

# H3k27ac-HiChIP in prostate cell lines identifies risk genes for prostate cancer susceptibility

Claudia Giambartolomei<sup>\*1,2</sup>, Ji-Heui Seo<sup>\*3,4</sup>, Tommer Schwarz<sup>6</sup>, Malika Kumar Freund<sup>5</sup>, Ruth Dolly Johnson<sup>7</sup>, Sandor Spisak<sup>3</sup>, Sylvan C. Baca<sup>3</sup>, Alexander Gusev<sup>3</sup>, Nicholas Mancuso<sup>8</sup>, Bogdan Pasaniuc<sup>\*2,5,6,9,10</sup>, Matthew L. Freedman<sup>\*3,4</sup>

\* These authors contributed equally.

1. Central RNA Lab, Istituto Italiano di Tecnologia, Genova, 16163, Italy
2. Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA
3. Department of Medical Oncology, The Center for Functional Cancer Epigenetics, Dana Farber Cancer Institute, Boston, MA 02215, USA
4. The Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Boston, MA 02215, USA
5. Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA
6. Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles, CA, USA
7. Department of Computer Science, University of California Los Angeles, Los Angeles, CA, USA
8. Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90032, USA
9. Department of Computational Medicine, UCLA
10. Johnson Comprehensive Cancer Institute at UCLA

## Abstract

Genome-wide association studies (GWAS) have identified more than 140 prostate cancer (PrCa) risk regions which provide potential insights into causal mechanisms. Multiple lines of evidence show that a significant proportion of PrCa risk can be explained by germline causal variants that dysregulate nearby target genes in prostate-relevant tissues thus altering disease risk. The traditional approach to explore this hypothesis has been correlating GWAS variants with steady-state transcript levels, referred to as expression quantitative trait loci (eQTLs). In this work, we assess the utility of chromosome conformation capture (3C) coupled with immunoprecipitation (HiChIP) to identify target genes for PrCa GWAS risk loci. We find that interactome data confirms previously reported PrCa target genes identified through GWAS/eQTL overlap (e.g., *MLPH*). Interestingly, HiChIP identified links between PrCa GWAS variants and genes well-known to play a role in prostate cancer biology (e.g., *AR*) that are not detected by eQTL-based methods. We validate these findings through CRISPR interference (CRISPRi) perturbation of the variant-containing regulatory elements for *NKX3-1* and *AR* in the LNCaP cell line. Our results demonstrate that looping data harbor additional information beyond eQTLs and expand the number of PrCa GWAS loci that can be linked to candidate susceptibility genes.

## Introduction

Prostate cancer (PrCa) is the most common cancer in men, the second leading cause of cancer deaths worldwide, and has a strong familial component<sup>1–4</sup>. Genome wide association studies (GWAS) have identified single nucleotide polymorphisms (SNPs) associated with increased risk of PrCa across 147 predominantly non-coding genomic loci<sup>5</sup>, collectively explaining 28.4% of familial risk<sup>6–9</sup>. With the vast majority of risk variants discovered by PrCa GWAS located outside protein-coding regions, the molecular mechanisms driving pathogenesis have only been

described for a handful of loci<sup>10–15</sup>. GWAS loci predominantly colocalize with tissue-specific regulatory elements thus supporting the hypothesis that risk variants exert their effects on disease by influencing the transcriptional levels of their target genes. For example, a substantial proportion of PrCa heritability lies in regions marked by H3K27ac<sup>16</sup>, a histone modification marking active enhancers and promoters<sup>17–19</sup>.

A standard approach to link variants to genes is through expression quantitative trait loci (eQTL) analysis, which identifies genotypes that correlate with transcript levels across individuals in a target tissue of interest<sup>20–22</sup>. Overlapping eQTLs with variants identified by GWAS is a powerful tool to prioritize target genes for further functional investigation<sup>20,21,23–28</sup>. However, eQTLs have several limitations. They are primarily studied under steady-state conditions, therefore eQTLs that are elicited under specific conditions are missed. For example, context-specific eQTLs have been identified after stimulation with interferon- $\gamma$  and bacterial lipopolysaccharides in monocytes<sup>29</sup>. Second, statistical power for eQTL identification is dependent on sample and effect size with many eQTLs being undetected (i.e. a false negative), particularly for hard-to-collect tissues and/or eQTLs in a rare cell type. Despite some PrCa GWAS loci mapping near well-known prostate cancer biology genes, such as *FOXA1*, *GATA2*, *AR*, and *MYC*, the absence of strong eQTL associations with these transcripts is notable. This suggests that traditional eQTL/GWAS overlap may fail to detect important susceptibility genes for PrCa.

Chromosome conformation capture (3C)-based assays have recently emerged as powerful tools to evaluate physical interactions between regulatory elements and target genes in the context of GWAS, including prostate cancer risk<sup>30–32</sup> and can be used as an orthogonal method to validate eQTL-based findings. Efficiencies of 3C-based techniques can be increased by enriching for interactions via chromatin immunoprecipitation against proteins, such as H3K27ac, a marker of active promoter/enhancer activity (HiChIP)<sup>33–35</sup> or through nucleic acid hybridization<sup>36</sup>.

Notably, across a number of models, enhancer-promoter contacts (loops) pre-exist at stimulation-responsive genes prior to the stimulus<sup>37-39</sup>. In other words, stimulation does not induce significant *de novo* looping even at genes that are transcriptionally responsive in differentiated cells. These observations raise the hypothesis that looping (measured even in the steady-state) can identify GWAS target genes beyond traditional GWAS/eQTL overlap; that is, looping could reveal eQTL-target gene relationships that are observable only under certain contexts (e.g., cell-type) and/or capture eQTLs with small, but important, effects.

We utilized high-resolution H3K27ac-HiChIP data in the LNCaP prostate cancer cell line to identify links between target genes and PrCa risk loci. We linked 99 out of 130 known susceptibility loci for PrCa with 665 genes and observed a significant overlap between eQTL-target gene pairs and loops. Notably, we identified looping between candidate PrCa causal variants from GWAS and genes with established roles in PrCa biology at eQTL-negative loci. We used CRISPR interference (CRISPRi) to functionally validate the enhancer-promoter interaction for *NKX3-1* and *AR*. Overall, our results confirm that 3C-based strategies can not only validate eQTLs, but can also discover important target gene links not detected by current GWAS/eQTL strategies.

## Results

### A high-resolution H3K27ac-HiChIP chromatin contact map for LNCaP

We performed H3K27ac HiChIP in LNCaP across 5 biological replicates and identified 126,280 loops (FitHiChIP, FDR<0.01, **Methods** and **Figure 1**). We called between 4,000 and 14,000 loops from a read depth ranging from 183,000 reads to 235,000 reads across individual replicates (**Table S1**). We observed that loop count and loop length depended on read depth. Replicates 1 and 5 had the highest number of high-quality uniquely mapped read pairs and final number of loops, and therefore the highest median loop length and paired end tag (PET) counts

**(Figures S1a and S1b).** The intersection of significantly called loops across the 5 replicates ranged from 20% to 60%. Once the individual replicates data was merged, each replicate shared 90% or more significantly called loops with the merged data (**Table S2**). Pairwise correlations of PET counts supporting each loop were significant across all replicates ( $p > 0.7$ ,  $p$ -value  $< 0.001$ , **Figure 2**). Merging data across all replicates substantially increased the number of significant loops to 126,280. The mean (median) loop length was 173kb (95kb), and the mean (median) number of paired-end tags (PETs) per loop was 46 (19) (**Table S1**). As expected, the number of loops decreased with increasing loop distance (**Figure S1b**). 17,690 genes (out of 27,063 RefSeq genes considered) had promoters overlapping at last one loop; each gene promoter overlaps a mean (median) of 7.9(6) loops per gene, with a mean (median) of 351 (174) PETs per gene. Gene connectivity, defined as the number of loops per gene promoter (or as total number of PETs per gene), is moderately correlated with gene expression activity (as assayed by RNA-seq in the same cell line, **Methods**) (Spearman  $\rho = 0.489$ ;  $p$ -value  $< 2.2e-16$ ) (**Figures S3a and S3b**).

Prior studies have demonstrated that looping interactions do not qualitatively change in response to a perturbation<sup>37-39</sup>. To assess whether loop formation correlates with differential expression across perturbations and states, we evaluated the looping status of genes in androgen stimulated cells and during tumorigenesis (**Methods**). We observed that genes that are destined to change under these conditions have pre-existing loops (e.g., upon androgen stimulation, genes with  $>1.5$ -fold changes in transcript levels after 4h have 10 loops on average whereas the average number of loops of any gene in LNCaP is 9, see **Table S3**; **Table S4**). This observation suggests that *de novo* loop formation is not a major mechanism underlying gene expression changes in this data.

## **H3K27ac-HiChIP loops are enriched for eQTLs in prostate tissue**

To evaluate if chromatin interactions are enriched in prostate eQTLs, we assessed the overlap of loops with eQTLs measured in two prostate datasets: Thibodeau eQTLs in prostate normal tissue (n=471)<sup>40</sup>, and TCGA eQTLs in prostate tumor tissue (n=378)<sup>41</sup>. Local eQTLs (within 3Mb of every gene) were called using standard approaches (**Methods**). First, we found that the top eQTL variant for each eGene is 2 times more likely to fall within a HiChIP anchor than a random SNP in a 3Mb region centered on the eGene (empirical p-value 0.0099, **Figure S4a**). Second, we found that loops with eQTLs in one anchor were more likely to loop to the promoter of their eGene than expected by chance (1.2-fold change compared to random, empirical p-value = 0.0099, **Methods**), across a range of distances from their target promoters (**Figures S4b** and **S4c**).

## H3k27ac-HiChIP loops link prostate cancer risk variants to target genes

Next, we quantified the enrichment of GWAS signals at the anchor regions of the loops using stratified linkage disequilibrium score (LDSC) regression analysis<sup>42</sup> integrating the H3K27ac-HiChIP loops with the largest PrCa GWAS to date<sup>5</sup>. Variants residing in loop anchors show high enrichment of prostate cancer heritability as compared to random variants in the genome overall (s-LDSC enrichment of 2.61, p=1.06E-09), and the enrichment is even greater when the variants are also falling within H3K27ac peaks (our calling allows for one anchor not to be acetylated thus yielding many loops with one non-acetylated anchor, see **Methods**) (s-LDSC enrichment of 14.06, p = 2.54e-5 for all loop types, **Table S5**). This confirms that H3K27ac-HiChIP looping localizes relevant PrCa GWAS signal. Interestingly, we observed a higher enrichment when restricting loops to specific categories (e.g., 18.72 for enhancer-promoter, p=8.93e-6, **Table S5**).

Next, we performed probabilistic fine-mapping of the prostate cancer GWAS at 130 of the 147 previously reported risk loci except 8q24 for which exhaustive fine-mapping is available<sup>11</sup>. We integrated 1,953 HiChIP loops with at least one anchor overlapping a PrCa credible causal variant in 104 risk loci regions. Overall, 2,016 PrCa credible causal variants linked to 665 genes across these 104 loci (**Figure 3, Table S6**). 48 (out of 104) regions have 3 or fewer HiChIP genes overlapping credible GWAS SNPs (**Table S7**). Of the 665 genes, 37 are eGenes in prostate tissue and have evidence of GWAS/eQTL overlap with PrCa risk variants either through TWAS and/or co-localization (**Figure 3, Table 1, Table S8**). One example is *MLPH* (**Figure 4a**), which is an eGene in prostate tissue and has been previously reported in normal prostate tissue from GTEx<sup>43</sup>. By contrast, 4 genes (*CEACAM21*, *MOB2*, *ASCL2*, *GDF7*) are eGenes and show evidence of TWAS/Colocalization but are not supported by any HiChIP loop (**Table S8**).

## Looping maps PrCa GWAS variants to known PrCa biology genes

We noted that many genes previously implicated in PrCa biology, such as *GATA2*, *AR*, *MYC*, and *NKX3-1* are nearby PrCa GWAS risk loci. Despite the clear role of these genes in PrCa, compelling evidence linking risk variants to these genes through expression-based methods is currently lacking. Interestingly, looping provides links from promoters of these genes to GWAS risk regions. We highlight three such genes *MYC*, *AR* and *NKX3-1*.

*MYC* has an uncontested role in cancer biology and has been associated with numerous cancer types through GWAS<sup>44-46</sup>, *MYC* has never been reported as an eGene. A study from Matejcic et al.<sup>11</sup> fine-mapped the prostate cancer susceptibility region at 8q24 where *MYC* is located and observed 174 variants in the 95% credible set. 169 of these candidate causal variants overlap one of 225 HiChIP loops, and the majority (152/169) link to the *MYC* promoter (**Figure 4b, Table S9**).

*AR* is critical for prostate cancer as PrCa is dependent on the actions of androgens and therefore on the function of the *AR* gene. More than half of primary tumors and almost all tumor metastases are associated with overexpressed or deregulated *AR*<sup>47,48</sup>, and mutations in the *AR* gene have been associated particularly with tumor progression<sup>49–51</sup>. We identified H3K27Ac HiChIP loops that link PrCa credible SNPs located in the X chromosome with the *AR* promoter, potentially pointing to enhancer regions important for PrCa (**Figure 4c**). Moreover, these variants have never been implicated as eQTLs.

*NKX3-1* is another interesting example since it is one of the most androgen responsive genes in the LNCaP cell line, it is involved in prostate development, and is recurrently mutated in advanced PrCa<sup>52–55</sup>. We measured looping before and after androgen stimulation in the LNCaP cell line and we observed that the number of *NKX3-1* promoter loops is largely unchanged (**Methods, Table S4 and Figure 4d**). For example, the expression changes by ~2 fold after 4(16) hours of androgen stimulation ( $\log FC = 2.23(1.88)$  p-value < 1E-50), while the total number of loops change minimally (from 14 loops to 11 and to 13, respectively for the two time points, **Table S4**). These data indicate that transcriptionally dynamic genes, which may represent context-dependent eQTL targets are discoverable through looping.

## Looping identifies germline-somatic interactions

We next investigated germline-somatic interactions by evaluating whether genes known to be somatically mutated in PrCa oncogenesis also show evidence of looping to germline PrCa GWAS. We identified a set of 119 prostate cancer–genes curated from large-scale PrCa studies that show evidence of somatically acquired mutations<sup>56–58</sup> (**Methods**). Interestingly these genes are on average closer to a PrCa credible causal variant when compared to other genes within a 3Mb region, providing additional evidence of their importance to PrCa (**Figure S5**, 30% of PrCa genes compared to 8% of all genes are within 100Kb of PrCa credible causal variant). Eighteen (out of 104) genes have HiChIP loops linking a credible causal variant to the promoter of the

gene (**Table 1**). Strikingly, none show significant colocalization GWAS/eQTL, with only 2 genes being eGenes in existing prostate transcriptomic data (**Figure 3**).

## Validation of enhancer-promoter contacts using CRISPRi

Next, we validated candidate HiChIP SNP-gene targets at PrCa two GWAS loci using CRISPRi in LNCaP cells. We selected two GWAS loci where anchors containing a candidate causal variant looped to genes that play clear roles in PrCa biology, *NKX3-1* and *AR* (**Figure 5**). As described above, these two loci have not been strongly implicated through eQTL- or COLOC/TWAS-based analyses. Putative regulatory enhancer regions containing candidate causal variants within a HiChIP anchor and overlapping DNase hypersensitivity sites (DHS) were identified and targeted. Guide RNAs (gRNAs) were designed and tested against distinct DHS peaks to evaluate the role of these enhancers on *NKX3-1* and *AR* expression (**Table S10**). Suppression of these enhancer regions containing candidate causal variants significantly reduced RNA levels in the target genes (**Figure 5, Table S11**).

## Discussion

A central issue driving post-GWAS studies is a mechanistic understanding of non-protein coding risk loci, which account for over 90% of GWAS variants. In this work, we outlined a systematic approach, based on chromosome conformation capture technology, to link GWAS PrCa risk SNPs to candidate target genes. We used H3K27Ac HiChIP methodology as a means to assay genome-wide enhancer-promoter chromatin interactions. 3C-based methods measure physical interactions and, thus complement other approaches, such as eQTLs, which are based on association between genotypes and transcript levels. eQTL studies can be confounded by LD and are dependent on sample size whereby interactome maps do not suffer from these limitations.

An additional limitation of large-scale expression-based studies is that they are based on steady-state transcript levels. By contrast, studies have demonstrated that looping is less dynamic in response to defined perturbations<sup>37,39,59</sup>. Stated another way, looping identifies the *potential* of an enhancer-promoter interaction to be active and is less informative as a quantitative readout of transcriptional levels of genes. This observation raises the provocative notion that looping can identify latent, stimulus- and context-dependent eQTLs and highlight important candidate genes without requiring experiments that directly measure these other conditions. This rationale is consistent with recent reports showing that steady-state eQTLs are insufficient to explain the majority of disease heritability<sup>60</sup>. Indeed, our results showed that important PrCa biology genes interact with risk loci that previously escaped detection through expression-based methods.

Other techniques can be used to assay enhancer-promoter interactions. Traditional Hi-C can assay all loops; however Hi-C can lack resolution in mapping the enhancer-promoter link if not sequenced to extremely high depths. Through enrichment based on immunoprecipitation of a target protein, H3K27Ac ChIP is an efficient way to detect enhancer-promoter loops and has been shown to identify a similar number of loops with ten-fold less sequencing compared to Hi-C<sup>61</sup>. ChIA-PET is another whole genome 3C-based assay able to detect protein-centric long range contacts, however this method still requires hundreds of millions of cells per experiment and is less efficient than ChIP<sup>61</sup>.

As with any method, there are limitations of ChIP. First, all 3C-based methods have limited power to confidently detect nearby genes. Second, the computational pipelines to analyze ChIP data are still in their infancy and future developments could affect results. ChIP loop-calling is dependent on H3K27ac ChIP-seq peak-calling used for loop anchors. In this study, we defined promoters based on the longest transcript from TSS from RefSeq, which reports the most representative initiation site across different cell types. Using this definition, novel genes

with as-yet unannotated start sites are missed. Furthermore, a 5-kb resolution of the HiChIP data analysis limits our definition of anchors, and there is the risk that the additional 5-kb padding added to anchors could result in decreased resolution of enhancer-promoter links. In this work we focused on H3K27ac histone modification. In future studies, assays using other modifications (e.g., H3K4me1, H3K4me2) could be used in combination to better refine enhancer and promoter regions <sup>62</sup>.

We demonstrate the benefit of using HiChIP to both validate eQTLs as well as prioritize genes for PrCa that are missed by eQTL-based methodologies. Moving forward, we propose to utilize the complementary techniques of eQTL-based methodologies and HiChIP to prioritize genes at GWAS loci. This work also creates an opportunity to create a unified model combining expression and interaction information to extract the strengths of both methods.

## Methods

**H3K27ac ChIP-Seq in LNCaP.** H3K27Ac ChIP using either LNCaP (CRL1740, ATCC) was performed as previously described <sup>51</sup>. 10 million cells were fixed with 1% formaldehyde at room temperature for 10 minutes and quenched. Cells were collected in lysis buffer (1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS and protease inhibitor (#11873580001, Roche) in PBS) <sup>63</sup>. Chromatin was sonicated to 300-800 bp using Covaris E220 sonicator (140PIP, 5% duty cycle, 200 cycle burst). H3K27ac Antibody (C15410196, Diagenode, 1:600 ratio) were incubated with 40 µl of Dynabeads protein A/G (Invitrogen) for at least 6 hours before immunoprecipitation with the sonicated chromatin overnight. Chromatin was washed with LiCl wash buffer (100 mM Tris pH 7.5, 500 mM LiCl, 1% NP-40, 1% sodium deoxycholate) 6 times for 10 minutes. Eluted sample DNA was prepared as the sequencing libraries using the ThruPLEX-FD Prep Kit (Rubicon Genomics). Libraries were sequenced using 150-base pair single reads on the Illumina platform (Illumina) at Novogene. For further analyses, we used the union of ChIPSeq narrow and broad peaks in regular media. This comprises 49,638 ChIPSeq peaks in LNCaP cells (length of peaks ranged from 146 to 129,126 bases). 43,335 out of 49,638 peaks overlap HiChIP anchors.

**Peak calling.** ChIP-seq and ATAC-seq were processed through the ChiLin pipeline <sup>64</sup>. Briefly, Illumina Casava1.7 software used for base calling and raw sequence quality and GC content were checked using FastQC (0.10.1). The Burrows–Wheeler Aligner (BWA, version 0.7.10) was used to align the reads to human genome hg19. Then, MACS2 (v. 2.1.0.20140616) was used for peak calling with a FDR q-value threshold of 0.01. Bed files and Bigwig files were generated using bedGraphToBigWig for H3K27Ac. The union of H3K27Ac narrow and broad peaks was used in the downstream analyses. The following quality metrics were assessed for each sample: 1. Percentage of uniquely mapped reads , 2. PCR bottleneck coefficient to identify potential over amplification by PCR, 3. FRIP (fraction of non-mitochondrial reads in peak

regions), 4. Peak Number, 5. Number of peaks with 10-fold and 20-fold enrichment over background, 6. Fragment size, 7. Percentage of the merged peaks with promoter, enhancer, intron, or intergenic, and 8. Peak overlap with DNaseI hypersensitivity sites. For datasets with replicates, ChiLin calculates the replicate consistency with two metrics: 1. Pearson correlation of ChIP-seq reads across the genome using UCSC software wigCorrelate after normalizing signal to reads per million and 2. percentage of overlapping peaks in the ChIP replicates.

**RNA-Seq in LNCaP.** 1 million of LNCaP cells were harvested for RNA seq. For the DHT treatment samples, LNCaP cells were grown in hormone depleted media for 3 days then treated with 10nM DHT for either 4h or 16h. Total RNA was collected from the cells using RNeasy kit (74104, Qiagen) with DNase I treatment (79254, Qiagen). The library preparations, quality control and sequencing on a HiSeq platforms with paired-end 150 bp (PE 150) were performed by Novogene. RNA-seq data were processed using the VIPER pipeline<sup>65</sup>. Reads were aligned to the hg19 human genome build with STAR<sup>66</sup>. FPKM values were calculated with Cufflinks<sup>67</sup> for 20,114 RefSEQ genes included in the VIPER repository. Differential expression (DE) analyses was performed with the DESeq2 R package<sup>68</sup>, using supervised analysis based on gene expression levels (counts from STAR) and cutoffs of FDR-adjusted p-value (padj) < 0.05 and log2 fold-change > 1. DHT treatment of LNCaP for 4 or 16 hours was compared to vehicle treatment (two replicates each). We called loops at FDR 1%, and annotated genes to loops, following the procedure described below. The total number of loops called in 4hrs and 16 hrs were 98,960 and 183,958, respectively. The total number of genes annotated in 4hrs and 16 hrs HiChIP loop data were 17,649 and 20,899, respectively. We considered only the genes in common between LNCaP and 4hr sample (Ngenes = 15630), and between LNCaP and 16hr sample (Ngenes = 16979).

**H3k27ac HiChIP in LNCaP.** HiChIP was performed mainly following the procedure described in<sup>61</sup>, trypsinized 10 million LNCaP cells were fixed with 1% formaldehyde at room temperature for

10 minutes and quenched. Sample was lysed in HiChIP lysis buffer and digested with MboI (NEB) for 4 hrs. After 1hr of biotin incorporation with biotin dATP, the sample was ligated using T4 DNA ligase for 4 hrs and, chipped with H3K27ac antibody (DiAGenode, C1541019) after chromatin. Reverse crossed IP sample was pulled down with streptavidin C1 beads (life tech) treated with Transposase (illumina) and was amplified with reasonable cycle numbers based on the qPCR using 5cycle-preamplified library. Library was submitted sequenced using 150-base pair end reads on the Illumina platform (Novogene). All of these libraries were generated using the 4-bp cutter MboI restriction fragment. The raw fastq files (paired-end data) were first trimmed to remove adaptor sequences using Trim Galore<sup>69</sup>. HiC-Pro version 2.9.0<sup>70</sup> was used to align the reads to the hg19 human genome, assign reads to MboI restriction fragments, remove duplicate reads. We used the following options: MIN\_MAPQ = 20, BOWTIE2\_GLOBAL\_OPTIONS = --very-sensitive --end-to-end --reorder, BOWTIE2\_LOCAL\_OPTIONS = --very-sensitive --end-to-end --reorder, GENOME\_FRAGMENT = MboI\_resfrag\_hg19.bed, LIGATION\_SITE = GATCGATC, LIGATION\_SITE = 'GATCGATC', BIN\_SIZE = '5000'. All other default settings were used. The HiC-Pro pipeline selects only uniquely mapped *valid* read pairs involving two different restriction fragments to build the contact maps.

We applied FitHiChIP version 5.1<sup>71</sup> for bias-corrected peak calling and DNA loop calling. FitHiChIP models the genomic distance effect using a spline fit, normalizes for coverage differences using regression, and computes statistical significance estimates for each pair of loci. The FitHiChIP loop significance model is used to determine whether interactions are significantly stronger than the random background interaction frequency. We used 49,638 regions from H3K27Ac LNCaP union of narrow and broad peaks as anchors to call loops. We used a 5Kb resolution and considered only interactions between 5kb-3Mb. We used the peak to all for the foreground, meaning at least one anchor to be in the H3K27 peak but it doesn't have

to be both. The corresponding FitHiChiP options specification is “IntType=3”. For the global background estimation of expected counts (and contact probabilities for each genomic distance), FitHiChiP can use either peak-to-peak (stringent) or peak-to-all loops (loose) loops for learning the background and spline fitting. We specified the suggested option to merge interactions close to each other to represent a single interaction when their originating bins are closer. The corresponding FitHiChiP options specifications are “UseP2PBackgrnd=0” and “MergeInt=1” (FitHiChiP(L+M)). We used the default FitHiChiP q-value < 0.01 to identify significant loops. For comparisons across replicates, we used the results not merged MergeInt=0, as suggested by the authors. The length considered was between 5kb and 3Mb. We explored reproducibility across replicates in the following way. We have processed 5 biological replicates separately using the FitHiChiP pipeline, as well as all replicates together in a dataset called “merged” made up of the combined reads across the replicates. We used the q-value < 0.01 cutoff to define high confidence loops, and compared the level of accuracy achieved by one replicate and the merged data versus our high confidence loops (i.e., the proportion of reference loops reported by one replicate that are captured at differing number of loop calls from other replicates, or from our combined library). The final number of significant loops (q-value < 0.01) considered in these analyses are those using background 0 and merged FitHiChiP settings.

**Mapping loops to enhancers and promoters.** For loop annotations, we first extended loop anchors by 5kb on either side. To identify potential gene targets, we defined promoter regions around the TSS (+/- 500 bases) for 27,063 genes using RefSeq hg19 ([http://genome.ucsc.edu/cgi-bin/hgTables?hgsid=694977049\\_xUU5i1QklJ50dj5miBt9wkAYuxN3&clade=mammal&org=&db=hg19&hgta\\_group=genes&hgta\\_track=knownGene&hgta\\_table=knownGene&hgta\\_regionType=genome&position=&hgta\\_outputType=selectedFields&hgta\\_outFileName=knownGene.gtf](http://genome.ucsc.edu/cgi-bin/hgTables?hgsid=694977049_xUU5i1QklJ50dj5miBt9wkAYuxN3&clade=mammal&org=&db=hg19&hgta_group=genes&hgta_track=knownGene&hgta_table=knownGene&hgta_regionType=genome&position=&hgta_outputType=selectedFields&hgta_outFileName=knownGene.gtf)); 27 genes were removed because of ambiguous positions. The longest transcript was used to

define TSS based on strand. To define enhancers, we used the subset of 49,638 regions from H3K27ac LNCaP in regular media (union of narrow and broad peaks). We then labeled the promoters and enhancer regions that overlap either right or left anchors, and considered a loop as E-E if both the anchors overlap an enhancer region; a loop as P-P if both the anchors overlap a promoter region; a loop as E-P if only one anchor overlaps a promoter and the other an enhancer region; a loop as E-O or P-O if one anchor overlaps a promoter or enhancer and the other overlaps a region without an H3K27ac or a TSS (+/- 500 bases).

**Correlation of gene expression with looping.** We considered only loops involving promoters (E-P, P-P, P-O). In cases where there were duplicated loops (supporting the same two anchors but in opposite directions), we summed the PETs read counts. For each gene, we considered two measures of gene connectivity: (i) the number of loops with one anchor overlapping the promoter (the opposing anchor can overlap a promoter, an enhancer, or neither (“Other”)), (ii) the sum of PET counts across all loops overlapping the promoter. We compute a Spearman correlation between the expression across genes with both measures of connectivity. For the expression of every 13,274 genes that have looping and expression data, we averaged the FPKM value across two LNCaP RNA-Seq replicates. FPKM was converted to TPM by dividing the each FPKM value by the sum of all FPKM values of the respective sample, and multiplying by 1e6. We divided the genes into ten expression bins of equal size. We compare this to the log of the counts of loops (**Figure S3**) and the log of the counts of PETs (**Figure S4B**). Spearman correlations were computed between average TPM and log of counts of loops or PETs. To estimate the amount that each loop/PET contributes to expression, we fitted a linear regression (with or without adjusting for H3k27ac): (1) expression (TPM) ~ loops/PETs ; (2) expression (TPM) ~ loops/PETs + H3k27ac. The H3k27ac level per gene is extracted by overlapping H3k27ac broad peaks scores with each gene promoter. If more than one peak overlaps the region, we sum across scores. We report the value of the coefficient in the model to represent

the per-loop/per-PET contribution to expression, Spearman correlation in each version of the model, and report corresponding p-values.

**Credible sets of causal variants from Prostate Cancer GWAS.** We used 147 SNPs previously reported<sup>5</sup> index SNPs to define regions for fine-mapping with GWAS summary statistics from <sup>72</sup> downloaded from [http://practical.icr.ac.uk/blog/?page\\_id=8164](http://practical.icr.ac.uk/blog/?page_id=8164) (N=79,148 cases and 61,106 controls) across 20,370,946 SNPs. SNPs were compared by chromosome and position to the 1000 Genomes phase 3, for allele and rsIDs matching. After filtering by MAF>=0.001, 15,818,179 SNPs remained, with 20,155 SNPs passing genome-wide significance (p<5e-08). A ~150kb window (adjusted manually) around the 147 index SNPs, which resulted in 137 regions after merging overlapping regions on chromosomes 4, 8, and 17 was considered. In 6 of the regions (chr18.51504059.51971031.rs8093601, chr1.9945583.10662959.rs636291, chr6.170305546.170805546.rs138004030, chr7.1694537.2194537.rs527510716, chr9.123429584.124429584.rs1571801, chr9.21791998.22291998.rs17694493), a pvalue of genome-wide significance (p<5e-08) was not reached, so we do not consider these regions further. Furthermore, in the MYC region (8q24, 127.6–129.0Mb) we used 174 SNPs included in the 95% credible set from the JAM fine-mapping (Supplementary Data 3 of Matejcic et al. <sup>11</sup> ). Therefore, the final number of PrCa risk loci considered was 130 (129 regions which we fine-mapped (described below), and one previously mapped MYC region). We used PAINTOR <sup>73</sup>, a Bayesian statistical method, with no functional annotations and specifying a maximum of 1 causal SNP, to fine-map 129 regions (excluding MYC which was previously fine-mapped). We then constructed a 95% credible set for the most likely causal variants by taking the cumulative sum of the posterior probability until a cumulative 95% posterior probability was reached. The final probable causal set contained 3,243 (3,069 from PAINTOR, 174 from JAM) SNPs across 130 PrCa risk regions. 104 PrCa

loci overlapped 1,953 LNCaP loop anchors and 665 genes, connecting 2,016 fine-mapped SNPs.

**SNP-heritability enrichment for loop types.** To estimate enrichment of PrCa risk heritability across functional categories, we ran stratified LDSC regression using PrCa GWAS summary statistics from <sup>72</sup>. First, we created custom bed tracks using H3K27ac peaks called from ChIPseq in LNCaP. We used the custom bed tracks to annotate SNPs genome-wide indicating their overlap. We computed annotation-specific LD scores using the above annotations, using the baseline model containing 53 functional annotation<sup>42</sup> and estimated functional enrichments of heritability using sLDSC <sup>74</sup>.

**eQTL in prostate tissues.** eQTL results on two prostate-specific datasets were used: Thibodeau eQTLs in prostate normal tissue <sup>40</sup>, TCGA eQTLs in prostate tumor tissue <sup>41</sup>. Thibodeau dataset contains gene expression from tumor-adjacent normal prostate tissue on 471 individuals <sup>40</sup>, on autosomal and X chromosomes. TCGA dataset contains gene expression on 378 prostate cancer samples, only on autosomal chromosomes <sup>41</sup>. We used MatrixEQTL <sup>75</sup> and RNA sequencing and genotype data to conduct the gene expression linear regression association for each study datasets. The cis-eQTLs were computed using a window of 3Mb from the TSS around the RefSeq genes, to match the HiChIP loop analysis. For the Thibodeau dataset, 51,232,032 SNP-eQTL pairs and 15,673 genes were tested, and a Bonferroni threshold of 0.05/51,232,032 was used to define “eQTLs” and “eGenes”. After the Bonferroni cut-off, there were 85,435 significant eQTLs and 4,747 eGenes. For the TCGA dataset, 219,256,418 SNP-eQTL pairs and 15,723 genes were tested, and a Bonferroni threshold of 0.05/219,256,418 was used to define “eQTLs” and “eGenes”. There were 86,555 significant eQTLs and 1,118 eGenes. Across both the datasets, we found 4,871 eGenes. To obtain the list of eQTL-eGene pairs, for each gene we selected the SNP with the most significant pvalue.

**Genes with somatically acquired mutations in PrCa.** Genes somatically mutated in prostate cancer were extracted from three publications: an exome sequencing study looking at both localized and advanced prostate cancer <sup>56</sup>, TableS4 “Known in prostate cancer and Recurrently altered in cancer” and TableS6 “cancer\_pathways\_mutation”); a study looking at recurrent alterations in primary (localized) prostate cancer <sup>57</sup>, TableS1C “SigMutated” and TableS1D “genes in wide peak” - we only took the gene highlighted in bold - and TableS1E “Fusions”); and an analysis describing recurrent alterations in metastatic prostate cancer <sup>58</sup>, TableS5 “SigMutated”). Together these sources identified 122 unique genes, of which 119 were in RefSeq and also considered in our analyses (missing MRE11A, FL1, WHSC1L1). 43 genes were within 3Mb of a credible risk SNP for PrCa.

**Gene expression data.** Genes lists associated with gene expression were retrieved from the following sources:

- 1) 908 Genes generally expressed in prostate tissue from GTEx <sup>43</sup> (TPM > 100 TPM). We downloaded GTEx median gene-level TPM by tissue from the GTEx website (GTEx\_Analysis\_2017-06-05\_v8\_RNASeQCv1.1.9\_gene\_median\_tpm.gct), and selected genes in RefSeq with a TPM>100.
- 2) 803 genes from tissue-specific expression (top 10% t.stat from <sup>76</sup>). We downloaded the file from ([https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC\\_SEG\\_Idscores/tstats/GTEx.tstat.ts](https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC_SEG_Idscores/tstats/GTEx.tstat.ts)) v) which contains t-statistics comparing each gene in the particular tissue to all other tissues using 24,842 genes in GTEx. We used only the 18,026 RefSeq genes. Then, we ordered the genes based on absolute value of the t statistics in prostate, and took the top 10%.
- 3) 2,384 genes from differential gene expression tumor/normal <sup>77</sup>. Differential gene regulation in prostate tissues was detected with GEPIA (<http://gepia2.cancer-pku.cn/#degenes>), which use the TCGA and GTEx projects databases to compare gene expression between tumor and

normal tissues under Limma, both under and over expressed. We used the default thresholds of logFC of 1 and qvalue cut-off of 0.01.

**Over-representation of eQTL-eGene links in HiChIP loops.** eGenes that overlapped HiChIP were considered: n=3,837 for Thibodeau, n=870 for TCGA. For every gene, we chose a random SNP from the same window that eQTL was computed from (i.e. 3Mb for Thibodeau and TCGA data), and computed how many times this overlaps a HiChIP anchor. We constructed 100 random control loop datasets by randomly flipping anchor1 (in 50% of the loops) or anchor2 (in the other 50%), and quantified eQTL-eGene that are supported by HiChIP loops in real data versus the random data. We find the empirical pvalue by comparing the proportion of eQTL-eGene supported by HiChIP loops in the actual versus the 100 artificial loop datasets.

**GWAS/eQTL overlap.** We tested for colocalization of GWAS and eQTL using COLOC with default parameters <sup>78</sup>. COLOC TCGA tested 7,001 genes, 11 genes (9 of the 130 regions) had PP4  $\geq 0.75$ . COLOC Thibodeau tested 7,148 genes, 42 genes (33 of the 130 regions) had PP4  $\geq 0.75$ . In total 46 unique genes had evidence of colocalization. 32 unique genes (3 in TCGA, 22 in Thibodeau, 7 in common across the two datasets) were also eGenes under a strict Bonferroni threshold (see **Methods** above). We used the multi-tissue TWAS for PrCa from <sup>79</sup>: This includes 892 significant TWAS associations (TWAS.P  $< 0.05 / 109170$ ) in 45 tissues covering 217 genes (N = 4,458), including normal and tumor prostate. Restricting to only the genes in RefSeq, this includes 651 TWAS associations in 170 genes. We restrict to prostate specific results, specifically the results in "TCGA.PRAD\_SP.TUMOR", "TCGA.PRAD.TUMOR", "GTEx.Prostate", which includes 190 signals and 74 unique genes, and 42 were also eGenes under a strict Bonferroni threshold. In total 101 unique genes had evidence from COLOC or TWAS, and 74 were also eGenes. 29 genes (COLOC) and 40 (TWAS) were also HiChIP genes (n=17,690), and 24 genes (COLOC) and 26 (TWAS) were also HiChIP genes looping to a 95% credible SNP (n=665) (**Table 1**).

**CRISPR/dCas9-mediated repression and gene expression analysis.** In order to create stable dCas9-KRAB expressing cell line LNCaP cells were infected with lenti-KRAB-dCas9-blast (Addgene, 89567) and selected with 6 µg/ml blasticidin for two weeks. sgRNAs were designed according to the "NGG" protospacer adjacent motive (PAM) restriction and gRNA efficiency score was calculated and ranked<sup>80</sup>. Non-human genome targeting negative control and HPRT1 promoter targeting positive control sgRNAs were also selected. sgRNA cassettes were synthesized (Integrated DNA Technologies) and cloned into lentiGuide-Puro (Addgene, 52963) vector. All sgRNA sequences are listed in **Table S10**. LNCaP cells stably expressing KRAB-dCas9 were then subsequently infected with sgRNA vectors and selected with 2 µg/ml puromycin for five days. For gene expression, qRT-PCR 500 ng total RNA (Macherey- Nagel) was reverse transcribed (High Capacity Reverse transcription kit, LifeTechnologies) and cDNA was diluted (20x). SYBR Green assay was performed on Light Cycler 480 instrument (2x Probe Master Mix, Roche). All primer sequences are listed in **Table S11**. Relative gene expression was calculated based on the ddCT method<sup>81</sup>. Each sample was measured by two biological and technical replicates. *GAPDH1* gene was used as housekeeping genes to normalize the samples.

**CRISPRi data and analysis.** The ddCT method<sup>81</sup> was used to determine relative gene expression alterations from CT values of the qPCR. Briefly, for each sample, the average of the housekeeping gene (*ACTB*) was used to calculate the dCT values for each of the three technical replicates, and the average dCT values were computed for each sample. After this, using the control sample average, the ddCT values were computed for each replicate. This shows the relative deviation of each sample from the control condition. The expression values for each condition was then computed using this formula:  $\text{exp} = 2^{\text{-ddCT}}$ . Finally, To compute the effects and the SEs, the averages of the expression values for each sample were combined

across the two biological replicates using a fixed effect meta-analysis. We then used a t-test to compare each sample to the control.

## **Data Availability**

We provide HiChIP interactome maps integrated with GWAS and eQTL information generated as a resource to the research community to investigate PrCa GWAS mechanisms.

## **Acknowledgements**

CG has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 754490 – MINDED project.

## **Author Contributions**

CG, JHS, TS, MKF, AG, NM, BP, MLF contributed to the idea and design of the study. CG, TS, MKF, RDJ, SCB, NM performed the analyses. JHS performed the HiChIP experiment. SS performed the CRISPRi experiment. CG, JHS, TS, MKF, AG, NM, BP, MLF drafted the manuscript. All authors provided critical input on the analyses and the drafted manuscript.

## **Competing Interests Statement**

We declare no conflict of interest.

## **References**

1. Benafif, S., Kote-Jarai, Z. & Eeles, R. A. A Review of Prostate Cancer Genome-Wide

Association Studies (GWAS). *Cancer Epidemiology Biomarkers & Prevention* vol. 27 845–857 (2018).

2. Dadaev, T. *et al.* Fine-mapping of prostate cancer susceptibility loci in a large meta-analysis identifies candidate causal variants. *Nat. Commun.* **9**, 2256 (2018).
3. Hjelmborg, J. B. *et al.* The Heritability of Prostate Cancer in the Nordic Twin Study of Cancer. *Cancer Epidemiology Biomarkers & Prevention* vol. 23 2303–2310 (2014).
4. Mucci, L. A. *et al.* Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA* vol. 315 68 (2016).
5. Schumacher, F. R. *et al.* Author Correction: Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **51**, 363 (2019).
6. Eeles, R. Prostate cancer genome-wide association study from 89,000 men using the OncoArray chip to identify novel prostate cancer susceptibility loci. *Journal of Clinical Oncology* vol. 34 1525–1525 (2016).
7. Olama, A. A. A. *et al.* Multiple novel prostate cancer susceptibility signals identified by fine-mapping of known risk loci among Europeans. *Human Molecular Genetics* vol. 24 5589–5602 (2015).
8. Eeles, R. *et al.* Prostate cancer meta-analysis from more than 145,000 men to identify 65 novel prostate cancer susceptibility loci. *Journal of Clinical Oncology* vol. 2017 1–1 (2017).
9. Olama, A. A. A. *et al.* Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nature Genetics* vol. 41 1058–1060 (2009).
10. Hua, J. T. *et al.* Risk SNP-Mediated Promoter-Enhancer Switching Drives Prostate Cancer through lncRNA PCAT19. *Cell* vol. 174 564–575.e18 (2018).
11. Matejcic, M. *et al.* Germline variation at 8q24 and prostate cancer risk in men of European ancestry. *Nat. Commun.* **9**, 4616 (2018).
12. Guo, H. *et al.* Modulation of long noncoding RNAs by risk SNPs underlying genetic predispositions to prostate cancer. *Nature Genetics* vol. 48 1142–1150 (2016).

13. Luo, Z., Rhie, S. K., Lay, F. D. & Farnham, P. J. A Prostate Cancer Risk Element Functions as a Repressive Loop that Regulates HOXA13. *Cell Reports* vol. 21 1411–1417 (2017).
14. Spisák, S. *et al.* CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. *Nature Medicine* vol. 21 1357–1363 (2015).
15. Gao, P. *et al.* Biology and Clinical Implications of the 19q13 Aggressive Prostate Cancer Susceptibility Locus. *Cell* vol. 174 576–589.e18 (2018).
16. Gusev, A. *et al.* Atlas of prostate cancer heritability in European and African-American men pinpoints tissue-specific regulation. *Nat. Commun.* **7**, 10979 (2016).
17. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics* vol. 102 717–730 (2018).
18. Laat, W. de, de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* vol. 502 499–506 (2013).
19. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–21936 (2010).
20. Mancuso, N. *et al.* Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nat. Commun.* **9**, 4079 (2018).
21. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* vol. 48 245–252 (2016).
22. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
23. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
24. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
25. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).

26. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet.* **13**, e1006646 (2017).
27. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
28. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).
29. Fairfax, B. P. *et al.* Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. *Science* vol. 343 1246949–1246949 (2014).
30. Taberlay, P. C. *et al.* Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res.* **26**, 719–731 (2016).
31. Rhie, S. K. *et al.* A high-resolution 3D epigenomic map reveals insights into the creation of the prostate cancer transcriptome. *Nat. Commun.* **10**, 4154 (2019).
32. Ramanand, S. G. *et al.* The landscape of RNA polymerase II-associated chromatin interactions in prostate cancer. *J. Clin. Invest.* **130**, 3987–4005 (2020).
33. Mumbach, M. R. *et al.* Enhancer connectome in primary human cells reveals target genes of disease-associated DNA elements. doi:10.1101/178269.
34. Jeng, M. Y. *et al.* Enhancer Connectome Nominates Target Genes of Inherited Risk Variants from Inflammatory Skin Disorders. *J. Invest. Dermatol.* **139**, 605–614 (2019).
35. Fang, R. *et al.* Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res.* **26**, 1345–1348 (2016).
36. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
37. D’Ippolito, A. M. *et al.* Pre-established Chromatin Interactions Mediate the Genomic

Response to Glucocorticoids. *Cell Systems* vol. 7 146–160.e7 (2018).

38. Greenwald, W. W. *et al.* Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nat. Commun.* **10**, 1054 (2019).
39. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* vol. 503 290–294 (2013).
40. Thibodeau, S. N. *et al.* Identification of candidate genes for prostate cancer-risk SNPs utilizing a normal prostate tissue eQTL data set. *Nat. Commun.* **6**, 8653 (2015).
41. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
42. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
43. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
44. Wasserman, N. F., Aneas, I. & Nobrega, M. A. An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. *Genome Res.* **20**, 1191–1197 (2010).
45. Sur, I. K. *et al.* Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* **338**, 1360–1363 (2012).
46. Ahmadiyeh, N. *et al.* 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 9742–9746 (2010).
47. Taylor, B. S. *et al.* Integrative Genomic Profiling of Human Prostate Cancer. *Cancer Cell* vol. 18 11–22 (2010).
48. Tindall, D. & Lonergan, P. Androgen receptor signaling in prostate cancer development and progression. *Journal of Carcinogenesis* vol. 10 20 (2011).
49. Brooke, G. N. & Bevan, C. L. The role of androgen receptor mutations in prostate cancer

progression. *Curr. Genomics* **10**, 18–25 (2009).

50. Bu, H. *et al.* Putative Prostate Cancer Risk SNP in an Androgen Receptor-Binding Site of the Melanophillin Gene Illustrates Enrichment of Risk SNPs in Androgen Receptor Target Sites. *Hum. Mutat.* **37**, 52–64 (2016).
51. Takeda, D. Y. *et al.* A Somatically Acquired Enhancer of the Androgen Receptor Is a Noncoding Driver in Advanced Prostate Cancer. *Cell* **174**, 422–432.e13 (2018).
52. Asatiani, E. *et al.* Deletion, methylation, and expression of the NKX3.1 suppressor gene in primary human prostate cancer. *Cancer Res.* **65**, 1164–1173 (2005).
53. Bowen, C. & Gelmann, E. P. NKX3.1 activates cellular response to DNA damage. *Cancer Res.* **70**, 3089–3097 (2010).
54. Bhatia-Gaur, R. *et al.* Roles for Nkx3.1 in prostate development and cancer. *Genes Dev.* **13**, 966–977 (1999).
55. Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215–1228 (2015).
56. Armenia, J. *et al.* The long tail of oncogenic drivers in prostate cancer. *Nat. Genet.* **50**, 645–651 (2018).
57. Schlomm, T. Re: The Molecular Taxonomy of Primary Prostate Cancer. *European Urology* vol. 69 1157 (2016).
58. Robinson, D. *et al.* Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell* **162**, 454 (2015).
59. Comoglio, F. *et al.* Thrombopoietin signaling to chromatin elicits rapid and pervasive epigenome remodeling within poised chromatin architectures. *Genome Res.* (2018) doi:10.1101/gr.227272.117.
60. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* (2020) doi:10.1038/s41588-020-0625-2.

61. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).
62. Mora, A., Sandve, G. K., Gabrielsen, O. S. & Eskeland, R. In the loop: promoter–enhancer interactions and bioinformatics. *Briefings in Bioinformatics* bbv097 (2015) doi:10.1093/bib/bbv097.
63. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* vol. 316 1497–1502 (2007).
64. Qin, Q. *et al.* ChiLin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. *BMC Bioinformatics* vol. 17 (2016).
65. Cornwell, M. *et al.* VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. *BMC Bioinformatics* **19**, 135 (2018).
66. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
67. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
68. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
69. Website. Krueger, F., 2012. Trim Galore! Available at: [http://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).
70. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
71. Bhattacharyya, S., Chandra, V., Vijayanand, P. & Ay, F. Identification of significant chromatin contacts from HiChIP data by FitHiChIP. *Nat. Commun.* **10**, 4221 (2019).
72. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).
73. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-

mapping studies. *PLoS Genet.* **10**, e1004722 (2014).

74. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

75. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* vol. 28 1353–1358 (2012).

76. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621 (2018).

77. Tang, Z. *et al.* GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* **45**, W98–W102 (2017).

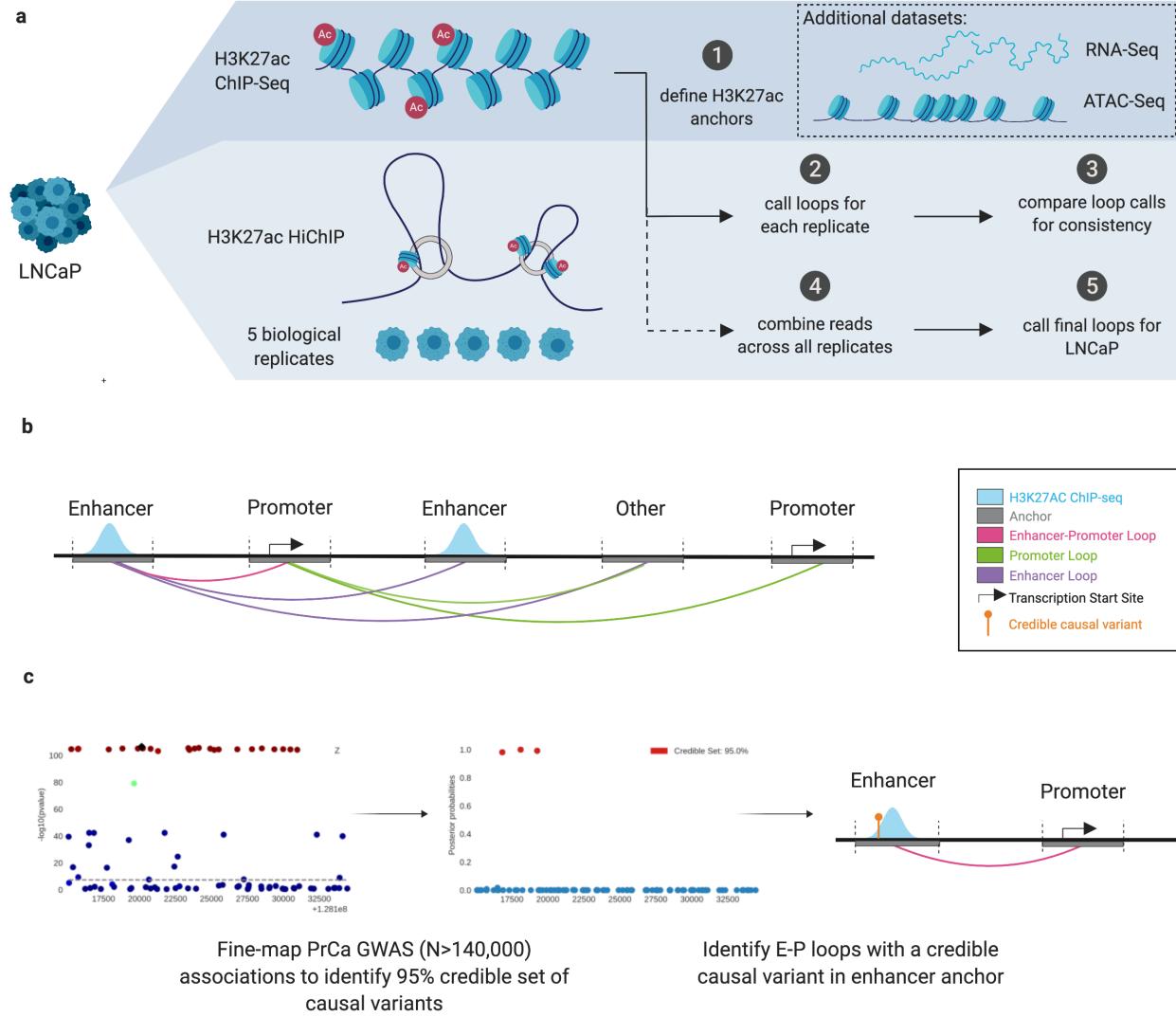
78. Giambartolomei, C. *et al.* A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538–2545 (2018).

79. Mancuso, N. *et al.* Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nat. Commun.* **9**, 4079 (2018).

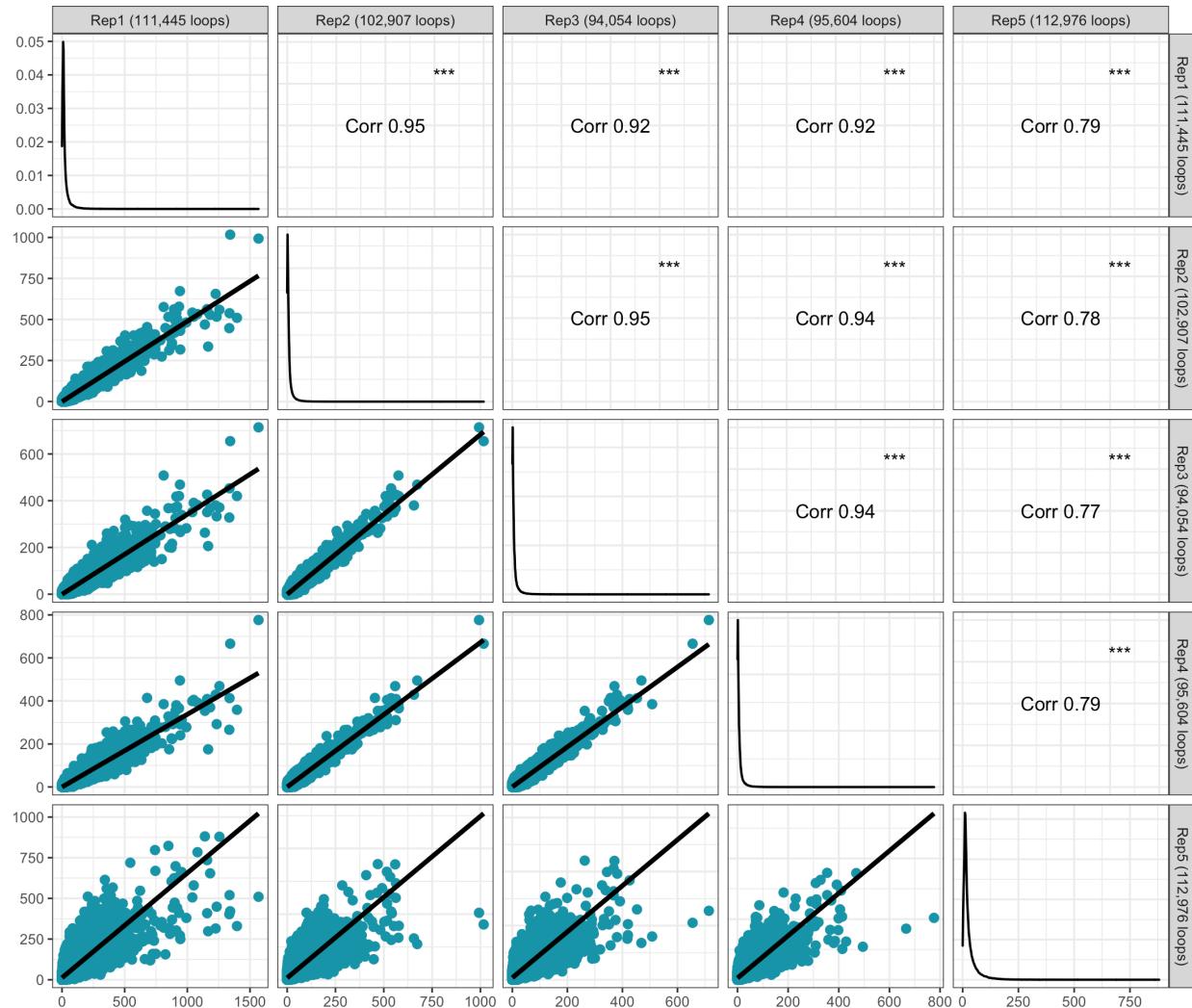
80. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).

81. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* **29**, e45 (2001).

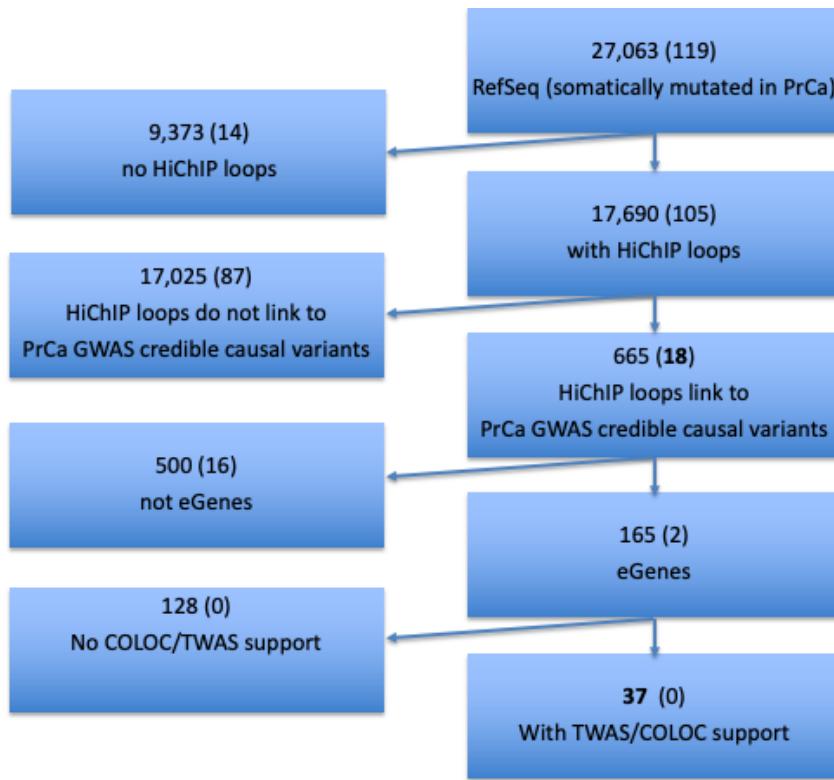
## Figure Legends



**Figure 1: Experimental design.** **a.** Description of experimental datasets and workflow of HiChIP analysis to call total LNCaP loops. **b.** Anchor and loop type definitions. **c.** Possible regulatory mechanism describing a causal variant in an enhancer anchor interacting with a promoter of a target gene.

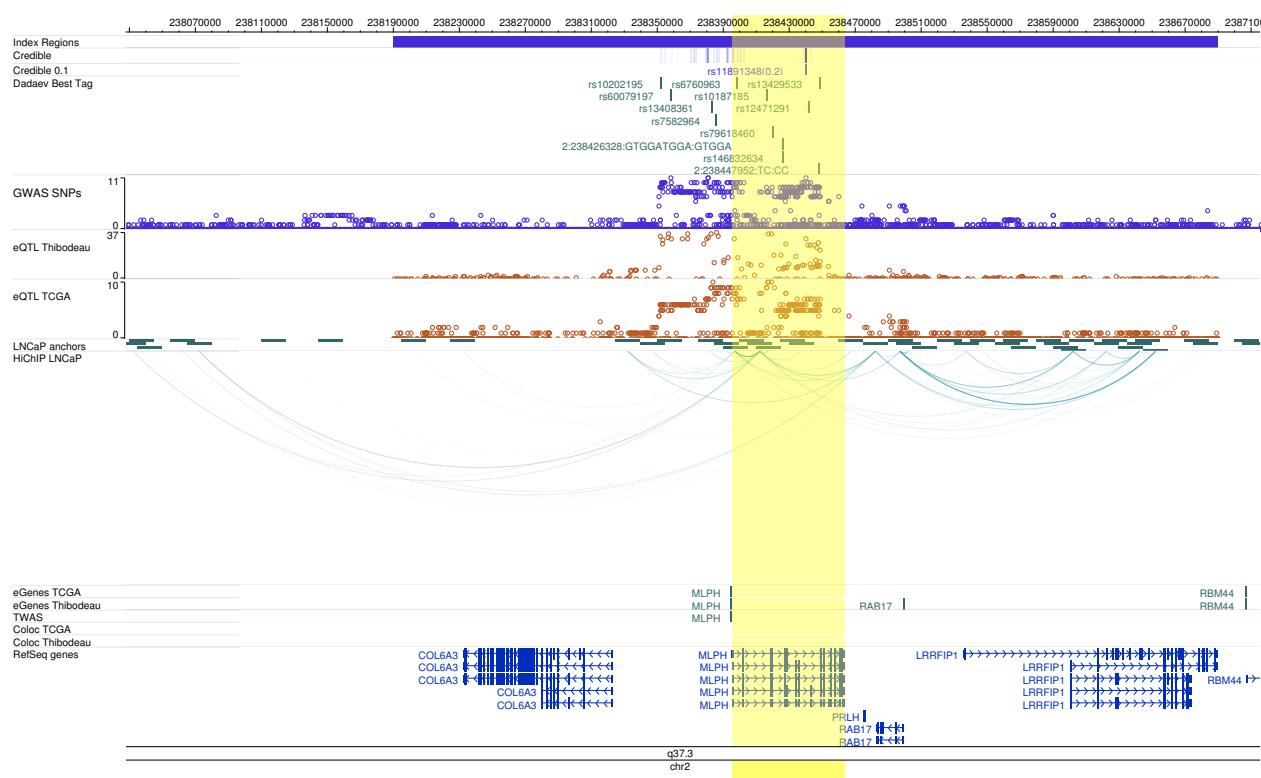


**Figure 2: Pairwise correlations of pair-end-tag (PET) counts at loops called significant in at least one of the LNCaP-HiChiP replicates.** Every dot represents a HiChiP loop called at FDR <0.01 in any of the replicates; the union of across replicates  $N = 114,142$  loops. We consider all the loops called at no FDR cut-off. Out of these, we then take the loops called at FDR 0.01 in any of the replicates (i.e. the union of loops across replicates,  $N = 114,142$  loops). Within the bottom left panels, the axis represent the number of pair-end-tags observed in the two compared replicates. The diagonal panel shows the distribution of each replicate's PET counts. Pearson correlation coefficients are shown on the top right panels. If the PET count is missing for a loop in one replicate but it is present in the other, it is treated as 0 in the correlation estimation. See Supplementary Figure 1 for a comprehensive comparison across replicates.

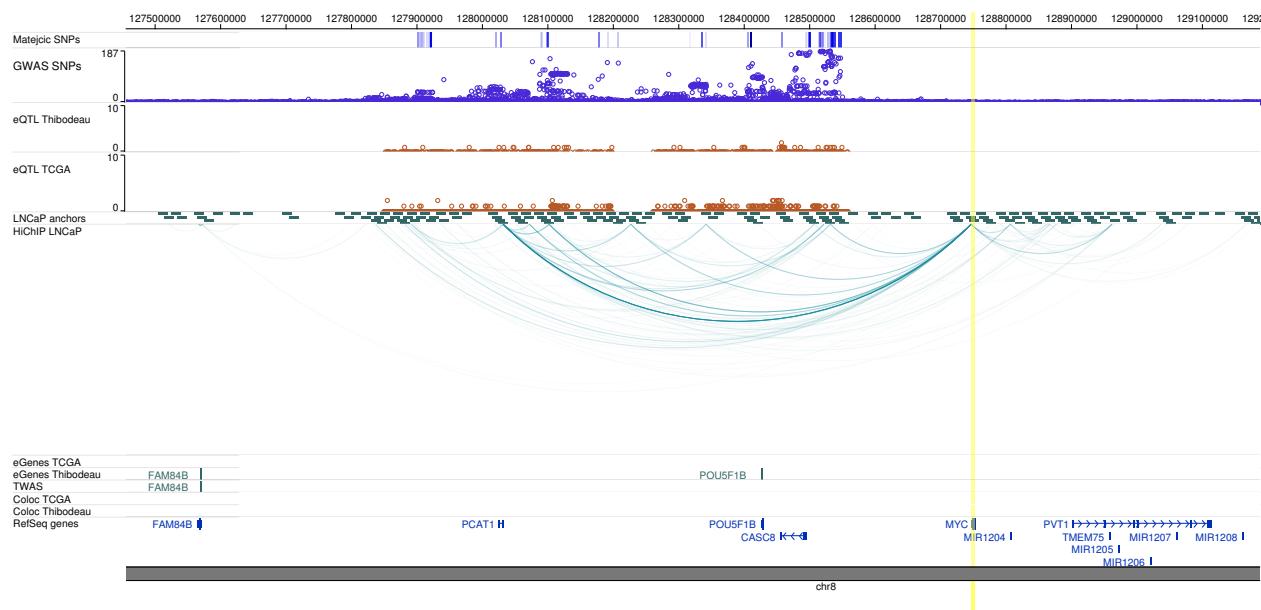


**Figure 3: Breakdown of genes with various forms of evidence for linked to PrCa risk variants.** Out of 27,063 genes in RefSeq, 17,690 show a HiChIP loop overlaying their promoter. 665 genes at 104 PrCa GWAS regions have a loop linking their promoter to a PrCa credible causal variant. 165 (out of 665) are also eGenes in tissues relevant to PrCa with 37 also showing evidence of co-localization and/or transcriptome-wide association. The numbers in parentheses showcases the breakdown of the 119 genes with evidence of somatically acquired mutations in prostate cancer (**Methods**).

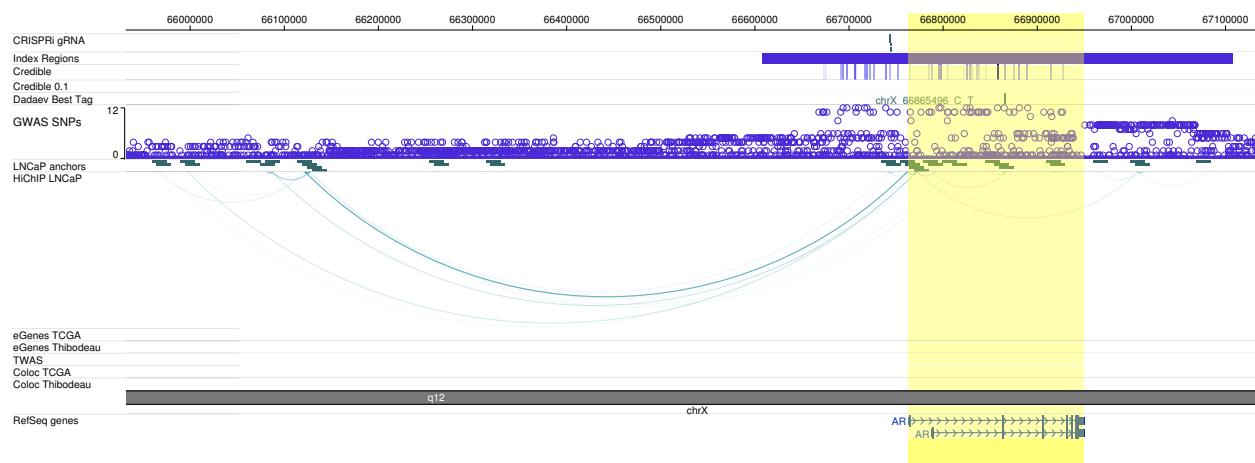
a.



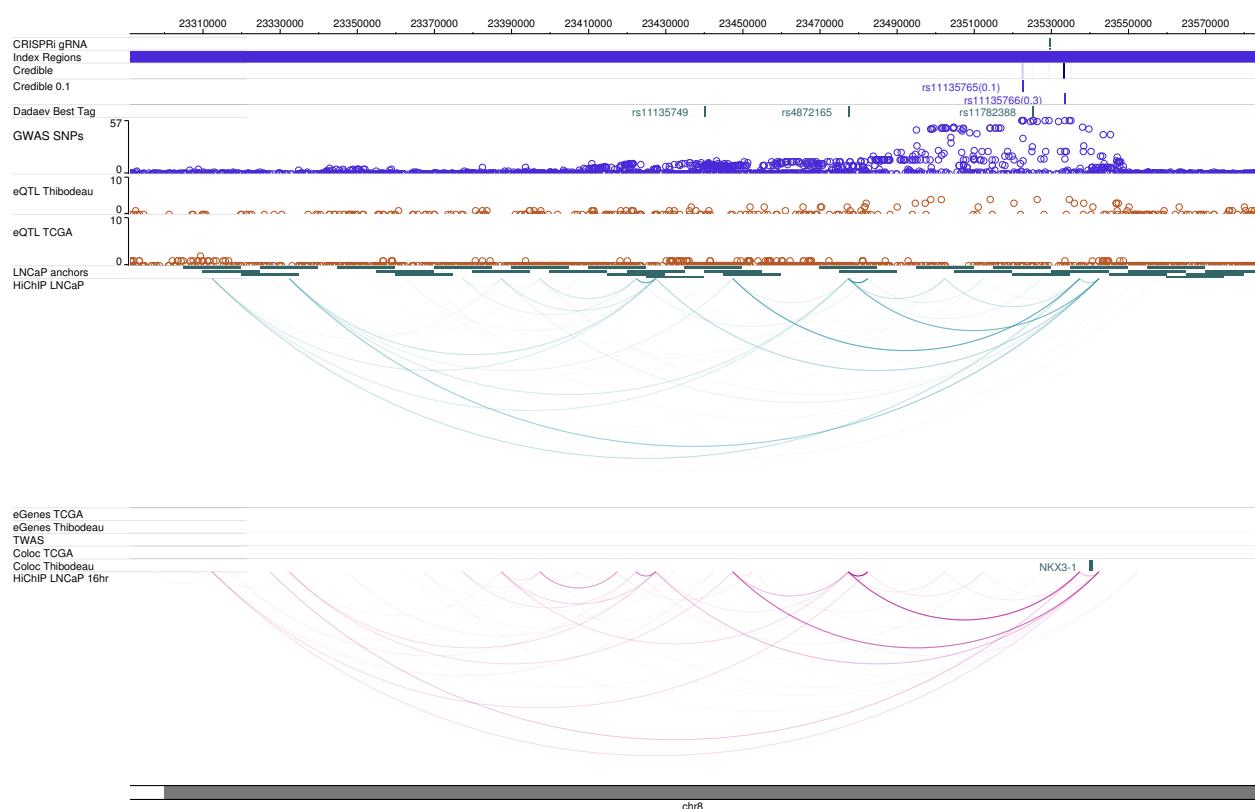
b.



C.



d.



**Figure 4.** Examples of genes supported by HiChIP loops and GWAS: **a.** *MLPH*, with evidence of eQTLs in prostate tissue using both prostate tissue datasets (Thibodeau and TCGA); **b.** *MYC* and **c.** *AR*, with no evidence of eQTLs. **d.** *NKX3-1*, with weak evidence of eQTLs in prostate tissue.

The tracks shown are listed below.

**Index Regions.** the fine-mapped region.

**Credible.** Position of the SNPs included in the 95% credible SNP set of being the most likely causal variants from PAINTOR analyses (Methods). The color of the track is deeper for SNPs with higher probability of being causal from 0 to 1.

**Matejcic SNPs.** Position of fine-mapped SNPs from Matejcic et al.<sup>11</sup>

**CRISPRi gRNA.** gRNA positions targeted for the CRISPRi experiment.

**Credible 0.1.** Position and name of SNPs that reach a posterior probability of being causal of 0.1.

**Dadaev Best Tag.** Position and name of SNPs within 341 "best tag" SNPs fine-mapped in ref. Dadaev et al.

**GWAS SNPs.** Each hollow circle represents a SNP; y-axis is the position and x-axis is the  $-\log_{10}(P)$  genome-wide association with PrCa risk.

**LNCaP anchors.** Regions of the genome containing HiChIP anchors.

**HiChIP LNCaP.** Loops from merged data of 5 replicates.

**eQTL TCGA.** Same as above track. Associations were run on gene expression and genotype data on 494 prostate cancer patients (ref. Pancan website).

**eGenes Thibodeau.** Highlighted genes that are eGenes in Thibodeau dataset.

**eGenes TCGA** Highlighted genes that are eGenes in TCGA dataset.

**eQTL Thibodeau.** Each hollow circle represents a SNP; y-axis is the position and x-axis is the associations between SNP and gene expression ( $-\log_{10}(pvalue)$ ) for SNPs within 3Mb window around a gene promoter. Associations were run using gene expression and genotype data from 471 samples from normal prostate tissue (ref. Thibodeau).

**eQTL TCGA.** Each hollow circle represents a SNP; y-axis is the position and x-axis is the associations between SNP and gene expression ( $-\log_{10}(pvalue)$ ) for SNPs within 3Mb window around a gene

promoter. Associations were run using gene expression and genotype data from 471 samples from normal prostate tissue (ref. Thibodeau).

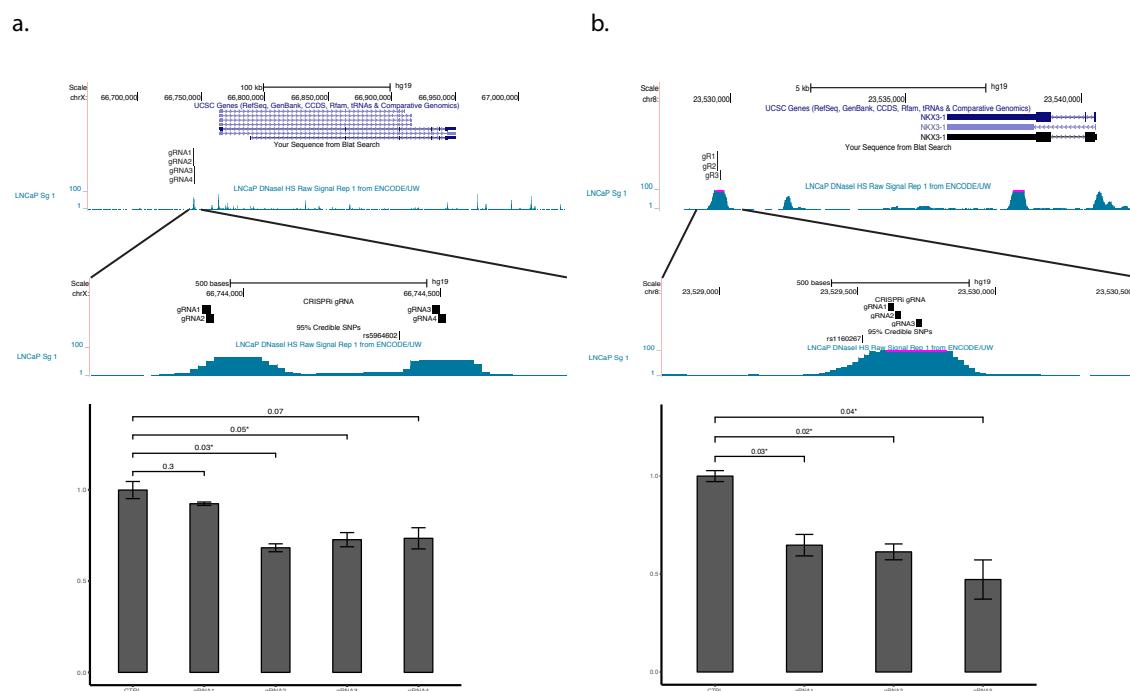
**TWAS.** Highlighted genes that have a significant TWAS signal in Prostate cancer.

**Coloc Thibodeau.** Highlighted genes that have a significant COLOC signal in Prostate cancer, using Thibodeau eQTL data.

**Coloc TCGA.** Highlighted genes that have a significant COLOC signal in Prostate cancer, using TCGA eQTL data.

**RefSeq genes.** Known human protein-coding and non-protein-coding genes taken from the 2019 RefSeq Release.

**HiChIP LNCaP 16hr.** HiChIP loops measured after 16 hours from androgen stimulation.



**Figure 5. CRISPRi functional validation of the AR and NKX3-1 loci. Functionally relevant enhancers were identified by integrating genetic (PrCa causal variants) and epigenetic datasets (DHS peaks and HiChIP anchors in LNCaP cell line).** a. AR gene genomic region, LNCaP DHS signals, 95% credible SNP set and gRNA positions targeting DHS peaks for the CRISPRi experiment. Bottom panel shows the suppression effect on the AR gene using four different gRNAs compared to the

non-human targeting negative control. Three out of the four tested gRNAs showed ~1 fold significant suppression on the AR gene expression. Columns demonstrate averages of two biological replicates, error bars represent standard errors. **b.** NKX3-1 genomic region, LNCaP DHS signals, 95% credible SNP set and gRNA positions targeting DHS peaks for the CRISPRi experiment. Bottom panel shows the suppression effect on the NKX3-1 gene using three different gRNAs compared to the non-human targeting negative control. All three tested gRNAs showed ~1 fold significant suppression on the NKX3-1 gene expression. Columns demonstrate averages of two biological replicates, error bars represent standard errors.

## Tables

Genes for which looping links their promoters to PrCa risk variants with extra evidence from eQTL/GWAS overlap or somatically-acquired in PrCa.	
COLOC positive only	<i>CASP10, COL2A1, COMMD7, CTSK, DNMT3B, HNF1B, LARP4B, MSMB, PPFIBP2, RAD9A, SNRPC</i>
TWAS positive only	<i>ACAT2, C10orf95, CNTROB, FAM84B, MLPH, NOL10, RGS17, SESN1, SPINT2, TSPO, USP20, USP39, ZGPAT</i>
COLOC and TWAS positive	<i>C9orf78, CTBP2, FAM57A, GEMIN4, HAUS6, KRT8, MBNL1, MMP7, MYO9B, NCOA4, PPP1R14A, UHRF1BP1, VPS53</i>
With somatically acquired mutations in PrCa	<i>AR, CCND1, CDKN1B, HD3, CIC, ERF, GATA2, KMT2D, MAP2K1*, MED12, MYC, NKX3-1, PTEN,</i>

	<i>RNF43, RPRD2*</i> , <i>SETDB1, TP53, ZMYM3</i>
--	---

**Table 1: Genes where HiChIP looping data links their promoters to PrCa risk variants.** Out of the 665 genes linked by looping to PrCa risk variants (Figure 3), we list genes with evidence of eQTL/GWAS overlap (37) or somatically-acquired mutations in PrCa (18); see Table S6 for all RefSeq genes. For 37 genes looping links their promoter to PrCa GWAS variants, and also show evidence of eQTL/GWAS overlap through colocalization (COLOC) and/or transcriptome-wide association (TWAS). 18 out of the 119 genes previously reported to have somatically acquired mutations in PrCa show loops linking their promoters to PrCa germline risk variants. The asterisk points to eGenes.