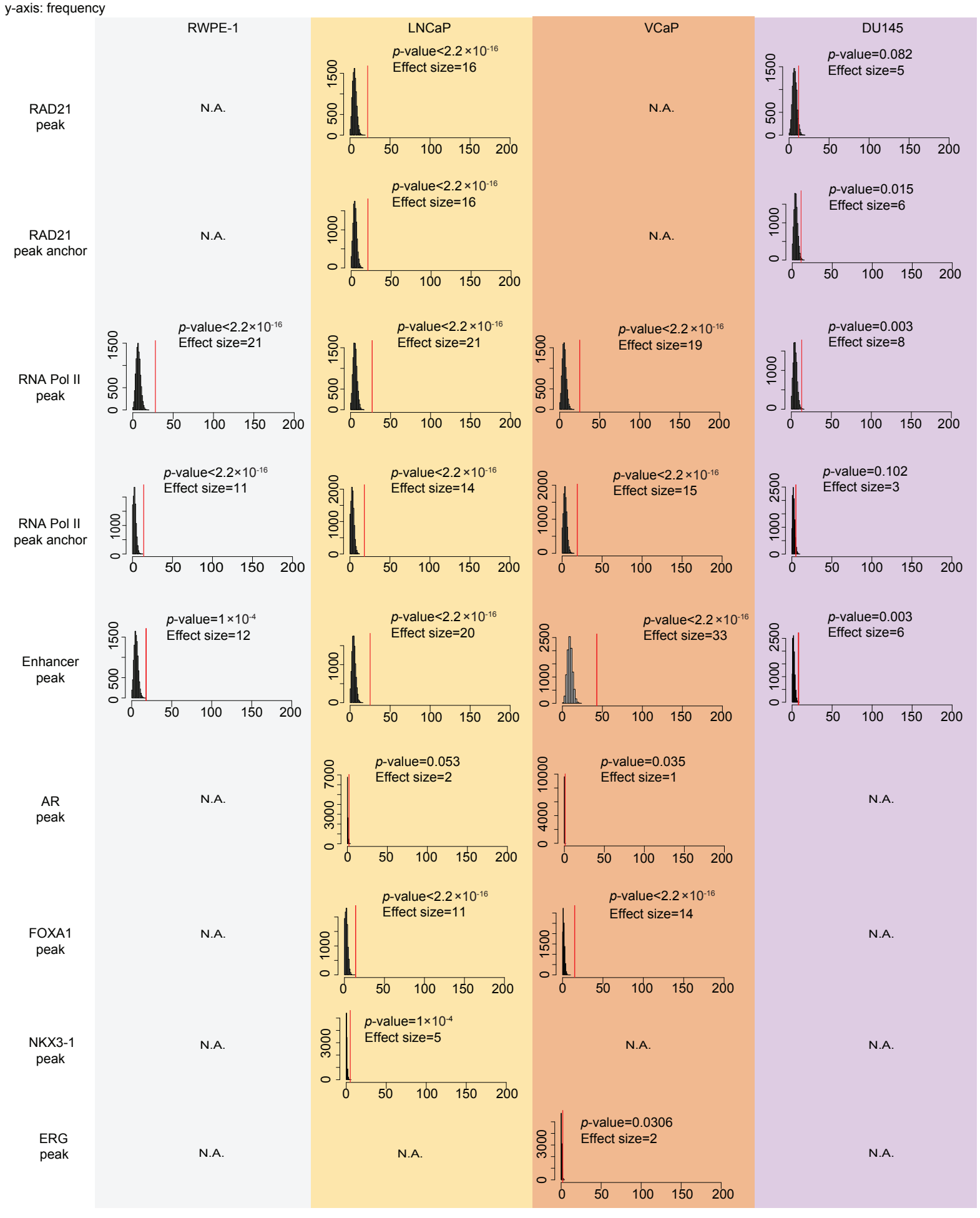
Figure S1, Page 1

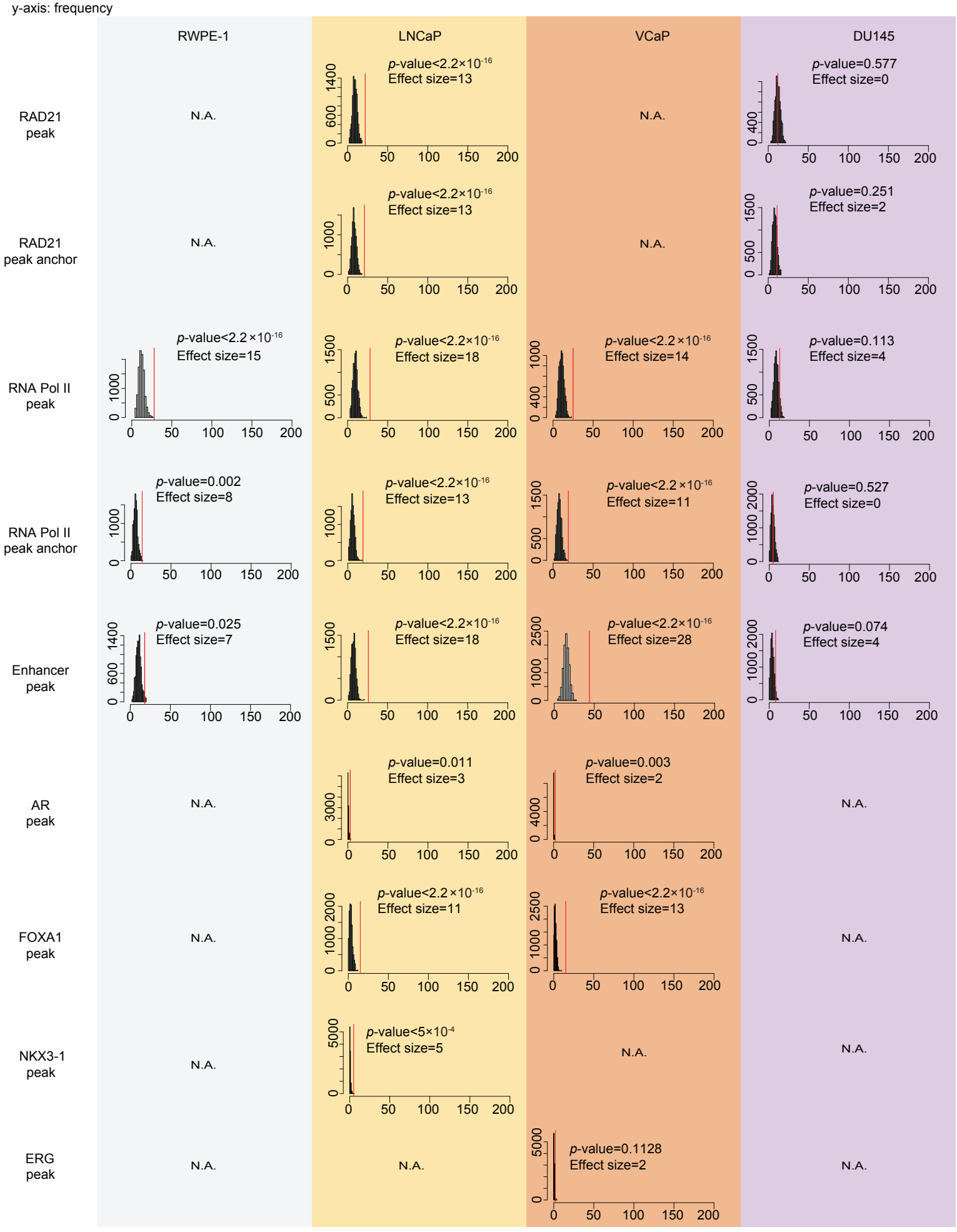
A

x-axis: overlapped PCa risk SNP number
y-axis: frequency



B

x-axis: overlapped PCa risk SNP number
y-axis: frequency



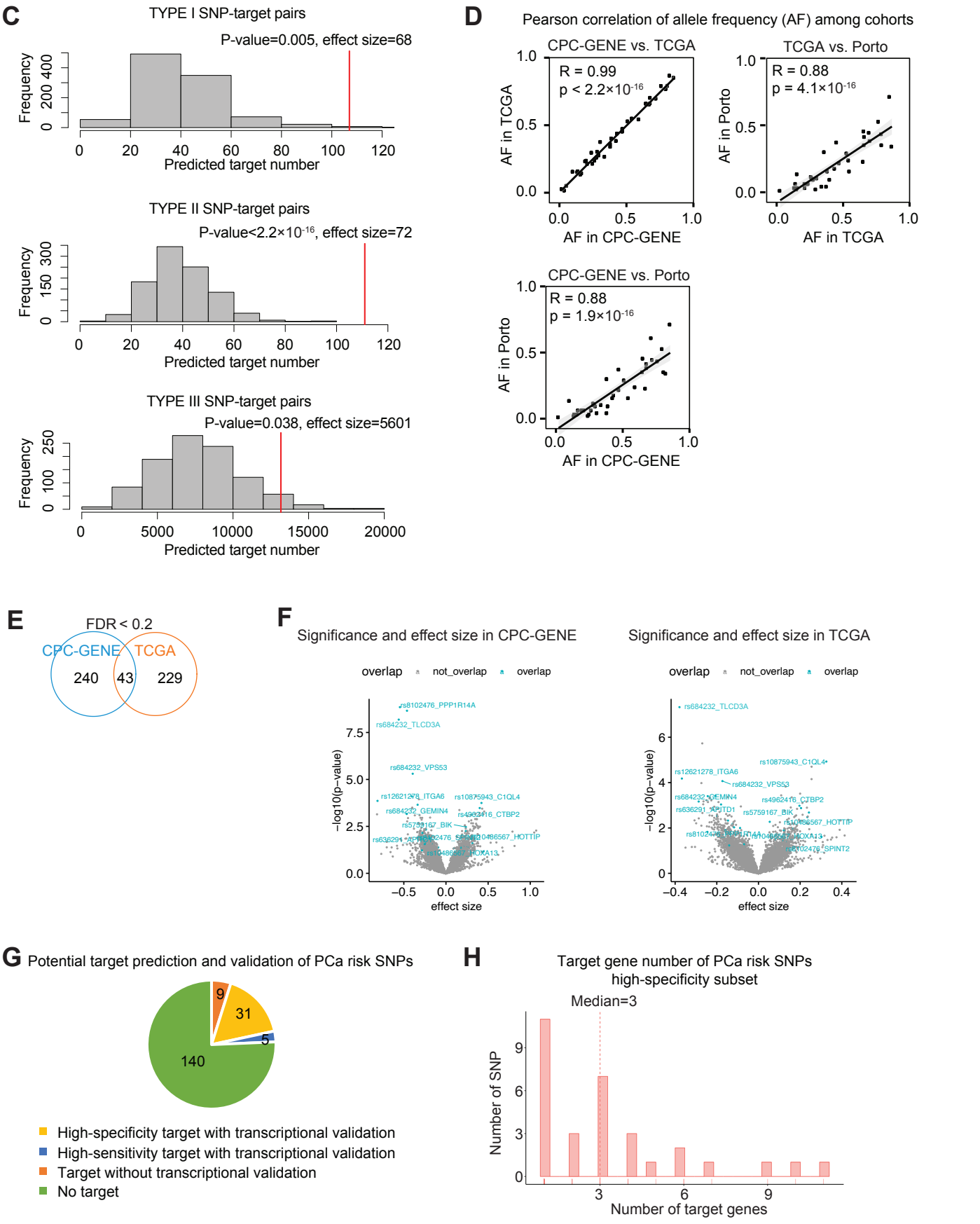

C
TYPE I SNP-target pairs
P-value=0.005, effect size=68
Frequency
Predicted target number
TYPE II SNP-target pairs
P-value<2.2×10-16, effect size=72
Frequency
Predicted target number
TYPE III SNP-target pairs
P-value=0.038, effect size=5601
Frequency
Predicted target number
D
Pearson correlation of allele frequency (AF) among cohorts
CPC-GENE vs. TCGA
R = 0.99
p < 2.2×10-16
AF in TCGA
AF in CPC-GENE
TCGA vs. Porto
R = 0.88
p = 4.1×10-16
AF in Porto
AF in TCGA
CPC-GENE vs. Porto
R = 0.88
p = 1.9×10-16
AF in Porto
AF in CPC-GENE
E
FDR < 0.2
CPC-GENE
TCGA
240
43
229
F
Significance and effect size in CPC-GENE
Significance and effect size in TCGA
overlap
not_overlap
overlap
−log10(p−value)
effect size
rs8102476_PPP1R14A
rs684232_TLCD3A
rs684232_VPS53
rs12621278_ITGA6
rs10875943_C1QL4
rs684232_GEMIN4
rs4962416_CTBP2
rs5759167_BIK
rs636291_APITD1
rs10486567_HOTTIP
rs10486567_HOXA13
rs684232_TLCD3A
rs10875943_C1QL4
rs12621278_ITGA6
rs684232_VPS53
rs684232_GEMIN4
rs4962416_CTBP2
rs636291_APITD1
rs5759167_BIK
rs10486567_HOTTIP
rs8102476_PPP1R14A
rs10486567_HOXA13
rs8102476_SPINT2
G
Potential target prediction and validation of PCa risk SNPs
9
31
5
140
High-specificity target with transcriptional validation
High-sensitivity target with transcriptional validation
Target without transcriptional validation
No target
H
Target gene number of PCa risk SNPs
high-specificity subset
Median=3
Number of SNP
Number of target genes

Figure S1. Related to Figure 1. Features of PCa risk SNPs.

(A) Enrichment analysis of PCa risk SNPs in peak regions versus random regions. Enrichment of PCa risk SNPs in RAD21 and RNA Pol II occupied regions (total peaks, anchor peaks), enhancers (H3K27ac marked peaks), AR, FOXA1, ERG and NKX3-1 occupied regions was estimated. The red line indicates the observed overlap between PCa risk SNPs and the peaks. The histograms illustrate the expected overlap between PCa risk SNPs and the randomly shuffled peaks from 10,000 simulations. Significant differences between the expected and observed values are indicated, and p-values were generated by empirical test. The X-axis and Y-axis represent the number of PCa risk SNPs, and the number of simulations, respectively. (B) Enrichment analysis of PCa risk SNPs versus randomly selected SNPs in peak regions. Enrichment of random SNPs in RAD21 and RNA Pol II occupied regions (total peaks, anchor peaks), enhancers (H3K27ac marked peaks), AR, FOXA1, ERG and NKX3-1 occupied regions was estimated. The red line indicates the observed overlap between PCa risk SNPs and the peaks. The histograms illustrate the expected overlap between random sampled SNPs and the peaks (from sequencing data) from 10,000 simulations. Significant differences between the expected and observed values are indicated, and p-values were generated by empirical test. The X-axis and Y-axis represent the number of PCa risk SNPs, and the number of simulations, respectively. (C) Simulation for target prediction by randomly shuffled interacted peak regions. The red line indicates the observed number of targets for PCa SNPs by utilizing ChIA-PET data. The grey background represents the number of targets for PCa SNPs by overlapping with simulated peak regions for 1,000 times. The X-axis and Y-axis represent the number of targets, and the number of simulations, respectively. P-values were generated by empirical test. (D) Allele frequency of PCa SNPs from genotyping the three cohorts. Pearson correlation efficient is shown for each comparison. (E) The scaled venn diagram shows the overlap of PCa SNP-target pairs with FDR < 0.2 in CPC-GENE and TCGA. (F) Volcano plots show the distribution of p-value and effect size of eQTL analysis in CPC-GENE (left) and TCGA (right) cohort. The overlapped SNP-targets with FDR < 0.2 in both cohorts are shown in turquoise. All individual examples shown in figures are labeled. (G) Pie chart shows the number of PCa risk SNPs with high-specificity targets (yellow), the number of PCa SNPs with high-sensitivity targets (blue), the number of PCa risk SNPs with potential targets but without validation in clinical cohorts (orange) and the number of PCa risk SNPs without potential targets (green). (H) Histogram plot shows the number of transcriptionally validated high-specificity target genes for individual PCa risk SNPs. X-axis shows the number of target genes and median value is indicated as dashed line.

Figure S2, Page 1

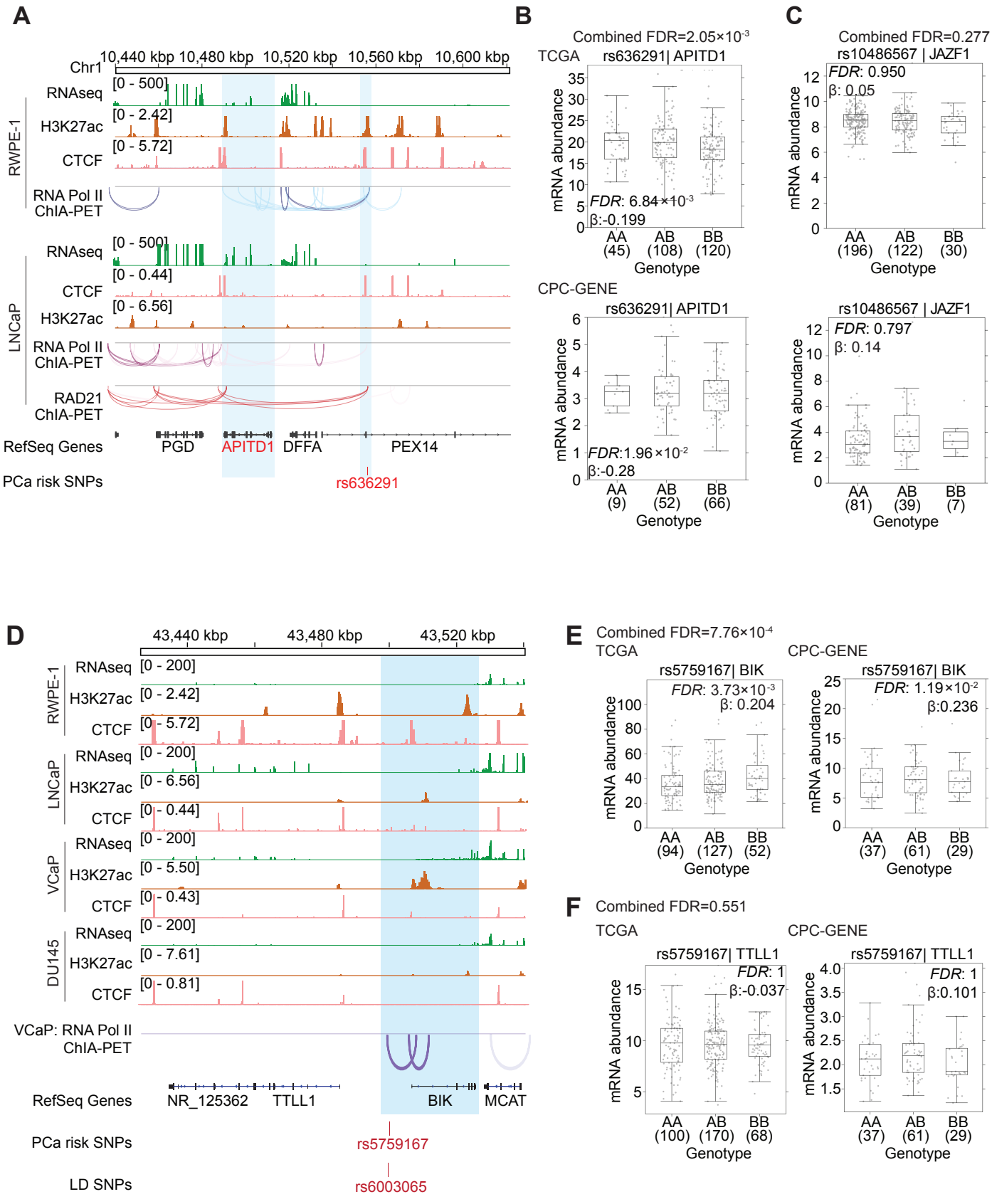


Figure S2, Page 2

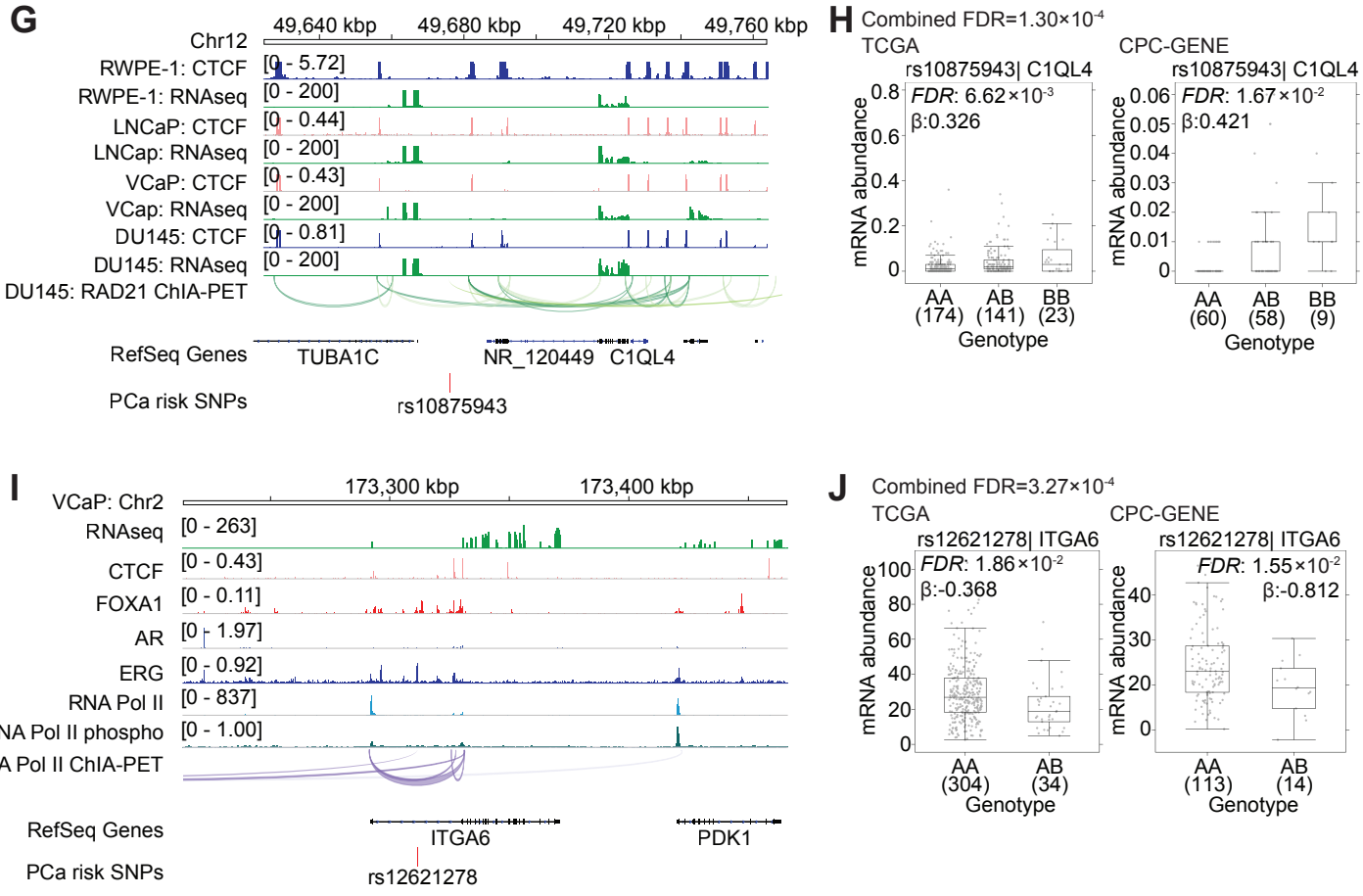


Figure S2. Related to Figure 2. Transcriptional regulation by the PCa risk SNPs rs636291, rs5759167, rs10875943 and rs12621278 in patient cohorts.

(A) Integrated genome view of the PCa SNP rs636291 and its adjacent regions. ChIA-PET, ChIP-Seq and RNA-Seq tracks for RWPE-1 and LNCaP cells are shown. (B) The mRNA abundance of *APITD1* in TCGA and CPC-GENE cohorts stratified by genotype. Boxplot represents median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range. mRNA abundance was measured in FPKM. Numbers next to genotypes reflect number of samples in each group. FDR was calculated by "BH" method, and β value were estimated from linear regression model. (C) The mRNA abundance of *JAZF1* in TCGA and CPC-GENE cohorts stratified by genotype. FDR was calculated by "BH" method, and β value were estimated from linear regression model. (D) Integrated genome view of the PCa SNP rs5759167 and its adjacent regions. ChIA-PET, ChIP-Seq and RNA-Seq tracks are shown. The LD SNP is listed below the tag SNP track. (E and F) The mRNA abundance of *BIK* and *TTLL1* in TCGA and CPC-GENE cohorts stratified by genotype. FDR was calculated by "BH" method, and β value were estimated from linear regression model. (G) Integrated genome view of the PCa SNP rs10875943 and its adjacent regions. ChIA-PET, ChIP-Seq and RNA-Seq tracks for RWPE-1, LNCaP, VCaP and DU145 cells are shown. (H) The mRNA abundance of *C1QL4* in TCGA and CPC-GENE cohorts stratified by genotype. Boxplot represents median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range. mRNA abundance was measured in FPKM. Numbers next to genotypes reflect number of samples in each group. FDR was calculated by "BH" method, and β value were estimated from linear regression model. (I) Integrated genome view of the PCa SNP rs12621278 and its adjacent regions. ChIA-PET, ChIP-Seq and RNA-Seq tracks for VCaP cells are shown. (J) The mRNA abundance of *ITGA6* in TCGA and CPC-GENE cohorts stratified by genotype. FDR was calculated by "BH" method, and β value were estimated from linear regression model.

Figure S3

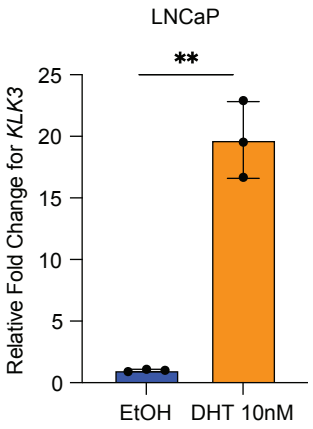
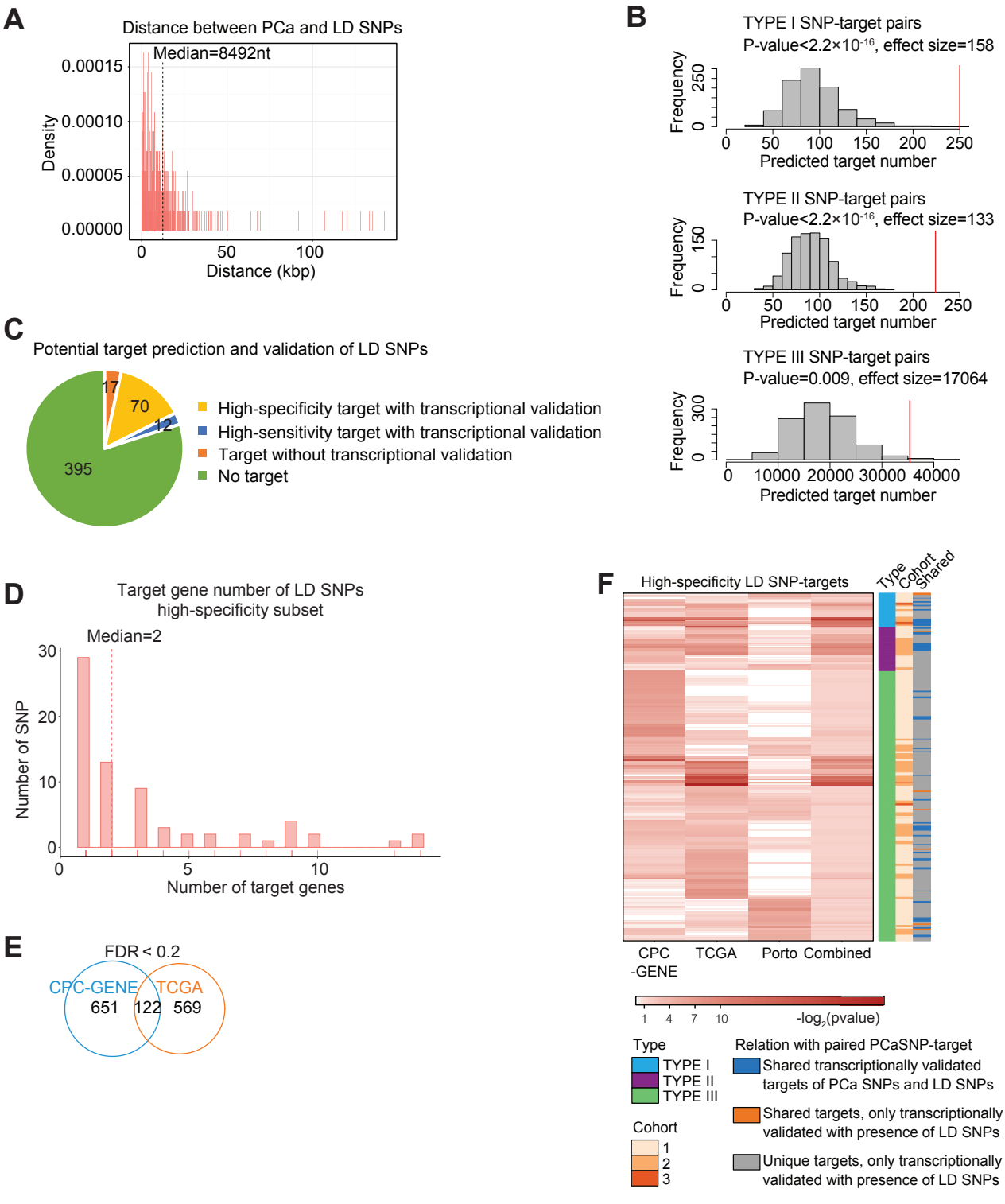

LNCaP
**
Relative Fold Change for KLK3
25
20
15
10
5
0
EtOH
DHT 10nM

Figure S3. Related to Figure 3. Positive control for androgen stimulation.

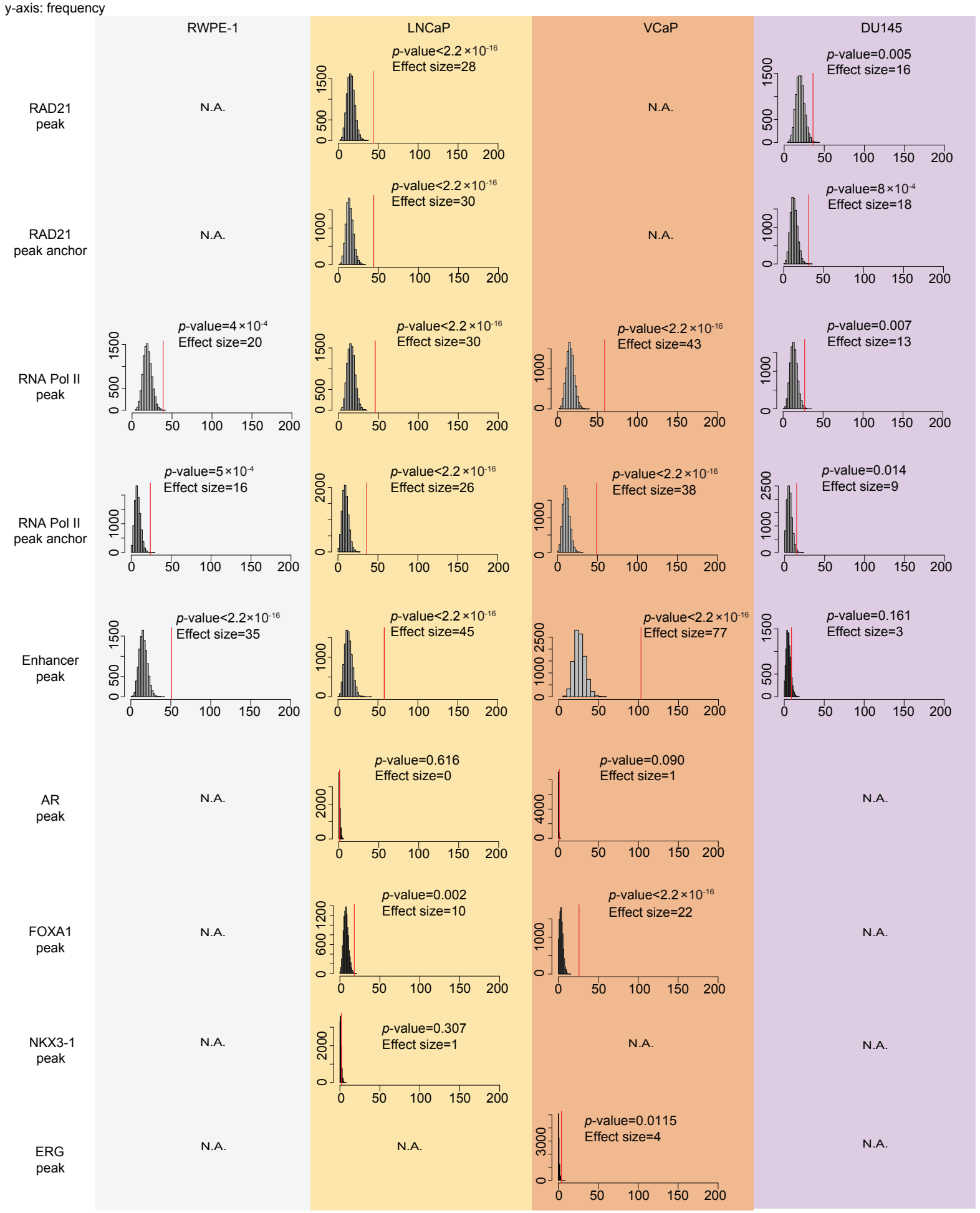
qPCR validation of the expression of *PSA (KLK3)* gene upon treatment of LNCaP cells with 10 nM DHT for 24 hrs. $^{**}P < 0.01$, by 2-tailed Student's *t* test. Error bars indicate the SD of 3 technical replicates.

Figure S4, Page 1



G

x-axis: overlapped PCa risk SNP number
y-axis: frequency



H

x-axis: overlapped PCa risk SNP number
y-axis: frequency

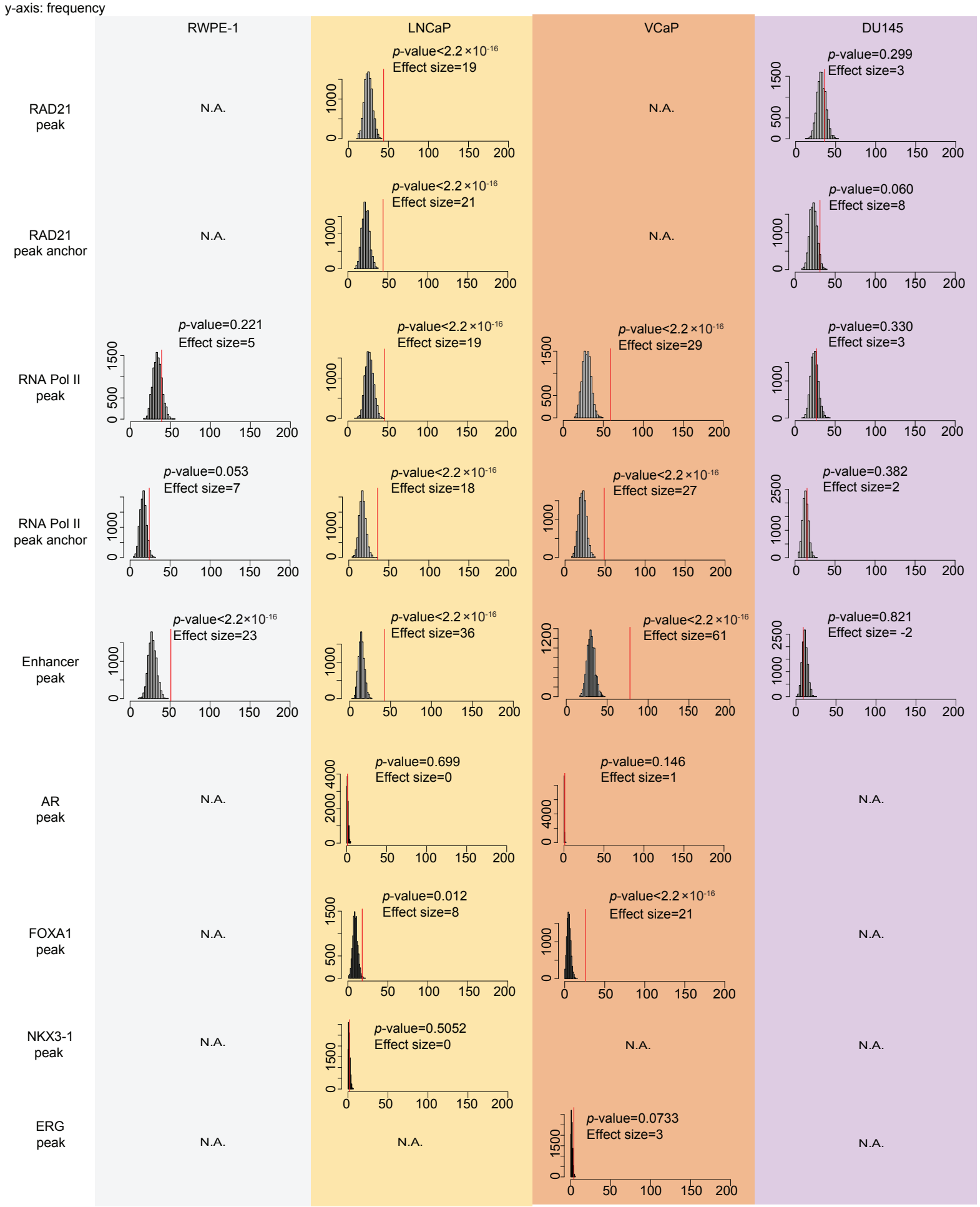


Figure S4, Page 4

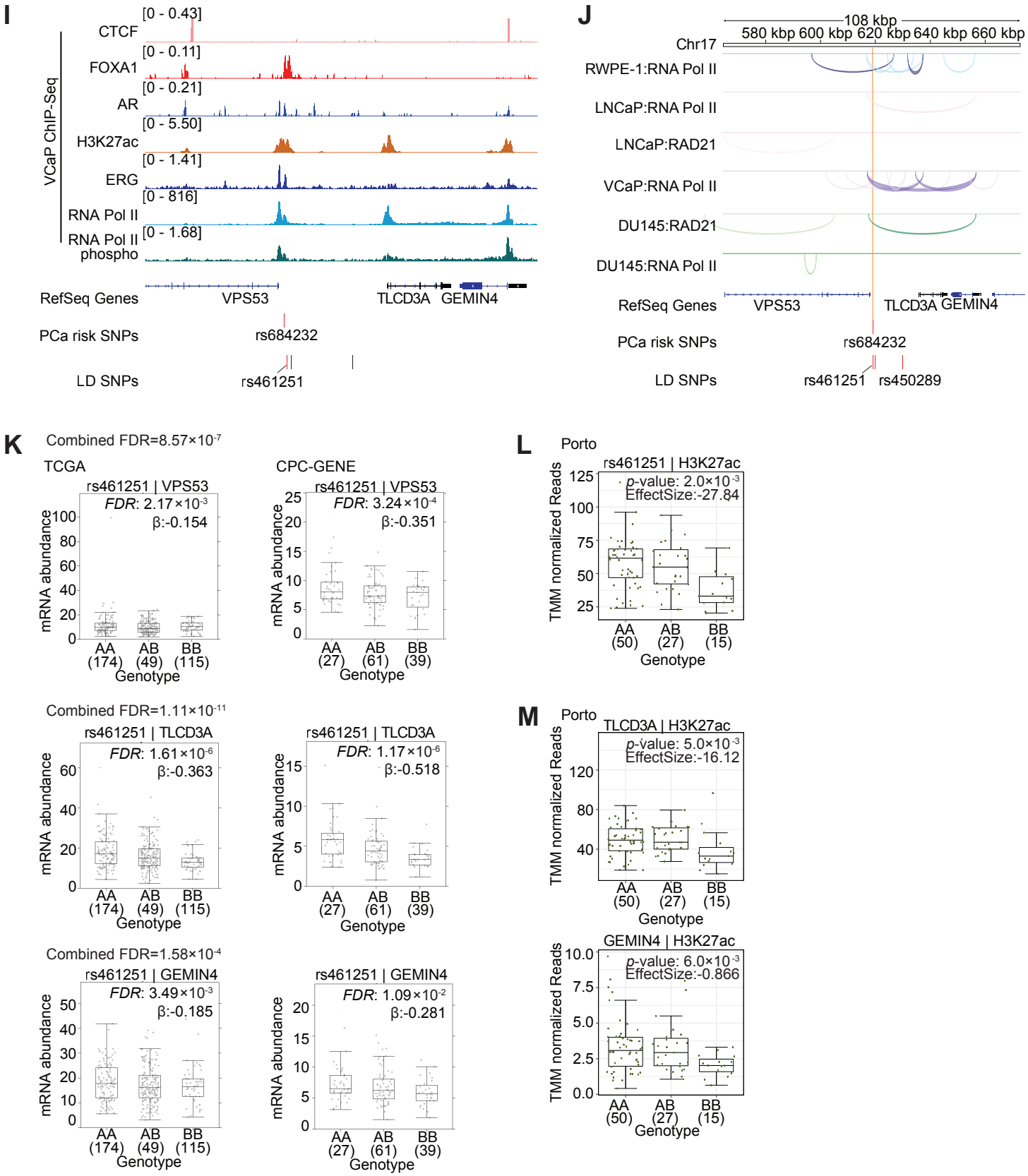
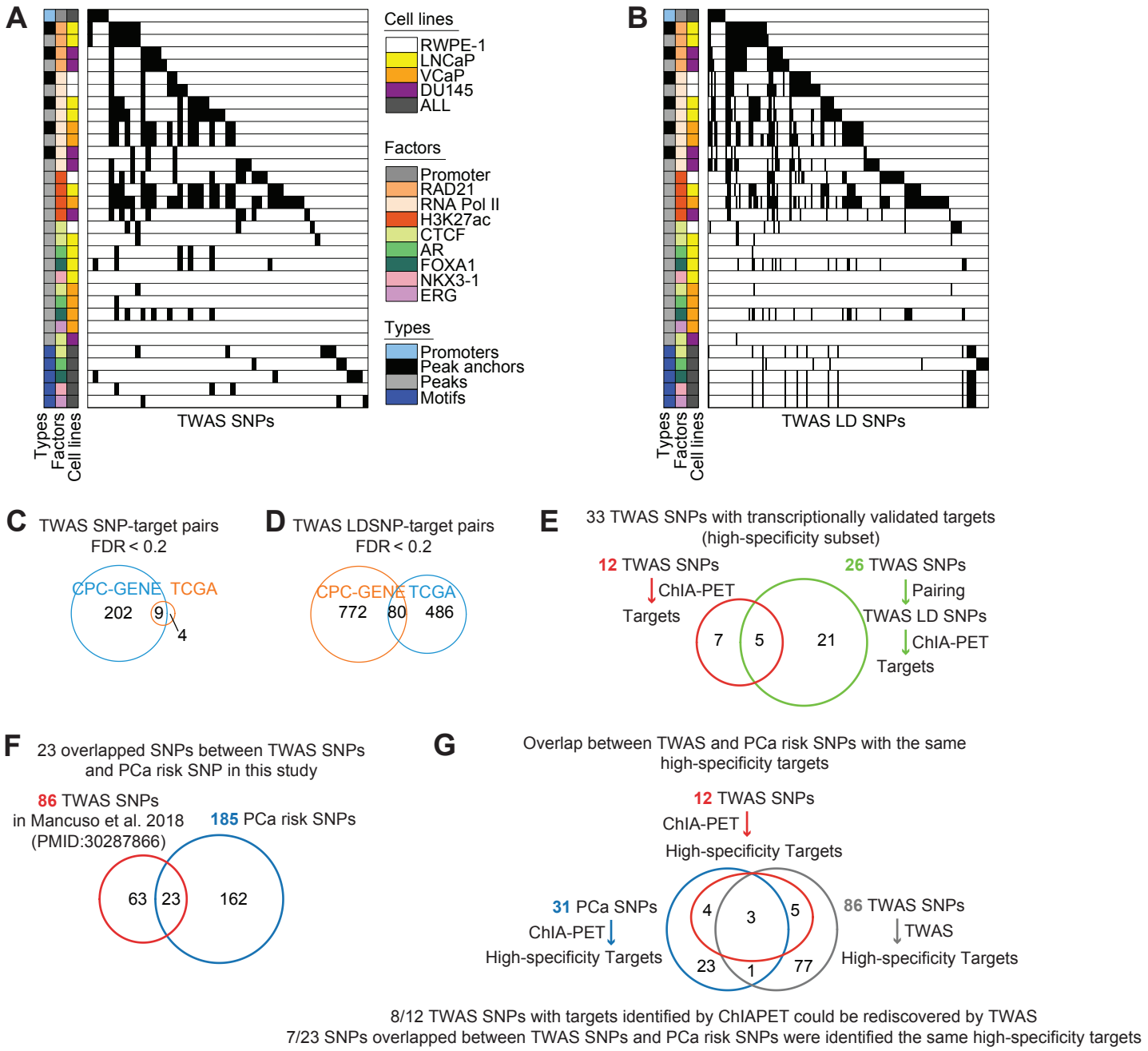


Figure S4. Related to Figure 5. Features of LD SNPs and validation of LD SNP-target pairs.

(A) Density plot shows the distance between PCa SNPs and LD SNPs. X-axis is the distance measured in kilo-base pair nucleotide unit. (B) Simulation for target prediction by randomly shuffled interacted peak regions. The red line indicates the observed number of targets for LD SNPs by utilizing ChIA-PET data. The grey background represents the number of targets for LD SNPs by overlapping with simulated peak regions for 1,000 times. The X-axis and Y-axis represent the number of targets, and the number of simulations, respectively. P-values were generated by empirical test. (C) Pie chart shows the number of LD SNPs with high-specificity targets (yellow), the number of LD SNPs with high-sensitivity targets (blue), the number of LD SNPs with potential targets but without validation in clinical cohorts (orange) and the number of LD SNPs without potential targets (green). (D) Histogram plot shows the number of transcriptionally validated high-specificity target genes for individual LD SNPs. X-axis shows the number of target genes and median value is indicated as dashed line. (E) The scaled venn diagram shows the overlap of LD SNP-target pairs with FDR < 0.2 in CPC-GENE and TCGA. (F) The heatmap (left) represents all high-specificity LD SNP-target pairs. The color intensity in the first three columns denotes the significance in the cohort, which is measured by p-value for β value from linear regression model. The color intensity in the fourth column indicates the FDR adjusted combined p-value. The annotation bars in the right for each row describe the interaction context, the number of validated cohorts and whether gene targets are shared with tag SNPs. (G) Enrichment analysis of LD SNPs in peak regions versus random regions. Enrichment of LD SNPs in RAD21 and RNA Pol II occupied regions (total peaks, anchor peaks), enhancers (H3K27ac marked peaks), AR, FOXA1, ERG and NKX3-1 occupied regions was estimated. The red line indicates the observed overlap between LD SNPs and the peaks. The histograms illustrate the expected overlap between LD SNPs and the randomly shuffled peaks from 10,000 simulations. Significant differences between the expected and observed values are indicated, and p-values were generated by empirical test. The X-axis and Y-axis represent the number of LD SNPs, and the number of simulations, respectively. (H) Enrichment analysis of LD SNPs versus randomly selected SNPs in peak regions. Enrichment of random SNPs in RAD21 and RNA Pol II occupied regions (total peaks, anchor peaks), enhancers (H3K27ac marked peaks), AR, FOXA1, ERG and NKX3-1 occupied regions was estimated. The red line indicates the observed overlap between LD SNPs and the peaks. The histograms illustrate the expected overlap between random sampled SNPs and the peaks (from sequencing data) from 10,000 simulations. Significant differences between the expected and observed values are indicated, and p-values are generated by empirical test. The X-axis and Y-axis represent the number of LD SNPs, and the number of simulations, respectively. (I-M) Transcriptional regulation by the SNP rs461251, which is in LD with the lead risk SNP rs684232. (I) Chromatin features of the LD SNP rs461251 locus in VCaP cells are described using ChIP-seq analysis. The SNP rs461251 is in LD with the tag SNP rs684232 (r^2>0.8). (J) Chromatin interactions indicate that *VPS53*, *TLCD3A* and *GEMIN4* are potential targets of rs461251. (K) rs461251 is significantly associated with mRNA abundance of *VPS53, TLCD3A* and *GEMIN4* in CPC-GENE and TCGA cohorts. Boxplot represents median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range. mRNA abundance was measured in FPKM. Numbers next to genotypes reflect number of samples in each group. FDR was calculated by “BH” method, and β value were estimated from linear regression model. (L) Boxplot shows H3K27ac signal intensity in the rs461251 locus stratified by genotype (Mann-Whitney test of recessive model). Y-axis is the number of H3K27ac ChIP-Seq read counts mapped to LD SNP rs461251 region, which is normalized by TMM method. (M) Boxplots show H3K27ac signal intensity in the promoter regions of *TLCD3A* and *GEMIN4* stratified by genotype (Mann-Whitney test of recessive model).

Figure S5



8/12 TWAS SNPs with targets identified by ChIAPET could be rediscovered by TWAS
7/23 SNPs overlapped between TWAS SNPs and PCa risk SNPs were identified the same high-specificity targets

Figure S5. Related to Figure 6. Features of TWAS SNPs, TWAS LD SNPs and validation of TWAS SNP-target pairs.

(A) Overlap between TWAS SNPs and genomic features. (B) Overlap between TWAS LD SNPs and genomic features. In the central heatmap, each column represents a SNP, and each row describes a genomic feature. Black color indicates overlap between SNP and genomic feature, while white color indicates no overlap. The individual genomic features are described by the three annotations (types, factors and cell lines). For types, sequences denote the manually defined genomic elements (promoter and motif sequence). Peaks represent antibody bound regions, and anchor peaks represent the sub-set of peaks that interact with other peaks in ChIA-PET analysis. The RWPE-1, LNCaP, VCaP and DU145 cell lines are presented in different colors. (C-D) The scaled venn diagrams show the overlap of TWAS SNP-targets (C) and LD SNP-targets (D) with FDR < 0.2 in CPC-GENE and TCGA. (E) 12 TWAS SNPs are assigned targets by ChIA-PET directly (red), 26 TWAS SNPs are assigned targets by pairing with LD SNPs then followed by ChIA-PET (green). (F) 23 SNPs overlap between the 86 TWAS SNPs used in Mancuso et al. 2018 (red) and 185 SNPs used in this study (blue). (G) Overlap between PCa risk SNPs with targets identified by ChIA-PET in this study (blue), TWAS SNPs with targets identified by ChIA-PET in this study (red), and TWAS SNPs with targets identified in Mancuso et al. 2018 (grey).

Figure S6

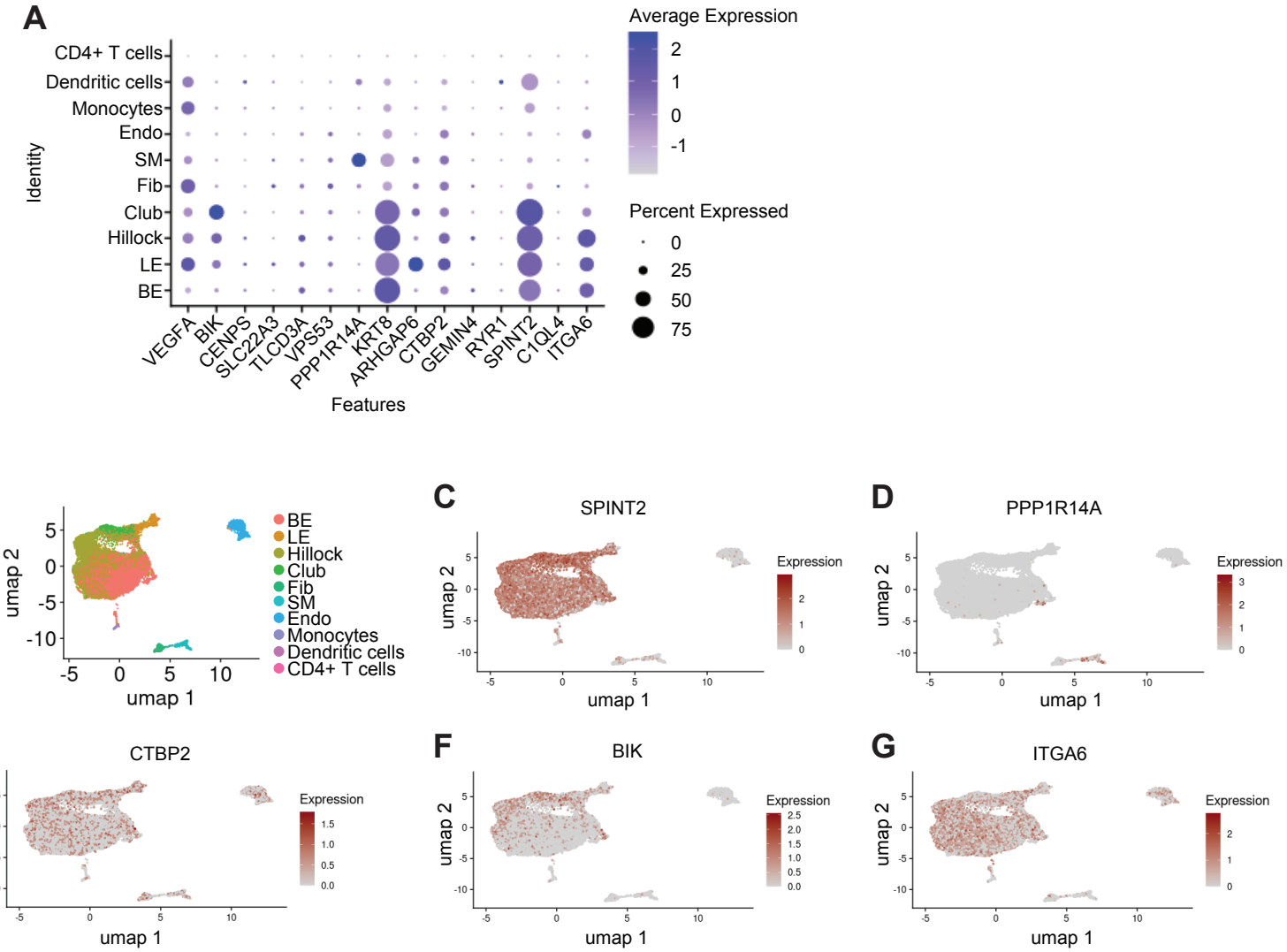

A
Average Expression
2
1
0
-1
Percent Expressed
0
25
50
75
CD4+ T cells
Dendritic cells
Monocytes
Endo
SM
Fib
Club
Hillock
LE
BE
Identity
VEGFA
BIK
CENPS
SLC22A3
TLCD3A
VPS53
PPP1R14A
KRT8
ARHGAP6
CTBP2
GEMIN4
RYR1
SPINT2
C1QL4
ITGA6
Features
B
BE
LE
Hillock
Club
Fib
SM
Endo
Monocytes
Dendritic cells
CD4+ T cells
umap 2
umap 1
C
SPINT2
Expression
D
PPP1R14A
Expression
E
CTBP2
Expression
F
BIK
Expression
G
ITGA6
Expression

Figure S6. Related to Figure 7. Single-cell RNA-Seq analysis of the transcriptomic targets of the PCa risk alleles.

(A) The dot plot shows the expression patterns of a subset of 15 high-specificity target genes (Fisher's combined FDR < 0.05) in 10 cell types of the prostate gland. The dot size represents the expression percentage in each cell type, and the dot color represents the expression level of the genes. (B) Color legends for individual cell-types in the two-dimensional umap plots (C-G) Cell-type specific expression patterns of selected genes in the single cell level: *SPINT2, PPP1R14A, CTBP2*, *BIK,* and *ITGA6*.

Supplementary Table S18. Previous works overlapping eQTL with PCa GWAS risk loci.

Publication	Overlapped PCa SNPs	Overlapped SNP-target pairs	Number of significant overlapped SNP-target pairs								PubMed ID
			High-specificity subset				High-sensitivity subset				
			TYPE I	TYPE II	TYPE III	Total	TYPE I	TYPE II	TYPE III	Total	
AlOlama et al. 2014	21	3	0	0	0	0	0	0	0	0	25217961
Grisanzio et al. 2012	4	1	0	0	0	0	0	0	0	0	22730461
Huang et al. 2014	1	1	0	1	0	1	0	1	0	1	24390282
Li et al. 2014	32	11	2	1	3	6	2	1	4	7	24907074
Mancuso et al. 2018	23	27	3	1	2	6	3	1	4	8	30287866
Penney et al. 2015	20	12	1	1	1	3	1	1	1	3	25371445
Thibodeau et al. 2015	119	27	5	3	5	13	5	3	7	15	26611117
Whitington et al. 2016	33	12	5	2	2	9	5	2	2	9	26950096
Xu et al. 2014	6	7	1	0	1	2	1	0	1	2	24022300

Description of Supplementary Table S18:

The table summarizes the number of overlapped SNPs between 185 PCa risk SNPs used in this study and other publications. And the number and classification of overlapped SNP-target pairs are also shown.

Supplementary Table S19. Novel SNP-target pairs identified for PCa risk SNPs reported in previous work.

Publication	Overlapped PCa SNPs	Novel SNP-target pairs	Number of significant novel SNP-target pairs								PubMed ID
			High-specificity subset				High-sensitivity subset				
			TYPE I	TYPE II	TYPE III	Total	TYPE I	TYPE II	TYPE III	Total	
AlOlama et al. 2014	21	1129	1	0	12	13	1	0	52	53	25217961
Grisanzio et al. 2012	4	1	0	0	0	0	0	0	1	1	22730461
Huang et al. 2014	1	103	0	0	2	2	0	0	9	9	24390282
Li et al. 2014	32	2911	2	0	15	17	2	0	114	116	24907074
Mancuso et al. 2018	23	1741	1	1	13	15	1	1	75	77	30287866
Penney et al. 2015	20	981	1	3	15	19	1	3	55	59	25371445
Thibodeau et al. 2015	119	8290	2	3	66	71	3	5	368	376	26611117
Whitington et al. 2016	33	3446	0	3	31	34	1	3	143	147	26950096
Xu et al. 2014	6	227	0	3	4	7	0	3	20	23	24022300

Description of Supplementary Table S19:

The table summarizes the number of overlapped SNPs between 185 PCa risk SNPs used in this study and other publications. And beyond the overlapped SNP-target pairs shown in Table S18, the number and classification of novel SNP-target pairs are also shown.

Supplementary Table S20. SNP-target pairs reported by previous study.

No.	SNP_target	Type	Subset	Publications
1	rs684232_FAM57A.exp	TYPE I	high-specificity	Whitington_2016;Thibodeau_2015;Mancuso_2018;Li_2014
2	rs651164_SLC22A3.exp	TYPE I	high-specificity	Thibodeau_2015;Mancuso_2018
3	rs8102476_SPINT2.exp	TYPE II	high-specificity	Thibodeau_2015;Penney_2015
4	rs6869841_STC2.exp	TYPE III	high-sensitivity	Thibodeau_2015
5	rs339331_RFX6.exp	TYPE II	high-specificity	Huang_2014
6	rs4962416_CTBP2.exp	TYPE II	high-specificity	Whitington_2016;Thibodeau_2015
7	rs684232_GEMIN4.exp	TYPE II	high-specificity	Whitington_2016;Thibodeau_2015;Mancuso_2018;Li_2014
8	rs10486567_HOXA13.exp	TYPE III	high-specificity	Whitington_2016
9	rs11672691_CEACAM21.exp	TYPE III	high-specificity	Whitington_2016;Thibodeau_2015;Mancuso_2018;Li_2014
10	rs12621278_ITGA6.exp	TYPE III	high-specificity	Thibodeau_2015;Penney_2015;Li_2014
11	rs636291_APITD1.exp	TYPE I	high-specificity	Whitington_2016
12	rs902774_KRT8.exp	TYPE I	high-specificity	Whitington_2016;Thibodeau_2015
13	rs10486567_HOTTIP.exp	TYPE III	high-specificity	Thibodeau_2015
14	rs651164_SLC22A2.exp	TYPE III	high-sensitivity	Mancuso_2018
15	rs684232_VPS53.exp	TYPE I	high-specificity	Whitington_2016;Thibodeau_2015;Mancuso_2018;Li_2014
16	rs6465657_BHLHA15.exp	TYPE III	high-specificity	Thibodeau_2015;Mancuso_2018
17	rs1983891_FOXP4.exp	TYPE III	high-specificity	Xu_2014;Thibodeau_2015
18	rs8102476_PPP1R14A.exp	TYPE I	high-specificity	Xu_2014;Whitington_2016;Thibodeau_2015;Penney_2015
19	rs10875943_C1QL4.exp	TYPE III	high-specificity	Li_2014

Description of Supplementary Table S20:

The table shows the subset of 19 SNP-target pairs reported in other publications before.

Supplementary Methods

Overlap of SNPs with genomic and peak regions

To get the significance of PCa SNP enrichment in peak or genomic regions, we used two ways of shuffling to generate the background distribution of SNP overlapping number.

Shuffling peak or genomic regions

Firstly, for all genomic regions (promoter, gene body and other region) and peak regions (RNA Pol II, RAD21, enhancer, AR and FOXA1), random sequence were shuffled from the whole genome (hg19) by *shuffle bedtools* for 10,000 times. This command kept the shuffled peak file with the same length and the same number as the original peak set file. Then, randomly shuffled genomic/peak regions were used to overlap with PCa SNP sites. Thirdly, 10,000 overlapped SNP number with the shuffled genomic/peak regions were generated to serve as the expected background in the histogram. Forth, p-value was calculated based on the observed overlapped SNP number and expected overlapped SNP number by the formula below (bootstrapping probability calculation):

$$\text{p-value} = \text{sum}(\text{abs}(\text{overlapNumber}) >= \text{abs}(\text{obsNumber})) / \text{length}(\text{overlapNumber}) \qquad \text{(i)}$$

obsNumber is the observed overlapped SNP number with real genomic/peak regions. OverlapNumber is overlapped SNP number between SNP and shuffled peak regions. Length is the shuffling times, here the number is 10,000. Abs means the absolute number. The effect size is calculated as the difference between observed overlapped SNP number and median value from the distribution of overlapped SNP number. Besides, the median number from the overlapping of SNP and shuffled genomic/peak regions were used as control to do the chi-square test.

Sampling SNPs

Firstly, GWAS SNPs were downloaded from https://www.ebi.ac.uk/gwas/api/search/downloads/alternative. As the coordinates of SNPs were hg38, we converted SNP coordinates from version hg38 to hg19 via online tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Secondly, we randomly selected SNPs from the whole GWAS SNP set for 10,000 times, with the same SNP number as PCa SNPs. Then, overlap 10,000 sampled SNP sets with the genomic/peak regions to generate the background of expected overlapped number. Lastly, the p-value was calculated using formula (i) above.

Shuffling interacted peak regions

As above, we shuffled interacted peak regions of RNA Pol II and RAD21 by *shuffle bedtools* for 1,000 times, to keep the shuffled interacted peaks with the same length and chromosome ID as sequenced peak region. Then, we used shuffled RNA Pol II and RAD21 interaction peak file to predict targets of PCa SNPs and LD SNPs for 1,000 times. Thirdly, 1,000 target number predicted by shuffled peak regions are pooled together as expected background. Lastly, the *p*-value is calculated as formula (i) to estimate the significance of difference between observed and expected target number.

Calculation of LD score

For PCa SNPs, linkage disequilibrium (LD) score was calculated by *PLINK* (--r^2) (http://pngu.mgh.harvard.edu/purcell/plink/) (1). All GWAS SNPs having $r^2 > 0.8$ with PCa risk SNPs were retained for down-stream analysis.

Prediction and classification of potential targets of PCa risk SNPs.

In this study, RNA Pol II and RAD21 ChIA-PET datasets were utilized to predict target genes. Firstly, all peak anchors were overlapped with SNPs; then the paired peak anchors which interacted with SNP-overlapped peak anchors were extracted as SNP-interacted peak anchors. Genes located within SNP-interacted peak anchors or located within interactions between SNP-overlapped peak anchors and SNP-interacted peak anchors were predicted as potential targets. According to SNP interacted positions with genes, potential genes were classified into three types: TYPE I, II and III. If the genes' promoters were located within the SNP-interacted peak anchor, the targets were classified as TYPE I. If the gene body regions overlapped with the SNP-interacted peak anchor, and genes' promoter region don't overlap with the SNP-interacted peak anchor, the targets were classified as TYPE II. If neither promoter or gene body overlapped with the SNP-interacted peak anchor, and genes were located inside interactions between SNP-overlapped peak anchors and SNP-interacted peak anchors, the genes were classified as TYPE III targets. In some cases, it is possible that a given SNP-target pair can be called either Type I, Type II or Type, III (e.g., analyses of different ChIA-PET datasets). In such a situation, Type I gets preference over Type II and Type III; Type II gets preference over Type III. The strategy of target prediction and validation for LD SNPs, TWAS SNPs and TWAS LD SNPs was as same as the pipeline described above for PCa risk SNPs.

Genotyping based on ChIP-seq data

For Porto cohort, we used ChIP-seq data from 94 prostate cancer patients to do the genotyping. All raw (.fastq) and processed peak (.bed) files were downloaded from GEO database (GSE120738), including 92 sets of H3K27ac, 76 sets of H3K27me3, 88 sets of AR and 56 sets of H3K4me3. Tophat was used to map raw fastqs to human genome (hg19) and picard was used to add readerGroup for each mapped bam file. Then the aligned bam files from each factor (H3K27ac, H3K4me3, H3K27me3, AR) were merged for each patient to increase the coverage and reduce allele bias due to allele specific binding. Next step, GATK (v4.x) HaplotypeCaller was used to do the genotyping at overlapping sites of interest for patients based on the merged bam files.

eQTL analysis of mRNA abundance and epigenetic marks binding

mRNA abundance in CPCG, TCGA and Porto cohorts

For each SNP-target pair, we tested for eQTLs first in the CPC-GENE, TCGA and Porto cohorts. We inverse rank normalized the mRNA abundance and fit a linear regression model treating genotypes additively, *i.e.* number of alternative alleles present (AA = 0, AB = 1, BB = 2). We included the first two genetic principal components to control for population structure and 10 PEER factors to control for technical noise in RNA sequencing (2). eQTLs

were considered validated as high-sensitivity if they met these two criteria: (A) p-value < 0.05 in one or more cohorts, and (B) the same direction of effect (β value) in all the cohorts. The raw p-values were corrected by Benjamini-Hochberg method. eQTLs were considered validated as high-specificity if they met these three criteria: (A) p-value < 0.05 in one or more cohorts, (B) the same direction of effect (β value) in all the cohorts, (C) Fisher's combined FDR < 0.2. The combined p-value was calculated by Fisher's method. The FDR was calculated by Benjamini-Hochberg method and was constructed for Type I, Type II, Type III SNP-target pairs separately.

Chromatin features in the Porto cohort

Consensus peak bed was generated by DiffBind R package based on the processed peak (.bed) files. The number of reads mapped within consensus peak was counted by coverageBed (bedtools). Then, the edgeR was used to normalize the overlapped read counts by TMM method for sequencing depth. Differential enrichment of H3K27ac or AR around SNP regions were quantified by Mann-Whitney test (recessive model).

Supplemental References

1. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D*, et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics **2007**;81:559-75
2. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nature protocols **2012**;7:500-7