nature genetics

Modulation of long noncoding RNAs by risk SNPs underlying genetic predispositions to prostate cancer

Haiyang Guo[1,18], Musaddeque Ahmed[1,18], Fan Zhang[2], Cindy Q Yao[3], SiDe Li[2], Yi Liang[1], Junjie Hua[1,4], Fraser Soares[1], Yifei Sun[2], Jens Langstein[1,5], Yuchen Li[1], Christine Poon[1], Swneke D Bailey[1], Kinjal Desai[6], Teng Fei[7], Qiyuan Li[8], Dorota H Sendorek[3], Michael Fraser[1], John R Prensner[9,17], Trevor J Pugh[1,4], Mark Pomerantz[7], Robert G Bristow[1,4], Mathieu Lupien[1,3,4], Felix Y Feng[10–13], Paul C Boutros[3,4,14], Matthew L Freedman[7,15], Martin J Walsh[2,16] & Housheng Hansen He[1,4]

Long noncoding RNAs (lncRNAs) represent an attractive class of candidates to mediate cancer risk. Through integrative analysis of the lncRNA transcriptome with genomic data and SNP data from prostate cancer genome-wide association studies (GWAS), we identified 45 candidate lncRNAs associated with risk to prostate cancer. We further evaluated the mechanism underlying the top hit, *PCAT1*, and found that a risk-associated variant at rs7463708 increases binding of ONECUT2, a novel androgen receptor (AR)-interacting transcription factor, at a distal enhancer that loops to the *PCAT1* promoter, resulting in upregulation of *PCAT1* upon prolonged androgen treatment. In addition, PCAT1 interacts with AR and LSD1 and is required for their recruitment to the enhancers of *GNMT* and *DHCR24*, two androgen late-response genes implicated in prostate cancer development and progression. PCAT1 promotes prostate cancer cell proliferation and tumor growth *in vitro* and *in vivo*. These findings suggest that modulating lncRNA expression is an important mechanism for risk-associated SNPs in promoting prostate transformation.

Recent large-scale projects such as the Encyclopedia of DNA Elements (ENCODE) and Functional Annotation of The Mammalian Genome (FANTOM) indicate that 75% of the human genome is transcribed into primary transcripts across different cell types, producing a range of noncoding RNAs[1,2]. Noncoding RNAs longer than 200 nt are classified as lncRNAs[3]. LncRNA genes are typically transcribed by RNA polymerase II, with the transcripts usually containing sequence from multiple exons and polyadenylated tails[3]. The recent description of more than 58,000 expressed lncRNAs in 27 tissue and cancer types makes these the most pervasive subclass of transcription in humans[4]. LncRNAs participate in a wide range of biological and cellular processes through mechanisms such as regulation of transcription and mRNA post-transcriptional processing[5–7]. Although a few lncRNAs have well-characterized functions, such as HOTAIR in promoting metastasis in breast cancer[8], MALAT-1 in inducing migration in lung cancer[9] and SChLAP1 in promoting aggressiveness in prostate cancer[10], the vast majority of this class of molecules remains functionally uncharacterized.

GWAS have identified thousands of SNPs associated with predisposition to disease[11]. Over 90% of these are located outside of the exons of protein-coding genes[4,12,13]. Characterizing the function of these trait-associated SNPs provides an opportunity to identify mechanisms driving these traits. Recent studies have shown that trait-associated SNPs are enriched in regulatory regions and can modulate transcription factor binding to these regions[14–19]. Efforts to interpret the functional consequences of noncoding SNPs have mainly focused on the regulation of protein-coding genes, although in a few cases lncRNA regulation was studied[20,21].

To systematically pinpoint lncRNAs involved in prostate cancer pathogenesis, we implemented a series of computational analyses using data from sources including chromatin accessibility, transcription factor binding, lncRNA expression and SNP genotyping. We identified 45 candidate lncRNAs regulated by noncoding SNPs in prostate cancer. Among these, we investigated *PCAT1*, a lncRNA located in the 8q24.21 gene desert. We demonstrate that *PCAT1* is regulated

[1]Princess Margaret Cancer Center, University Health Network, Toronto, Ontario, Canada. [2]Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [3]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [4]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [5]German Cancer Research Center (DKFZ), Heidelberg, Germany. [6]Department of Genetics, Norris Cotton Cancer Center, Dartmouth Medical School, Lebanon, New Hampshire, USA. [7]Department of Medical Oncology, Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [8]Medical College of Xiamen University, Xiamen, China. [9]Michigan Center for Translational Pathology, University of Michigan Medical School, Ann Arbor, Michigan, USA. [10]Department of Radiation Oncology, University of California at San Francisco, San Francisco, California, USA. [11]Department of Urology, University of California at San Francisco, San Francisco, California, USA. [12]Department of Medicine, University of California at San Francisco, San Francisco, California, USA. [13]Helen Diller Family Comprehensive Cancer Center, University of California at San Francisco, San Francisco, California, USA. [14]Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada. [15]Eli and Edythe L. Broad Institute, Cambridge, Massachusetts, USA. [16]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [17]Present address: Department of Pediatrics, Boston Children's Hospital, Boston, Massachusetts, USA. [18]These authors contributed equally to this work. Correspondence should be addressed to H.H.H. (hansenhe@uhnresearch.ca).

Received 26 January; accepted 6 July; published online 15 August 2016; doi:10.1038/ng.3637

by a SNP located in its enhancer region. Encoded by an androgen late-response gene, PCAT1 interacts with AR and lysine-specific demethylase 1 (LSD1) to promote prostate cancer cell growth.

RESULTS

Prostate cancer risk SNPs are enriched in regulatory regions

Prostate cancer GWAS have identified multiple common risk-associated SNPs (referred to as tag SNPs from here on) in different populations, cumulatively explaining 33% of familial risk[22]. We gathered the coordinates of 122 tag SNPs from the literature and the GWAS catalog (National Human Genome Research Institute–European Bioinformatics Institute; **Fig. 1a** and **Supplementary Table 1**). Of the 122 tag SNPs, 105 were significantly ($P \leq 1 \times 10^{-8}$) associated with prostate cancer in individuals of European ancestry, constituting most of the total tag SNPs (**Fig. 1a**). Recent studies suggest that causal variants are more likely to be SNPs in linkage disequilibrium (LD) with tag SNPs than the tag SNPs themselves[18,23]. We thus generated a list of SNPs that were in strong LD with each tag SNP using 1000 Genomes Project Phase I genotyping data. Using the LD cutoff of $r^2 \geq 0.8$, we identified 4,867 SNPs in LD with the tag SNPs (referred to as LD SNPs) across African, Asian and European groups (**Fig. 1a**). This resulted in a total of 4,989 LD and tag SNPs, which we refer to as risk SNPs from here on. These SNPs spanned 18, 5 and 75 independent risk-associated loci in African, Asian and European groups, respectively (**Fig. 1a**). Together, these risk loci constitute the associated variants set (AVS) for prostate cancer.

We next evaluated whether this AVS was enriched in selected genomic features using variant set enrichment (VSE) analysis

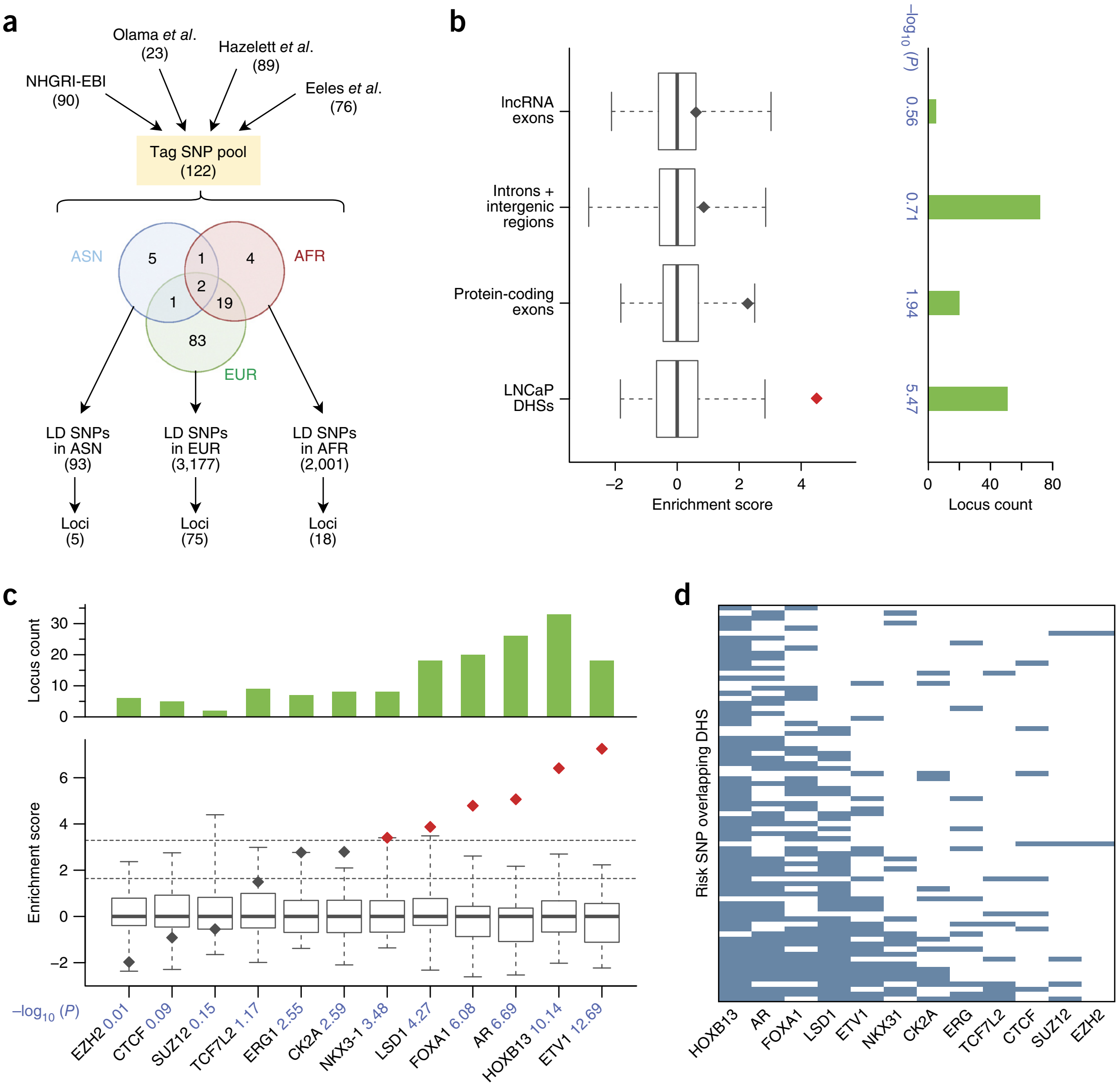

Figure 1 Genomic distribution of prostate cancer risk-associated SNPs. (**a**) Schematic of LD SNP compilation from available GWAS-derived studies[11,19,22,58]. AFR, African; ASN, Asian; EUR, European. (**b**) Enrichment analyses of the prostate cancer AVS in different genomic regions. Box-and-whisker plots show enrichment over null distributions—100 matching random AVSs generated from the pool of tag SNPs present on the GWAS array (Illumina Human OmniExpress). The whiskers in each box plot indicate the range of the data, and the bar inside the box corresponds to the median enrichment score. Each diamond denotes enrichment of the prostate cancer AVS in the respective genomic feature. The red diamond indicates significant enrichment relative to the null distribution. (**c**) Enrichment analyses of the prostate cancer AVS in the cistromes of different transcription factors. The lower dashed line represents the unadjusted *P*-value threshold and the upper dashed line denotes the Bonferroni-corrected *P*-value threshold for significance. (**d**) Occupancy matrix of transcription factor cistromes across DHSs.

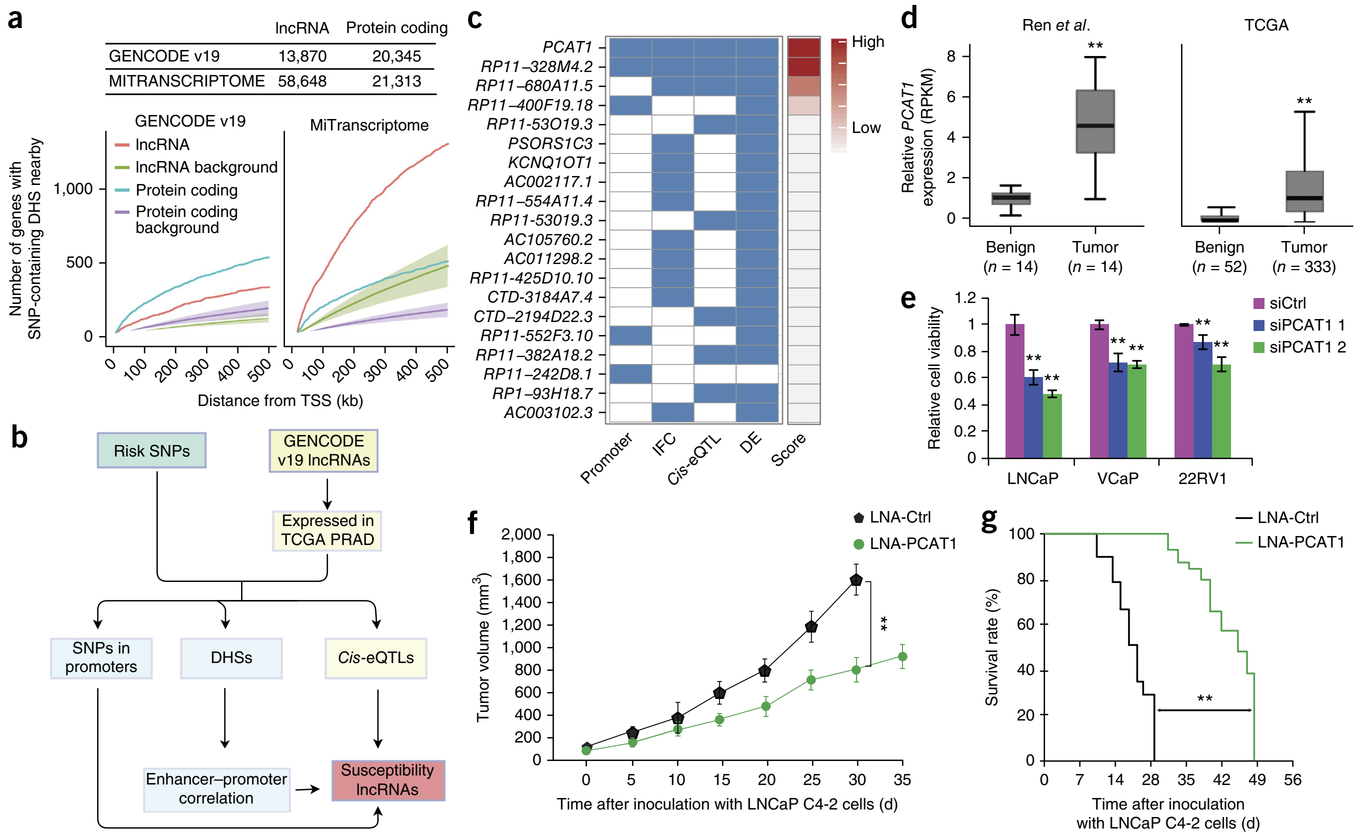

Figure 2 Identification of lncRNAs associated with prostate cancer risk. (**a**) DHSs encompassing one or more risk-associated SNPs are closer to protein-coding and lncRNA genes than genomic background. Shading corresponds to 95% confidence intervals. (**b**) A simplified diagram of the study approach. (**c**) List of the top 20 candidate lncRNA genes. Blue and white correspond to the presence or absence of each factor: risk SNP in promoter (Promoter), high intercellular functional correlation (IFC), eQTL with one or more risk SNPs (*Cis*-eQTL) and differential expression (DE). The score column shows a weighted summary of the binary states of the four factors. (**d**) Expression of *PCAT1* in benign and tumor samples from two different studies[59]. The whiskers in each box plot indicate the range of the data, and the bar inside the box denotes the median enrichment score. (**e**) siRNA-mediated knockdown of PCAT1 significantly reduces proliferation of LNCaP, 22RV1 and VCaP cells. The bar plots show relative cell proliferation on day 6 after treatment with control siRNA (siCtrl) or two different siRNAs targeting PCAT1 (siPCAT1 1 and 2). Error bars, s.d. from four technical replicates. (**f**,**g**) Tumor growth and survival analysis in LNCaP C4-2 xenografts. Locked nucleic acid RNA (LNA-RNA) was used to deplete endogenous *PCAT1* expression in LNCaP C4-2 cells. (**f**) Tumor volume was measured after tumor excision; error bars, s.d. from tumors in 12 mice per group. (**g**) Survival data for male mice ($n = 12$ per group) bearing LNCaP C4-2 cells were obtained from another batch of tumor inoculation. P values were calculated by Student's t test in **d**, **f** and **g** and by one-way ANOVA in **e**: $*P < 0.05$, $**P < 0.01$.

(Online Methods)[18]. The AVS was significantly enriched in DNase I–hypersensitive sites (DHSs) ($P = 3.4 \times 10^{-6}$) but not in the exons of lncRNA or protein-coding genes (**Fig. 1b**). We next extended our analysis to include all known transcription factor binding sites (cistromes) in prostate cancer cells[24] and observed that the prostate cancer AVS had significant over-representation ($P < 0.05$) of cistromes for AR and AR cofactors such as ETV1, HOXB13, FOXA1, LSD1 and NKX3-1 (**Fig. 1c**). Among all the DHSs that contained at least one risk SNP, more than 50% were occupied by one or more transcription factors (**Fig. 1d** and **Supplementary Fig. 1a**,**b**). Enrichment of risk-associated loci in open chromatin and the cistromes of functionally related transcription factors was further reflected when we analyzed the distribution of risk loci across different epigenetic modifications in LNCaP human prostate cancer cells. The prostate cancer risk loci were significantly enriched in regions with active epigenetic modifications, including acetylation of histone H3 at lysine 27 (H3K27ac) and monomethylation of histone H3 at lysine 4 (H3K4me1), but not in regions with the repressive modification trimethylation of histone H3 at lysine 27 (H3K27me3), in LNCaP cells (**Supplementary Fig. 1c**). Enrichment of prostate cancer risk loci in open chromatin and transcription factor cistromes is consistent with previous observations that common trait-associated variants are more likely to influence RNA abundance than to alter the exons of protein-coding genes[14,18]. Furthermore, when we evaluated the distribution of available fine-mapping data, we observed that SNPs significantly associated with prostate cancer[25–27] were highly enriched in DHSs and regions with active histone modifications in LNCaP cells (**Supplementary Fig. 1d**–**g**).

Risk-SNP-containing DHSs are proximal to lncRNAs

Of the 4,989 risk-associated SNPs, 343 overlapped with 265 DHSs in LNCaP cells. These DHSs were closer to lncRNAs or protein-coding genes than were randomly permutated DHSs that contained SNPs not associated with risk in prostate cancer (**Fig. 2a**). We conducted the same analysis using MiTranscriptome[4], which is more up to date and has a larger number of genes than GENCODE (v19)[28], and we observed a similar trend (**Fig. 2a**). Because of the higher number of lncRNAs annotated in MiTranscriptome, the number of lncRNAs in close proximity to risk-SNP-containing DHSs was much larger than that of protein-coding genes, suggesting that noncoding risk-associated SNPs may have equal or higher chances

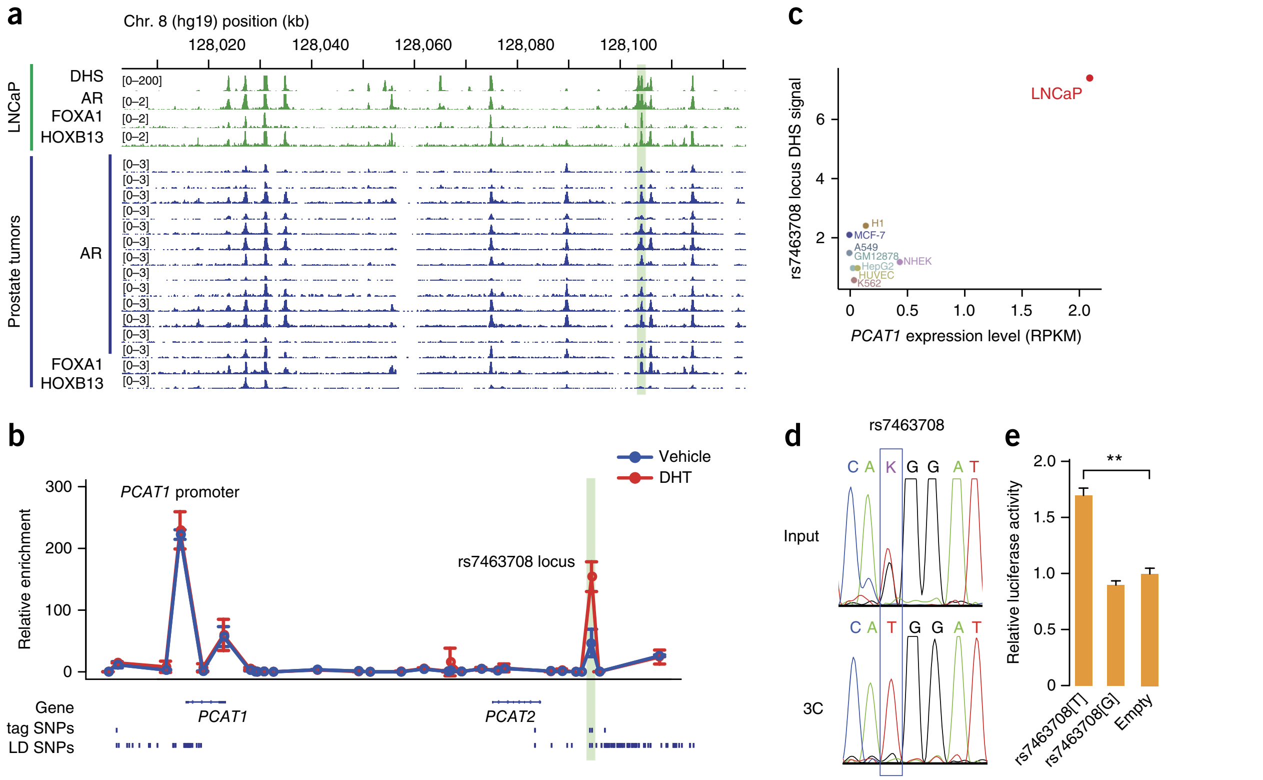

Figure 3 A risk-SNP-containing enhancer loops to the *PCAT1* promoter. (**a**) AR, FOXA1 and HOXB13 ChIP-seq and DHS signal in LNCaP cells and prostate tumor samples. The green shaded region highlights the rs7463708 locus. (**b**) Quantification of a 3C assay of the *PCAT1* genomic region. The data represent relative frequencies of interaction between the anchor region near the *PCAT1* TSS and PstI digestion sites (circles). All PstI digestion sites in this region were tested. Relative enrichment of interaction was determined by 3C–qPCR and normalized to interaction with a BAC control library. Error bars, s.d. from three technical replicates. (**c**) Expression level of *PCAT1* plotted against signal intensity for the rs7463708-containing DHS across nine cell lines. (**d**) Sanger sequencing of 3C input DNA and 3C amplicons from the *PCAT1*–rs7463708 locus. The blue box highlights risk SNP rs7463708. "K" represents nucleotides T and G. (**e**) Luciferase assay using plasmids containing the rs7463708 locus with either the risk or non-risk allele of rs7463708. The pGL3-Promoter plasmid (empty) was used as a baseline control. Luciferase signal was normalized to *Renilla* signal. Error bars, s.d. from three technical replicates. *P* values were calculated by Student's *t* test: ***P* < 0.01.

of modulating lncRNA genes than protein-coding genes to have causal roles in prostate cancer.

Integrative analysis identifies 45 candidate lncRNAs

To identify lncRNA genes regulated by noncoding risk-associated SNPs, we conducted an integrative analysis of risk-SNP-containing regulatory regions, lncRNA expression and SNP genotyping data (**Fig. 2b**). We retrieved RNA-seq data for benign prostate tissue and primary prostate cancer tumors from The Cancer Genome Atlas (TCGA) and analyzed expression of both lncRNA and protein-coding genes from GENCODE v19 annotation[29]. In line with previous studies[30,31], lncRNA genes were expressed at lower levels than protein-coding genes (**Supplementary Fig. 2a–d**). Of the 13,828 lncRNA genes that could be mapped and for which expression levels have been estimated, 7,375 were expressed in at least 50% of the total samples (**Supplementary Fig. 2a**). We further filtered out lncRNAs with expression levels lower than the median expression level for these 7,375 lncRNAs and considered the remaining 3,521 to be expressed in prostate cancer (**Supplementary Fig. 2e**). Using the same expression level cutoff, we retained 16,345 protein-coding genes that were considered to be expressed.

We further obtained available genotyping and copy number variation (CNV) data for the same set of samples from TCGA. The Affymetrix SNP 6.0 array platform used to genotype TCGA prostate cancer samples includes 243 risk-associated SNPs representing 52 independent prostate cancer risk loci. We performed a *cis*–expression quantitative trait locus (*cis*-eQTL) analysis for lncRNA and protein-coding genes that were located within 500 kb of these 243 SNPs after adjusting the expression values of lncRNA and protein-coding genes to account for the effect of CNV (Online Methods)[32]. We observed 18 lncRNAs whose expression was significantly associated ($P < 0.05$, false discovery rate (FDR) < 0.1) with 15 prostate cancer risk loci (**Supplementary Table 2**).

DHSs that contain one or more risk SNPs have the potential to regulate nearby lncRNA genes by looping to their promoter regions. Such association between DHSs, or enhancers, and lncRNA gene promoters can be predicted using intercellular functional correlation (IFC) analysis, which is a cross-correlation analysis, as previously described[33]. The foundation of this approach, as employed by others[34–37], is to evaluate the signal intensities of DHSs or other genomic measurements across multiple cell lines. Any two genomic coordinates with strongly correlated signal intensities for DHSs or other measurements are more likely to physically interact with one another[34]. Upon computing the correlation of DHS signal intensities for all DHSs containing at least one risk-associated SNP and lncRNA gene promoters, 27 lncRNAs were predicted to be strongly associated (Pearson's $r \geq 0.7$) with 31 DHSs that were within 500 kb of the lncRNA gene transcription start sites (TSSs) (**Supplementary Fig. 3**).

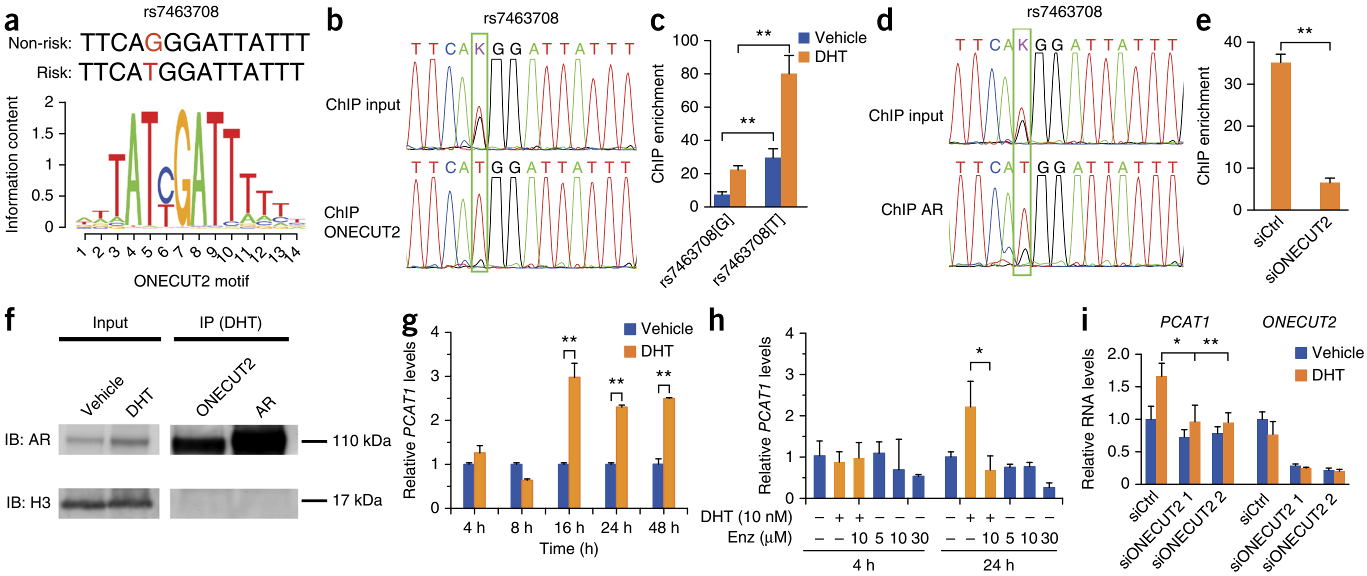

Figure 4 The prostate cancer risk SNP rs7463708 modulates ONECUT2 and AR binding at the *PCAT1* enhancer. (**a**) The T risk allele of rs7463708 creates a stronger ONECUT2 motif. (**b**,**c**) ONECUT2 preferentially binds to the T risk allele of rs7463708, as determined by ChIP–PCR followed by Sanger sequencing (**b**) and allele-specific ChIP–qPCR (**c**). (**d**) AR preferentially binds to the T risk allele of rs7463708, as determined by ChIP–PCR followed by Sanger sequencing. (**e**) AR binding at the rs7463708 locus was reduced by ONECUT2 knockdown (siONECUT2), as determined by ChIP–qPCR. (**f**) ONECUT2 and AR immunoprecipitation (IP) followed by immunoblot (IB) analysis. (**g**) *PCAT1* expression was induced by DHT stimulation in LNCaP cells after 16 h. (**h**) *PCAT1* upregulation induced by DHT in LNCaP cells at 24 h was blocked by the AR antagonist enzalutamide (Enz). (**i**) *PCAT1* expression was reduced after ONECUT2 knockdown. Error bars in **c**, **e** and **g**–**i**, s.d. from three technical replicates. *P* values were calculated by Student's *t* test in **c**, **e**, **g** and **h** and by one-way ANOVA in **i**: $*P < 0.05$, $**P < 0.01$.

The *cis*-eQTL-based SNP–lncRNA association together with the physical interaction prediction analyses provides strong support for regulation of lncRNA genes by prostate cancer risk loci. In addition, several risk-associated SNPs were identified in the promoter regions of five expressed lncRNA genes (*PCAT1*, *RP11-400F19.18*, *RP11-242D8.1*, *RP11-552F3.10* and *RP11-328M4.2*). To nominate lncRNAs as being under the regulation of risk loci, we considered the following factors to construct a comprehensive list of lncRNAs involved in prostate cancer susceptibility: (i) the distance between the lncRNA gene TSS and the risk-SNP-containing DHS (required to be less than 500 kb); (ii) association of the lncRNA with a risk SNP either by a *cis*-eQTL-based method or association with a DHS containing a risk SNP by the cross-correlation prediction, or presence of a risk SNP in the promoter region of the lncRNA gene; and (iii) expression of the lncRNA. Following this approach, we identified 45 candidate lncRNAs associated with 50% of the prostate cancer risk loci (**Supplementary Table 2**).

Although all 45 lncRNAs, including *KCNQ1OT1* and *H19* (**Supplementary Table 2**), may have merit for future study, we implemented a scoring system to prioritize validation of the predicted lncRNAs for initial characterization (Online Methods). The top scoring lncRNA gene was *PCAT1*, with support from *cis*-eQTL, cross-correlation, promoter alteration and expression analyses (**Fig. 2c**).

PCAT1 is located in the 8q24.21 gene desert region, which harbors ten loci associated with prostate cancer risk. The closest protein-coding gene in this region is *MYC*, which is 200 kb away from the nearest risk locus (**Supplementary Fig. 4**). Most previous studies have focused on relating risk SNPs to *MYC*, and an enhancer containing rs6983267 was reported to interact with the *MYC* promoter[38]. However, most risk SNPs in this region remain unexplored. There are ten lncRNA genes located in this gene desert (**Supplementary Fig. 4**). Among these, *PCAT1* had a relatively higher level of expression in normal prostate tissues and was significantly upregulated in primary prostate tumors (**Fig. 2d**). Consistent with a previous report[39], suppression of PCAT1 using two separate small interfering RNAs (siRNAs) diminished cell proliferation in LNCaP, VCaP and 22RV1 human prostate cancer cells (**Fig. 2e** and **Supplementary Fig. 5**). Notably, knockdown of PCAT1 dramatically reduced tumor growth of xenografts derived from LNCaP C4-2 cells and significantly prolonged the survival of mice bearing tumors (**Fig. 2f,g**).

The cross-correlation analysis predicted interaction between the *PCAT1* promoter and a DHS containing the risk SNP rs7463708 (**Supplementary Fig. 4**). In addition, our *cis*-eQTL analysis predicted an association between *PCAT1* and an independent risk locus defined by the tag SNP rs10086908 (**Supplementary Fig. 4**). The rs10086908 risk locus contains 28 SNPs in strong LD: 3 of these are located in the *PCAT1* promoter DHS, 6 are located in *PCAT1* exons, and the rest are located in intronic and intergenic regions with no or weak DHS signal (**Supplementary Fig. 4**). Because of the complexity of the rs10086908 risk locus, our functional validation was focused on the rs7463708 risk locus, which contains only three SNPs and is predicted to influence *PCAT1* expression only in a *cis*-regulatory manner.

rs7463708 modulates activity of *PCAT1* enhancer

A locus denoted as region 2 at 8q24 (chr. 8: 128.14–128.28 Mb, hg18) is strongly associated with prostate cancer in men of Japanese and African ancestry[40–43]. A fine-mapping and resequencing analysis of this region identified two SNPs, rs72725879 and rs7463708, to be most significantly associated with prostate cancer risk in Japanese individuals[44]. These two SNPs define a small risk locus with one additional LD SNP, rs1456315. We evaluated the association of these three SNPs with prostate cancer progression and observed a trend that the risk allele was associated with worse biochemical relapse rates (**Supplementary Fig. 6a**). This locus overlapped a strong DHS with

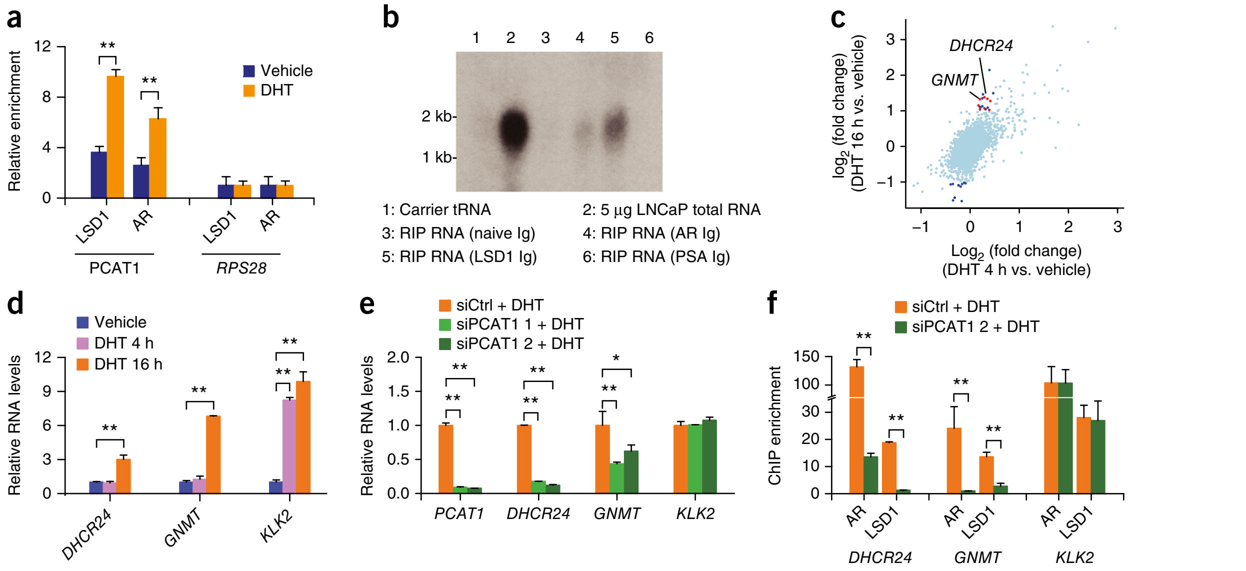

Figure 5 PCAT1 interacts with AR and LSD1 to regulate *GNMT* and *DHCR24* expression. (**a**) A RIP–qPCR assay in LNCaP cells shows that PCAT1 interacts with AR and LSD1. *RPS28* was used as a negative control. (**b**) RNA blotting of RNA precipitated by RIP assay in LNCaP cells shows that PCAT1 interaction with AR and LSD1. (**c**) Identification of androgen late-response genes. Expression data were retrieved from GSE11428. Blue points correspond to genes dysregulated after 16 h but not 4 h of DHT stimulation (activation cutoffs: absolute $\log_2$-transformed fold change (DHT 16 h versus vehicle) >1 and absolute $\log_2$-transformed fold change (DHT 4 h versus vehicle) <0.5). Red points correspond to late-response genes with AR and LSD1 binding within ±10 kb of the TSS. (**d**) qRT–PCR in LNCaP cells validates that *DHCR24* and *GNMT* are upregulated after 16 h but not 4 h of androgen stimulation. *KLK2* is an androgen early-response gene and was used as a control for androgen response. (**e**) Expression of *GNMT* and *DHCR24* was repressed by PCAT1 knockdown in LNCaP cells. (**f**) AR and LSD1 occupancy at *GNMT* and *DHCR24* enhancers was reduced upon knockdown of PCAT1 in LNCaP cells. Error bars in **a** and **d**–**f**, s.d. from three technical replicates. *P* values were calculated by Student's *t* test in **a** and **f** and by one-way ANOVA in **d** and **e**: $*P < 0.05$, $**P < 0.01$.

binding of multiple transcription factors, including AR, FOXA1 and HOXB13, in both prostate cancer cell lines and tumor samples (**Fig. 3a** and **Supplementary Fig. 6b**). Whereas rs1456315 and rs72725879 were located at the edge of the DHS, rs7463708 was located close to the center and is the only heterozygous SNP in this DHS in LNCaP cells (**Supplementary Fig. 6b**,**c**). The rs7463708-containing DHS is located 78 kb downstream of the *PCAT1* TSS, and the cross-correlation analysis identified strong correlation of DHS signal between this DHS and the *PCAT1* promoter, suggesting a long-range interaction (**Supplementary Fig. 4**).

To validate this interaction, we performed chromosome conformation capture (3C) using digestion by PstI enzyme. The 3C technique assesses whether a candidate fragment (for example, the *PCAT1* promoter) physically interacts with a region of interest (for example, the rs7463708 locus). Of all 28 PstI sites in this region, the rs7463708 locus had the strongest interaction with the *PCAT1* promoter upon dihydrotestosterone (DHT) stimulation (**Fig. 3b** and **Supplementary Fig. 7a**). Furthermore, *PCAT1* expression and the DHS signal in the rs7463708 locus were the highest in LNCaP cells among all other cell lines with matching chromatin accessibility and expression information (**Fig. 3c** and **Supplementary Fig. 7b**,**c**). This suggests that the rs7463708-containing DHS is specifically activated in LNCaP cells and may regulate prostate-specific *PCAT1* expression[39].

rs7463708 comprises a T risk allele and a G non-risk allele. The interaction with the *PCAT1* promoter under DHT stimulation was specific to the risk allele in LNCaP cells (**Fig. 3b**,**d**). We therefore further evaluated the enhancer activity of the rs7463708 locus, for both the G and T alleles, by luciferase reporter assay. Whereas the reporter vector containing the non-risk G allele of rs7463708 produced luciferase activity similar to that from the control vector, the reporter vector containing the T risk allele demonstrated significantly higher luciferase activity (**Fig. 3e**). Disruption of the rs7463708 locus by CRISPR/Cas9 reduced interaction between the *PCAT1* promoter and the rs7463708 locus and resulted in decreased expression of *PCAT1* (**Supplementary Fig. 7d**,**e**). It is worth noting that the rs7463708 locus and *PCAT1* promoter are located inside a putative topologically associating domain (TAD) in human embryonic stem cells (**Supplementary Fig. 8**)[45]. A TAD is formed when two distal genomic regions physically interact, often through a cohesin–CTCF complex, and insulate transcriptional regulation within the loop[46]. We observed that the CTCF-binding sites near the boundary of the TAD encompassing the *PCAT1*–rs7463708 interaction are conserved across tissues, and similar CTCF-mediated looping was validated in MCF-7 breast cancer and K562 leukemia cells (**Supplementary Fig. 8**). These results indicate that rs7463708 modulates the activity of an enhancer that regulates *PCAT1* transcription within a conserved TAD.

Motif analysis (Online Methods) indicated that SNP rs7463708 overlaps with the binding motif of ONECUT, and the motif had significantly higher preference for the T risk allele of rs7463708 (**Fig. 4a**). Of the three ONECUT family members, ONECUT2 is the only one expressed in prostate cancer (**Supplementary Fig. 9a**). In line with the motif analysis, ONECUT2 preferentially bound to the risk allele of rs7463708, as determined by both Sanger sequencing of ChIP–PCR amplicon and allele-specific ChIP–qPCR fragments in LNCaP cells (**Fig. 4b**,**c**). In addition, DHT stimulation increased ONECUT2 and AR binding to this locus (**Fig. 4c** and **Supplementary Fig. 9b**). Interestingly, although the AR motif is unaltered, AR binding was preferentially enriched by the risk allele of rs7463708 (**Fig. 4d** and **Supplementary Fig. 9c**), whereas silencing of ONECUT2 significantly decreased AR binding (**Fig. 4e**). This suggests an interaction

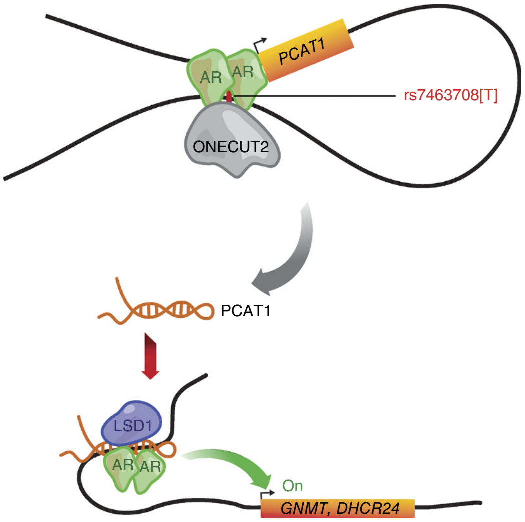

Figure 6 Graphical representation of the regulation and function of *PCAT1* in prostate cancer. The prostate cancer risk-associated T allele at rs7463708 results in increased expression of *PCAT1* through modulating ONECUT2 and AR binding at a distal enhancer. PCAT1 recruits AR and LSD1 to *GNMT* and *DHCR24* enhancers to activate transcription of these genes.

between ONECUT2 and AR, which was confirmed by immunoprecipitation analysis (**Fig. 4f**).

To further evaluate the involvement of AR in *cis* regulation of *PCAT1*, we estimated the expression of *PCAT1* in LNCaP cells upon androgen stimulation and observed a significant increase in expression at the late time point (16 h; **Fig. 4g**). This induction could be blocked by pretreatment of LNCaP cells with the androgen receptor antagonist enzalutamide or by silencing of ONECUT2 (**Fig. 4h**,**i**). Taken together, these data suggest that rs7463708 modulates ONECUT2 and AR binding to regulate *PCAT1* transcription.

PCAT1 mediates genomic occupancy of the AR–LSD1 complex

Having demonstrated that *PCAT1* is an androgen late-response gene, we assessed whether *PCAT1* is involved in the androgen signaling pathway, a key pathway in prostate cancer pathogenesis[47]. LncRNAs have previously been found to interact with protein complexes to regulate target gene transcription and expression[6,48]. To determine whether PCAT1 interacts with the AR protein complex, we performed RNA immunoprecipitation (RIP) assays against AR and LSD1, a major regulator of AR transcriptional activity[49,50]. The RIP–qPCR assays demonstrated binding of PCAT1 to both AR and LSD1 in an androgen-dependent manner (**Fig. 5a** and **Supplementary Fig. 10**). RNA blotting analysis of RNA precipitated from RIP assays further confirmed the interaction of PCAT1 with AR and LSD1 (**Fig. 5b**). To test whether PCAT1 RNA could pull down AR and LSD1, we fused a streptomycin-binding RNA aptamer to PCAT1 lncRNA by CRISPR/Cas9-mediated knock-in and analyzed proteins eluted from streptomycin-coated beads via multidimensional protein identification technology (MudPIT). MudPIT analysis identified multiple PCAT1-binding proteins, and AR and LSD1 as well as their interacting proteins were among the top factors (**Supplementary Table 3**). These results suggest that PCAT1 may participate in the regulation of other androgen late-response genes by interacting with the AR–LSD1 complex.

To test this hypothesis, we first identified genes that are upregulated by androgen at 16 h, but not at 4 h, in LNCaP cells[51] (**Fig. 5c**). To identify AR and LSD1 direct targets, we filtered out genes without AR- and LSD1-binding sites within 10 kb of the TSS. This analysis identified eight androgen late-response genes; six of these were verified by qRT–PCR assays (**Fig. 5c**,**d** and **Supplementary Fig. 11a**). Knockdown of PCAT1 using siRNAs decreased the abundance of three of these six genes (**Fig. 5e** and **Supplementary Fig. 11b**), and overexpression of PCAT1 confirmed the regulation of two genes, *GNMT* and *DHCR24* (**Supplementary Fig. 11c**). *GNMT* is reported as a cancer susceptibility gene[52] and is strikingly upregulated in a subset of primary prostate tumors in comparison with normal prostate tissues (**Supplementary Fig. 11d**). *DHCR24* encodes an oxidoreductase that has been implicated in prostate cancer progression[53].

AR and LSD1 regulate the transcription of target genes through interaction with chromatin[49,50]. We identified AR- and LSD1-binding sites near *GNMT* and *DHCR24* (**Supplementary Fig. 11e**–**j**). To evaluate the role of PCAT1 in AR and LSD1 interaction with chromatin, we performed ChIP against AR and LSD1 in LNCaP cells that were treated with siRNAs targeting PCAT1. Knockdown of PCAT1 resulted in significant decrease in AR and LSD1 binding to the sites near the *GNMT* and *DHCR24* TSSs (**Figs. 5f** and **6**, and **Supplementary Fig. 11k**). Interestingly, this effect was not observed at AR- and LSD1-binding site near *KLK2* (**Fig. 5f**), an androgen early-response gene, suggesting that PCAT1 regulation of AR and LSD1 chromatin interaction is site specific. Furthermore, our chromatin isolation by RNA purification (ChIRP) assay confirmed the binding of PCAT1 at enhancers for *GNMT* and *DHCR24*, but not *KLK2* (**Supplementary Fig. 11l**,**m**). These data suggest that PCAT1 regulates *GNMT* and *DHCR24* expression through the recruitment of AR and LSD1 to enhancers.

DISCUSSION

The molecular mechanisms underlying the causal actions and biological effects of prostate cancer risk-associated SNPs are largely unknown. As with most other diseases and traits, these risk-associated SNPs map to primarily noncoding regions of the genome. Here we show that prostate cancer risk-associated SNPs are enriched in open chromatin regions and cistromes of transcription factors that are involved in the androgen signaling pathway. The potential functional SNPs located in open chromatin regions are in close proximity to both protein-coding genes and lncRNAs. Because of the recent explosion in the number of lncRNAs characterized, the number of lncRNAs neighboring these noncoding SNPs far exceeds that of protein-coding genes, providing a strong rationale to investigate the functional link between the noncoding SNPs and lncRNAs.

The integrative approach taken in this study resulted in detection of 45 susceptibility lncRNAs in prostate cancer. Although we validated only the top ranked lncRNA, *PCAT1*, because of limitation of scope, the rest of the lncRNAs also need to be analyzed in depth. The inclusion in the list of several already known important lncRNAs, including *H19* and *KCNQ1OT1*, further validates our approach. However, despite its comprehensiveness, the approach may have underestimated the number of lncRNAs involved in prostate cancer predisposition. The expression of lncRNAs and the DHS signals in their promoter regions are low, which reduced the sensitivity of the cross-correlation and *cis*-eQTL analyses. Moreover, the SNP array used for genotyping TCGA samples does not include all SNPs that are in strong LD with the associated SNPs; this restricts the *cis*-eQTL analysis to only the risk SNPs that are present on the array. These factors may have contributed to the weaker agreement between *cis*-eQTL and cross-correlation analyses across the nominated lncRNAs, although a strong agreement is not expected because the cross-correlation analysis predicts direct associations whereas *cis*-eQTL analysis may also predict indirect associations[32]. These limitations can be overcome as more and better quality data become available.

Our approach implicates *PCAT1* as a top prostate cancer susceptibility lncRNA that is affected by risk SNP rs7463708 located in an enhancer region 78 kb away. This risk SNP was also identified through a fine-mapping study and therefore may be the causal variant at this locus[44]. We observed that the T risk allele of rs7463708 creates a stronger binding site for the transcription factor ONECUT2, which subsequently interacts with AR at this enhancer region. To our knowledge, the identification of ONECUT2 as an interacting factor of AR is new, and this warrants further analysis. ONECUT2 has previously been reported to regulate epithelial–mesenchymal transition, migration and invasion of colorectal cancer cells[54]; however, its function in prostate cancer, especially in the androgen signaling pathway, needs to be further explored. The SNP rs7463708 overlaps with another lncRNA gene, *PRNCR1*, which is not annotated in GENCODE v19 or MiTranscriptome, but is in RefSeq[55]. *PRNCR1* expression is not detectable from LNCaP and TCGA prostate adenocarcinoma (PRAD) poly(A)-selected RNA-seq data , and this lncRNA is thus excluded from our analysis. However, this does not eliminate the possibility of a functional relationship between this SNP and *PRNCR1*, and this warrants further analysis. In addition to distal *cis* regulation of *PCAT1* by rs7463708, we identified another risk locus, rs10086908, as being associated with *PCAT1* expression. Nine SNPs in this locus are positioned across the *PCAT1* promoter and exons, indicating a more complicated regulation of this lncRNA by this locus. Future efforts are warranted to explore the underlying causal mechanisms of association between the rs10086908 locus and the *PCAT1* lncRNA. Taken together, our analysis identified two separate risk loci that converge upon the lncRNA *PCAT1*, suggesting its importance in prostate transformation.

PCAT1 has been implicated in prostate cancer progression[39]. It has been reported to suppress BRCA2 expression and regulate MYC stabilization in the cytoplasm in prostate cancer[56,57]. Our study highlights a previously unknown function of PCAT1 in the androgen signaling pathway in the nucleus. PCAT1 interacts with AR and LSD1 upon prolonged androgen treatment. In addition, PCAT1 is required for the recruitment of AR and LSD1 to binding sites near *GNMT* and *DHCR24*, two androgen late-response genes in prostate cancer. Considering the importance of BRCA2, MYC and androgen signaling in prostate cancer, these data suggest that PCAT1 may promote prostate tumorigenesis through multiple pathways.

In conclusion, our integrative analysis of the lncRNA transcriptome with genomicdata, epigenomic profiles and genetic predispositions has identified a set of lncRNAs that are implicated in susceptibility to prostate cancer, which should be prioritized for future analysis in tumorigenicity assays. We suggest that similar studies should be performed to characterize risk-associated lncRNAs in all cancer types.

URLs. ENCODE data download page at the UCSC Genome Browser, http://genome.ucsc.edu/ENCODE/downloads.html; HaploReg, http://www.broadinstitute.org/mammals/haploreg/haploreg.php; PLINK, http://pngu.mgh.harvard.edu/purcell/plink/; TCGA Research Network, http://cancergenome.nih.gov/; VSE R package, https://cran.r-project.org/web/packages/VSE/index.html.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Brown, X.S. Liu, D. Borges-Rivera and R.A. Young for discussion, as well as the Princess Margaret Genomic Centre for high-throughput sequencing support. The work was supported by the Princess Margaret Cancer Foundation (to H.H.H., M.L., T.J.P. and R.G.B.), the Canada Foundation for Innovation and the Ontario Research Fund (CFI32372 to H.H.H. and CFI32383 to T.J.P.), NSERC discovery grant (498706 to H.H.H.), the WICC Ontario 20th Anniversary Prostate Cancer Innovation Grant of the CCS (703800 to H.H.H.), and a CIHR transitional operating grant (142246 to H.H.H.) and Movember Rising Star awards from Prostate Cancer Canada (RS2014-01 to P.C.B., RS2014-04 to M.L. and RS2016-1022 to H.H.H.). H.H.H. holds an ICS-IG Maud Menten New Principal Investigator Prize (ICS-145381) and an OMIR Early Researcher Award. M.L. holds a young investigator award from the OICR and a new investigator salary award from the CIHR. R.G.B. is a recipient of a Canadian Cancer Society Research Scientist Award. P.C.B. is supported by a Terry Fox Research Institute New Investigator Award and a CIHR New Investigator Award. M.L.F. is supported by the Prostate Cancer Foundation (Challenge Award) and an NIH grant (R01CA193910). J.H. is a CIHR Graduate Student Fellowship recipient. S.D.B. is a CIHR Postdoctoral Fellowship recipient. M.J.W. is supported by grants 5RO1CA154903 and 5RO1HL103967 from the NIH. We acknowledge the ENCODE Consortium and the ENCODE production laboratories that generated the data sets provided by the ENCODE Data Coordination Center used in the manuscript. The results shown here are in whole or part based upon data generated by the TCGA Research Network.

AUTHOR CONTRIBUTIONS

H.G., M.A. and H.H.H. designed the studies and wrote the manuscript. H.G., Y. Liang, J.H. and J.L. performed the experiments with help from S.L., Y.S., C.P., T.F., K.D., J.R.P., F.S. and Y. Li. M.A., H.G. and H.H.H. conducted the data analysis with help from F.Z., C.Q.Y., D.H.S., P.C.B., R.G.B., M.L.F., S.D.B., Q.L., T.J.P., M.P., F.Y.F., M.L.F. and M.J.W. M.L., M.L.F., F.Y.F., J.R.P., M.P., M.J.W. and P.C.B. revised the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
2. FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
3. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
4. Iyer, M.K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
5. Geisler, S. & Coller, J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.* **14**, 699–712 (2013).
6. Rinn, J.L. & Chang, H.Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145–166 (2012).
7. Prensner, J.R. & Chinnaiyan, A.M. The emergence of lncRNAs in cancer biology. *Cancer Discov.* **1**, 391–407 (2011).
8. Gupta, R.A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076 (2010).
9. Schmidt, L.H. *et al.* The long noncoding MALAT-1 RNA indicates a poor prognosis in non–small cell lung cancer and induces migration and tumor growth. *J. Thorac. Oncol.* **6**, 1984–1992 (2011).
10. Prensner, J.R. *et al.* The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat. Genet.* **45**, 1392–1398 (2013).
11. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP–trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
12. Zhang, X., Bailey, S.D. & Lupien, M. Laying a solid foundation for Manhattan—'setting the functional basis for the post-GWAS era'. *Trends Genet.* **30**, 140–149 (2014).
13. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
14. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
15. Huang, Q. *et al.* A prostate cancer susceptibility allele at 6q22 increases *RFX6* expression by modulating HOXB13 chromatin binding. *Nat. Genet.* **46**, 126–135 (2014).
16. Zhang, X., Cowper-Sal lari, R., Bailey, S.D., Moore, J.H. & Lupien, M. Integrative functional genomics identifies an enhancer looping to the *SOX9* gene disrupted by the 17q24.3 prostate cancer risk locus. *Genome Res.* **22**, 1437–1446 (2012).
17. Yao, L., Tak, Y.G., Berman, B.P. & Farnham, P.J. Functional annotation of colon cancer risk SNPs. *Nat. Commun.* **5**, 5114 (2014).

18. Cowper-Sallari, R. *et al.* Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.* **44**, 1191–1198 (2012).
19. Hazelett, D.J. *et al.* Comprehensive functional annotation of 77 prostate cancer risk loci. *PLoS Genet.* **10**, e1004102 (2014).
20. Meyer, K.B. *et al.* A functional variant at a prostate cancer predisposition locus at 8q24 is associated with *PVT1* expression. *PLoS Genet.* **7**, e1002165 (2011).
21. Kim, T. *et al.* Long-range interaction and correlation between *MYC* enhancer and oncogenic long noncoding RNA CARLo-5. *Proc. Natl. Acad. Sci. USA* **111**, 4173–4178 (2014).
22. Al Olama, A.A. *et al.* A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat. Genet.* **46**, 1103–1109 (2014).
23. McClellan, J. & King, M.C. Genetic heterogeneity in human disease. *Cell* **141**, 210–217 (2010).
24. Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.* **12**, R83 (2011).
25. Al Olama, A.A. *et al.* Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat. Genet.* **41**, 1058–1060 (2009).
26. Amin Al Olama, A. *et al.* Multiple novel prostate cancer susceptibility signals identified by fine-mapping of known risk loci among Europeans. *Hum. Mol. Genet.* **24**, 5589–5602 (2015).
27. Han, Y. *et al.* Integration of multiethnic fine-mapping and genomic annotation to prioritize candidate functional SNPs at prostate cancer susceptibility regions. *Hum. Mol. Genet.* **24**, 5603–5618 (2015).
28. Harrow, J. *et al.* GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
29. Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
30. Popadin, K., Gutierrez-Arcelus, M., Dermitzakis, E.T. & Antonarakis, S.E. Genetic and epigenetic regulation of human lincRNA gene expression. *Am. J. Hum. Genet.* **93**, 1015–1026 (2013).
31. Cabili, M.N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
32. Li, Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633–641 (2013).
33. Bailey, S.D. *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat. Commun.* **2**, 6186 (2015).
34. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
35. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
36. Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* **24**, 1–13 (2014).
37. Sheffield, N.C. *et al.* Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res.* **23**, 777–788 (2013).
38. Ahmadiyeh, N. *et al.* 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with *MYC*. *Proc. Natl. Acad. Sci. USA* **107**, 9742–9746 (2010).
39. Prensner, J.R. *et al.* Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.* **29**, 742–749 (2011).
40. Haiman, C.A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* **39**, 638–644 (2007).
41. Robbins, C. *et al.* Confirmation study of prostate cancer risk variants at 8q24 in African Americans identifies a novel risk locus. *Genome Res.* **17**, 1717–1722 (2007).
42. Takata, R. *et al.* Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. *Nat. Genet.* **42**, 751–754 (2010).
43. Han, Y. *et al.* Prostate cancer susceptibility in men of African ancestry at 8q24. *J. Natl. Cancer Inst.* **108**, djv431 (2016).
44. Chung, S. *et al.* Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. *Cancer Sci.* **102**, 245–252 (2011).
45. Dowen, J.M. *et al.* Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374–387 (2014).
46. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
47. Shen, M.M. & Abate-Shen, C. Molecular genetics of prostate cancer: new prospects for old challenges. *Genes Dev.* **24**, 1967–2000 (2010).
48. Tsai, M.C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689–693 (2010).
49. Cai, C. *et al.* Lysine-specific demethylase 1 has dual functions as a major regulator of androgen receptor transcriptional activity. *Cell Rep.* **9**, 1618–1627 (2014).
50. Metzger, E. *et al.* LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription. *Nature* **437**, 436–439 (2005).
51. Wang, Q. *et al.* Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. *Cell* **138**, 245–256 (2009).
52. Song, Y.H., Shiota, M., Kuroiwa, K., Naito, S. & Oda, Y. The important role of glycine *N*-methyltransferase in the carcinogenesis and progression of prostate cancer. *Mod. Pathol.* **24**, 1272–1280 (2011).
53. Romanuik, T.L. *et al.* LNCaP Atlas: gene expression associated with *in vivo* progression to castration-recurrent prostate cancer. *BMC Med. Genomics* **3**, 43 (2010).
54. Sun, Y. *et al.* MiR-429 inhibits cells growth and invasion and regulates EMT-related marker genes by targeting Onecut2 in colorectal carcinoma. *Mol. Cell. Biochem.* **390**, 19–30 (2014).
55. Pruitt, K.D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42**, D756–D763 (2014).
56. Prensner, J.R. *et al.* PCAT-1, a long noncoding RNA, regulates BRCA2 and controls homologous recombination in cancer. *Cancer Res.* **74**, 1651–1660 (2014).
57. Prensner, J.R. *et al.* The long non-coding RNA PCAT-1 promotes prostate cancer cell proliferation through cMyc. *Neoplasia* **16**, 900–908 (2014).
58. Eeles, R. *et al.* The genetic epidemiology of prostate cancer and its clinical implications. *Nat. Rev. Urol.* **11**, 18–31 (2014).
59. Ren, S. *et al.* RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Res.* **22**, 806–821 (2012).

ONLINE METHODS

Data acquisition. The following ChIP-seq data for LNCaP cells were retrieved from public databases: ERG ChIP-seq data (GSE14097)[60]; CTCF ChIP-seq data (GSE38684)[61]; CK2A ChIP-seq data (GSE58607)[62]; H3K27me3, SUZ12 and EZH2 ChIP-seq data (GSE39459)[63]; ETV1 ChIP-seq data (GSE47120)[64]; H3K27ac and TCF7L2 ChIP-seq data (GSE51621)[19]; POLII, AR and NKX3-1 ChIP-seq data (GSE28264)[65]; H3K4me2 and H3K4me3 ChIP-seq data (GSE20042)[66]; and AR, FOXA1 and HOXB13 ChIP-seq data in human prostate tumor samples (GSE56288)[67]. The RNA-seq data for Asian population were retrieved from EMBL-EBI (E-MTAB-567)[59]. CTCF ChIP-seq and ChIA-PET data were retrieved from the UCSC ENCODE download portal.

Cell culture and treatment. The LNCaP, 22RV1 and VCaP cell lines were obtained from the American Type Culture Collection (ATCC). All prostate cancer cell lines were cultured as recommended by ATCC. No mycoplasma contamination was detected in these cell lines by MycoAlert Mycoplasma Detection kit (LT07-118, Lonza). LNCaP cells were serum starved for 48 h followed by 4 or 16 h of stimulation with DHT (10 nM). The AR antagonist enzalutamide was added 2 h before DHT treatment.

siRNA transfection. siRNAs targeting PCAT1 and control siRNA were purchased from GE Dharmacon. siRNAs targeting ONECUT2 and control siRNA were purchased from Thermo Fisher. Lipofectamine RNAiMAX transfection reagent (13778150, Thermo Fisher) was used for siRNA transfection following the manufacturer's instructions. The siRNA target sequences are listed in **Supplementary Table 4**.

Preparation of lentiviral vectors and infection. Lentiviral vectors of pLenti6-PCAT1 and pLenti6-LacZ were generated by the laboratory of F.Y.F. (ref. 56). Lentiviral particle production and infection were performed as described previously[68]. In brief, lentiviral vectors were cotransfected with psPAX2 and pMD2G vectors into HEK293T cells. Supernatants were collected at 24 and 48 h after transfection and stored at −80 °C. For infection, 5×10^4 cells per well were seeded in six-well plates and infected with lentiviral supernatant on the following day.

Chromatin immunoprecipitation. ChIP assays were performed using LNCaP cells treated with DHT for 24 h. Protein A (88845, Thermo Fisher) and G (88847, Thermo Fisher) Dynabeads were mixed at a 1:1 ratio and preincubated with antibodies 3 h before immunoprecipitation. LNCaP cells were cross-linked by 1% formaldehyde for 10 min, and the reaction was quenched with 125 mM glycine. After washing with cold PBS, nuclear fractions were extracted by 10 ml of LB1 buffer (50 mM HEPES-KOH, pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% IGEPAL CA-630 and 0.25% Triton X-100) for 10 min at 4 °C. Nuclear fractions were then pelleted and resuspended in 10 ml of LB2 buffer (10 mM Tris-HCl, pH 8.0, 200 mM NaCl, 1 mM EDTA and 0.5 mM EGTA) at 4 °C for 5 min. Nuclear fractions were pelleted again and resuspended in LB3 buffer (10 mM Tris-HCl, pH 8, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% sodium deoxycholate, 0.5% N-lauroylsarcosine and protease inhibitor cocktail). Nuclear fractions were sonicated in a water bath sonicator (Diagenode Bioruptor) to generate chromatin fragments ~300 bp in length. 0.1 volume of 10% Triton X-100 was added to chromatin lysate. Chromatin lysate was cleared by centrifugation, and 10% of the supernatant was taken as input DNA. The remaining chromatin lysate was divided equally between tubes with antibody-conjugated beads and rotated at 4 °C overnight. The antibodies used for ChIP assays were to AR (sc-13062, Santa Cruz Biotechnology), ONECUT2 (ab181229, Abcam) and LSD1 (ab17721, Abcam). The beads were washed in RIPA buffer, and elution buffer (0.1 M $NaHCO_3$, 1% SDS and proteinase K) was used to reverse cross-linking of DNA–protein complexes at 65 °C for 8–16 h. DNA was purified by phenol-chloroform extraction and subjected to qPCR. All primers used in this study are listed in **Supplementary Table 4**.

Allele-specific ChIP–qPCR. ChIP–qPCR was performed as previously described[66]. The allele-specific enrichment of AR and ONECUT2 was assessed by real-time-based mismatch amplification mutation assays. For allele-specific ChIP–qPCR, primers were designed as previously described to discriminate between two alleles[18]. The ultimate 3′ base of the forward primer was designed to be allele specific, and the penultimate 3′ bases of both the forward primer and reverse primer was designed to be a mismatch. The primer sequences are listed in **Supplementary Table 4**.

RNA immunoprecipitation. LNCaP cells were fixed by formaldehyde for 10 min at 37 °C, and the cross-linking was quenched by glycine. Cells were collected after washing with cold PBS. Nuclear fractions were extracted and subjected to sonication. ChIP-grade antibodies were used to precipitate RNA from nuclear extracts. RNase inhibitor (03335399001, Roche Life Science) was added to prevent RNA degradation. The antibodies used in the RIP assays were to AR (sc-13062, Santa Cruz Biotechnology), LSD1 (ab17721, Abcam), YY1 (sc-281, Santa Cruz Biotechnology) and hnRNPL (sc-32317, Santa Cruz Biotechnology). The primer sequences for the RIP assays are listed in **Supplementary Table 4**.

Chromosome conformation capture assays. The 3C assay was performed as previously described[16,38]. In brief, 5×10^6 LNCaP cells were fixed with 1% formaldehyde and lysed with cell lysis buffer. Nuclear extracts were digested with 400 U PstI (R0140L, New England BioLabs) at 37 °C overnight. Digested chromatin was diluted in ligation buffer (750 µl of 10% Triton X-100, 750 µl of 10× NEB ligation buffer, 75 µl of 10 mg/ml BSA, 5,925 µl of distilled water and 4,000 U T4 DNA ligase) and incubated at 16 °C overnight. DNA fragments were extracted with phenol-chloroform and subjected to PCR amplification using the primers listed in **Supplementary Table 4**. A BAC clone (RPCI-11-524D19) containing human genomic region 8q24.21 used for the 3C assay was purchased from Empire Genomics and used for 3C control template preparation. The ligation product from PstI-digested BAC DNA was used to verify primer efficiency and to normalize 3C interaction frequency.

Enhancer deletion. An sgRNA targeting rs7463708 was cloned into the lentiCRISPRv2 vector (52961, Addgene), and the vector was packaged as lentivirus in HEK293T cells. LNCaP cells were infected by either negative-control lentivirus or rs7463708 locus deletion lentivirus and then selected with puromycin. The sequence of the sgRNA is listed in **Supplementary Table 4**.

Enhancer reporter assays. The rs7463708-containing DHS region selected on the basis of DNase-seq data in LNCaP cells was amplified from LNCaP genomic DNA and inserted upstream of an SV40 promoter in the pGL3-Promoter vector (E1761, Promega). Site-directed mutagenesis was performed to obtain either the T or G allele at rs7463708. The primers used for reporter construction are listed in **Supplementary Table 4**. For enhancer assays, 4×10^4 LNCaP cells were seeded into 24-well tissue culture plate. Cells were cotransfected with both reporter plasmid and pRL-TK *Renilla* Luciferase Control Vector (E2241, Promega) using Lipofectamine 2000 reagent (12566014, Thermo Fisher). Cells were collected using passive lysis buffer (E194A, Promega), and luciferase activity was determined using the Dual-Luciferase Reporter Assay System (E1910, Promega). The luminescent signals from experimental samples were normalized to those from control samples to obtain relative luciferase activities. The data are presented as means ± s.d. from at least three replicate wells.

ChIRP assays. ChIRP assays were carried out as described previously[69]. Briefly, antisense DNA tiling probes targeting PCAT1 were designed using the free online Biosearch Technologies Stellaris FISH Probe Designer. Twelve PCAT1 probes with a 3′ biotinTEG modification were purchased from Sigma-Aldrich. Probe sequences are listed in **Supplementary Table 4**. For one ChIRP assay, 40 million LNCaP cells were collected and then cross-linked by 1% glutaraldehyde for 10 min at room temperature. After washing by PBS, lysis buffer (50 mM Tris-HCl, pH 7.0, 10 mM EDTA and 1% SDS) with fresh protease inhibitor, PMSF and RNase inhibitor was added to resuspend the cell pellet. Cell lysates were sonicated using a Bioruptor sonicator (B01060001, Diagenode) with 30 s on/45 s off pulse intervals until lysates turned clear. RNA input and DNA input were taken from the sonicated samples before hybridization. Two volumes of hybridization buffer and 100 nM probes were added to cell lysates. Hybridization was carried out at 37 °C for 4 h with rotation. After hybridization, 100 µl of Streptavidin Magnetic C1 beads (65001, Thermo

Fisher) was added to each hybridization reaction, and the tubes were incubated at 37 °C for 30 min with shaking. Beads–probe–RNA complexes were captured by DynaMag-15 magnetic strip and washed five times with 1 ml of wash buffer at 37 °C. After washing, 10% of the beads complex was used for RNA isolation and 90% of the beads complex was used for DNA purification.

Immunoprecipitation and immunoblotting. Forty microliters of 1:1 mixture of protein A (88845, Thermo Fisher) and protein G (88847, Thermo Fisher) Dynabeads was washed with RIPA buffer (50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 1 mM EDTA, 1% CA-630, 0.5% sodium deoxycholate and protease inhibitor cocktail) and then incubated together with antibody to AR (sc-13062, Santa Cruz Biotechnology) or ONECUT2 (ab181229, Abcam) at 4 °C overnight. LNCaP cells were lysed using RIPA buffer. Cell lysates were added to bead–antibody complex and incubated with rotation at 4 °C overnight. Precipitated proteins were eluted from beads by adding 2× SDS–PAGE loading buffer and boiled for 10 min. The immunoprecipitated proteins were then subjected to immunoblotting. Protein samples from immunoprecipitation were separated by SDS–PAGE (12%) and transferred onto PVDF membranes. After incubation with primary antibodies, membranes were incubated with horseradish peroxidase (HRP)-conjugated protein G (ab7460, Abcam) to prevent interference of IgG heavy chain. The immunoreactive proteins were visualized by ECL substrate (34096, Thermo Fisher). The antibodies used for immunoblot analysis were to AR (sc-13062, Santa Cruz Biotechnology; 1:1,000 dilution) and histone H3 (ab1791, Abcam; 1:2,000 dilution).

Mouse xenograft studies with LNCaP C4-2 prostate adenocarcinoma cells. Tumor xenografts, generated using LNCaP C4-2 cells obtained from ATCC, were conducted under arrangement with Charles River Laboratories. The justification for the use of human tumor cell xenografts in immune-deficient mice was to determine the growth of human prostate adenocarcinoma (LNCaP C4-2) cells under a biologically relevant tissue microenvironment to assess native tumor cell growth. In brief, mouse xenograft transplants were established directly from LNCaP C4-2 cells manipulated by depleting *PCAT1* transcripts using locked nucleic RNAs (purchased through Exiqon). Briefly, 2 × 10^6 LNCaP C4-2 cells were recovered, placed in 1× PBS and inoculated by subcutaneous engraftment in the hind rump of male immunodeficient NOD-SCID mice (Charles River Laboratories) from 8–10 weeks of age. Monitoring of tumor growth was performed at least twice per week using calipers. The growth of the subcutaneous tumors was followed by means of caliper measurements and specific volume. Because of the predicted variation in tumor cell growth with LNCaP C4-2 cells, it was necessary to use a sample size of 12 mice per experimental group to achieve statistical significance. Animal survival was determined on the basis of tumor sizes reaching maximal volumes allowable (1,500 mm^3) under the Icahn School of Medicine at Mount Sinai Institutional Animal Care and Use Committee (IACUC). All cells planned for inoculation into mice require the authentication of being free of all mycoplasma contamination. Typically, qPCR profiling of mycoplasma contamination was performed as a service by the Center for Comparative Medicine and Surgery (CCMS) before subcutaneous inoculation. All experimental procedures were approved by the Icahn School of Medicine at Mount Sinai's IACUC under protocol LA09-00445 following strict adherence to the guidelines approved for animal use by the NIH.

Processing prostate cancer risk SNPs and variant set enrichment analyses. The 122 prostate cancer risk-associated tag SNPs were collected from GWAS analysis studies[11,19,58]. The SNPs in LD with tag SNPs were identified using PLINK v1.07 software (see URLs)[70]. The LD files for the 1000 Genome Project (Phase I release 2010-11-23) were obtained from the complete package of the stand-alone tool LocusZoom[71].

VSE is a computational method to compute the enrichment or depletion of GWAS-derived genetic predispositions for a disease across defined genomic regions (available as an R package from the CRAN repository)[18]. The exon regions for all protein-coding genes and lncRNAs were extracted from GENCODE v19 annotation. The overlapping exon regions for each gene type were merged to avoid overestimation. LNCaP DHSs were obtained from ENCODE and expanded to 500 bp. The binding sites for various transcription factors were obtained from different sources and processed uniformly using MACS[72]. The genomic distribution of the SNPs was visualized using Circos[73].

lncRNA expression analysis. To identify expressed lncRNAs in prostate tumors, we obtained mapped RNA-seq data for primary tumor samples and normal prostate samples from TCGA. The expression data for 14 tumor–normal sample pairs in the Asian population were obtained from Ren *et al.*[59] and mapped against hg19 using TopHat2. The mapped data were quantified for gene expression for GENCODE v19 annotation using the module featurecounts in the R package Rsubread[74]. The raw counts for each sample were normalized, and the expression level and differential expression were determined using the R package EdgeR[75]. The selection of expressed genes and the frequency of their expression were estimated using custom Perl scripts.

***Cis*-eQTLs.** We obtained matched RNA-seq, CNV, methylation and SNP genotyping data for prostate adenocarcinoma (PRAD) tumor samples from the TCGA Data Portal. Normalization of the expression levels for all genes was calculated by adapting the strategy described by Li *et al.*[32] using a custom Perl script. The expression data for each gene were transformed to be normally distributed while preserving the relative rankings. This approach is similar to the approach taken by GTEx and proposed by MatrixEQTL[76]. For the genotyping data, we filtered out SNPs that overlapped with our risk SNPs, and we removed outlier SNPs that had a minor allele frequency <0.05. We first performed *cis*-eQTL analysis with all expressed protein-coding genes and lncRNAs (19,866) using the R package MatrixEQTL[76]. The FDR for each association test between SNPs and TSSs within 500 kb was calculated using a total locus–gene association test within the distance.

Intercellular functional correlation. The 500-bp LNCaP DHSs (the same as those used for VSE analysis) overlapping at least one risk SNP were obtained using the intersect command of the BEDTools software suite (version 2.23.0)[77]. These DHSs distal from promoter regions are referred to as 'risk enhancers'. Promoter regions were defined by a region of 500 bp upstream and downstream of the start site of each transcript for each gene in GENCODE v19 annotation. The Pearson's correlation coefficient between the average DHS signal intensity of each risk enhancer and nearby lncRNA gene promoters was computed using a custom Perl script based on the principle first described by Thurman *et al.*[34]. In short, the principle describes that any two genomic regions with highly correlated DNase I signals across different cell lines are likely to be physically associated. The DNase-seq signal tracks for hg19 in bigwig format for 77 cell lines were downloaded from the UCSC ENCODE download page (see URLs)[78]. These 77 cell lines are among the 79 that were used by Thurman *et al.* and were grouped into 32 clusters as described in the original article[34]. A cutoff of 0.7 was used for Pearson's correlation coefficient.

Comparative enhancer analysis. An in-house Perl program computed the maximum DHS signal in the available DNase-seq data for the genomic ranges provided and detected which DHSs had significantly higher signal in a particular cell line. Uniform DNase-seq signals were obtained from the UCSC ENCODE download page (see URLs). The genomic ranges provided were the 500-bp DHSs for LNCaP cells that overlapped with at least one risk SNP, that is, risk enhancers, as used throughout the study. Enrichment was calculated by two-tailed one-sample *t* test. The RNA-seq data for the available matching cells (H1, MCF-7, A549, GM12878, HepG2, HUVEC, NHEK and K562) were obtained from the UCSC ENCODE download page in fastq format. The RNA-seq data for LNCaP cells were generated in our laboratory. The fastq files were mapped against human genome assembly hg19 using TopHat2 and annotated against GENCODE v19 using the module featurecounts in R package Rsubread[70]. Expression was calculated in RPKM.

Prioritization of susceptible lncRNA candidates. The candidate lncRNAs identified as associated with susceptibility to prostate cancer were ranked to nominate the best hit for validation and functional analyses. We developed a weighted scoring system that combined binary state for presence or absence of each factor in consideration: risk SNP in promoter, positive enhancer–promoter looping, eQTL with one or more risk SNPs and differential expression. The algorithm computes a composite score for each lncRNA based on the state

doi:10.1038/ng.3637

of each factor, while adding 1.5 times more weight to the promoter and reducing the weight of differential expression by 0.5 times owing to their different potentials in detecting risk-associated lncRNAs. The algorithm also takes the distance between a candidate lncRNA and a risk SNP and the expression status of the lncRNA into consideration by filtering out weakly expressed lncRNAs and lncRNAs that were more than 500 kb from the associated risk SNP locus.

Topologically associated domain analysis. TAD and hESC CTCF looping data were obtained from Ji *et al.*[79]. Raw CTCF ChIP-seq data were obtained from ENCODE[78]. CTCF looping data for MCF-7 and K562 cells were obtained from ENCODE ChIA-PET data.

Survival analysis. Analysis was performed in a cohort of patients with prostate cancer ($n = 127$), each of which had complete survival information available, matched tumor and normal (blood) whole-genome sequencing profiles from frozen specimens and analyzed as part of the Canadian Prostate Cancer Genome Network (CPC-GENE). These patients were used to evaluate the association between the composite SNP status and biochemical relapse (BCR; a proxy for prostate-cancer-specific survival). A Cox proportional hazard regression model was fit for the composite SNP group, which is defined as patients with the AA genotype for rs7463708, AA genotype for rs72725879 and BB genotype for rs1456315. The R package survival (v2.38-3) was used to fit the model and check for the proportional hazard assumption. The prognostic value of the composite SNP group was visualized using Kaplan–Meier curves implemented in the lattice (v0.20-33) and latticeExtra (v0.6-26) packages for R.

Motif analysis. The effect of rs7463708 on transcription factor binding motifs was analyzed using HaploReg v4.1 (see URLs)[80] coupled with our in-house algorithm to calculate statistical significance. The rs7463708 SNP and its flanking sequence overlap with six motifs, but only the ONECUT motif was significantly affected by rs7463708.

Code availability. VSE has been developed as an R package and is publically available (see URLs).

60. Yu, J. *et al.* An integrated network of androgen receptor, Polycomb, and *TMPRSS2–ERG* gene fusions in prostate cancer progression. *Cancer Cell* **17**, 443–454 (2010).
61. Bert, S.A. *et al.* Regional activation of the cancer genome by long-range epigenetic remodeling. *Cancer Cell* **23**, 9–22 (2013).
62. Basnet, H. *et al.* Tyrosine phosphorylation of histone H2A by CK2 regulates transcriptional elongation. *Nature* **516**, 267–271 (2014).
63. Xu, K. *et al.* EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent. *Science* **338**, 1465–1469 (2012).
64. Chen, Y. *et al.* ETS factors reprogram the androgen receptor cistrome and prime prostate tumorigenesis in response to PTEN loss. *Nat. Med.* **19**, 1023–1029 (2013).
65. Tan, P.Y. *et al.* Integration of regulatory networks by NKX3-1 promotes androgen-dependent prostate cancer survival. *Mol. Cell. Biol.* **32**, 399–414 (2012).
66. He, H.H. *et al.* Nucleosome dynamics define transcriptional enhancers. *Nat. Genet.* **42**, 343–347 (2010).
67. Pomerantz, M.M. *et al.* The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nat. Genet.* **47**, 1346–1351 (2015).
68. Guo, H. *et al.* Chemokine receptor CXCR2 is transactivated by p53 and induces p38-mediated cellular senescence in response to DNA damage. *Aging Cell* **12**, 1110–1121 (2013).
69. Chu, C., Qu, K., Zhong, F.L., Artandi, S.E. & Chang, H.Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA–chromatin interactions. *Mol. Cell* **44**, 667–678 (2011).
70. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
71. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
72. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
73. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
74. Liao, Y., Smyth, G.K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**, e108 (2013).
75. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
76. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
77. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
78. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
79. Ji, X. *et al.* 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell* **18**, 262–275 (2016).
80. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).