Supplemental figure legends

Figure S1. Related to Figure 1. (A) Saturation analysis for evaluating ChIA-PET data quality. The left and middle panels represent RNA Pol II ChIA-PET from this study. The right panel represents RAD21 ChIA-PET from ENCODE. By employing a down-sampling strategy, the total ChIA-PET sequencing reads were randomly portioned into 10%, 20%, …, 80%, 90% sub-samples representing varying sequencing depth, which were then individually used to do the interaction-calling separately. Then, interactions detected from sub-samples were compared with interactions detected from the total sequencing reads. (B) Length of interactions across the four cell lines. Three criteria of the PETs (>= 1, 2 or 3) are shown in pink, blue and green. The y axis is distance transformed by log10. The blue, red, and green lines connecting the violins signify the median, mean, and Huber M-estimator, respectively. The black dotted line distinguishes TAD interactions (less than or equal to 1MB) from longer interactions. (C) The number of chromatin interactions between different genomic elements across four cell lines, including interaction types between promoter, gene body, other region and enhancers. The color density shows the ratio of each type of interactions. (D) Distributions of PET numbers along with the distance between peak anchors in the four cell lines. The length of PET is transformed to log10bp. Anova test was used for multiple groups comparison. Boxplot represents median, 0.25 and 0.75 quantiles with lines at 1.5x interquartile range. Distribution of enhancers per gene (E) and genes per enhancer (F) in the four cell lines. (G) The number of enhancers for each cell line (yellow) and the number of enhancers that display RNA Pol II interactions with gene promoters (orange). (H) Correlation between transcript abundance and enhancers per gene. The transcript abundance was measured by FPKM transformed by log2. The correlation efficient was calculated by spearman correlation and p-value is labeled. (I) The interactions are well fitted as NB distribution in four cells. The PETs scores were normalized as the raw pairs among per 10k region peak per 10 million PETs. The distributions of chromatin interactions were fitted as NB distribution by using Matlab fitdist toolbox. The fitting goodness was tested by using K-S test. (J) Motif analysis of promoter and enhancer regions in the four cell lines. Upper panel: motif enrichment of enhancer regions in the four cell lines. The type of motif sequences and p-values are labelled. Lower panel: motif enrichment of promoter regions in the four cell lines.

Figure S2. Related to Figure 1. Interaction comparison. The top, left panel displays all shared and unique interactions. The middle-left and bottom-left panels divide these interactions into “short” and “long”, respectively. The analysis was repeated excluding interactions that only have one PET (middle columns) and excluding interactions that have only one or two PETs (right columns). (B-C) Enriched pathways and gene ontology (GO) terms based on RNA-seq and ChIA-PET data, upon comparison with the benign RWPE-1 cells. (B) Enriched GO terms for differential gene expression. (C) Enriched GO terms for differential chromatin interactions. Blue: comparison between DU145 and RWPE-1; red:

comparison between LNCaP and RWPE-1; yellow: comparison between VCaP and RWPE-1.

Figure S3. Related to Figure 2. Integrated genome view of *HOXB13* gene and its adjacent regions in LNCaP cells. (A) The data tracks represent RNA-seq, ChIP-seq for CTCF, FOXA1, AR, H3K27ac and RNA Pol II, ChIA-PET for RNA Pol II and RAD21, and HiC data in the LNCaP cell line. (B) ChIA-PET contact heatmap representing RNA Pol II (left) and RAD21 (right) associated chromatin interactions for the *HOXB13* gene and neighborhood regions. The *HOXB13* gene is shown in light green color.

Figure S4. Related to Figure 3. (A) Immunoblot analysis to determine the effect of treatment with various doses of BETi, JQ1 for 24 hours on the steady-state levels of AR in LNCaP and VCaP cells. MYC is used as a positive control and actin is the loading control. The antibody against AR recognizes its N-terminus. (B) Immunoblot representing the steady-state levels of AR in four cell lines using antibodies recognizing its N- and C-terminus, respectively. (C) Amplification status analysis of *AR*, AR enhancers peaks, *MYC* and *FOXA1* in mCRPC patient samples by ddPCR. (D) Correlation of all 6 AR enhancer peaks amplification versus AR gene copy number amplification in mCRPC patient samples. (E) Correlation of all 6 AR enhancer peaks gain versus AR gene copy number gain in mCRPC patient samples. (F) Correlation between *AR* copy number and *AR* gene expression ($R = 0.43$, $P = 4x10^{-6}$). (G) Integrated genome view of a genomic region (chrX:66389092-66738319) using whole genome sequencing data of a CRPC sample. A 349227 bp deletion is discovered which is supported by 139 reads out of total 7775 reads in the region.

Figure S5. Related to Figure 3. Transcriptional regulation of *FOXA1* gene and its neighborhood genes. (A-B) Integrated genome view of *FOXA1* and its adjacent regions based on data of RNA-seq, ChIP-seq of CTCF, FOXA1, AR, H3K27ac and RNA Pol II, ChIA-PET of RNA Pol II are shown in LNCaP and VCaP cell lines, respectively. Additionally, phospho RNA Pol II and ERG ChIP-seq is described for VCaP cells. The *FOXA1* gene is highlighted in light-blue color. (C) Immunoblot analysis showed Increased expression of FOXA1 in LNCaP and VCaP cell lines. Actin is the loading control.

Figure S6. Related to Figure 5. Integrated genome view of *KLK3* gene and its neighborhood regions representing RNA-seq, CTCF, FOXA1, AR, H3K27ac and RNA Pol II ChIP-seq and RNA Pol II ChIA-PET data from LNCaP (A) and VCaP cells (B), respectively. In addition, phospho RNA Pol II and ERG ChIP-seq is described for VCaP cells. The *KLK3* gene and up-stream regions are highlighted in light blue color.

Figure S7. Related to Figure 5. Integrated genome view of gene *KLK3* and its adjacent regions based on H3K27ac ChIP-seq data. H3K27ac ChIP-seq in 96

datasets show universal peaks in the constitutively active distal enhancer of *KLK3*. The *KLK3* gene and up-stream regions are highlighted in light-blue color.

Figure S8. Related to Figure 6. (A-B) Integrated genome view of gene *MYC* and its adjacent regions representing RNA-seq, CTCF, FOXA1, AR, H3K27ac and RNA Pol II ChIP-seq, and RNA Pol II ChIA-PET are shown for LNCaP (A) and VCaP cells (B), respectively. In addition, phospho RNA Pol II and ERG ChIP-seq is described for VCaP cells. The MYC gene locus is highlighted in light-blue color. SNP sites located in *MYC* neighborhood are shown. (C-D) Integrated genome view of *EZH2* locus and its adjacent regions based on data from RNA-seq, ChIP-seq of CTCF, FOXA1, AR, H3K27ac and RNA Pol II, ChIA-PET of RNA Pol II are shown in LNCaP (C) and VCaP (D) cells, respectively. Additionally, phospho RNA Pol II and ERG ChIP-seq is described for VCaP cells. The EZH2 gene locus is highlighted in light-blue color. (E) Integrated genome view of gene *MYC* and its adjacent regions based on H3K27ac ChIP-seq data in a panel of cell lines. The *MYC* gene region is highlighted in light-blue color. (F) This scatterplot shows the correlation between RNA expression level of *MYC* and the maximum enhancer peak score within the region shown in panel E (chr8: 127,800kb-129,400kb). All samples which have expressions of *MYC* (TPM>1) are retained in this plot, and the enhancer score is measured by the maximum ChIP-seq peak height. The correlation efficient is calculated by Pearson correlation and p-value is labeled. (G) Integrated genome view of gene *EZH2* and its adjacent regions based on H3K27ac ChIP-seq data in a panel of cell lines. The *EZH2* gene region is highlighted in light-blue color. (H) This scatterplot shows the correlation between RNA expression level of *EZH2* and the maximum enhancer peak score within the region shown in panel G (chr7: 148,400kb-149,400kb). All samples which have expression of *EZH2* (TPM>1) are included in this plot, and the enhancer score is measured by the maximum ChIP-seq peak height. The correlation efficient is calculated by Pearson correlation and p-value is labeled.

Figure S9. Related to Figure 9. (A) siRNA mediated knockdown of *FAM57A*, *GEMIN4* and *VPS53* genes resulted in a modest increase in the viability of LNCaP cells as shown by Cell titer glow analysis conducted 7 days post siRNA treatment. (B) Validation of siRNA-based gene knockdown in LNCaP cells. qRT-PCR analysis of the expression of *VPS53*, *GEMIN4* and *FAM57A* genes with the treatment of siRNA as indicated. (* $P<0.05$, ** $P<0.01$, *** $P<0.001$, **** $P<0.0001$, by two-tailed Student's t-test; Error bars, standard deviation of 3 technical replicates).

Figure S1 (Page 1)

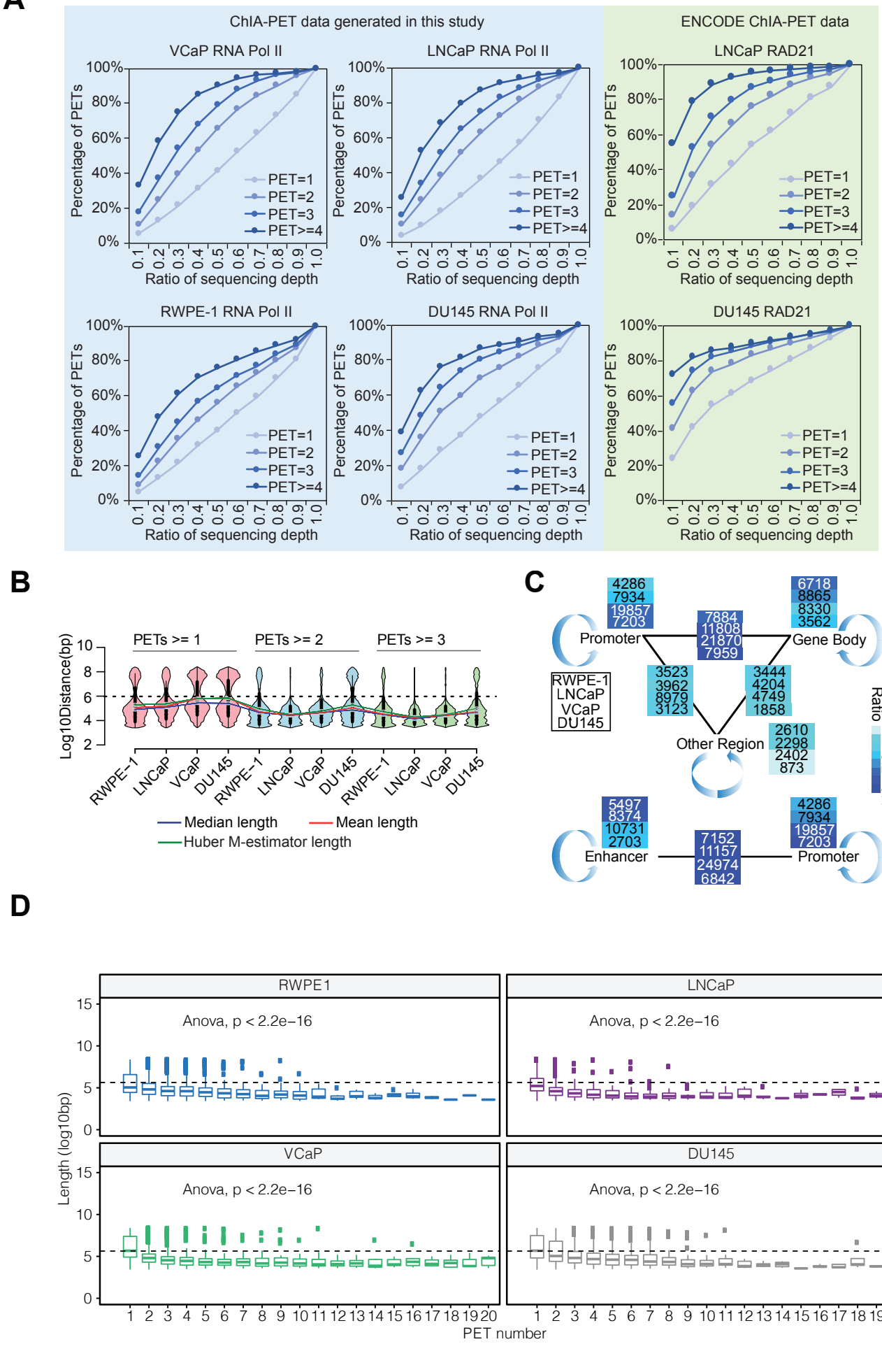
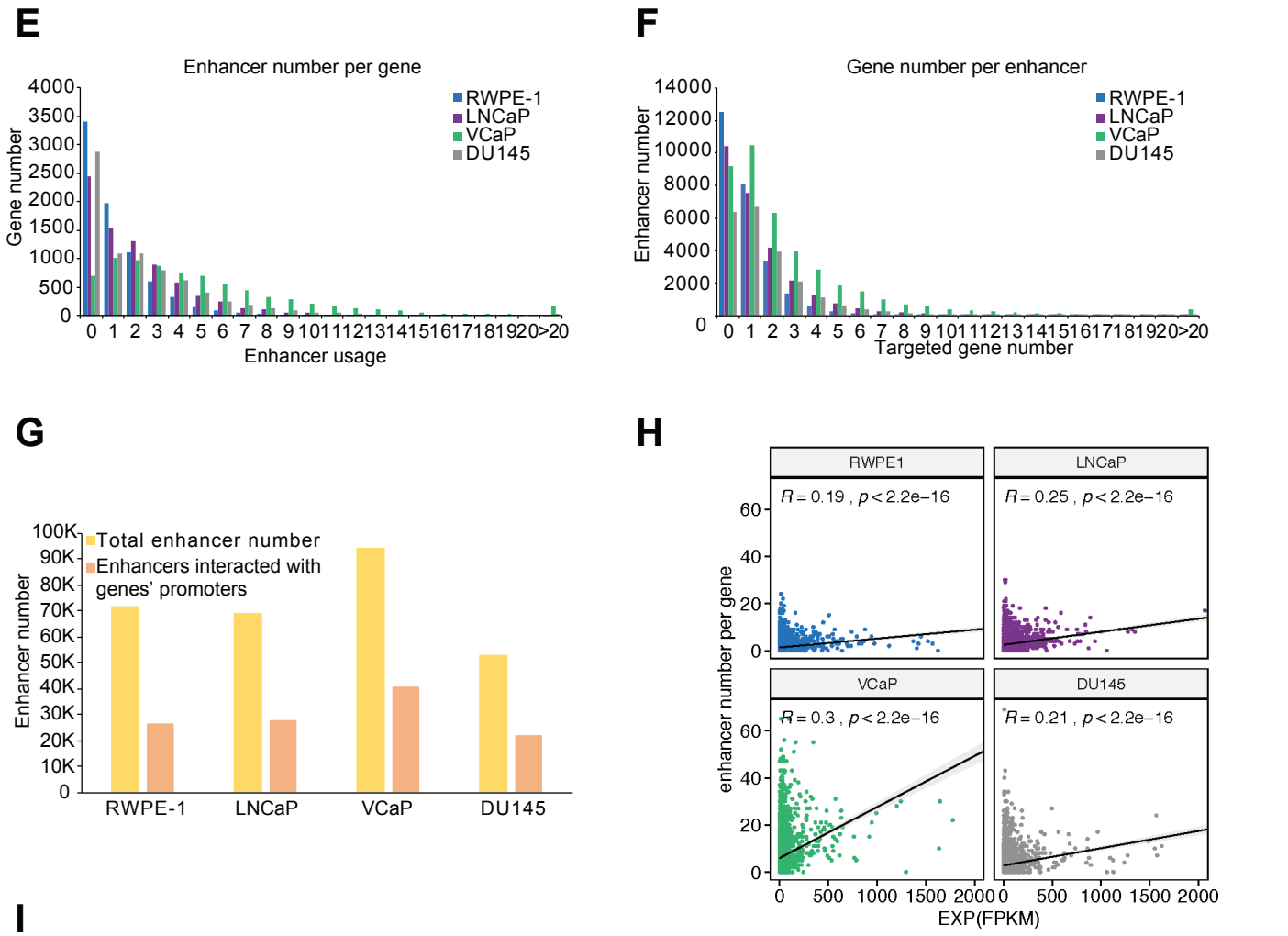
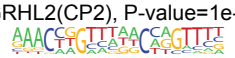
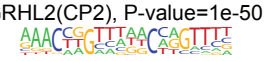


Figure S1 (Page 2)



J

RWPE-1	LNCaP	VCaP	DU145
Enhancer	Enhancer	Enhancer	Enhancer
Fra1(bZIP), P-value=1e-514	FOXA1(Forkhead), P-value=1e-285	CTCF(Zf), P-value=1e-323	Fra1(bZIP), P-value=1e-414
			
p53(p53), P-value=1e-58	NF1(CTF), P-value=1e-74	FOXA1(Forkhead), P-value=1e-305	CTCF(Zf), P-value=1e-27
			
TEAD4(TEA), P-value=1e-44	HOXB13(Homeobox), P-value=1e-54	GRHL2(CP2), P-value=1e-167	ETV1(ETS), P-value=1e-24
			
Fli1(ETS), P-value=1e-31	GRHL2(CP2), P-value=1e-50	NF1(CTF), P-value=1e-156	TEAD(TEA), P-value=1e-24
			
ERG(ETS), P-value=1e-30	EHF(ETS), P-value=1e-48	ETV1(ETS), P-value=1e-120	
			
KLF5(Zf), P-value=1e-30	Jun-AP1(bZIP), P-value=1e-48	HOXB13(Homeobox), P-value=1e-106	
			
CTCF(Zf), P-value=1e-22	CTCF(Zf), P-value=1e-32	Fli1(ETS), P-value=1e-104	
			
	Fli1(ETS), P-value=1e-22	GRE(NR), P-value=1e-71	
			

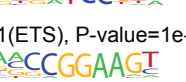
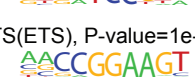
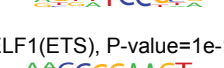
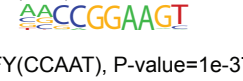
RWPE-1	LNCaP	VCaP	DU145
Promoter	Promoter	Promoter	Promoter
Atf3(bZIP), P-value=1e-21	Elk4(ETS), P-value=1e-38	Elk4(ETS), P-value=1e-57	NFY(CCAAT), P-value=1e-29
			
Elk4(ETS), P-value=1e-20	ELF1(ETS), P-value=1e-35	ETS(ETS), P-value=1e-47	Fli1(ETS), P-value=1e-25
			
ELF1(ETS), P-value=1e-13	NFY(CCAAT), P-value=1e-25	NFY(CCAAT), P-value=1e-37	YY1(Zf), P-value=1e-23
			
NFY(CCAAT), P-value=1e-11	YY1(Zf), P-value=1e-24	YY1(Zf), P-value=1e-37	ELF1(ETS), P-value=1e-22
			
BMYB(HTH), P-value=1e-6	BMYB(HTH), P-value=1e-12	TFE3(bHLH), P-value=1e-7	Sp1(Zf), P-value=1e-14
YY1(Zf), P-value=1e-5	NRF(NRF), P-value=1e-8	NRF(NRF), P-value=1e-6	Atf3(bZIP), P-value=1e-11
	TFE3(bHLH), P-value=1e-7	Sp1(Zf), P-value=1e-5	E2F4(E2F), P-value=1e-10
			NRF(NRF), P-value=1e-7

Figure S2

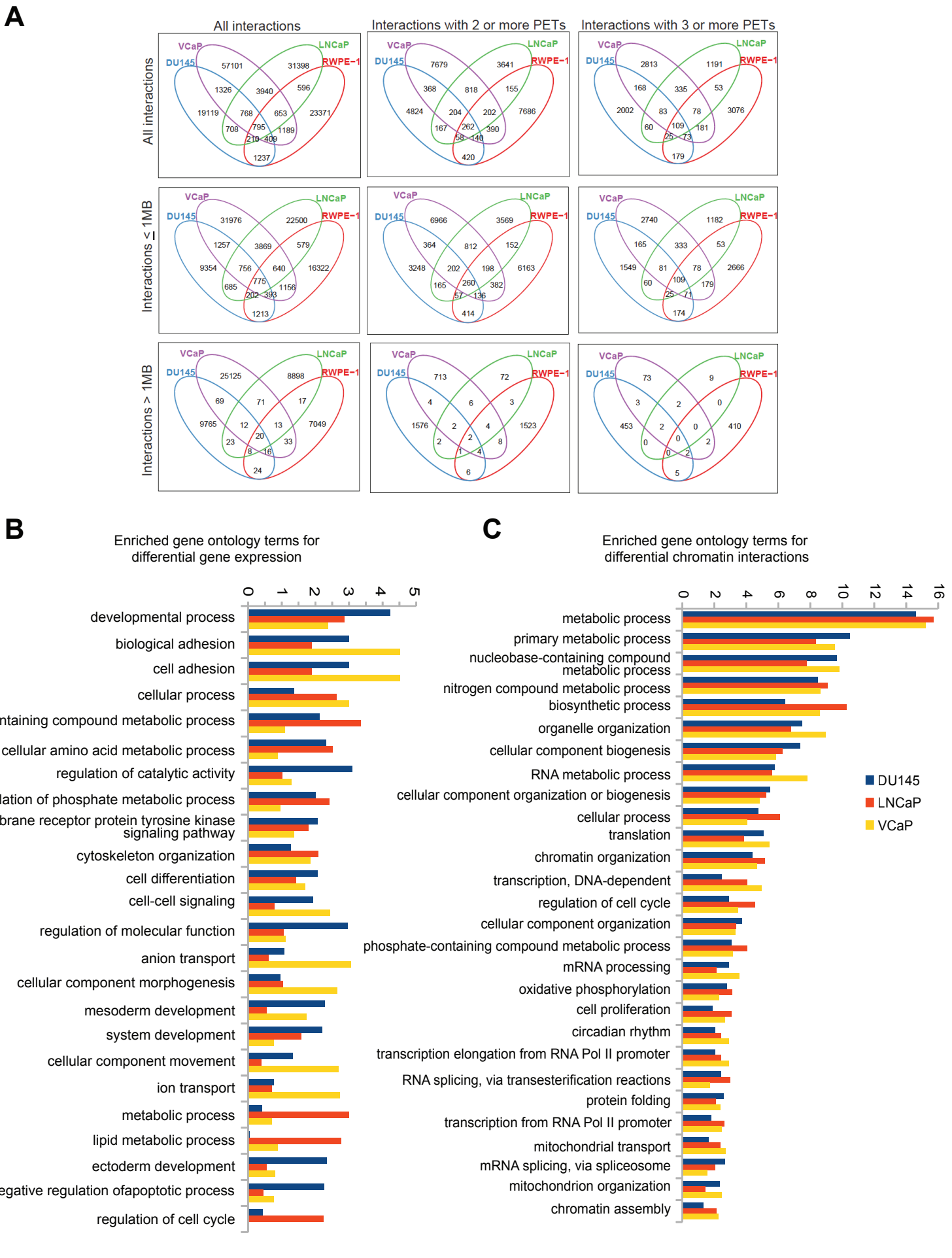

A
All interactions
Interactions with 2 or more PETs
Interactions with 3 or more PETs
All interactions
Interactions < 1MB
Interactions > 1MB
VCaP
LNCaP
DU145
RWPE-1
57101
31398
1326
3940
596
19119
768
653
23371
708
795
1189
210
409
1237
7679
3641
368
818
155
4824
204
202
7686
167
262
390
58
140
420
2813
1191
168
335
53
2002
83
78
3076
60
109
181
25
73
179
31976
22500
1257
3869
579
9354
756
640
16322
685
775
1156
202
393
1213
6966
3569
364
812
152
3248
202
198
6163
165
260
382
57
136
414
2740
1182
165
333
53
1549
81
78
2666
60
109
179
25
71
174
25125
8898
69
71
17
9765
12
13
7049
23
20
33
8
16
24
713
72
4
6
3
1576
2
4
1523
2
2
1
4
8
6
73
9
3
2
0
453
2
0
410
0
0
0
2
5
B
Enriched gene ontology terms for differential gene expression
0
1
2
3
4
5
developmental process
biological adhesion
cell adhesion
cellular process
phosphate-containing compound metabolic process
cellular amino acid metabolic process
regulation of catalytic activity
regulation of phosphate metabolic process
transmembrane receptor protein tyrosine kinase signaling pathway
cytoskeleton organization
cell differentiation
cell-cell signaling
regulation of molecular function
anion transport
cellular component morphogenesis
mesoderm development
system development
cellular component movement
ion transport
metabolic process
lipid metabolic process
ectoderm development
negative regulation ofapoptotic process
regulation of cell cycle
C
Enriched gene ontology terms for differential chromatin interactions
0
2
4
6
8
10
12
14
16
metabolic process
primary metabolic process
nucleobase-containing compound metabolic process
nitrogen compound metabolic process
biosynthetic process
organelle organization
cellular component biogenesis
RNA metabolic process
cellular component organization or biogenesis
cellular process
translation
chromatin organization
transcription, DNA-dependent
regulation of cell cycle
cellular component organization
phosphate-containing compound metabolic process
mRNA processing
oxidative phosphorylation
cell proliferation
circadian rhythm
transcription elongation from RNA Pol II promoter
RNA splicing, via transesterification reactions
protein folding
transcription from RNA Pol II promoter
mitochondrial transport
mRNA splicing, via spliceosome
mitochondrion organization
chromatin assembly
DU145
LNCaP
VCaP

Figure S3

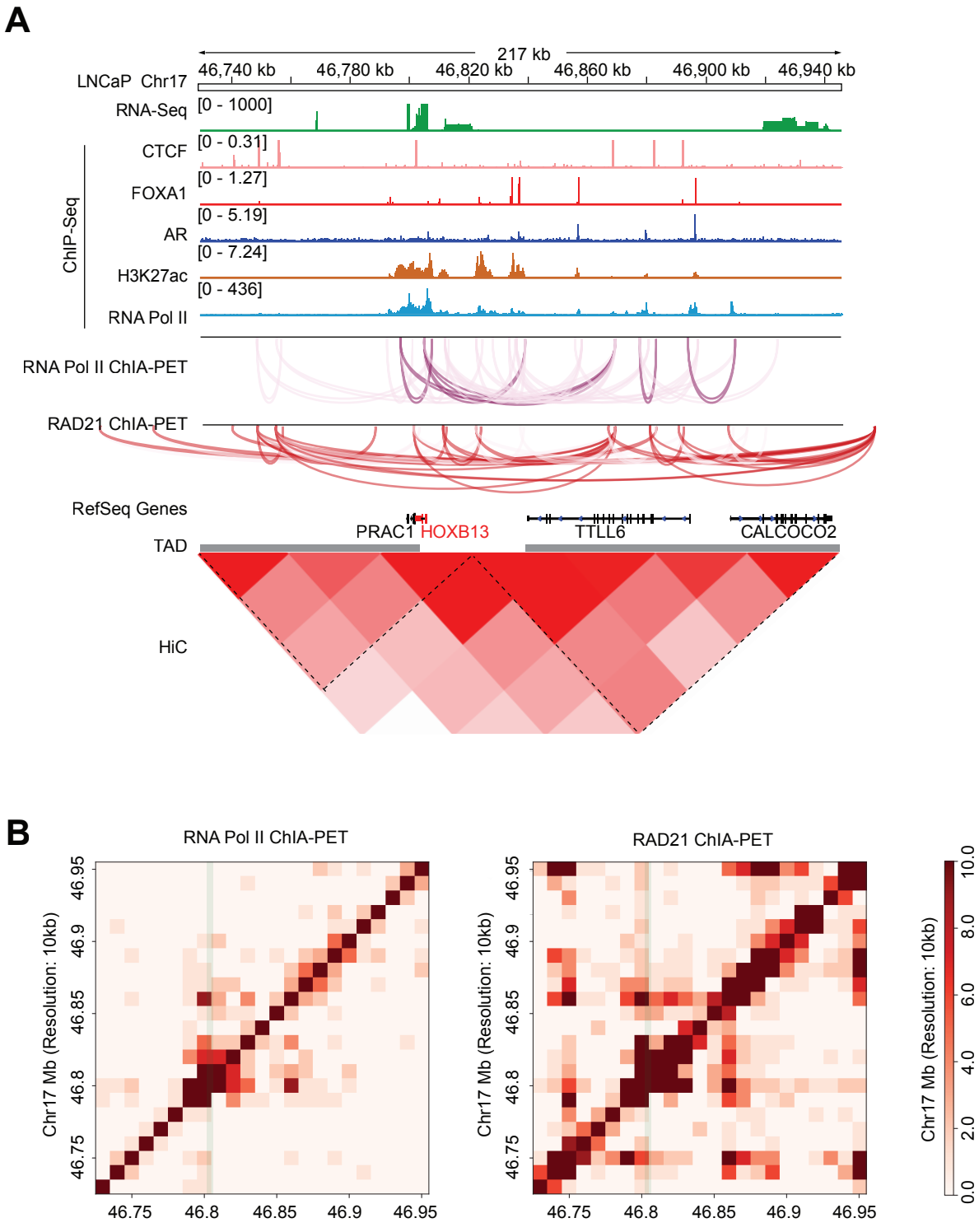
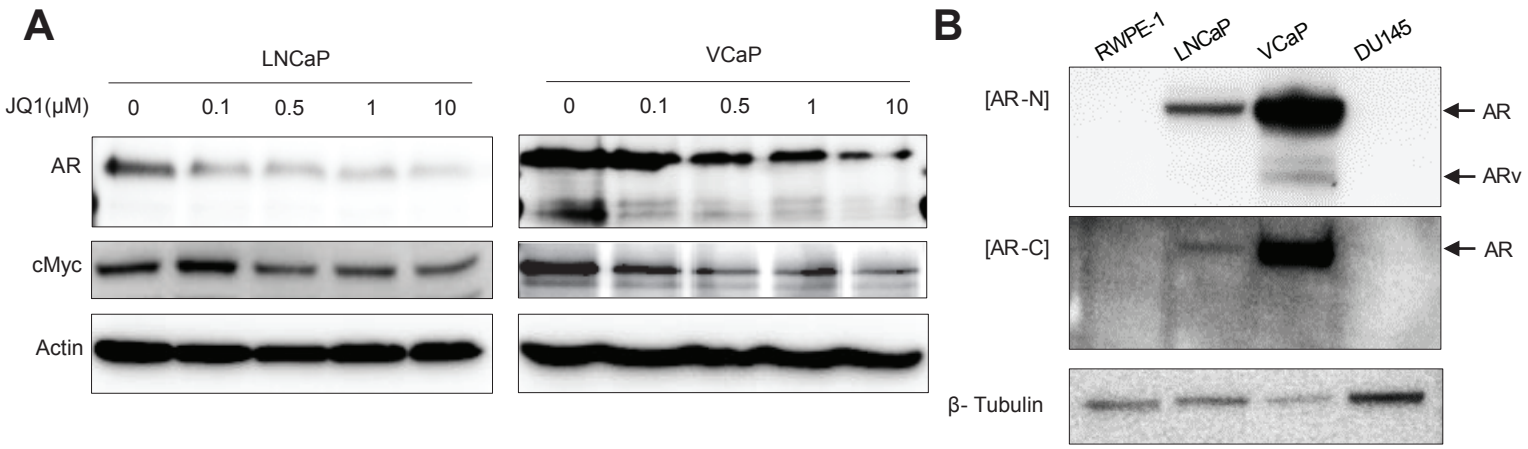
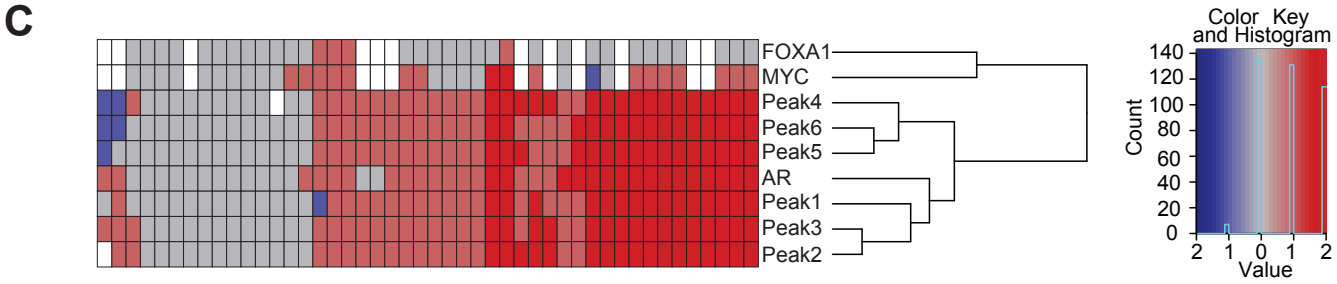
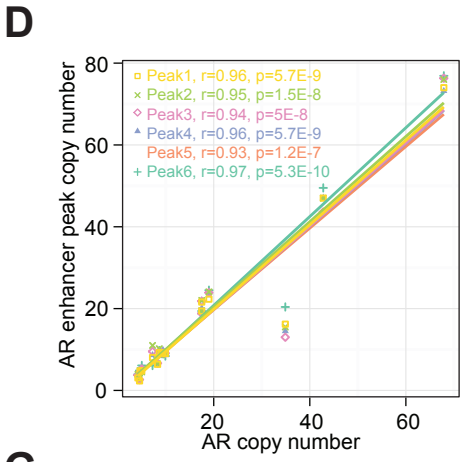
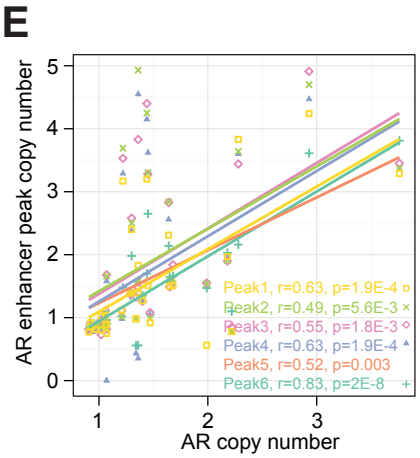

A
217 kb
LNCaP Chr17
46,740 kb
46,780 kb
46,820 kb
46,860 kb
46,900 kb
46,940 kb
RNA-Seq
[0 - 1000]
ChIP-Seq
CTCF
[0 - 0.31]
FOXA1
[0 - 1.27]
AR
[0 - 5.19]
H3K27ac
[0 - 7.24]
RNA Pol II
[0 - 436]
RNA Pol II ChIA-PET
RAD21 ChIA-PET
RefSeq Genes
PRAC1
HOXB13
TTLL6
CALCOCO2
TAD
HiC
B
RNA Pol II ChIA-PET
RAD21 ChIA-PET
Chr17 Mb (Resolution: 10kb)
46.75
46.8
46.85
46.9
46.95
0.0
2.0
4.0
6.0
8.0
10.0

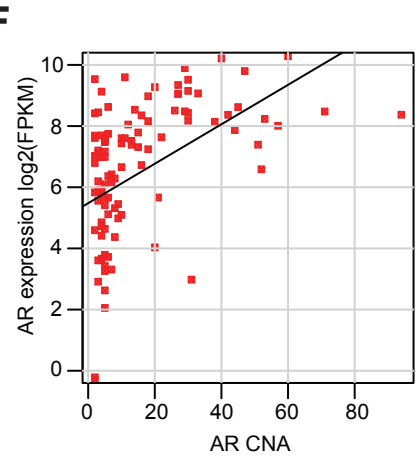
Figure S4


A
LNCaP
VCaP
JQ1(μM)
0
0.1
0.5
1
10
AR
cMyc
Actin
B
RWPE-1
LNCaP
VCaP
DU145
[AR-N]
[AR-C]
AR
ARv
β- Tubulin


C
FOXA1
MYC
Peak4
Peak6
Peak5
AR
Peak1
Peak3
Peak2
Color Key and Histogram
Count
Value


D
AR enhancer peak copy number
AR copy number
Peak1, r=0.96, p=5.7E-9
Peak2, r=0.95, p=1.5E-8
Peak3, r=0.94, p=5E-8
Peak4, r=0.96, p=5.7E-9
Peak5, r=0.93, p=1.2E-7
Peak6, r=0.97, p=5.3E-10


E
AR enhancer peak copy number
AR copy number
Peak1, r=0.63, p=1.9E-4
Peak2, r=0.49, p=5.6E-3
Peak3, r=0.55, p=1.8E-3
Peak4, r=0.63, p=1.9E-4
Peak5, r=0.52, p=0.003
Peak6, r=0.83, p=2E-8


F
AR expression log2(FPKM)
AR CNA

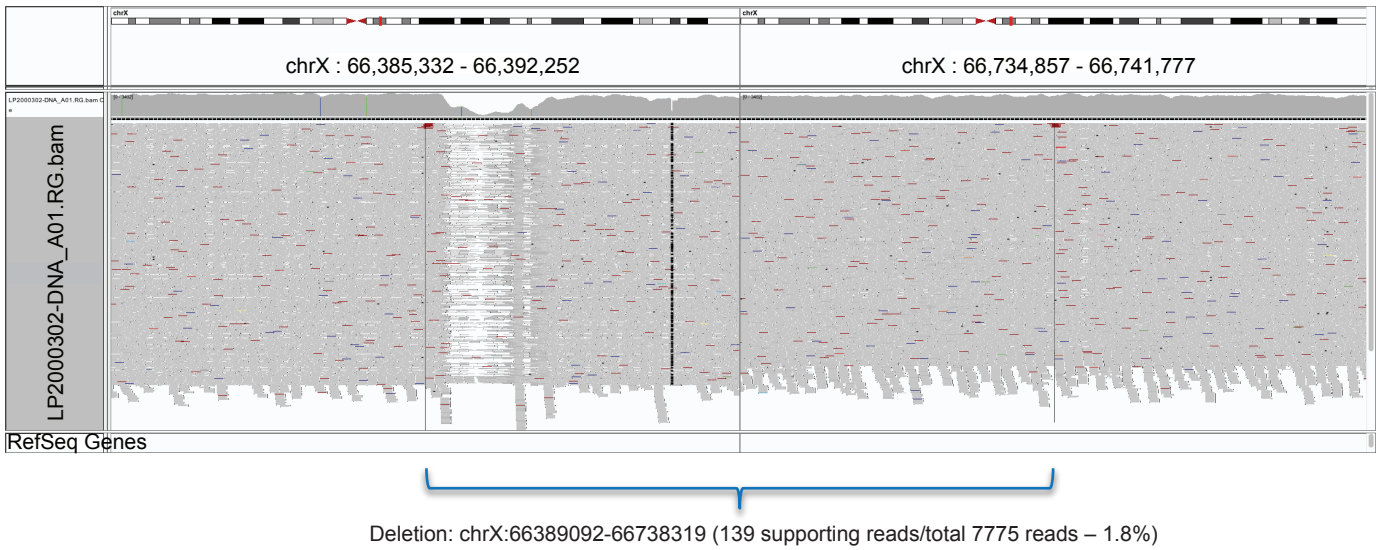
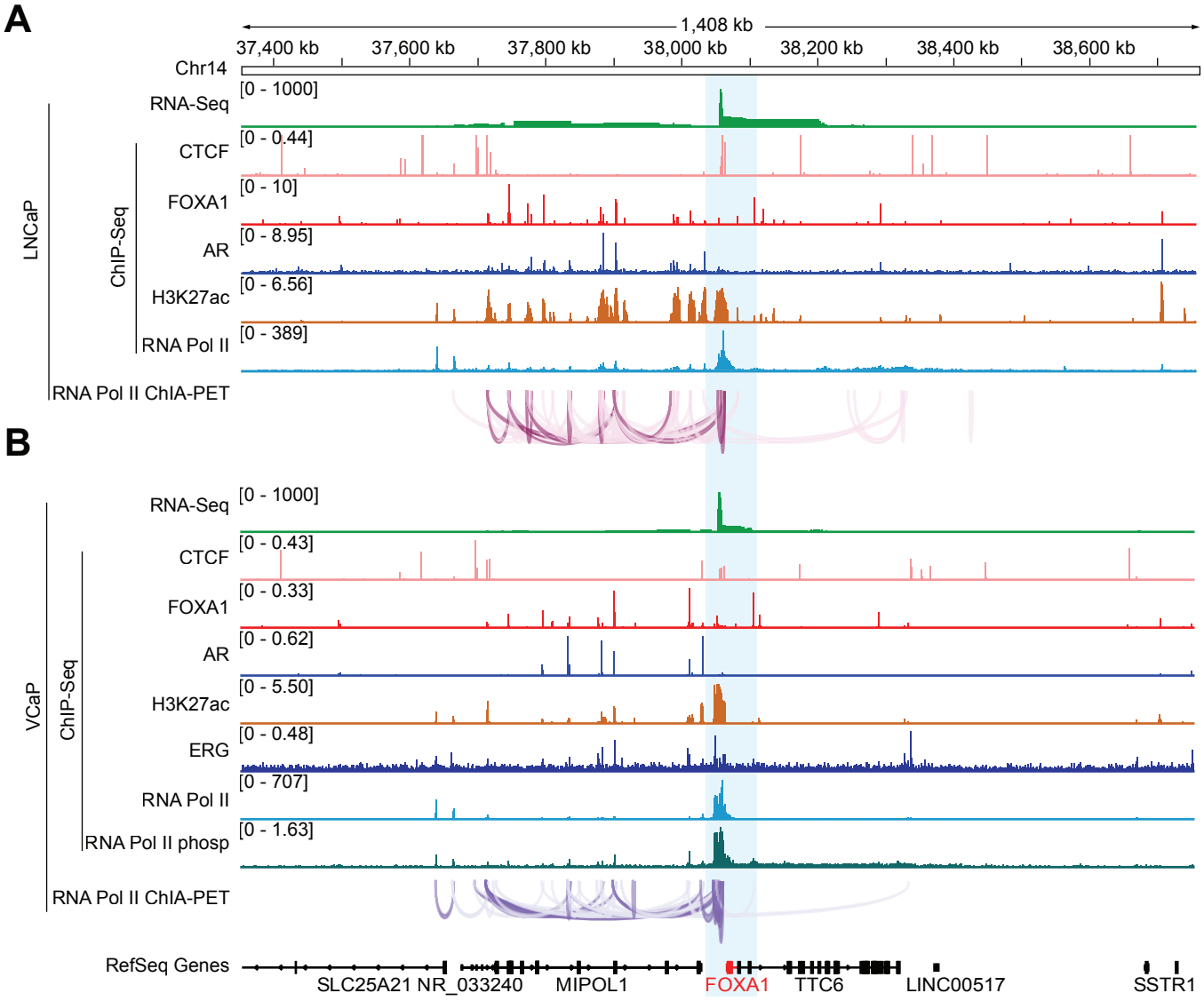

G
chrX : 66,385,332 - 66,392,252
chrX : 66,734,857 - 66,741,777
LP2000302-DNA_A01.RG.bam
RefSeq Genes
Deletion: chrX:66389092-66738319 (139 supporting reads/total 7775 reads – 1.8%)

Figure S5


A
1,408 kb
37,400 kb
37,600 kb
37,800 kb
38,000 kb
38,200 kb
38,400 kb
38,600 kb
Chr14
LNCaP
ChIP-Seq
RNA-Seq
[0 - 1000]
CTCF
[0 - 0.44]
FOXA1
[0 - 10]
AR
[0 - 8.95]
H3K27ac
[0 - 6.56]
RNA Pol II
[0 - 389]
RNA Pol II ChIA-PET
B
VCaP
ChIP-Seq
RNA-Seq
[0 - 1000]
CTCF
[0 - 0.43]
FOXA1
[0 - 0.33]
AR
[0 - 0.62]
H3K27ac
[0 - 5.50]
ERG
[0 - 0.48]
RNA Pol II
[0 - 707]
RNA Pol II phosp
[0 - 1.63]
RNA Pol II ChIA-PET
RefSeq Genes
SLC25A21
NR_033240
MIPOL1
FOXA1
TTC6
LINC00517
SSTR1

C

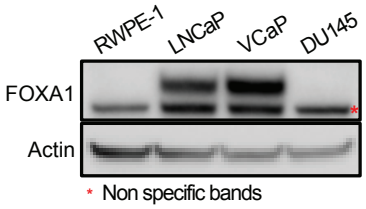

RWPE-1
LNCaP
VCaP
DU145
FOXA1
Actin
* Non specific bands

Figure S6

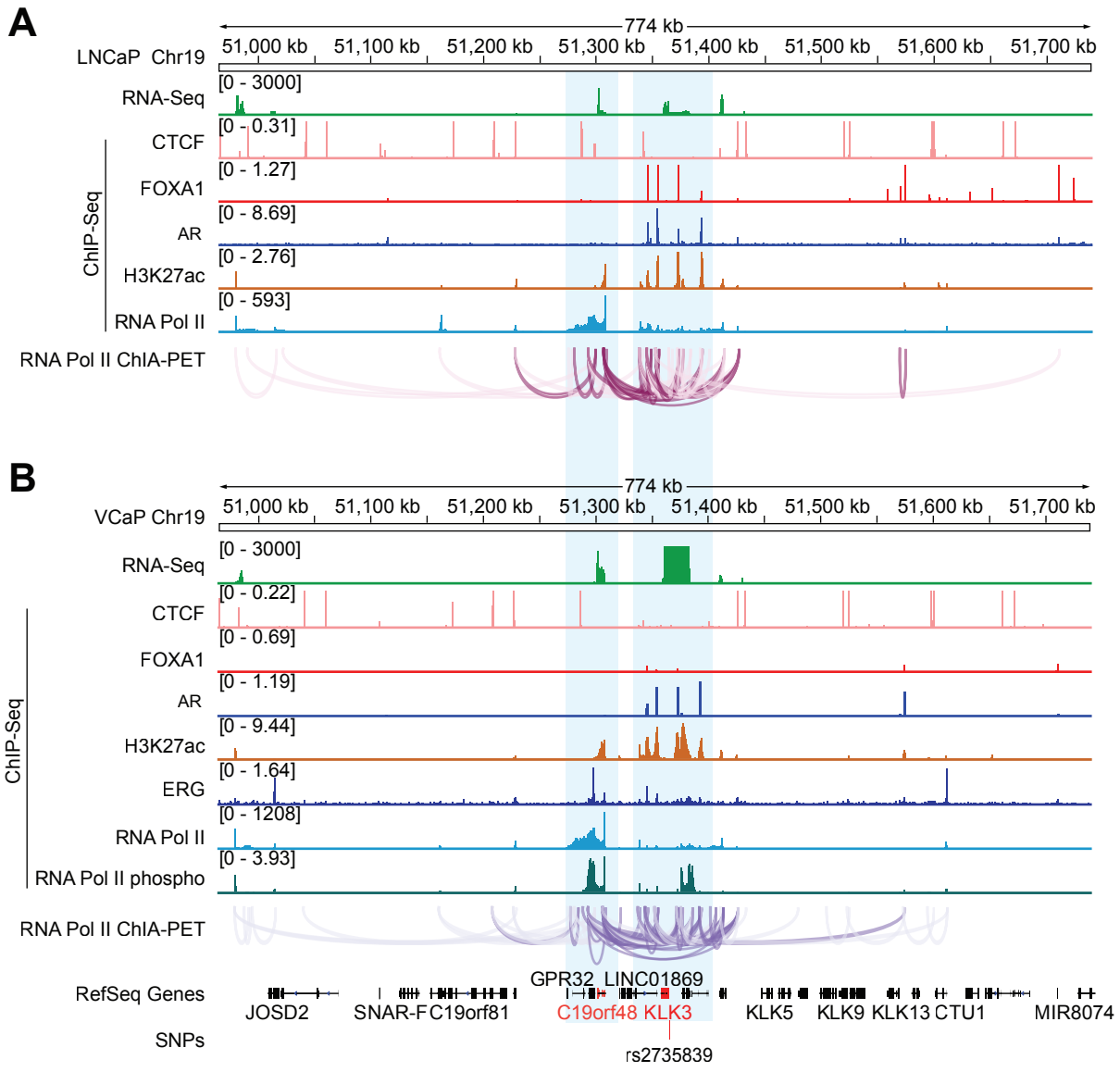

A
774 kb
LNCaP Chr19
51,000 kb
51,100 kb
51,200 kb
51,300 kb
51,400 kb
51,500 kb
51,600 kb
51,700 kb
RNA-Seq
[0 - 3000]
CTCF
[0 - 0.31]
FOXA1
[0 - 1.27]
AR
[0 - 8.69]
H3K27ac
[0 - 2.76]
RNA Pol II
[0 - 593]
ChIP-Seq
RNA Pol II ChIA-PET
B
774 kb
VCaP Chr19
51,000 kb
51,100 kb
51,200 kb
51,300 kb
51,400 kb
51,500 kb
51,600 kb
51,700 kb
RNA-Seq
[0 - 3000]
CTCF
[0 - 0.22]
FOXA1
[0 - 0.69]
AR
[0 - 1.19]
H3K27ac
[0 - 9.44]
ERG
[0 - 1.64]
RNA Pol II
[0 - 1208]
RNA Pol II phospho
[0 - 3.93]
ChIP-Seq
RNA Pol II ChIA-PET
RefSeq Genes
GPR32
LINC01869
JOSD2
SNAR-F
C19orf81
C19orf48
KLK3
KLK5
KLK9
KLK13
CTU1
MIR8074
SNPs
rs2735839

Figure S7


774 kb
Chr19
51,000 kb
51,200 kb
51,400 kb
51,600 kb
Ascending aorta, ENCFF002VCM [0 - 96]
SK-N-SH, ENCFF020DAV [0 - 47]
Spleen, ENCFF036NIB [0 - 252]
MCF-7, ENCFF051GGA [0 - 282]
Thyroid gland, ENCFF051GYP [0 - 33]
Neutrophil, ENCFF065VTJ [0 - 605]
Upper lobe of left lung, ENCFF071QUR [0 - 52]
SK-N-SH, ENCFF073KYI [0 - 303]
Endothelial cell of umbilical vein, ENCFF089DEG [0 - 77]
Uterus, ENCFF090IUI [0 - 113]
GM23338, ENCFF132AIF [0 - 50]
Stomach, ENCFF134NHD [0 - 159]
Fibroblast of lung, ENCFF138ZMZ [0 - 41]
Body of pancreas, ENCFF147LSN [0 - 24]
A549, ENCFF153QRX [0 - 362]
Sigmoid colon, ENCFF156QWL [0 - 117]
Adrenal gland, ENCFF162SNB [0 - 7.65]
HEK293, ENCFF175HIN [0 - 182]
CD14-positive monocyte, ENCFF195PCP [0 - 353]
Adrenal gland, ENCFF217RXL [0 - 55]
Hepatocyte, ENCFF227OHE [0 - 128]
DND-41, ENCFF233SGJ [0 - 187]
Gastrocnemius medialis, ENCFF241KYN [0 - 218]
Adrenal gland, ENCFF256CHZ [0 - 103]
Osteoblast, ENCFF261BXV [0 - 132]
MCF-7, ENCFF276WAM [0 - 12]
KOPT-K1, ENCFF277BMG [0 - 444]
Esophagus muscularis mucosa, ENCFF282WYM [0 - 115]
MM.1S, ENCFF284AZM [0 - 237]
Peyer's patch, ENCFF284VHL [0 - 102]
Body of pancreas, ENCFF301QHM [0 - 177]
Mammary epithelial cell, ENCFF305KUC [0 - 107]
KMS-11, ENCFF312KWS [0 - 233]
Esophagus squamous epithelium, ENCFF320HXP [0 - 92]
Smooth muscle cell, ENCFF323VHZ [0 - 98]
Thoracic aorta, ENCFF344YDV [0 - 78]
Fibroblast of dermis, ENCFF350EFY [0 - 87]
Thyroid gland, ENCFF367IJJ [0 - 13]
A549, ENCFF384KPD [0 - 349]
HCT116, ENCFF401XSL [0 - 153]
ACC112, ENCFF420AVK [0 - 158]
GM12878, ENCFF440GZA [0 - 45]
HCT116, ENCFF445BLD [0 - 173]
RWPE2, ENCFF453JGO [0 - 225]
Ascending aorta, ENCFF472VCY [0 - 82]
OCI-LY3, ENCFF530OSB [0 - 273]
Body of pancreas, ENCFF554AUP [0 - 213]
B cellENCFF561, APY [0 - 182]
Tibial nerve, ENCFF563ABT [0 - 51]
Neural cell, ENCFF574HAY [0 - 244]
Bipolar neuron, ENCFF593QNT [0 - 708]
Coronary artery, ENCFF609WYJ [0 - 82]
22Rv1, ENCFF615PSS [0 - 3263]
Karpas-422, ENCFF618INJ [0 - 96]
RefSeq Genes
LINC01869
LRRC4B C19orf48 KLK3 KLK7 MIR8074
774 kb
Chr19
51,000 kb
51,200 kb
51,400 kb
51,600 kb
22Rv1, ENCFF620QMB [0 - 5804]
OCI-LY1, ENCFF643VAM [0 - 61]
RWPE1, ENCFF648UKZ [0 - 4605]
Gastrocnemius medialis, ENCFF649TBC [0 - 158]
Thoracic aorta, ENCFF652JBU [0 - 34]
DOHH2, ENCFF656KBE [0 - 217]
PC-9, ENCFF670FVV [0 - 208]
Neuroepithelial stem cell, ENCFF678YWF [0 - 213]
Right lobe of liver, ENCFF680QZX [0 - 125]
Fibroblast of arm, ENCFF704JGE [0 - 125]
HeLa-S3, ENCFF706HLO [0 - 723]
SU-DHL-6, ENCFF710FJI [0 - 143]
Heart left ventricle, ENCFF729BAZ [0 - 138]
Vagina, ENCFF736SWP [0 - 74]
Panc1, ENCFF750VPZ [0 - 25]
Prostate, ENCFF753JCJ [0 - 78]
Keratinocyte, ENCFF757GRD [0 - 54]
OCI-LY7, ENCFF764QRU [0 - 167]
H1-hESC, ENCFF766NYI [0 - 62]
RWPE2, ENCFF770MSW [0 - 322]
Transverse colon, ENCFF780IFS [0 - 79]
Skeletal muscle myoblast, ENCFF788PRN [0 - 58]
Right atrium auricular region, ENCFF790FPZ [0 - 106]
A673, ENCFF794NZR [0 - 261]
Epithelial cell of prostate, ENCFF823OTM [0 - 2094]
K562, ENCFF840LLW [0 - 115]
Neural progenitor cell, ENCFF842EXW [0 - 68]
Astrocyte, ENCFF854VET [0 - 176]
Transverse colon, ENCFF856JEH [0 - 81]
Breast epithelium, ENCFF865HUX [0 - 96]
HepG2, ENCFF866RVV [0 - 150]
Gastroesophageal sphincter, ENCFF870AFA [0 - 85]
SK-N-MC, ENCFF874TEV [0 - 248]
VCaP, ENCFF884RIP [0 - 2967]
Thyroid gland, ENCFF908ZKL [0 - 65]
Myotube, ENCFF919CDP [0 - 88]
Transverse colon, ENCFF939NAG [0 - 39]
PC-3, ENCFF940SZQ [0 - 260]
Gastrocnemius medialis, ENCFF957CMX [0 - 355]
Fibroblast of lung, ENCFF977TER [0 - 31]
Cardiac muscle cell, ENCFF987RQQ [0 - 350]
C4-2B, ENCFF994HZA [0 - 2040]
RefSeq Genes
LINC01869
LRRC4B C19orf48 KLK3 KLK7 MIR8074

Figure S8 (Page 1)

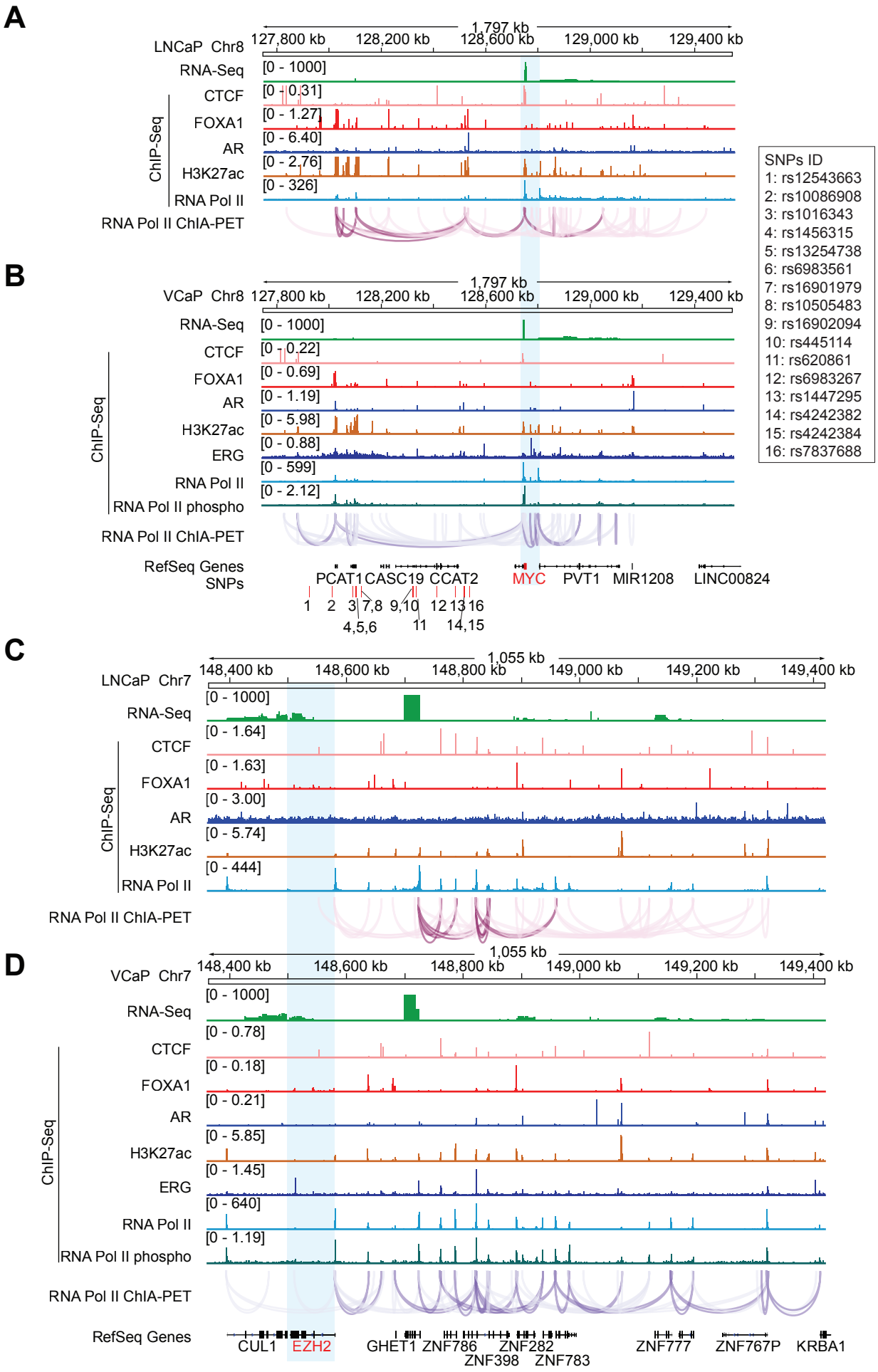


Figure S8 (Page 2)

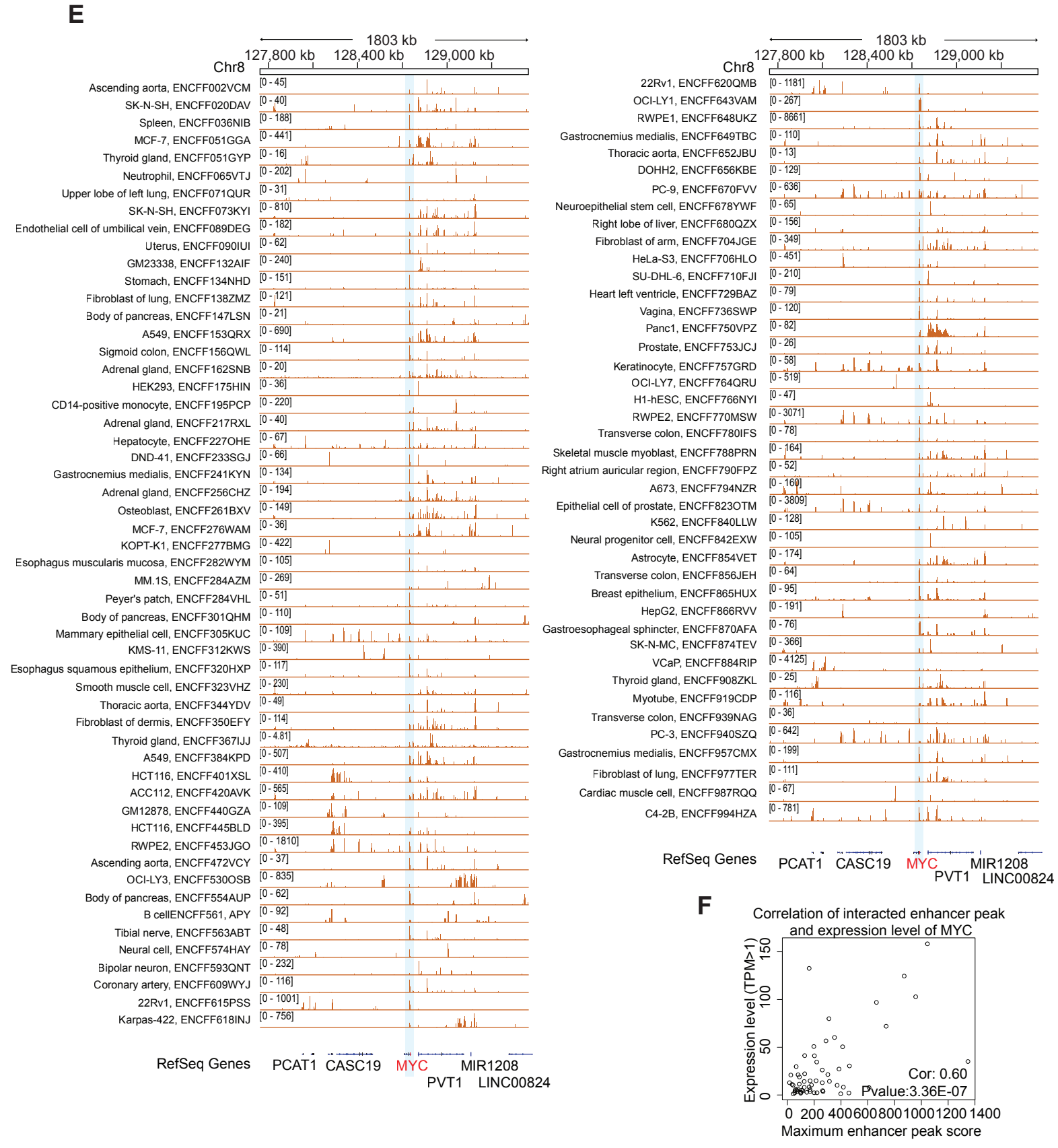


Figure S8 (Page 3)

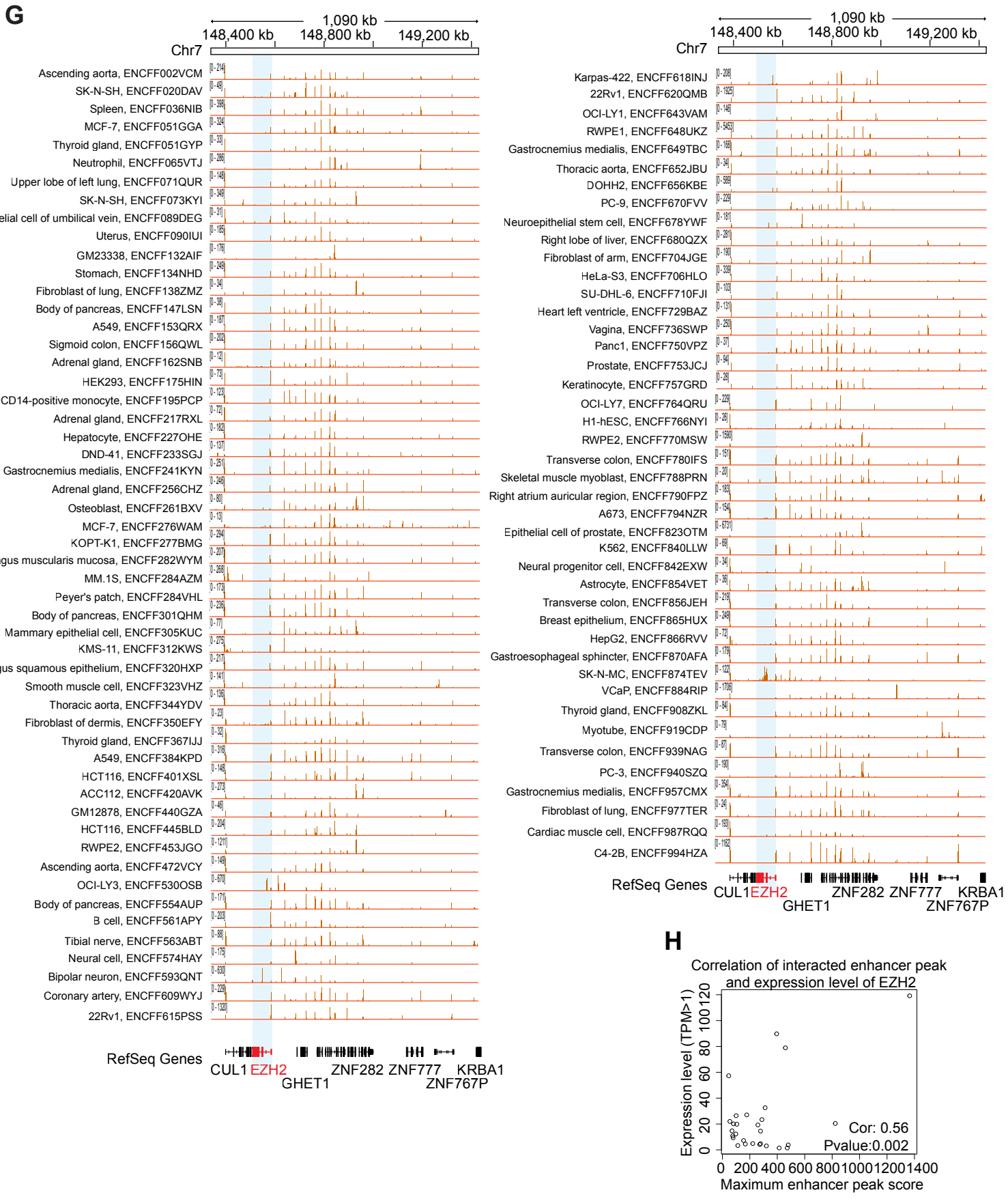
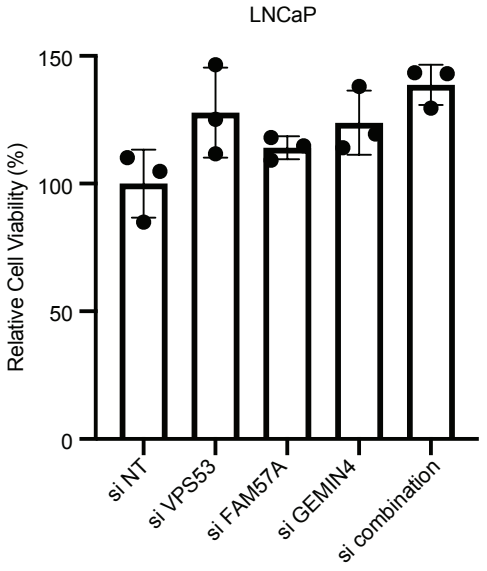


Figure S9

A


LNCaP
Relative Cell Viability (%)
150
100
50
0
si NT
si VPS53
si FAM57A
si GEMIN4
si combination

B

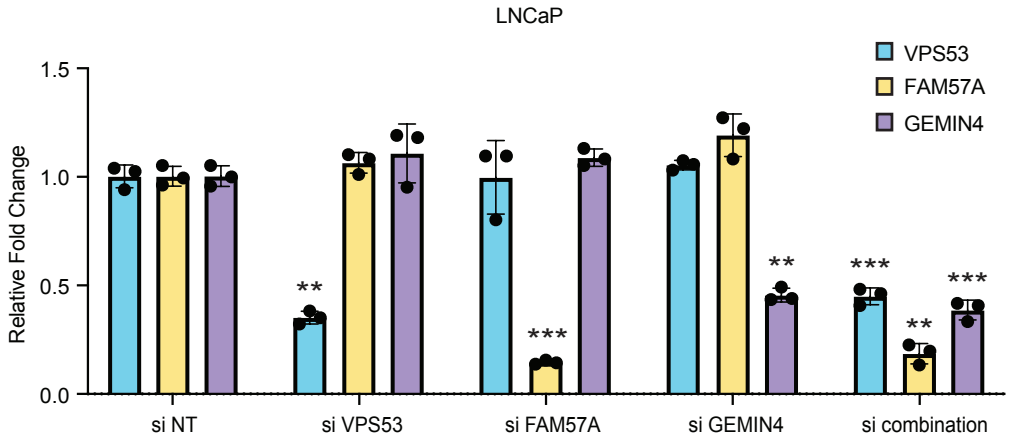

LNCaP
VPS53
FAM57A
GEMIN4
Relative Fold Change
1.5
1.0
0.5
0.0
**
***
**
***
**
***
si NT
si VPS53
si FAM57A
si GEMIN4
si combination

Table S1. Related to Figures 1 and 2. Metadata table for the datasets used in this study.

Experiment	Cell line	Molecule	Source	GEO/ENCODE ID
RNA-seq	DU145	/	This study	GSE121022
ChIP-seq	DU145	H3K27ac	This study	GSM3424003
ChIP-seq	DU145	CTCF	This study	GSM3424002
ChIAPET	DU145	RNA Pol II	This study	GSM3423999
ChIAPET	DU145	RAD21	ENCODE	ENCSR672RHL
RNA-seq	LNCaP	/	This study	GSE121022
ChIP-seq	LNCaP	CTCF	GEO	GSM947528
ChIP-seq	LNCaP	AR	This study	GSM3424005
ChIP-seq	LNCaP	FOXA1	GEO	GSM2219864
ChIP-seq	LNCaP	H3K27ac	GEO	GSM1902615
ChIP-seq	LNCaP	H3K4me3	GEO	GSM945240
Dnase-seq	LNCaP	DNase I	GEO	GSM816637
ChIAPET	LNCaP	RNA Pol II	This study	GSM3423998
ChIAPET	LNCaP	Rad21	ENCODE	ENCSR011ITK
HiC	LNCaP	/	ENCODE	ENCSR346DCU
RNA-seq	RWPE-1	/	This study	GSE121022
ChIP-seq	RWPE-1	CTCF	ENCODE	ENCFF391XEK
ChIP-seq	RWPE-1	H3K27ac	GEO	GSM1541009
ChIAPET	RWPE-1	RNA Pol II	This study	GSM3424000
RNA-seq	VCaP	/	This study	GSE121022
ChIP-seq	VCaP	CTCF	GEO	GSE84432
ChIP-seq	VCaP	AR	GEO	GSM1328945
ChIP-seq	VCaP	FOXA1	GEO	GSM2058886
ChIP-seq	VCaP	H3K27ac	ENCODE	ENCFF190ZWT
ChIP-seq	VCaP	ERG	GEO	GSM1328979
ChIP-seq	VCaP	RNA Pol II phospho	GEO	GSM1328964
ChIAPET	VCaP	AR	GEO	GSM1327093
ChIAPET	VCaP	RNA Pol II	This study	GSM3423997

Table S2-1. Related to Figures 1 and 2. Metadata table for ChIA-PET sequenced in this study.

Cell Line	Total PETs[1]	Valid PETs[2]	Output PETs[3]	Uniquely Mapped PETs[4]	Total Broad Peaks	Self PETs[5]	PETs between Peaks	Intra PETs[6]	Inter PETs[7]
RWPE-1	278,678,565	256,081,201	213,698,640	90,123,783	86,711	10,235,232	184,671	47,469	137,202
LNCaP	342,749,728	244,073,222	152,299,081	135,848,887	65,335	12,972,501	159,283	51,365	107,918
VCaP	254,392,059	188,211,704	135,372,820	117,548,587	60,656	15,665,876	492,546	88,374	404,172
DU145	231,264,122	202,773,594	165,934,884	72,043,490	58,097	7,658,637	243,953	38,130	205,823

[1]Total PETs: total sequenced raw Pair End Tags.

[2]Valid PETs: total PETs after linker trimming that kept for alignment, regardless of whether they have linkers or not. (this is used for alignment).

[3]Output PETs: Successfully aligned on hg19 reference genome after removing unmapped reads and low MAPQ reads.

[4]Uniquely Mapped PETs: Uniquely mapped PETs after removing PCR duplicates. (This is used for following peak calling and interaction analysis).

[5]Self PETs: Self-ligation PET (two ends of the same DNA fragment).

[6]Intra PETs: intra-chromosomal PETs.

[7]Inter PETs: inter-chromosomal PETs.

Table S2-2. Related to Figures 1 and 2. Metadata table for ChIP-seq sequenced in this study.

Sample ID	Target	Cell line	Total reads	Mapped hg19 reads	Mapped hg19 ratio
DU145_INPUT	None	DU145	55,199,482	37,950,546	68.75%
DU145_H3K27ac	H3K27ac	DU145	44,248,954	35,526,292	80.29%
DU145_CTCF	CTCF	LNCaP	52,829,386	42,818,743	81.05%
LNCaP_INPUT	None	LNCaP	53,984,265	34,778,710	64.42%
LNCaP_AR	AR	LNCaP	83,915,780	55,429,089	66.05%

Table S2-3. Related to Figures 1 and 2. Metadata table for RNA-seq data sequenced in this study.

Cell line	Replicate	Total reads	Mapped rRNA reads	Mapped rRNA ratio	Retained reads	Mapped genome reads	Mapped genome ratio
*DU145	1	134,038,110	10,743,305	8.02%	123,294,805	113,512,312	92.07%
DU145	2	23,796,797	313,557	1.32%	23,483,240	22,513,624	95.87%
RWPE-1	1	1,228,613	35,375	2.88%	1,193,238	1,031,576	86.45%
RWPE-1	2	43,783,623	2,296,860	5.25%	41,486,763	39,111,293	94.30%
*RWPE-1	3	117,596,563	6,435,787	5.47%	111,160,776	102,784,548	92.46%
LNCaP	1	24,266,367	313,951	1.29%	23,952,416	22,311,856	93.15%
*LNCaP	2	100,996,513	1,359,705	1.35%	99,636,808	93,931,367	94.30%
LNCaP	3	36,843,667	282,480	0.77%	36,561,187	29,593,164	80.90%
*VCaP	1	119,025,099	5,905,928	4.96%	113,119,171	98,423,765	87.01%
VCaP	2	22,201,556	203,535	0.92%	21,998,021	19,614,776	89.17%

*Data shown in IGV track.

rRNA represents for ribosome RNAs.

Table S5. Related to Figure 2. Peak overlapping of RAD21, RNA Pol II, H3K27ac and CTCF for LNCaP and DU145 cells. The P-value of an overlapped number was calculated by using hypergeometric test, one-side testing of bigger than the overlapped number.

Cell Name	**Peak Name**	**RNA Pol II**		**RAD21**	
		RNA Pol II (All Peaks 62563)	RNA Pol II (Intra-interaction Peaks 33336)	RAD21 (All Peaks 65485)	RAD21 (Intra-interaction Peaks 55294)
LNCaP	CTCF (76292)	19952, P(X≥19952)=0.99	13147, P(X≥13147)=1	**36122, P(X≥36122)=8.41E-04**	**34109, P(X≥34109)=1.08E-13**
	RAD21 (65485)	27029, P(X≥27029)=1	17640, P(X≥17640)=0.99	---	---
	H3K27ac (62723)	**33649, P(X≥33649)=1.89E-08**	**23013, P(X≥23013)=2.91E-71**	23258, P(X≥23258)=1	21138, P(X≥21138)=1
		RNA Pol II (All Peaks 58097)	RNA Pol II (Intra-interaction Peaks 22063)	RAD21 (All Peaks 90274)	RAD21 (Intra-interaction Peaks 51326)
DU145	CTCF (36152)	9542, P(X≥9542)=1	5541, P(X≥5541)=1	**28936, P(X≥28936)=9.18E-139**	**26739, P(X≥26739)=0**
	RAD21 (90274)	25084, P(X≥25084)=1	14612, P(X≥14612)=1	---	---
	H3K27ac (50364)	**28448, P(X≥28448)=7.14E-09**	**20095, P(X≥20095)=0**	23138, P(X≥23138)=1	19664, P(X≥20095)=1

Table S6. Related to Figure 3. Baseline characteristics for those patients with aCGH performed.

	All patients N=27
Age at diagnosis, years Mean (SD)	62.6 (8.2)*
Histology, N (%)	
Adenocarcinoma	23 (85)
NA	4 (15)
Gleeson, N (%)	
≤7	7 (26)
>7	14 (52)
NA	6 (22)
Presenting PSA, ug/l	
Median (IQR)	110.5 (16.7-565.0)^
Metastatic at diagnosis, N (%)	
No	9 (33)
Yes	14 (52)
NA	4 (15)
Primary therapy, N (%)	
Radical	8 (30)
Palliative	17 (63)
NA	2 (7)
Biopsy site, N (%)	
Bone marrow	9 (33)
Lymph node	11 (41)
Liver	2 (7)
Soft tissue	1 (4)
Prostate	4 (15)

NA – not available, N - number, SD – standard deviation, IQR – interquartile range.

*Data not available for two patients

^3 PSA values missing

Table S8. Related to Figure 3. AR enhancers assays

Assay ID	Reporter 1 Dye	Reporter 1 Concentration (µM)	Forward Primer Concentration (µM)	Reverse Primer Concentration (µM)	Forward Primer Sequence	Reverse Primer Sequence	Reporter 1 Sequence	Amplicon Size
Peak1_CDAAAHV	FAM	5	18	18	CTGGATGGACCTTTGGGTTATCTTT	GCCCACTCCTATACACTTCTTCACT	CTTGGAGAGGAGGTAAATATG	80
Peak2_CDDJXNP	FAM	5	18	18	GGCAGATGGACAGGCAAGTTT	CCCACTCTCAGCCAGTTCTG	ACGTAAAGCAAGGCCACATAAA	73
peak3_CDFVKUJ	FAM	5	18	18	CTGTAGTTGACTGCAGAGGGAATTT	CATTCTTCTCTGTGCACTGTCTAGT	CTAGGAACCTCCCCAAATACCT	82
peak4_CDMFW39	FAM	5	18	18	TGTGTTTCTGGTTTTGTTAGTTATCTATTGCT	TCCTGACCCACAGAAATTGTGAAAA	AAGGTGCTAAGTTGTGAATAATTT	107
peak5_CD9HH46	FAM	5	18	18	CAATGCAACATGTTACAAATCCTTCTGA	GGTTCAGGCCCATTGATTTGG	CTTCTGGACGTGAGTGATAGTC	93
peak6_CDWCWJT	FAM	5	18	18	GCGCAGCCTATTTACATGATTTCA	GGTGATGATAGAGAGATAGTGGACCAA	AAGGTCAACAAGATAAGGTAATCTC	101

Table S9. Related to Figure 9.

siRNAs used in the study are:

siRNAs			
Gene	Gene Accession	Catalogue	Sequence
Non-targeting (pool)	N/A	D-001810-10	UGGUUUACAUGUCGACUAA UGGUUUACAUGUUGUGUGA UGGUUUACAUGUUUUCUGA UGGUUUACAUGUUUUCCUA
Androgen Receptor (pool)	NM_000044	L-003400-00-0005	GAGCGUGGACUUUCCGGAA UCAAGGAACUCGAUCGUAU CGAGAGAGCUGCAUCAGUU CAGAAAUGAUUGCACUAUU
FAM57A (pool)	NM_024792	L-014448-02-0020	CGAUCAACUAGGAUGAAUU ACUACAGGGAAGCGUGAAA CCAAAUUGUGCCCUGGGUA GUAGGUAGUUAUUGAUCGU
GEMIN4 (pool)	NM_015721	L-020756-01-0020	CGUUGACACUUCUGCCGAA UACCUGUGCGGGAGCGAAA GGUCAUAGUGCACCCGGAA AUGCAGAACUAGAGCGCUU
VPS53 (pool)	NM_001128159	L-017048-00-0020	GCAAUUAGAUCACGCCAAA GGACGUAGUAUCAGGCAUU UAUUCGAACUGUUGUAAGA GGAUGUAAGUCUGAUUGAA

Table S10. Related to Figure 9.

Quantitative reverse transcription polymerase chain reaction (qRT-PCR) primers used in the study are:

QPCR primers		
Gene	Forward(5'-3')	Reverse(5'-3')
FAM57	CGAACCAGAGACCAGAACCG	CTGAATCAGAACCCTGCCCA
GEMIN4	TCTTCTAAGTCGGCCTCAGC	CATGTTCCTCCTGGCTGGATT
VPS53	ACCAGGCGAAGACAATACGG	GTCTGTGCAGCCTTCACTCT
Androgen Receptor (AR)	AGTCAATGGGCAAAACATGG	TTGTGTCAAAAGCGAAATGG
GAPDH	GAGAGAAACCCGGGAGGCTA	TCACCTTCCCCATGGTGTCT

Methods

ChIP-seq

The cells were crosslinked using 1% formaldehyde and quenched by 1.25M glycine. The cells were lysed, sonicated, and used in a Chromatin Immunoprecipitation (ChIP) reaction (HighCell# ChIP kit protein G (Diagenode Inc., NJ) as per the manufacturer's recommendation using the CTCF (3418S, Cell signaling Technologies) and H3K27ac (39133, Active Motif) antibodies. The ChIP processed DNA was then reverse crosslinked and purified using the IPure kit (Diagenode Inc., NJ). The ChIP DNA was quantified and used for construction of next generation sequencing libraries (KAPA Hyper Prep, KAPA Biosystems, MA). The ChIP sequencing (ChIP-Seq) libraries, after appropriate size selection using Agencourt AMPure XP (Beckman Coulter Life Sciences, IN), were sequenced with single-end reads of 75bp on Illumina NextSeq 500.

ChIP-seq data analysis

First, we did quality control for ChIP-seq datasets by FastQC (version 0.11.5) [Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data.], and then trimmed adaptor sequences from raw reads by FASTX_Toolkit (version 0.0.13.2, command fastx_clipper, http://hannonlab.cshl.edu/fastx_toolkit/). After adaptor trimming, reads shorter than 18 nt were discarded (adaptor sequence is "GATCGGAAGAGCACACGTCTGAA"). Second, ChIP-Seq reads were aligned to human reference genome (hg19 version) by bowtie (version 1.0.0) (1), using the parameter as -v 1 -M 1 -m 2 --best -S. Considered reads have multiple alignments in the genome, we only retained reads that have no more than two alignments and reported the best aligned position by "--best". Third, macs2 was used to call peaks from aligned ChIP-Seq reads, with default parameters (2). Integrated Genomics Viewer (IGV, v2.3.79) (3) was used to visualize ChIP-seq peak data.

RNA-seq

RNA was extracted from the cell lines used for ChIA-PET using the RNeasy Mini kit (QIAGEN, MD). The total RNA was analyzed using the RNA Screen Tape Station (Agilent, CA) and quantified using the ThermoFisher Qubit (Waltham, MA). One microgram of total RNA from each sample was used to prepare libraries using the KAPA mRNA Hyper prep kit (KAPA Biosystems, CA). The quality of the libraries was assessed using the Tape Station High Sensitivity DNA Screentape (Agilent, CA) and sequenced with single read of 75 bp. Library clustering was performed on a cBot with Illumina HiSeq PE Cluster kit v3 (San Diego, CA). The libraries were sequenced as single end 75 basepair reads on an Illumina NextSeq 500 (San Diego, CA) with an Illumina HiSeq SBS kit v3 (San Diego, CA). Base calling and quality scoring were performed using Real-Time Analysis (version 1.18.64) (San Diego, CA) and FASTQ file generation and de-multiplexing using CASAVA(San Diego, CA).

RNA-Seq data analysis

Quality Control

We did quality control for RNA-Seq datasets of four cell lines by FastQC. Then, we removed reads that mapped to human ribosome RNA sequences using bowtie.

Alignment

Retained reads from the previous step were processed to map to human reference genome (version hg19) by Tophat2 (version 2.0.10) (4), with parameter "--read-mismatches 2 --segment-mismatches 0 -g 1 --report-secondary-alignments". RNA-Seq libraries were constructed with strand information (first-stranded). DNA duplications generated by PCR were removed by samtools (command rmdup) (5).

Differential expression

Express (version 1.4.1) was used to calculate read counts for each gene (6). As express can only calculate read counts on transcript level, the estimated gene counts were summed across all the isoforms of the gene. Then the read count matrix files were used as input file for DESeq2 to identify differentially expressed genes (7). Significant differentially expressed genes were selected following the criteria that the absolute value of log2foldchange is no less than 1, and adjusted p-value is less than 0.05.

Definition of regulatory elements

For our analysis, several different genomic elements were defined, including promoters, enhancers, gene bodies and other regions. Promoters are defined as regions from upstream 500nt (-500nt) to downstream 250nt (+250nt) of the transcription start site (TSS) of a gene. Gene bodies are defined by subtracting promoter regions from whole gene regions based on annotation file of hg19. Other regions are genomic regions that are neither promoters not gene bodies. Combining ChIP-seq data of H3K27ac of four cell lines, enhancers are defined as H3K27ac peak regions, discarding peaks which have overlaps with promoters.

Correlation analysis

Correlation co-efficiency was calculated between RNA expression level and chromatin interactions. RNA expression level was measured by FPKM transformed by log2 calculated based on RNA-Seq data, while chromatin interactions were measured by PET numbers calculated from ChIA-PET data of four cell lines. As there are two or three biological replicates for RNA-seq datasets in each cell line, we took the mean FPKM of replicates as the expression level of each gene. Cor.test in R was used to calculate the correlation co-efficiency and p-value.

Motif enrichment analysis

Motif sequences enriched in promoter regions and enhancer regions were identified by homer (version 4.7), with default parameters (8). Significantly enriched motifs were selected following the thresholds described below:

I. For motifs enriched in enhancer regions: p-value < 1e-20, while for motifs enriched in promoter regions: p-value < 1e-5.

II. Motifs are ranked by p-value.

III. Motifs with the same sequence type, like bZIP and Zf, are further curated to avoid redundancy (showing the one with the most significant p-value).

Data Visualization

The Integrative Genomics Viewer (IGV) is used to visualize the track files for different gene loci (3). HiC-Plotter was used to draw chromatin interaction heatmap (9).

Predicting potential gene target of enhancers, AR binding peaks, FOXA1 binding peaks and PCa risk SNP loci by using ChIA-PET data

The potential gene targets of cis-elements and trans-acting factors were predicted on the basis of their relationship with ChIA-PET interaction anchor peaks. In this study, potential gene targets of enhancers, AR binding peaks, FOXA1 binding peaks and PCa risk SNP loci are provided in supplementary tables S3, S7 and S8.

Potential gene targets of enhancers

For identifying the potential targets of enhancers, we filtered the RNA Pol II ChIA-PET interaction anchor peaks that overlapped with H3K27ac ChIP-seq peaks. Next, the H3K27ac peaks were assigned their paired peaks on the basis of PET interactions. The genes whose promoter regions (from -500 nt to +250nt from the TSS) overlapped with the paired peak regions were considered as potential targets of corresponding enhancers.

Potential gene targets of AR and FOXA1 binding peaks

Potential targets of AR and FOXA1 binding peaks in LNCaP and VCaP cells were also predicted by following the same strategy as above, only replacing the H3K27ac ChIP-Seq data with AR or FOXA1 ChIP-Seq data. Potential target genes identified by integrating RNA Pol II ChIA-PET with AR or FOXA1 binding peaks were classified as ChIA-PET plus ChIP-seq group (Group 1). Genes that are nearest to AR or FOXA1 binding peaks were classified as AR or FOXA1 ChIP-seq group (Group 2) (by utilizing tool closestBed in bedtools). To balance the number of different gene groups, genes in Group 2 were randomly selected with comparable number as Group 1. Also, comparable number of all annotated genes (GENCODE v19) were randomly selected as control group (Group 3). The AR-FOXA1 co-occupied regions that overlapped with the H3K27ac mark were designated as AR-FOXA1 co-occupied enhancers, and genes located in the paired RNA Pol II anchor peaks were assigned as their targets.

Potential gene targets of PCa risk SNP loci

We filtered the ChIA-PET interaction anchor peaks that overlapped with PCa risk SNP loci. The genes located in the paired anchor-peaks were extracted as targets of the risk SNP loci.

Patients and tissue samples

Patients were identified from a population of men treated at the Royal Marsden NHS Foundation Trust who developed castration resistant prostate cancer. Patients with adequate DNA from formalin-fixed, paraffin-embedded biopsies to perform array CGH were selected. Biopsies were obtained from bone marrow (9), lymph node (11), liver (2), soft tissue (1) and prostate (4). All patients had given written informed consent and were enrolled in institutional protocols approved by the Royal Marsden NHS Foundation Trust Hospital (London, UK) ethics review committee (reference no. 04/Q0801/60). A total of 27 patients had sufficient DNA for testing. Demographic and clinical data for each patient were retrospectively collected from the hospital electronic patient record system.

Array Comparative genomic hybridization (aCGH)

Briefly 10 ng of tumor DNA and a pool reference of 20 male normal DNA were amplified using GenomePlex® Complete Whole Genome Amplification (WGA) Kit (Sigma-Aldrich, MO, USA), according to manufacturer procedures**.** 500ng of amplified tumor DNA was fluorescently labeled with Cy5, and the reference male DNA labeled with Cy3 using SureTag Complete DNA Labeling Kit (Agilent Technologies CA, USA). Labeled DNA was purified and hybridized utilizing the Agilent SurePrint G3 Human array CGH Microarray Kit, 2x400K. Slides were scanned using the SureScan microarray scanner and ratios of tumor DNA to reference male reference DNA were determined using the Feature extraction software and followed by CytoGenomics Software from Agilent Technologies. Smoothed $\log_2$ ratio values < -0.25 were categorized as losses; those > 0.25 as gains; and those in between as unchanged. Amplifications were defined as smoothed $\log_2$ ratio values ≥ 1.2 and homozygous deletions as the segment $\log_2$ ratio values ≤ -1.2.

Digital PCR

Digital PCR was performed on a QX100 droplet digital PCR system (Bio-Rad). Copy Number Assays were performed with a FAM-AR Taqman Copy Number Assay, assay ID : Hs 00037560_cn and a reference copy number Vic-assay for RNAse P (cat 4403326) (Thermofisher). Copy Number Assay for AR enhancer peaks were FAM-Taqman custom designed probes (see table below). PCR reactions were prepared with 5ng DNA and Bio-Rad 2x ddPCR supermix (no dUTP) for probes (Cat number 1863010) in a total volume of 20 µl, and partitioned into ~14,000 droplets per sample in a QX100 droplet generator according to manufacturer's instructions. Emulsified PCR reactions were run on a 96 well plate on an Eppendorf Nexus GSX1 thermal cycler incubating the plates at 95°C for 10 min followed by 40 cycles of 95°C for 15 sec and 60°C for 60 sec, followed by 10 min incubation at 98°C. The temperature ramp increment was 2.5°C/sec for all steps. Plates were read on a Bio-Rad QX100 droplet reader and digital PCR analysis was performed using QuantaSoft v1.3.2.0 software from Bio-Rad to assess the number of positive droplets. At least two negative control wells with no DNA were included in every run.

Whole genome sequencing (WGS) and Copy number alteration (CNA) analysis

For the WGS experiments, tumor DNA was extracted from seven 10-µm-thick frozen sections after confirmation of tumor cell presence on a 4-µm-thick hematoxylin and eosin (H&E)–stained slide. Germline DNA was extracted from saliva. The QIAmp DNA Mini Kit (#51304; Qiagen) was used for DNA isolation from both

fresh/frozen sections and buccal swabs. Library preparation was performed using the TruSeq Nano DNA Library Prep Kit. Sequencing was performed using the Illumina HiSeq 2000 System. Tumor and matched germline DNA samples were profiled with a standard WGS protocol. Paired-end sequencing reads were aligned to the human reference genome (GRCh37/hg19) using Illumina Isaac Genome Alignment Software embedded in Illumina pipeline version v2.0.1(10). The absolute gene copy number in tumor was derived from NGS data based on comparing genome coverage between matched tumor-normal samples. CNV analysis used CNV-seq (11) which applied a robust statistical model that allows confidence assessment of observed copy number ratios and adjust confounders to achieve absolute gene copy number in tumor.

SNP data analysis

The 122 germline SNPs were obtained from previous study of prostate cancer (12). The hg19 genome was separated into bins/blocks with the size of the mean length of all peaks in each cell line. Based on overlapping the RNA Pol II peaks, all blocks were subsequently classified into peak-blocks and non-peak blocks. For each cell line, peak-blocks were defined as the number of peaks, and non-peak blocks were calculated by subtracting the number of peaks from the total number of blocks. A bootstrapping probability was calculated by random sampling of genome blocks. Each run, the blocks were randomly sampled as same number as peak-blocks and count the number of blocks that have SNPs. Then the empirical p-value was calculated for enrichment analysis based on in total of 10000 runs.

CPCG cohort

We interrogated a previously published cohort of 128 intermediate risk prostate cancer samples that underwent RNA-Seq, described previously (13). Genotype calls were extracted from WGS of matched blood samples using GATK (v.3.4.0+), described previously (14).

TCGA cohort

The TCGA PRAD was used a validation cohort (15). Only samples with concordance between SNP6 array genotypes and whole exome sequencing (WXS) blood samples calls > 80% were retained (348 samples). Genotypes were imputed using the Sanger Imputation Service – pre-phasing using Shapeit2 (16), imputation using PBWT (17) and the Haplotype Reference Consortium (release 1.1) panel (18).

Germline-RNA associations (eQTL) discovery and validation

For each SNP-target pair, we tested for eQTLs first in the CPCG discovery cohort and then validated significant associations in the TCGA cohort. We fit a linear regression model treating genotypes additively, *i.e.* number of alternative alleles present (AA = 0, AB = 1, BB = 2), with principal components 1 and 2 to control from population stratification. False discovery adjustment was applied over all SNP-target pairs and significant associations were defined as FDR < 0.05. eQTLs were considered validated if they had FDR < 0.05 in the TCGA validation cohort and the directionality of the coefficient was the same in both cohorts (*i.e.* $\text{sign}(\beta_{\text{discover}}) =$

sign($\beta_{validation}$)). To assess the impact of ERG status on rs684232:FAM57A, rs684232:VPS53 and rs684232:GEMIN4 eQTLs, CPCG samples were subsetted into ERG positive and negative based on the presence of the TMPRSS2-ERG fusion, as described previously (14). We quantified each eQTL within positive and negative individuals separately using the same linear regression model described above. mRNA abundance differences between individuals positive and negative for *TMPRSS2-ERG* fusion with the same genotype were quantified using Mann-Whitney test.

Differential H3K27ac binding

Peak bed files and raw fastqs for H3K27ac (n = 92), H3K27me3 (n=76), AR (n=88) and H3K4me3 (n=56) were downloaded from the Gene Expression Omnibus (GSE120738) (19). Raw fastqs were aligned using bwa (v.0.7.15) and the aligned bam files from each target were merged for each patient (*i.e.* H3K27ac, H3K27me3, H3K4me3 and AR bams from the same patient were merged). This merging step increased the coverage as well as reduced allele bias due to allele specific binding. Using the merged bams, patients were genotyped at overlapping sites of interest using GATK (v3.4.0+) HaplotypeCaller. Using the peak bed files, consensus peak bed was generated by using DiffBind R package. Bedtools coverageBed was used to overlap the read counts with the consensus peaks. The overlapped read counts were normalized for sequencing depth using TMM method by edgeR. Differential H3K37ac binding at rs684232 was quantified using Mann-Whitney test considering a recessive model.

Immunoblot analysis

Cells were harvested and lysed using RIPA (150mM NaCl, 5mM EDTA, 50mM Tris-Cl, 1% NP-40, 0.5% Sodium deoxycholate, 0.1% SDS) buffer supplemented with proteinase inhibitor (cOmplete EDTA free protease inhibitor tablet, cat no 11873580001, Roche). After determining the protein concentration of lysates, samples were separated on 4 - 12% gradient gels (Invitrogen, NW04122). To detect full length and variants of AR protein, the immunoblots were performed using anti-AR (Sigma, A9853; this antibody binds to the N-terminus of AR), anti-AR (Sigma, SAB5500007; this antibody binds to the C-terminus of AR), anti-Actin (Sigma, AC-74), anti-beta tubulin (SantaCruz, sc9104), and anti-cMyc (Cell Signaling, 5605S) antibodies. To compare endogenous EZH2 and FOXA1 proteins, anti-EZH2 (active motif, 39901), anti-FOXA1 (Cell Signaling, 58613) were used.

Luciferase reporter assays

The ~1100 base pair promoter/enhancer region containing the wild-type or risk allele rs684232 was cloned into pGL2 luciferase reporter plasmid (Promega) and verified by Sanger sequencing. LNCaP or VCaP cells were seeded on 24-well plates and transfected with the indicated plasmids. The Fire fly luciferase / Renilla luciferase activities were measured using the Dual-Glo luciferase assay system as described in the manufacturer's protocol (Promega, E2940).

siRNA knockdown and Gene expression studies

Small interfering RNA (siRNA) were purchased from Horizondiscovery. siRNAs transfection was carried out by using Lipofectamine RNAiMAX transfection Reagent (Thermo Fisher Scientific, #13778-150) and total RNA was extracted using RNeasy Plus Mini Kit (QIAGEN, #74134) at 72 hr post transfection. Reverse transcription and quantitative PCR were performed as previously described (20).

CellTiter-Glo assay

LNCaP cells were seeded into multiple 96-well plates one day before the transfection with siRNA. To determine the number of viable cells in the culture, CellTiter-Glo reagent (Promega, G7570) was added to each well (7 days post siRNA transfection) and mixed by orbital shaker according to the manufacturer's instructions. Plates were incubated for 10 min at room temperature, then luminescence was measured using a microplate reader (TECAN).

Supplemental References

1. Langmead B, Trapnell C, Pop M, and Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10(3):R25.
2. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008;9(9):R137.
3. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol.* 2011;29(1):24-6.
4. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 2013;14(4):R36.
5. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25(16):2078-9.
6. Roberts A, and Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature methods.* 2013;10(1):71-3.
7. Love MI, Huber W, and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.
8. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell.* 2010;38(4):576-89.
9. Akdemir KC, and Chin L. HiCPlotter integrates genomic data with interaction matrices. *Genome Biol.* 2015;16:198.
10. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics.* 2013;29(16):2041-3.
11. Xie C, and Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics.* 2009;10:80.

12. Guo H, Ahmed M, Zhang F, Yao CQ, Li S, Liang Y, et al. Modulation of long noncoding RNAs by risk SNPs underlying genetic predispositions to prostate cancer. *Nat Genet.* 2016;48(10):1142-50.
13. Chen S, Huang V, Xu X, Livingstone J, Soares F, Jeon J, et al. Widespread and Functional RNA Circularization in Localized Prostate Cancer. *Cell.* 2019;176(4):831-43 e22.
14. Fraser M, Sabelnykova VY, Yamaguchi TN, Heisler LE, Livingstone J, Huang V, et al. Genomic hallmarks of localized, non-indolent prostate cancer. *Nature.* 2017;541(7637):359-64.
15. Cancer Genome Atlas Research N. The Molecular Taxonomy of Primary Prostate Cancer. *Cell.* 2015;163(4):1011-25.
16. Delaneau O, Marchini J, and Zagury JF. A linear complexity phasing method for thousands of genomes. *Nature methods.* 2011;9(2):179-81.
17. Durbin R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics.* 2014;30(9):1266-72.
18. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet.* 2016;48(10):1279-83.
19. Stelloo S, Nevedomskaya E, Kim Y, Schuurman K, Valle-Encinas E, Lobo J, et al. Integrative epigenetic taxonomy of primary prostate cancer. *Nat Commun.* 2018;9(1):4900.
20. Li X, Baek G, Ramanand SG, Sharp A, Gao Y, Yuan W, et al. BRD4 Promotes DNA Repair and Mediates the Formation of TMPRSS2-ERG Gene Rearrangements in Prostate Cancer. *Cell reports.* 2018;22(3):796-808.