSP10 |
| UBIAD1 | BCL2L1 | CASP7 | REL | H2AFZ | LEPR | APPBP2 | PPP2CA | NKX3-1 |
| FRZB | ITGA4 | KLK15 | VAV3 | FOXO3 | VEGFC | ADAM17 | | |
| Figure S6P | | | | | | | | |
| ABR | CADPS | DAZL | FAM5C | IMPA1 | MAPK10 | NEK1 | PDZRN3 | RALGAPA1 |
| ACPP | CAND1 | DDR2 | FBXL5 | ITPR1 | MAPK9 | NFAT5 | PGM3 | RBL2 |
| ADD1 | CBFA2T2 | DHX29 | FGF13 | ITPR2 | MAX | NPTN | PIK3R3 | RBPJ |
| ANGPT1 | CCR9 | DIP2C | FLNB | KDM4B | MB | NR5A2 | PKN2 | RERE |
| ANK3 | CD46 | DLG2 | FTO | KDM5A | MED13 | NUMA1 | POU1F1 | RGS5 |
| ANKRD12 | CDC16 | DMXL1 | GAB1 | KDM6A | MIOS | NUP93 | PPP1R3D | RPGRIP1 |
| ANXA3 | CDC42EP3 | EFHA1 | GALNT3 | KIAA0368 | MIPEP | NUP98 | PPP2R2A | RPS6KA5 |
| AQR | CDH1 | ELK4 | GLRA3 | KIF5B | MITF | P4HB | PRKACB | RSBN1 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ARHGAP26 | CDK8 | ELL | GMDS | LCP1 | MLLT3 | PA2G4 | PRKCA | RYR2 |
| ARHGAP29 | CDS1 | ELL2 | GNAQ | LIFR | MPL | PAK2 | PRKCD | SEMA3C |
| ARHGDIA | CHD9 | EPHA4 | GNAS | LPHN2 | MRPS27 | PBX1 | PRKD1 | SETBP1 |
| ARL6IP5 | CNTN5 | ERBB3 | GOT2 | LPP | MTMR6 | PCCA | PRUNE2 | SLC19A1 |
| ATP2C1 | COBL | ERCC3 | GRIK1 | LRBA | MYCBP2 | PCGF2 | PSMD7 | SLC25A36 |
| ATXN1 | COBLL1 | ETFDH | GRIK2 | LUZP2 | MYH10 | PCLO | PTK2 | SLC25A44 |
| AUH | CSNK1A1 | EYA1 | GRM3 | MAD1L1 | MYO6 | PDE3B | PTPN13 | SLC30A4 |
| BCL9 | CUX2 | EYA2 | GSK3B | MAGI2 | MYRIP | PDE4D | PTPRB | SPG11 |
| BTD | CXADR | FAM172A | HS2ST1 | MAP3K1 | NAV3 | PDLIM5 | PTPRF | SVIL |
| CACNB2 | DAAM1 | FAM179B | HSPA1L | MAP3K5 | NCOA1 | PDS5B | PURA | SYCP2 |
| TEAD1<br>USP15 | THSD7A<br>USP34 | TLE1<br>USP46 | TLE3<br>UTRN | TMEM87A<br>VPS13B | TNK2<br>WASF3 | TXNIP<br>YLPM1 | UBC<br>ZBTB20 | UNC13B<br>ZHX3 |
| ZMPSTE24 | ZNF189 | ZNF318 | | | | | | |

Pathway analysis of antagonist-responsive location associated genes.

| Signal Transduction Pathway | P-Value | Gene list |
|---|---|---|
| VEGF signaling | 3.95e-5 | PRKCA, PRKACB, GSK3B, MAPK9, PAK2, ITPR1, MAP3K5, MAPK10, CSNK1A1, PRKCD, MAP3K1, CDK8, PTK2 |
| AKT-Bad signaling | 1.44e-4 | PRKCA, PRKACB, GSK3B, MAPK9, PAK2, MAP3K5, MAPK10, CSNK1A1, |

| | | |
|---|---|---|
| | | PRKCD, MAP3K1, CDK8, PTK2 |
| EGFR1 signaling | 5.08e-3 | PRKCA, TNK2, GAB1, PRKD1, ELK4, PIK3R3, RPS6KA5, PKN2, MAP3K1 |

## Supplementary Experimental Procedures

### Cell Culture

The androgen-dependent prostate cancer cell line LNCaP were obtained from the American Type Culture Collection, and cultured in RPMI complete medium, respectively. For hormone responsive experiments, LNCaP cells were maintained in phenol red-free medium with 5% charcoal-stripped FBS for 3 days, and then treated with vehicle and different ligands.

### Prostate Tissues of Patients

The malignant prostate tumors and paired NATs were obtained from The Ohio State University (OSU) Wexner Medical Center under the approval of OSU Institutional Review Board. Radical prostatectomy specimens were snap frozen in liquid nitrogen immediately after surgical excision. Confirmation of malignant tumor and NAT identity was performed by a urologic pathologist. All tissues were stored at -80ºC until they were used. Detailed clinical information is provided in Table S4.

### Antibodies

Antibodies used were anti-AR (N20) from Santa Cruz Biotechnology (Santa Cruz, CA), anti-FoxA1 (ab23738), anti-H3K4me1 (ab8895), anti-H3K4me2 (ab7766), anti-H3K4me3 (ab8580), anti-H3K9me3 (ab8898), and anti-H3K27me3 (ab6002) from Abcam (Cambridge, MA).

**ChIP-seq**

ChIP assay was performed as previously described (Chen et al, 2011; Wang et al, 2009). The ChIP-enriched DNA was PCR amplified to create a DNA library that was sequenced by Illumina Genome Analyzer II (GA II) at the OSU Comprehensive Cancer Center (CCC) sequencing core.

**Native MNase ChIP-seq**

Native chromatin from LNCaP cells was digested with MNase for 10-20 min to generate mainly mononucleosomes that were immunoprecipitated with anti-histone mark antibodies. The histone mark-associated DNA was amplified and sequenced by GA II at the OSUCCC sequencing core.

**ChIP-exo**

All adaptor sequences were changed to be compatible with the Illumina sequencing platform. T4 DNA polymerase, T4 PNK, and Klenow DNA Polymerase were used together for end polishing. The ligation step was performed with less reducing agent. Protein A Dynal magnetic beads were washed using modified RIPA buffer (50 mM Tris-HCL pH 7.8, 1 mM EDTA, 0.25% Na Deoxycholate, 1% NP-40, 0.5 M LiCl). The library was amplified with only 10 or 12 PCR cycles, and prepared without gel-based size selection. Paired-end sequencing was performed using Illumina HiSeq 2000 at the OSUCCC sequencing core. For tissue ChIP-exo, the frozen samples were trimmed and chopped into small pieces on ice, and then fixed immediately in 1% formaldehyde for 20 min at room temperature. After additional incubation with glycine for 10 min, the pellets were washed and homogenized in 1x PBS containing proteinase inhibitor. With

sufficient lysis, the samples were then processed according to ChIP-exo procedures described above.

**ChIP-exo peak and border analysis**

Different from the ChIP-seq procedure, ChIP-exo assay uses exonuclease to digest the DNA fragments to a relatively fixed distance from the target protein-binding region. The very 5'-end of the DNA fragment represents the first several bases that are protected by the target protein. Many current ChIP-seq peak finding programs employ a tag shifting algorithm and calculate the shift distance using bimodal distributions modeling tags mapped to the two strands of reference sequences or extend the tag to a certain length to cover the entire DNA fragment, and therefore do not fit with ChIP-exo data, because the application of exonuclease eliminates the need for tag shifting. The other challenge in the analysis of ChIP-exo data is that *cis*-regulatory elements are usually bound by multiple different transcription factors simultaneously in one transcription complex within a few hundred bp sized genomic regions. The resolution of standard peak finding programs is not high enough to identify two adjacent ChIP-exo borders separated by less than 20 bp. The Asymmetric Laplace model has been shown to better capture the distribution of the tags surrounding a target site, providing high resolution detection of DNA methylation level in MBD-seq analysis (Lan et al, 2011a). Therefore, we tested and applied a Mixture Asymmetric Laplace Distribution (MALD) model to increase the resolution of border detection and to capture multiple transcription factor binding sites within a single *cis*-regulatory element more accurately. The detailed procedure is discussed below:

*Initial scan for candidate regions*

The genome is divided into 50 bp bins. Regions with a higher tag count than a predetermined threshold weighted by the genome region amplification index *GRAI* (Lan et al, 2011a) were identified as candidate regions.

*Parameter estimation of the Asymmetric Laplace Distribution ALD*

The density function for an ALD is,

$$f_{\delta,\sigma,\kappa}(x)=\frac{\sqrt{2}}{\sigma}\cdot\frac{\kappa}{1+\kappa^2}\begin{cases}\exp\left(-\frac{\sqrt{2}\kappa}{\sigma}(x-\delta)\right), & for \quad x\geq\delta\\ \exp\left(\frac{\sqrt{2}}{\sigma\kappa}(x-\delta)\right), & for \quad x<\delta\end{cases} \qquad \textbf{1)}$$

1. The maximum likelihood estimators of the ALD are calculated as follows (Kotz et al, 2002; Kotz et al, 2001),
2. Given sample size *n*, determine $x_m$, $1\leq m\leq n$ that minimize the function

$$H(\delta)=2\ln\left(\sqrt{\alpha(\delta)}+\sqrt{\beta(\delta)}\right)+\sqrt{\alpha(\delta)}\sqrt{\beta(\delta)} \qquad \textbf{2)}$$

where $\alpha(\delta)=\frac{1}{n}\sum_{j=1}^{n}(x_j-\delta)^+$, $\beta(\delta)=\frac{1}{n}\sum_{j=1}^{n}(x_j-\delta)^-$,

$$(x_j-\delta)^+=\begin{cases}x_j-\delta, & x_j\geq\delta\\ 0, & x_j<\delta\end{cases}$$

$$(x_j-\delta)^-=\begin{cases}\delta-x_j, & x_j\leq\delta\\ 0, & x_j>\delta\end{cases}$$

$x_j$ is the *j*th element of *x* and $x_m$ is the target element which minimizes function $H(\delta)$.

3. Calculate estimators as,

$$\hat{\delta}=x_m \qquad \textbf{3)}$$

$$\hat{\sigma} = \sqrt{2}\sqrt[4]{\alpha(x_m)}\sqrt[4]{\beta(x_m)}\left(\sqrt{\alpha(x_m)} + \sqrt{\beta(x_m)}\right) \quad \textbf{4)}$$

$$\hat{\kappa} = \sqrt[4]{\alpha(x_m)} / \sqrt[4]{\beta(x_m)} \quad \textbf{5)}$$

*EM algorithm finds parameters that maximize the likelihood of tags within candidate regions*

Within a candidate region, multiple borders may exist. The position and the density of the borders are estimated by maximizing the likelihood of the given tags using the MALD,

$$\hat{\theta}_{mle} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \ln f(x_i \mid \theta) \quad \textbf{6)}$$

Where *n* is the number of tags and *f* is the probability density function of the MALD.

$C_i$ determines which component of the mixture is an independent observation $x_i$ originated. Thus the *k*th component can be written as,

$$X_i \mid (C_i = k) \sim MALD(\delta_k, \sigma_k, \kappa_k)$$

Let $\tau_k$ denote the portion of the *k*th component. σ and κ are treated as constants with the assumption that these parameters are identical for all components and can be estimated based on aggregation of high confidence borders. The unknown parameters to be estimated are,

$$\theta = (\tau, \delta)$$

The log likelihood function is,

$$\ln L(\theta; x, c) = \ln \sum_{i=1}^{n} \sum_{k=1}^{m} \tau_k f(X_i; \delta_k) \quad \textbf{7)}$$

Where *n* is the number of observations and *f* is the probability density function of ALD.

The following formula is used to update the parameters using Bayes' theorem (Redner & Walker, 1984),

$$\tau_k^{s+1} = \frac{1}{n} \sum_{i=1}^{n} S_{k,i}^{s} \quad \textbf{8)}$$

$\delta_k^{s+1} = X_{p(n):n}$, so that $\sum_{i=1}^{p(n)-1} S_{k,i}^s < \frac{\kappa^2 \sum_{i=1}^{n} S_{k,i}^s}{1+\kappa^2} \le \sum_{i=1}^{p(n)} S_{k,i}^s$ **9)**

where *i* is now the ranked index of *x*, so that $x_1 \le x_2 \le \ldots\ldots \le x_n$, *s* denotes the iteration step,

$S_{k,i}^s = \frac{\tau_k^s p_k\left(x_i \mid \delta_k^s\right)}{\sum_{j=1}^{m} \tau_j^s p_j\left(x_i \mid \delta_j^s\right)}$ is the conditional probability that $x_i$ is from *k*th component at step *s*.

Note that formula 9 is an empirical approximation of the MLE.

*Selecting the best mixture model*

Bayesian Information Criterion (*BIC*) is used to determine the number of components (borders) within a candidate region.

$BIC(M_l) = 2\ln L(X, M_l) - d\log N$ **10)**

where $M_l$ is the model being tested, $L(X, M_l)$ is the log likelihood that the given tags is from model $M_l$, *d* is the number of free parameters in model $M_l$ and N is sample size, i.e. tag number in a given candidate region.

**ChIP-exo Motif Analysis**

The enriched DNA motifs were defined by a multi-phase cross-validation procedure. Genomics Suite *v*6.6 (Partek) and MEME Suite *v*4.9 (Bailey et al, 2009) were used to find the candidate motifs. Initial motif candidates were generated using default program settings (one instance per sequence, less than 40 bp of border extension). Motifs were then clustered with the Pearson correlation coefficient. Exo signal was measured to define border patterns and classify motifs. A set of overrepresented motifs was then used to correct border extension according to the enriched motif position. Motif discovery was repeated twice. Motifs with *E* <1e-10 and counts >2,500 or

10% of sequences were retained as reliable predictions for the next round of analysis. Finally, we identified ARBEs satisfying the following extensible criteria: (1) Motif similarity compared to ARBEs in the TF binding databases or between core motifs defined in AR ChIP-exo data; (2) At least one common protected border exists upstream and downstream of the strand-specific motif; and (3) Same distance from borders to the most conserved nucleotides in variable motifs. For those core motifs that did not meet criteria (1), we also performed motif comparison and clustering based on criteria (2) and (3) using the exo-defined matrix. For identifying non-canonical full-length ARBEs, extended-ARBE clustering and ranking were performed using the Pearson correlation coefficient with the full length ARBE matrix (canonical full-length $q$<0.001, non-canonical full- length $q$<0.1).

**Reporter Gene Assays**

Putative ARBE-containing regions (~70 bp, sequences are listed in Table S2) were subcloned into pGL4.24 vector (Promega). Transfection was performed using Lipofectamine 2000 (Invitrogen, Carlsbad, CA). Cells were treated with or without 100 nM DHT or 10 nM R1881, 10 μM bicalutamide or 10 μM enzalutamide for 16-24 hr and then harvested. Luciferase activity was measured using a Dual-Luciferase Reporter Assay kit and GloMax™ Systems (Promega).

**Clustering Analysis**

To conduct clustering analysis, we first generated the wiggle tracks of the histone mark-H3K4me2 in LNCaP cells treated with vehicle or DHT for 4 hr by compiling the raw dataset (GSE20042 (He et al, 2010)). For each extended ARBE site (+/-500 bp, the middle of ARBE

sites (6 or 7 bp) is averaged into one position), we calculated the coverage depth on each nucleotide position, and normalized these values by the total signal summed from the whole peak region, thus casted them to [0, 1] boundary for comparison. Genomics Suite was applied to generate the partitional clustering with the Pearson Dissimilarity Distance Function.

**Conservation Analysis**

For each type of ARBE, we calculated the phyloP score on each nucleotide position extended +/- 10 bp from both ends of each ARBE using the 46-way vertebrate phyloPCons metric for measurement.

**SNP analysis**

Monte Carlo simulations were performed to compare the SNP frequencies of precisely defined ARBEs with the genome background regions. The dbsnp137 database was used for SNP frequency computation (45,755,572 SNPs). For each of the 4 types of precisely defined ARBEs, we first randomly selected the matched number of 15 bp sequences from the genome with replacement, which had the same distribution over chromosomes, and then computed the SNP frequencies at each base position as well as the average. This process was repeated 100,000,000 times to generate the null distributions of the SNP frequencies of randomly generated sequence sets. Then the empirical *P*-values were computed. To assess the SNP-correlation of precisely defined ARBEs in prostate cancer, we extracted the prostate cancer associated regions (PCaR) from the genome using the catalog of published genome-wide association studies (see URL: http://www.genome.gov/gwastudies/). The catalog updated on Sept. 29, 2012 was used, which

contains 9,162 SNPs. 176 SNPs were reported to be associated with prostate cancer, prostate cancer (gene x gene interaction), or prostate cancer mortality, which gave 111 unique cytogenetic regions covering 15.5% of the genome. These regions were defined as our PCaR, while the complement was classified as being outside of PCaR. We compared the SNP frequencies of these motif subsets to the matched background also. Matched numbers of sequences were randomly selected from PCaR and outside PCaR accordingly. 1,000,000 times of simulation were conducted for each subset.

**Electrophoretic Mobility Shift Assay (EMSA)**

EMSA probes were synthesized by Integrated DNA Technologies (IDT). The Gelshift Chemiluminescent EMSA kit (Active Motif, Carlsbad, CA) was used according to the manufacturer's protocols. 4 µg of nuclear extracts for each condition were incubated with probes for 20 min on ice. The gel was run at 100V for 1 h and then transferred to nylon membrane. After incubation with streptavidin-HRP conjugate, the membrane was exposed to film and developed.

**Immunohistochemistry**

Immunohistochemical analysis was conducted on formalin-fixed paraffin embedded tissue sections. After rehydration, antigen retrieval was performed by boiling the slides in citrate buffer (Antigen Retrieval Citra Plus, BioGenex, San Ramon, CA) for 30-35 min. Endogenous peroxidase was inhibited by incubation in peroxide blocking solution (Dako Cytomation, Carpinteria, CA) for 15 min followed by 60 min incubation with an anti-AR antibody (Rabbit

polyclonal PG-21, Upstate). After three washes with wash buffer (Super Sensitive Wash Buffer, Biogenex, Fremont, CA), sections were incubated for 30 min with a matched labeled polymer-HRP (Dako Cytomation, Carpinteria, CA). Color was developed by 10 min incubation with DAB chromogen solution (Dako Cytomation, Carpinteria, CA). Slides were counterstained with Hematoxylin and mounted. Negative controls include omission of primary antibody. Slides were scanned at 40X magnification.

**Differential Enrichment Analysis**

The BELT program (Lan et al, 2011b) was used to identify different ligand-stimulated locations in LNCaP cells, and enhanced AR binding locations in malignant prostate tumors or NATs. BELT compared the read densities of enriched locations in agonist/vehicle, antagonist/vehicle, and in tumor/NAT ChIP-exo to those of the same locations in NAT/tumor ChIP-exo using Fisher's exact test with a *P*-value cutoff of 0.05.

**Non-poly(A) Selection, Directional, Ligation-Free Paired-End RNA-seq**

All sizes of rRNA were removed by using the Ribo-Zero rRNA removal reagents (Epicentre Biotechnologies, Madison, WI). The library preparation was performed with the ScriptSeq RNA-Seq Library Preparation kit (Epicentre Biotechnologies) following the manufacturer's instructions. The di-tagged cDNA molecules were amplified by 8 cycles of PCR. Non-size selected libraries were then sequenced for paired-ends.

**RNA-seq Analysis**

edgeR (Robinson et al, 2010) was applied to conduct the differential expression analysis based on the count information. The differentially expressed genes were detected by the reasonable cut-off (*P*-value < 0.01) between DHT treated and control data.

**RNA interference and cell proliferation assay**

siRNA transfection and cell proliferation assay were performed as previously described (Chen et al, 2011; Wang et al, 2009).

**Expanded References**

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37:** W202-208

Barwick BG, Abramovitz M, Kodani M, Moreno CS, Nam R, Tang W, Bouzyk M, Seth A, Leyland-Jones B (2010) Prostate cancer genes associated with TMPRSS2-ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts. *Br J Cancer* **102:** 570-576

Chen Z, Zhang C, Wu D, Chen H, Rorick A, Zhang X, Wang Q (2011) Phospho-MED1-enhanced UBE2C locus looping drives castration-resistant prostate cancer growth. *EMBO J* **30:** 2405-2419

Denayer S, Helsen C, Thorrez L, Haelens A, Claessens F (2010) The rules of DNA recognition by the androgen receptor. *Mol Endocrinol* **24:** 898-913

Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL (2004) Gene expression profiling predicts clinical outcome of prostate cancer. *J Clin Invest* **113:** 913-923

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 (2010) Nucleosome dynamics define transcriptional enhancers. *Nat Genet* **42:** 343-347

Kotz S, Kozubowski TJ, Podgorski K (2002) Maximum likelihood estimation of asymmetric Laplace parameters. *Ann I Stat Math* **54:** 816-826

Kotz S, Kozubowski TJ, Podgórski K (2001) *The Laplace distribution and generalizations : a revisit with applications to communications, economics, engineering, and finance,* Boston: Birkhäuser.

Lan X, Adams C, Landers M, Dudas M, Krissinger D, Marnellos G, Bonneville R, Xu M, Wang J, Huang TH, Meredith G, Jin VX (2011a) High resolution detection and analysis of CpG dinucleotides methylation using MBD-Seq technology. *PLoS One* **6:** e22226

Lan X, Bonneville R, Apostolos J, Wu W, Jin VX (2011b) W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. *Bioinformatics* **27:** 428-430

Nakagawa T, Kollmeyer TM, Morlan BW, Anderson SK, Bergstralh EJ, Davis BJ, Asmann YW, Klee GG, Ballman KV, Jenkins RB (2008) A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy. *PLoS One* **3:** e2318

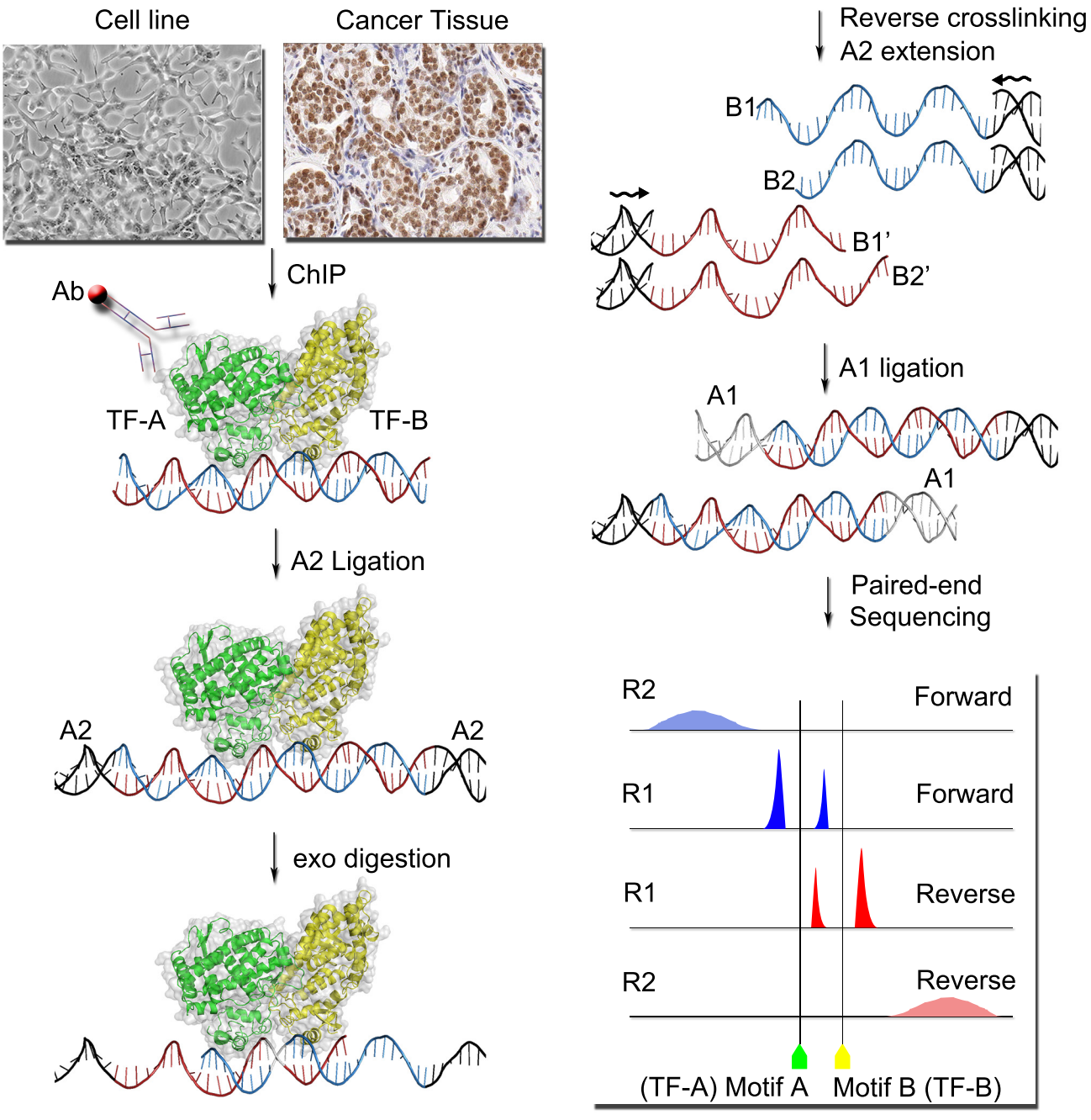
Redner RA, Walker HF (1984) Mixture Densities, Maximum-Likelihood and the Em Algorithm. *Siam Rev* **26:** 195-237

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139-140

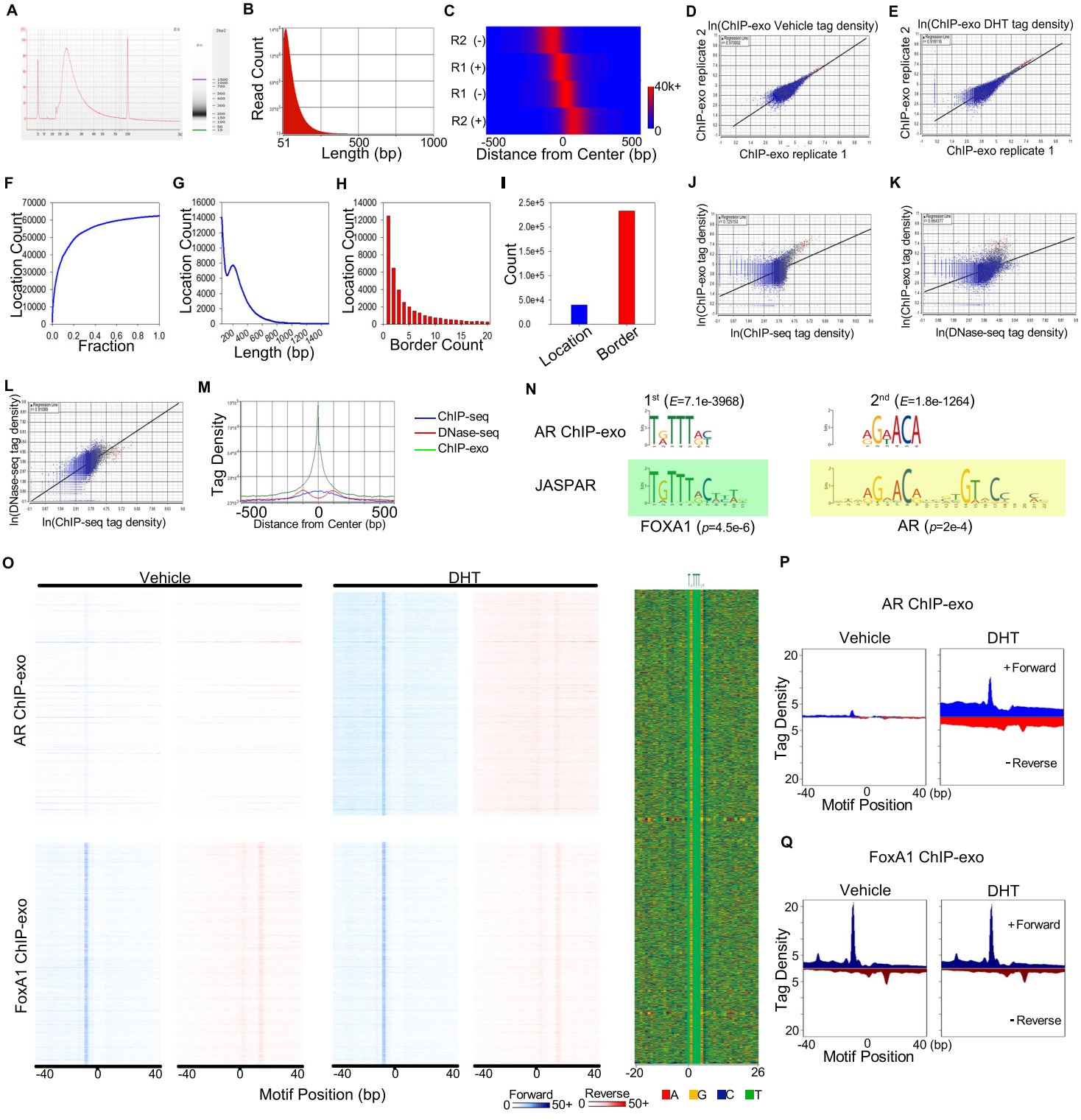
Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B, Antipin Y, Mitsiades N, Landers T, Dolgalev I, Major JE, Wilson M, Socci ND, Lash AE, Heguy A, Eastham JA, Scher HI, Reuter VE, Scardino PT, Sander C, Sawyers CL, Gerald WL (2010) Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18:** 11-22

Wang Q, Li W, Zhang Y, Yuan X, Xu K, Yu J, Chen Z, Beroukhim R, Wang H, Lupien M, Wu T, Regan MM, Meyer CA, Carroll JS, Manrai 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 (2009) Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. *Cell* **138:** 245-256

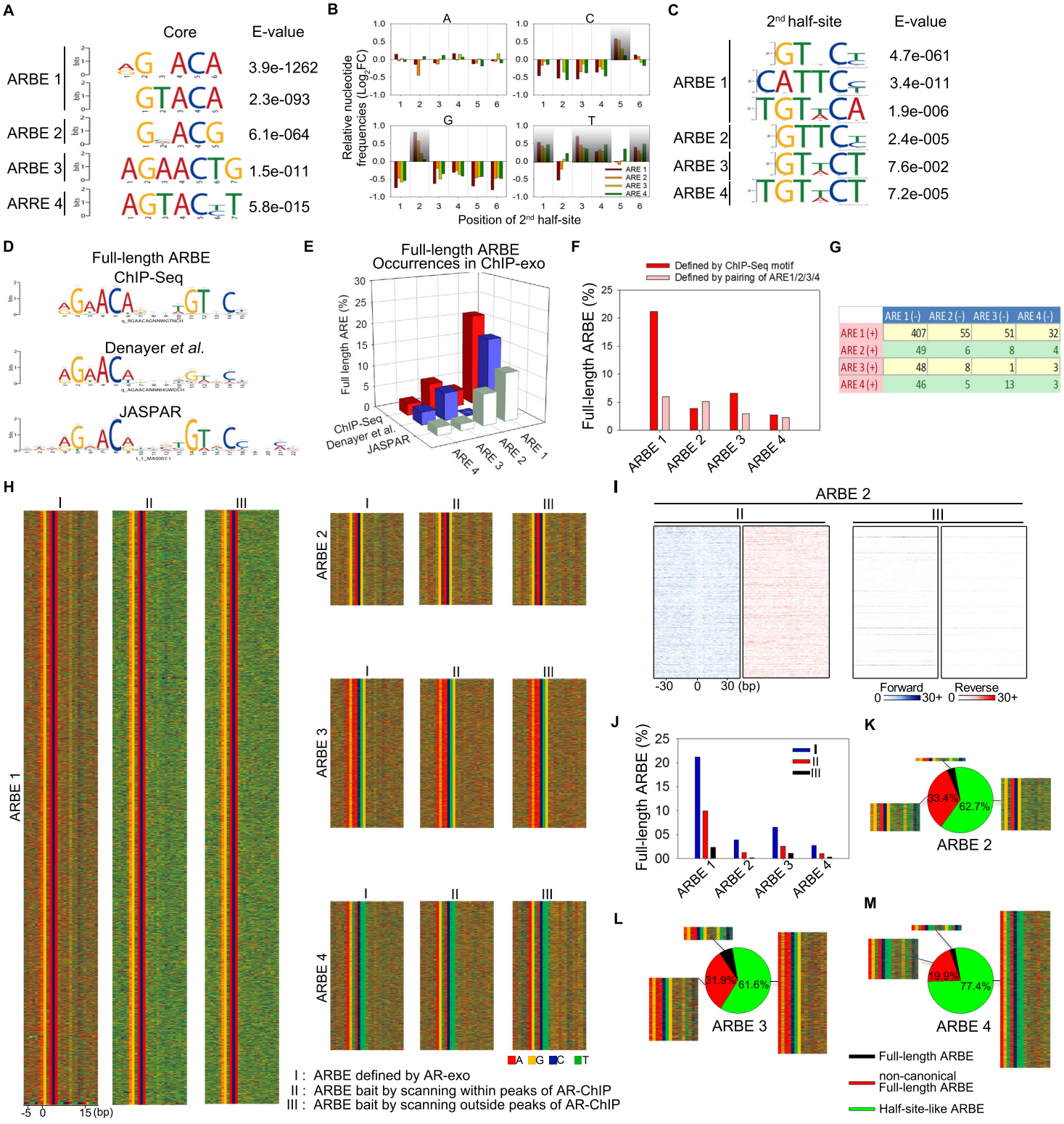
**Chen_Figure S1**



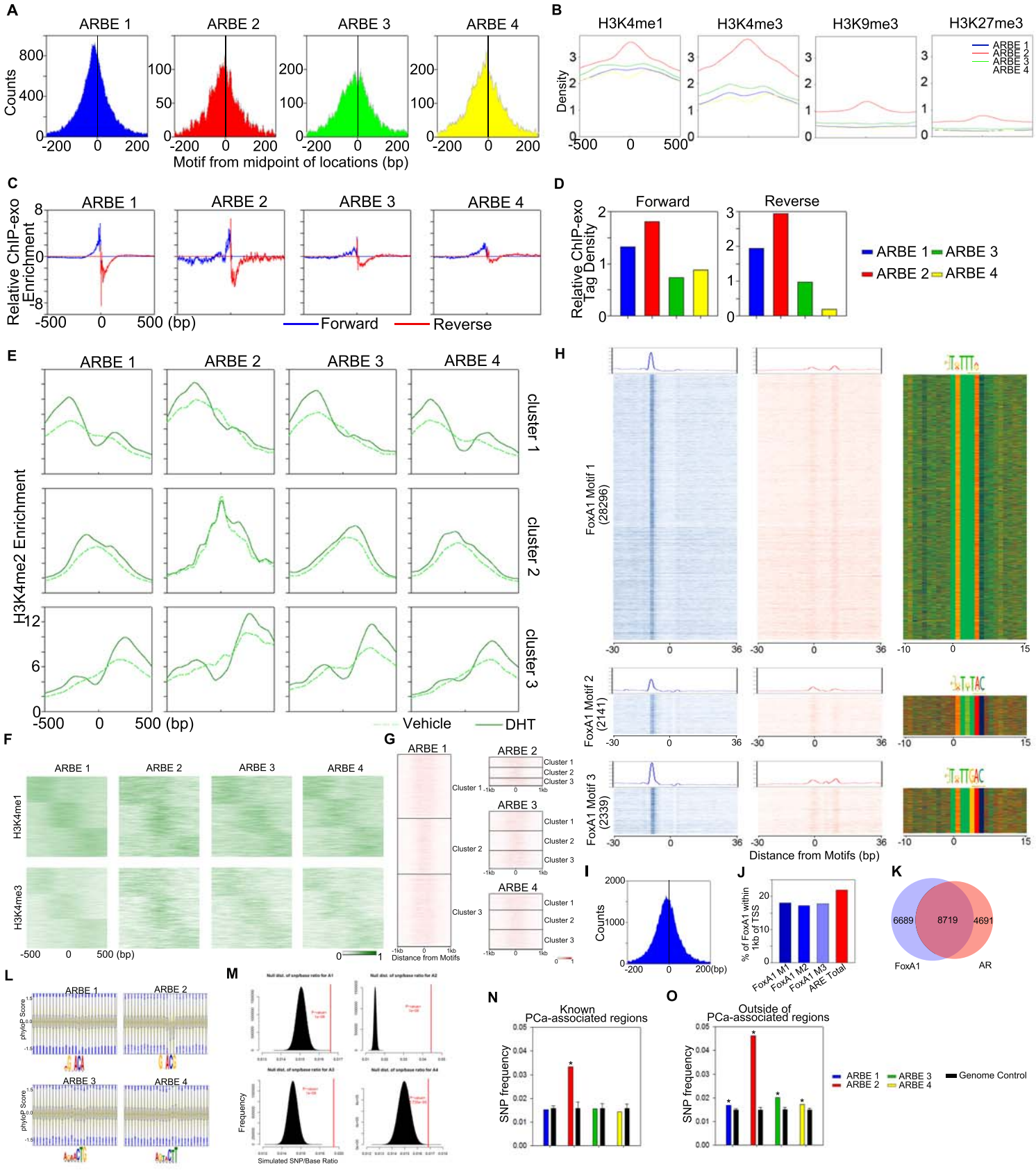
# Chen_Figure S2



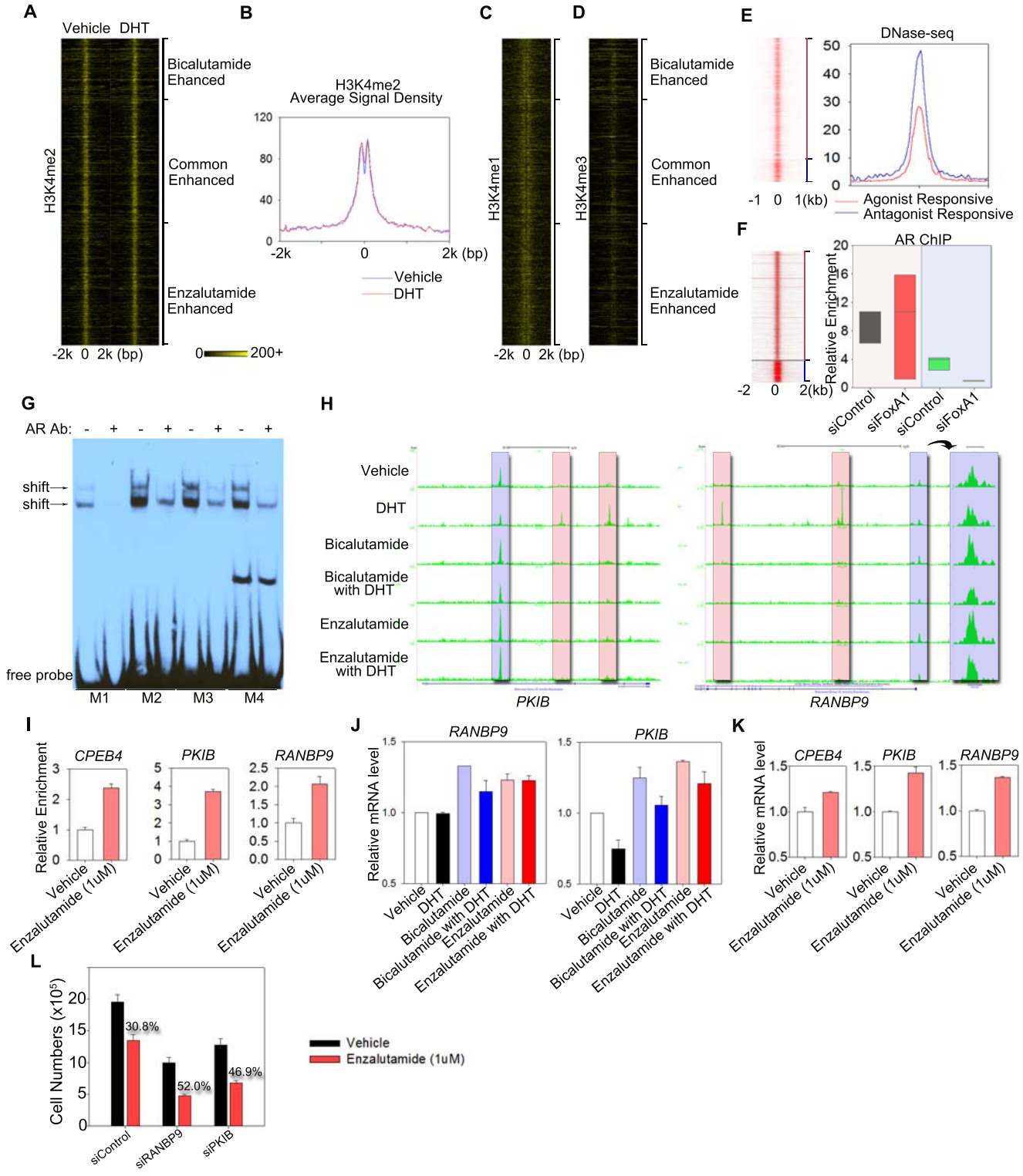
# Chen_Figure S3



# Chen_Figure S4



# Chen_Figure S5



# Chen_Figure S6