Title

Genetic determinants of chromatin reveal prostate cancer risk mediated by context-dependent gene regulation

Permalink

https://escholarship.org/uc/item/00v406xb

Journal

Nature Genetics, 54(9)

ISSN

1061-4036

Authors

Baca, Sylvan C
Singler, Cassandra
Zacharia, Soumya
et al.

Publication Date

2022-09-01

DOI

10.1038/s41588-022-01168-y

Peer reviewed

Published in final edited form as:

Nat Genet. 2022 September ; 54(9): 1364–1375. doi:10.1038/s41588-022-01168-y.

Genetic determinants of chromatin reveal prostate cancer risk mediated by context-dependent gene regulation

Sylvan C. Baca[1,2,3], **Cassandra Singler**[4], **Soumya Zacharia**[1,2], **Ji-Heui Seo**[1,2], **Tunc Morova**[5], **Faraz Hach**[5], **Yi Ding**[6], **Tommer Schwarz**[6], **Chia-Chi Flora Huang**[5], **Jacob Anderson**[7], **André P. Fay**[1], **Cynthia Kalita**[1,8], **Stefan Groha**[1,3], **Mark M. Pomerantz**[1,2], **Victoria Wang**[9,10], **Simon Linder**[11,12], **Christopher J. Sweeney**[1], **Wilbert Zwart**[11,12], **Nathan A. Lack**[5,13], **Bogdan Pasaniuc**[6,14,15,16], **David Y. Takeda**[4], **Alexander Gusev**[1,3,8,*], **Matthew L. Freedman**[1,2,3,*]

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

[2]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA, USA

[3]The Eli and Edythe L. Broad Institute, Cambridge, MA, USA

[4]Laboratory of Genitourinary Cancer Pathogenesis, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA

[5]Vancouver Prostate Centre University of British Columbia, Vancouver, BC, Canada

[6]Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA

[7]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA

[8]Division of Genetics, Brigham & Women's Hospital, Boston, MA, USA

[9]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA

[10]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

[11]Division of Oncogenomics, Oncode Institute, The Netherlands Cancer Institute, Amsterdam, The Netherlands

[12]Laboratory of Chemical Biology and Institute for Complex Molecular Systems, Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

[*]These authors jointly supervised this work. Correspondence should be directed to M.L.F (mfreedman@partners.org) or A.G. (Alexander_Gusev@dfci.harvard.edu).

Author contributions

S.C.B, A.G., and M.L.F conceived of the study. S.C.B analyzed the data and wrote the manuscript under the joint supervision of M.L.F. and A.G. A.F. performed ChIP-seq experiments. J.-H. S. generated and C.K. analyzed allelic imbalance data from LNCaP. T.M. and Y.D. analyzed SNP-STARR-seq data under the supervision of N.L. and B.P. C.S. performed CRISPRi experiments under the supervision of D.Y.T. S. Z. assisted with analysis of ChIP-seq data. S.G. assisted with genotype imputation. S. L. and W. Z provided prostate cancer ChIP-seq data. M.M.P. and V.W. analyzed ADT GWAS data preformed on samples and data provided by C.J.S. J.A. assisted with implementation of the CWAS pipeline.

Competing interest

The authors declare no competing interests.

Code availability

Scripts to reproduce analyses from this study are available at https://github.com/scbaca/cwas, https://github.com/scbaca/chip_imputation, and https://doi.org/10.5281/zenodo.6666796.

[13]School of Medicine, Koç University, Istanbul, Turkey

[14]Department of Computational Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA USA

[15]Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

[16]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

Abstract

Many genetic variants affect disease risk by altering context-dependent gene regulation. Such variants are difficult to study mechanistically using current methods that link genetic variation to steady-state gene expression levels, such as expression quantitative trait loci (eQTLs). To address this challenge, we developed the cistrome-wide association study (CWAS), a framework for identifying genotypic and allele-specific effects on chromatin that also associate with disease. In prostate cancer, CWAS identified regulatory elements and androgen receptor binding sites that explained the association at 52 out of 98 known prostate cancer risk loci and discovered 17 additional risk loci. CWAS implicated key developmental transcription factors in prostate cancer risk that are overlooked by eQTL-based approaches due to context-dependent gene regulation. We experimentally validated associations and demonstrated the extensibility of CWAS to additional epigenomic datasets and phenotypes, including response to prostate cancer treatment. CWAS is a powerful and biologically interpretable paradigm for studying variants that influence traits by affecting transcriptional regulation.

Genome-wide association studies (GWAS) have identified hundreds of thousands of genetic variants associated with human traits and diseases. The majority of these variants map to regulatory elements and confer risk by affecting transcription of nearby genes[1–7]. Determining how non-coding genetic variants contribute to diseases and complex phenotypes has proven difficult[8–11]. To address this challenge, large-scale efforts have catalogued many thousands of *cis*-acting expression quantitative trait loci (eQTLs)[12–14]. At these loci, the genotype of a single nucleotide polymorphism (SNP) correlates with steady-state expression of a nearby gene (eGene). eQTLs can identify genes that mediate risk[15–18] and are present at 40–50% of disease-associated genomic loci by some estimates[14,19].

The utility of eQTLs for mechanistically characterizing genetic risk variants is limited by several factors. eQTLs that are relevant for complex phenotypes are often context-dependent[12–14]. Such eQTLs are not observable at steady-state in bulk, differentiated tissues, but only in certain cell types, at specific developmental stages, or in response to stimuli[20–25]. Steady-state eQTLs are depleted near genes that are likely to contribute to complex phenotypes, including transcription factors, developmental genes, and highly conserved or essential genes[26]. Consequently, steady-state *cis*-eQTLs explain only 11% of the heritability for an average trait by a recent estimate[26,27], or up to 25% when transcription is profiled in disease-relevant tissues[28].

Many eQTLs influence gene expression through effects on chromatin – for instance, by altering regulatory element activity[29–32]. Increasingly, studies have analyzed the effect of risk-associated genetic variants on chromatin itself, rather than the more distal readout of gene expression[30,33–35]. Analogous to eQTLs, chromatin QTLs (cQTLs) are SNPs whose genotype correlates with chromatin state, as characterized by histone modifications, TF binding, or chromatin accessibility[36–39]. In a complementary manner to cQTLs, allelic imbalance (AI) in epigenomic data – differential representation of heterozygous SNP alleles in sequencing reads – can also identify variants that affect chromatin state[23,24,35,40,41]. The use of cQTLs and AI for understanding trait heritability is limited, however, by the lack of (1) large panels of reference epigenomes from relevant tissues and (2) a unified framework for integrating these data into GWAS.

Here, we describe a biologically and statistically principled approach for identifying variants that contribute to phenotypes through effects on the cistrome (genome-wide profiles of histone modifications and TF binding sites). We introduce the cistrome-wide association study (CWAS), which identifies the genetic determinants of TF binding and chromatin activity and associates genetically predicted chromatin signal with the trait using GWAS summary statistics.

We performed a CWAS of prostate cancer, one of the most heritable and common cancers[42]. We find that heritable variation in the cistrome of the androgen receptor (AR) – a critical TF in prostate cancer pathogenesis, treatment, and progression – mediates risk at 21% of prostate cancer risk loci. In addition, 45% of prostate cancer risk loci can be explained in part by genetic variation regulatory element activity, as measured by H3K27 acetylation (H3K27ac). CWAS annotates disease mechanisms at GWAS risk loci that are difficult to discover through eQTL-based analyses. CWAS implicated prostate developmental genes in prostate cancer risk that lack robust eQTLs, likely due to complex regulation and/or context-dependent expression.

Overview of the methods

We developed a systematic approach that links genetic variation in TF binding or chromatin state to trait variation (Fig. 1A). We leverage the growing number of chromatin immunoprecipitation and DNA sequencing (ChIP-seq) datasets from genetically distinct individuals to create epigenomic reference panels (Fig. 1B). A limitation of existing ChIP-seq datasets is that most lack SNP genotypes necessary for studying genetic-epigenetic interactions. We therefore created and benchmarked an approach to impute genotypes from ChIP-seq data with high accuracy[43] (Extended Data Fig. 1 and Supplementary Note). We identify genetic determinants of epigenomic features (e.g., AR binding or H3K27 acetylation) by jointly relating allelic imbalance and peak intensity to the genotypes of nearby SNPs. These models identify SNPs that correlate with epigenomic peak intensity. Integrating this information with summary statistics from GWAS, we identify peaks whose genetic determinants are associated with the trait of interest. The result is a cistrome-wide association study (CWAS) that identifies peaks whose genetically predicted activity is associated with risk of a trait or disease (Fig. 1B).

cis-SNP determinants of regulatory element activity

We utilized data from two recent studies of prostate cancer epigenomes, which performed ChIP-seq for TFs and histone modifications across a combined 163 individuals of predominantly European ancestry[44,45] (Tables S1-2, Extended Data Fig. 2). The dataset comprises 131 ChIP-seq experiments for AR and 176 for H3K27ac. Because these samples have not been subjected to genotyping, we used ChIP-seq reads to impute high-accuracy germline genotypes at ~5.5 million SNPs with a minor allele frequency of ≥ 5% [43,46] (Extended Data Fig. 1 and Supplemental Note).

By analyzing both allelic imbalance (AI) and cQTLs in large epigenomic reference panels, we detected widespread *cis*-genetic regulation of chromatin by common SNPs. A combined test for significant cQTL activity or AI[47] identified 4,243 AR binding sites (ARBS; 9% of total) and 13,569 H3K27ac peaks (17% of total) where the genotype of nearby SNP (cQTL) correlated with the intensity of a peak (“cPeak”) or was significantly imbalanced in ChIP-seq reads (Fig. 2A). AR cQTL activity and AI, which are measured independently, correlated in magnitude and direction ($\rho = 0.80$, $p < 2.2 \times 10^{-16}$), confirming a shared underlying effect in the population (Fig. 2B). This effect size concordance is similar to a larger study of gene expression AI and cQTLs[48]. Measuring both AI and cQTLs increased the number of peaks under detectable genetic control by roughly over 50% compared to either measure alone (Fig. 2C). Genetically determined H3K27ac peaks overlapped with only 41% of AR peaks (Fig. 2D), indicating that TF and H3K27ac ChIP-seq data captured overlapping but distinct genetic regulation. cQTLs overlapped significantly with eQTLs from an independent GTEx study and demonstrated correlated effects on chromatin and gene expression (Extended Data Fig. 3 and Supplemental Note).

cQTL SNPs tended to reside in or near peaks: 50% of AR cQTLs and 35% of H3K27ac cQTLs were within 10kb of the corresponding peak center (Fig. 2E; Extended Data Fig. 4). Ten percent of AR cQTLs fell within 200 base-pairs of the peak center, suggesting that these SNPs directly affect binding of core TF machinery. Accordingly, 450 heterozygous SNPs within binding motifs of AR and its cofactors demonstrated AI, with AR preferentially binding to the allele that is more similar to the consensus binding motif (Fig. 2F), bolstering the functional validity of these QTLs. Nonetheless, 16% of AR cPeaks did not contain a SNP, consistent with distal *cis*-genetic regulation.

Integrative genetic models of cistromes

Given the distinct contributions of AI and cQTLs (Fig. 2C), we created integrative models combining both features to capture genetic determinants of AR binding and regulatory element activity. We modeled total and allele-specific peak intensity[24,49] as a function of all nearby SNP genotypes (Fig. 3A), cross-validating our models on held-out samples. To allow for the possibility that multiple SNPs affect peak intensity, we considered sparse linear models that combine effects from multiple SNPs within 25kb of a peak[50], an interval that contained 84% of the top 5% of AR cQTLs by significance (Extended Data Fig. 4; Supplemental Note). Five-fold cross validation demonstrated that 5,580 out of 48,948 AR peaks (11%) and 17,199 out of 73,475 H3K27ac peaks (23%) showed significant correlation between the trained SNP model and peak intensity in held out samples, after correction for multiple hypothesis testing ($q < 0.05$; Tables S3-4). The variants incorporated by our models tended to also influence gene expression (Supplementary Note)

We validated allele-specific regulatory activity *in vitro* using an enhancer reporter assay for six H3K27ac peaks (Fig 3B; Methods). In addition, suppression of genetically determined ARBS in LNCaP prostate cancer cells using CRISPRi suppressed the expression of genes linked to these ARBS by H3K27ac HiChIP loops. For instance, suppression of a 14kb-upstream ARBS markedly reduced *TMPRSS2* expression (Fig 3C–D; Table S5), consistent with a report that this ARBS contains a *TMPRSS2* eQTL[51]. Similarly, suppression of a genetically determined ARBS decreased expression of its candidate gene based on HiChIP connectivity (BMPR1B; 134KB away) with no effect on the gene containing the ARBS (PDLIM5; Fig 3E–F; Table S5). These data indicate that our genetic models capture SNPs that influence gene expression through effects on regulatory elements (Fig. 1A) and highlight how chromatin conformational data can match cQTL ARBS to the genes they control.

Prostate cancer CWAS

Our genetic models of ARBS and regulatory elements revealed disease heritability that is likely mediated through effects on these epigenetic features. We performed a cistrome wide association study (CWAS) to associate genetically predicted peak intensity with prostate cancer risk, using summary statistics from a prostate cancer GWAS of 140,306 individuals[52]. Analogous to the framework for a transcriptome wide association study (TWAS)[50], this approach imputes the genetic component of total and allele-specific peak intensity into populations profiled by GWAS. By utilizing summary statistics from GWAS studies, CWAS takes advantage of the large size of GWAS studies without requiring participant-level information.

CWAS identified 74 ARBs (out of 5,580 ARBS with genetic models) and 199 H3K27ac peaks (out of 17,199) that were significantly associated with prostate cancer risk after Bonferroni correction for multiple hypotheses tested (Fig. 4A; Tables S6-7). CWAS association explained >90% of the GWAS signal for 41% of AR CWAS regions and 52% of H3K27ac CWAS regions (Fig. 4B–C and Extended Data Fig. 5A-B; Table S6-7). For instance, a single intragenic ARBS within *LMTK2* accounted for the significant GWAS association at this region (Fig. 4C). Similarly, H3K27 acetylation at five CWAS peaks near *BIK* and *TTLL12* explained nearby GWAS associations (Fig. 4C). In other regions, residual association remained after conditioning on CWAS peaks, suggesting additional mechanisms, more complex regulation, or incomplete tagging[9] (Extended Data Fig. 5C).

AR and H3K27ac CWAS identified 27 significant “novel” peak-trait associations across 17 regions without a nearby genome-wide significant GWAS SNP (Extended Data Fig. 6; Table S6-7). CWAS enabled these discoveries by limiting hypothesis testing to SNPs with a high prior likelihood of affecting phenotypes – *i.e.*, testing tens of thousands of genetically determined epigenomic features, as opposed to millions of unselected SNPs. Tested peaks are expected to be enriched for true positive associations, given that prostate cancer risk variants were highly enriched in cQTL ARBS and regulatory elements (Extended Data Fig. 7). Importantly, GWAS associations were confirmed in 12 out of 17 regions with novel CWAS associations after this manuscript was prepared in a larger GWAS study incorporating an additional ~94,000[52]. This finding indicates that CWAS identifies associations that fall short of GWAS significance but are detectable with larger sample sizes.

We verified the robustness and extensibility of CWAS by applying it to a previously reported blood cell ChIP-seq dataset, identifying 12,903 H3K27ac peak – trait associations across 12 blood-related phenotypes (Supplementary Note).

CWAS identifies associations at eQTL$^-$ loci

CWAS uncovered many chromatin-prostate cancer risk associations at eQTL-negative loci, where genetic effects on steady-state gene expression are not observed. We compared CWAS associations to results from TWAS (an integrative analysis of eQTL-trait associations) that used reference gene expression data from 45 tissues (4,448 individuals) including benign prostate tissue and prostate cancer[50,53]. Some CWAS peaks colocalized with genes identified by TWAS, such as *MLPH* and *MSMB/NCOA4*[51,53,54], but many did not. To compare the relative contributions of TWAS and CWAS in accounting for GWAS risk loci, we defined a set of high-confidence TWAS and CWAS hits where the standardized effect size Z^2 is greater than 90% of Z^2 for the top GWAS SNP. At these sites, a CWAS peak or a TWAS gene account for most of the GWAS association signal, allowing risk to be linked a specific regulatory element or gene.

Compared to TWAS, CWAS nearly doubled the number of GWAS risk loci that could be annotated with plausible risk mechanisms. We defined 98 prostate cancer risk regions by merging ±1Mb windows centered on genome-wide significant SNPs. Of these regions, 52 (53%) contained a high confidence AR or H3K27ac CWAS peak (N=21 and N=44, respectively) compared to 34 (35%) that contained a TWAS gene (Fig. 5A). Critically, at 28 regions (29%), CWAS detected a high-confidence peak association in the absence of a high-confidence TWAS gene association. Thus, CWAS implicated regulatory elements at the 53% of prostate cancer GWAS risk regions, including many regions that lacked a robust association with steady-state gene expression.

We considered why CWAS detected chromatin-prostate cancer associations in TWAS-negative (TWAS-) regions despite using substantially smaller reference panels than TWAS. A potential reason is that genetic variation affects the steady-state cistrome more consistently than it affects transcription. Consistent with this finding, *cis*-SNPs explained a significantly greater portion of the heritability of AR and H3K27ac total peak intensity $\left(h_{g-total}^2\right)$ than the heritability of gene expression levels ($p = 5 \times 10^{-171}$ and $p = 9 \times 10^{-279}$ for AR and H3K27ac, respectively; Fig. 5B–C). Accordingly, SNP genotypes correlated more robustly with regulatory element activity than with gene expression at many risk loci, including *TMPRSS2* and *NKX3-1* (Extended Data Fig. 8). This finding suggests that consistency of genetic effects on steady-state chromatin measurements improves the performance of CWAS models over TWAS.

An additional explanation for CWAS associations at TWAS- loci is that steady-state chromatin measurements capture context-dependent genetic determinants of transcription. To test this hypothesis, we measured allelic imbalance in chromatin accessibility (ATAC-seq), H3K27ac, and gene expression data in LNCaP cells at baseline and after 16h of androgen stimulation. We identified 760 transcripts that demonstrated imbalance with stimulation but not at baseline (Fig. 5D). These genes were enriched for nearby H3K27ac and ATAC-seq peaks with imbalance in the absence of stimulation (OR 2.3 and 2.6 for ATAC-seq and H3K27ac, respectively, $p < 2.2 \times 10^{-16}$; Fig. 5D). Thus, effects on expression that are only apparent with stimulation are preceded by genetic effects on nearby regulatory elements at steady-state, as observed previously in immune cells[39].

Several additional observations support this conclusion. First, tissue- and context-dependent regulatory elements were enriched for steady-state cQTLs compared to eQTLs. We considered eQTLs and cQTLs that overlap with accessible chromatin across 733 tissue samples representing 438 cell types and states[55]. eQTLs tended to localize to chromatin that is accessible in multiple tissues and conditions, while AR and H3K27ac cQTLs overlapped chromatin with more context- or tissue-restricted accessibility ($p = 7 \times 10^{-7}$ and 4×10^{-4}, for eQTLs vs. AR and H3K27ac cQTLs, respectively; Fig. 5E).

Second, for many genes with prostate-restricted expression (quantified by the z-score for expression in prostate compared to all other tissues), *cis*-SNPs did not correlate with transcript levels but robustly correlated with activity of nearby regulatory elements. We binned genes by quantiles of prostate-specific expression[13]. Then, for each bin we counted genes with a TWAS model (in prostate tissue or prostate cancer) and genes with a nearby CWAS model. Genes with increasingly prostate-enriched expression – where power to detect eQTLs should be high due to higher expression levels – were less likely to be modeled by TWAS, but more likely to harbor nearby ARBS or regulatory elements with CWAS models (Fig. 5F–G).

Third, consistent with prior work[56], we found that TWAS models were depleted among genes with the highest degree of regulation, as assessed by the enhancer domain score (EDS; Fig. 5H–I). In contrast, high-EDS genes were the most likely to have nearby CWAS models (Fig. 5H–I). A known limitation of steady-state eQTLs is that they are depleted around highly regulated (high-EDS) genes, which include TFs, developmental genes, and genes involved in disease pathogenesis[26,57]. This principle may explain the ability of CWAS to annotate prostate cancer risk in TWAS- regions. Prostate cancer risk regions with a CWAS association but no TWAS association (CWAS+/TWAS-) had significantly higher EDS scores than CWAS+/TWAS+ regions (Fig. 5J), suggesting that these regions were not captured by TWAS due to more complex regulation. CWAS+/TWAS- regions were enriched for TF genes, which are depleted for eQTLs[58], and contained key prostate developmental genes such as *NKX3-1*, *KLF5*, and *HOXB13* (Fig. 5K). Collectively, these results support a model where disease risk is mediated by context-dependent eQTLs that are not observable from steady-state expression, but can be identified in steady-state chromatin (Fig. 5L).

CWAS implicates developmental genes in prostate cancer

The advantages of chromatin models described above allowed CWAS to implicate genes involved in prostate development and oncogenesis that have not been mechanistically tied to prostate cancer GWAS associations. Several such genes, including *MYC*[59], *KLF5*[60], *NKX3-1*[61], *CCND1*[62], *HOXB13*[63], and *GATA2*[64], physically interacted with CWAS ARBS and/or H3K27ac peaks, as assessed by H3K27ac Hi-ChIP (Fig. 6A–F). Conditioning GWAS SNP associations upon the genetically predicted peak intensity left little or no residual GWAS significance in these regions, suggesting that regulatory element activity accounts for prostate cancer heritability at these sites.

Importantly, the above genes have not been tied to prostate cancer heritability by robust eQTLs and TWAS associations. These genes demonstrated low *cis*-SNP heritability of steady-state expression measurements, a likely reason they were not detected by TWAS (Fig. 6G). In contrast to gene expression, several peaks associated with the genes above were highly heritable with respect to *cis*-SNPs (Fig. 6G). Notably, disruption of the CWAS ARBS ~220kb centromeric to *MYC* containing the variant rs11986220 was recently shown to impair *MYC* expression, proliferation, and tumorogenesis in a cell-line dependent manner[65]. This finding supports the hypothesis that this ARBS contributes to prostate cancer risk. Thus, CWAS implicated biologically plausible prostate developmental genes and proto-oncogenes that have been overlooked by analyses based on steady-state expression.

CWAS annotates additional AR-driven phenotypes

We applied AR CWAS to additional phenotypes (Fig. 7A) and implicated ARBS in diseases and traits known to be driven by androgen signaling. We identified known and novel regions (N=45) associated with testosterone levels among male UK Biobank participants[66] (Fig. 7B; Table S8). The most significant ($p = 3 \times 10^{-28}$) is an ARBS that contacts *JMJD1C*, a gene with roles in testis development and steroid hormone metabolism[67] that has been associated with testosterone levels in prior GWAS[68]. Additional CWAS ARBS interacted with genes implicated by GWAS, including *SHBG*, which encodes sex hormone binding globulin. For 7 of these peaks (16%) a significant GWAS association was not detectable within 1Mb. These novel hits included an intergenic ARBS contacting the promoter of *YAP1* (Fig. 7C) a gene involved in steroid hormone biosynthesis[69].

Separately, CWAS of benign prostate hypertrophy (BPH) – another androgen-mediated disease – identified two ARBS associated with this disease (Fig. 7D; Table S9). The most significantly associated ARBS ($p = 2 \times 10^{-8}$) was in an intergenic region that physically interacts with the *FGFR2* promoter in LNCaP cells (Fig. 7E) and benign prostate tissue based on Hi-C data[70]. *FGFR2* encodes a receptor highly expressed in prostate stroma that is implicated in the development of BPH[71]. The other BPH-associated ARBS localized to an intergenic enhancer of the prostate lineage TF gene *NKX3-1*[45], which has not been implicated in BPH previously. These results demonstrate that CWAS identifies ARBS that account for heritability at known and previously unknown risk loci for androgen-related phenotypes.

We reasoned that the enhanced statistical power of CWAS would enable the study of heritability among small populations that are inadequately powered for GWAS. To this end, we applied CWAS to identify genetic determinants of response to androgen deprivation therapy (ADT) among 687 patients with metastatic prostate cancer[72,73]. No SNPs were associated with ADT response by GWAS at genome-wide significance threshold of $p < 5 \times 10^{-8}$. To increase power, we applied CWAS to regions within 1Mb of the 200 most significant SNPs, with Bonferroni correction for 475 tested ARBS. This approach nominated an intronic ARBS in *NAALADL2* that was significantly associated with time to progression on ADT ($p = 7.8 \times 10^{-5}$; hazard ratio 1.29; 95% CI 1.13 – 1.46) Fig. 7F; Table S10). Expression of *NAALADL2* has been associated with increased grade and stage of prostate cancer, as well as earlier recurrence[74,75]. Notably, a prior GWAS of prostate cancer aggressiveness identified an association at this gene ($p = 4.18 \times 10^{-8}$)[75]. This finding highlights the power of CWAS for studying therapeutic resistance and other features of interest in small but well-annotated groups such as clinical trial cohorts.

Discussion

We present the cistrome wide association study (CWAS), a principled and statistically powerful approach for associating the genetic determinants of regulatory element activity with trait heritability. Applying CWAS to prostate cancer implicated AR binding in 21% of all prostate cancer GWAS risk regions and regulatory element activity in an additional 32%, adding substantially to the number of prostate cancer risk loci that are annotated with plausible mechanisms. Genetic variation in one or a few ARBS accounted for prostate cancer risk at many loci identified by GWAS, such as regulatory elements near *MYC, TMPRSS2*, *GATA2*, and *NKX3-1*. We experimentally validated the predicted effect of cQTLs on gene expression for six regulatory elements and demonstrated that CWAS ARBS regulate candidate prostate cancer risk genes *TMPRSS2* and *BMPR1B*. AR CWAS also implicated AR binding sites and nearby genes in BPH, serum testosterone levels, and response to prostate cancer treatment.

CWAS is complementary to TWAS/eQTL-based approaches, which may miss associations involving genes with complex regulation and context-dependent expression[57,58]. These genes were depleted for genetic models of expression based on *cis*-SNPs, but contained the most nearby genetic models of AR binding or regulatory element activity. Strikingly, CWAS identified epigenome-trait association in the absence of a high-confidence transcriptome-trait (TWAS) association at 29% of prostate cancer risk regions. Compared to TWAS+ prostate cancer risk regions, genes in CWAS+/TWAS- regions were subject to more complex regulation and were enriched for transcription factors. This attribute allowed us to implicate key prostate developmental genes and proto-oncogenes in prostate cancer genetics that have largely been overlooked because their expression levels at steady state are highly regulated correlate poorly with *cis*-SNPs.

We hypothesize that cQTLs in CWAS+/TWAS- prostate cancer risk regions are context-dependent eQTLs. These variants may affect gene expression in specific tissues or cellular conditions that are relevant to prostate cancer, but their effects are obscured at steady state. The *NKX3-1* enhancer provides an example. Mutation of rs1160267 – a cQTL within the enhancer – modestly affects NKX3-1 expression at steady state, but this effect is amplified with androgen stimulation[76]. Context-dependent eQTLs frequently alter chromatin "priming" in the absence of stimuli required to elicit effects on gene expression[39], potentially explaining how the effects of these variants are visible in steady-state chromatin. Our androgen stimulation experiments provide additional evidence of this phenomenon. Transcripts with androgen-induced allelic imbalance tend to harbor nearby regulatory elements that are already imbalanced in the absence of stimulation.

Our approach has several limitations. First, epigenomic peak intensity may correlate with, but not mediate risk. Pleiotropic effects of variants that alter chromatin but affect risk through an independent mechanism are plausible and future studies will be required to determine their prevalence. A second limitation is that epigenomic reference panels from many individuals do not yet exist for most tissues and TFs, especially for populations of non-European ancestry. Ongoing efforts to perform epigenomic profiling on genetically diverse tissues will advance the utility of this approach further.

The strategy we describe charts a path for future analyses to uncover mechanistic insights into the thousands of variant-trait associations that lack explanatory steady-state eQTLs. While we focused on prostate cancer and AR, CWAS can be applied in a vast range of contexts. Because transcriptional biology often underlies complex phenotypes, CWAS should be a powerful and generalizable approach to ascertaining mechanisms of trait and disease heritability. Chromatin conformational data can be used to link risk-associated regulatory elements to genes. Importantly, our method for imputing genotypes from ChIP-seq data allows CWAS to leverage existing ChIP-seq datasets that lack genotyping information. Finally, the increased power for discovery afforded by CWAS unlocks the ability to study the genetics of human disease in smaller populations of interest, such as patients enrolled in clinical trials.

Methods

ChIP-seq peak calling

ChIP-seq fastq files from ref[44] were downloaded from SRA using SRA toolkit fastq dump v 2.10.0. For uniformity, only the first read in a pair was used for paired-end sequencing datasets. Epigenomic datasets previously generated by our group were processed as described[45,77]; these data are also available in GEO under accession numbers GSE130408 and GSE161948. ChIP-seq reads were aligned to the human genome build hg19 using the Burrows-Wheeler Aligner (BWA) version 0.7.17[78]. Non-uniquely mapping and duplicate reads were discarded. MACS v2.1.1.20140616[79] was used for ChIP-seq peak calling with a q-value (FDR) threshold of 0.01. ChIP-seq data quality was evaluated by a variety of measures, including total peak number, FrIP (fraction of reads in peak) score, number of high-confidence peaks (enriched > ten-fold over background), and percent of peak overlap with DHS peaks derived form the ENCODE project. IGV v2.8.2[80] was used to visualize normalized ChIP-seq read counts at specific genomic loci. Overlap of ChIP-seq peaks and genomic intervals was assessed using BEDTools v2.26.0. Peaks were considered overlapping if they shared one or more base-pairs. Fisher's test for overlap was performed using the BEDTools fisher command.

Genotype imputation

We imputed genotypes at 5,495,776 autosomal SNPs present at minor allele frequency > 5% in the Haplotype Reference Consortium (HRC) v1.191[46]. Bam files from epigenomic datasets were merged for each individual using SAMtools merge and run through STITCH v1.6.2[43] with the following parameters: k=10, ngen=1240, niterations=40, method=diploid (https://hub.docker.com/r/stefangroha/stitch_gcs/tags). The imputation reference panel contained haplotypes of 2,505 individuals in the 1000 Genomes Project Phase 3[81].

To ensure that individual bam files were correctly assigned to an individual, we used the mpileup and call functions from bcftools v1.9 to call genotypes at 100,000 SNPs and bcftools gtcheck function to test pairwise correlation of homozygous SNP across all files. Samples were clustered based on correlation. Six bam files out of 581 that clustered in a cluster of a different individual were excluded from the analysis.

24 samples were subject to genotyping with Infinium Global Screening Array-24, version 1.0 (Illumina) at the Broad Institute Genomic Services, Cambridge, MA. The Pearson correlation coefficient of allele dosages between imputed and array-based genotypes was evaluated using the R function cor(). A receiver operating characteristic curve was constructed comparing the true positive fraction vs. false positive fraction across cutoffs for genotype dosages.

These steps are implemented in a pipeline available at https://github.com/scbaca/chip_imputation.

Genetic models of epigenomic features

Total and allele-specific peak intensity for H3K27ac and AR were modeled based on *cis*-SNP genotypes in the following steps, which are incorporated into a Snakemake[82] workflow available at https://github.com/scbaca/cwas.

Consensus peak calling—We create a consensus set of H3K27ac and AR by dividing the genome into 50bp windows and including any window with peaks in > 5% of samples. Windows were buffered by 100bp and merged to create a set of 48,948 AR peaks and 81,150 H3K27ac peaks.

Allelic imbalance analysis—ChIP-seq reads were analyzed for imbalance of heterozygous SNP alleles using stratAS[40] (https://github.com/gusevlab/stratAS). Several upstream steps were performed to boost power and accuracy of allelic imbalance detection. Imputed SNP genotypes were phased with Eagle2[83] using the Sanger Imputation Service (https://imputation.sanger.ac.uk/). Heterozygous SNPs were filtered for mapping bias via the WASP pipeline[84] and allele-specific read counts were tabulated using ASEReadCounter from the Genome Analysis Toolkit v3.8103[85].

Briefly, stratAS identifies allelic imbalance by modeling the reads from heterozygous SNPs with a beta-binomial distribution. At each ChIP-seq peak, stratAS takes advantage of haplotype phasing to sum read counts from nearby heterozygous SNP alleles on the same haplotype for each individual. stratAS models the reads from individual *i* overlapping heterozygous germline SNP j as: $R_{alt,i} \mid R_{ref,i}\ BetaBin(\pi_j, \rho_{ij})$, where π is the mean allelic ratio and ρ is a locally-defined, per-individual sequence read correlation parameter reflecting over-dispersion.

Copy number profiles were estimated from off-target ChIP-seq reads with CopywriteR[86] and used in the modeling of the over-dispersion parameter ρ, in order to account for over-dispersion in regions of cancer-associated copy number alterations. ρ is estimated for each individual from all heterozygous read-carrying SNPs across ten declines of estimated copy number levels stratAS params.R script, with the following options: --min_snps 50, --min_cov 5, --group 10.

We tested variants with ≥ 20 informative reads within consensus AR and H3K27ac peaks defined above for imbalance. The following additional parameters were set for the stratas.R script: --max_rho 0.2, --window −1, min_cov 1, and --fill_cnv TRUE.

Allelelic imbalance p-values were FDR-adjusted with the qvalue R package (v2.18). Peaks were considered significantly imbalanced if they contained one or more SNPs with imbalance at $q < 0.05$.

Imbalanced SNPs in TF binding motifs—Homer v4.10 was used to identify the most significantly enriched motifs *de novo* among a random selection of 10,000 AR consensus peaks. Imbalanced heterozygous SNPs were tested for overlap with one of these motifs for either allele. Where heterozygous SNPs overlapped, the difference in PWM score between reference and alternate alleles was compared to the allele fraction of reference vs. alternate alleles.

cQTL detection—QTLtools v1.2[87] was used for cQTL detection. Rpkm for each sample at AR and H3K27ac consensus peaks was calculated for each bam file using QTLtools quan with the following flags: --filter-mismatch 5 --filter-mismatch-total 5 --filter-mapping-quality 30. Peaks with a summed rpkm < 10 across all samples were discarded. A covariate matrix was constructed using QTLtools pca --scale --center. Permutation-based p-values[87] for SNP-peak pairs within a 1Mb window were assessed for cQTLs with QTLtools cis (--normal --permute 1000) after regressing out the first 6 principal components of the peak rpkm covariate matrix. We plotted the distribution of distances between these cPeak-cQTL pairs. After finding that the majority of cQTLs SNPs were within 25kb of the corresponding peak, we also took a focused approach and calculated nominal p-values for *cis*-snp pairs within 25kb, forgoing permutation, which was often not possible for at a distance of 25kb due to a limited numbers of peaks for permutation. These p-values were adjusted by FDR correction and included in downstream analysis where $q < 0.05$.

For peaks that were tested for both allelic imbalance and cQTLs, combined significance was assessed by combining p-values from the two tests combined using Stouffer's method[47,88].

cQTL peak enrichment analysis—Enrichment of eQTL SNPs in cPeaks was tested by permutation. We counted the number of eQTLs for each tissue type overlapping AR or H3K27ac cPeaks and divided this number by the total base-pairs covered by these peaks.

We then performed this process on 5,000 same-sized samplings of the complete set of AR or H3K27ac peaks to generate a null distribution. We reported the ratio of peak territory containing a cQTL SNPs in the observed versus simulated data to calculate enrichment and a one-sided p-value. We also calculated enrichment compared to random background by repeating this process using random intervals matched to cPeaks for size, number, and chromosome.

CWAS model construction

Conventional TWAS models train a predictor of gene expression. Here we extended these models to additionally incorporate allele-specific information and a chromatin phenotype (similar to recent models proposed in the context of statistical fine-mapping[24] and gene expression[49]). For a given chromatin peak, we take as input the following: a vector of total chromatin activity y_{total}, with each row containing an individual; the vector of allelic chromatin activity $y_{allelic}$, defined as $\log\left(N_p / N_m\right)$ where N_* is the total number of reads mapping to the heterozygous variants of the maternal/paternal haplotype, and undefined otherwise; and the matrices of phased maternal and paternal haplotypes $\boldsymbol{H}_p$ and $\boldsymbol{H}_m$, with individuals as rows and variants within the locus window as columns, containing 0/1 indicators for reference or alternative alleles. We note that maternal or paternal haplotypes can be defined arbitrarily as long as the definition is consistent between the phased genotyped and the allelic reads. In model 1 ("cQTL model"), the relationship between total chromatin activity and genotype is modelled $y_{total} \sim \boldsymbol{X}_{total} + \epsilon$ where $\boldsymbol{X}_{total} = \boldsymbol{H}_p + \boldsymbol{H}_m$ and corresponds to the 0/1/2 allelic dosage for each sample and variant. This model is identical to the models used for conventional TWAS prediction. In model 2 ("allelic imbalance model"), following ref[24] and ref[49], the relationship between allelic chromatin activity and haplotype is modelled as $y_{allelic} \sim \boldsymbol{X}_{allelic} + \epsilon$, where = – and corresponds to the −1/0/1 allele phase. Lastly, in model 3 ("combined model"), we define a "combined" model as $\begin{bmatrix} \tilde{y}_{total} \\ \tilde{y}_{allelic} \end{bmatrix} \sim \begin{bmatrix} \widetilde{\boldsymbol{X}}_{total} \\ \widetilde{\boldsymbol{X}}_{allelic} \end{bmatrix} + \epsilon$, where the twiddle over a variable indicates scaling the columns to zero mean and unit variance. Each model was then fit using LASSO penalized regression to learn genotype to phenotype predictor weights W across all variants included in the model (previous work has shown that LASSO models preform comparably to other penalization schemes[89]). Predictive accuracy was evaluated by five-fold cross validation and quantified as the Pearson correlation to the true y_{total} or $y_{allelic}$ phenotype. All other model parameters (specifically the LASSO penalty) were fit by nested cross-validation within each training fold.

This analysis is implemented using stratAS with the --predict flag, with --window set to 25kb to include SNPs within 25kb of the peak center.

CWAS analysis

Integrative models of cQTL and AI were built as described above for each consensus AR or H3K27ac peak based on genotypes of *cis*-SNPs within 25kb (the number of significant models was largely insensitive to the window size, see Supplemental Note). We selected the model type with the most significant cross-validation p-value for each peak, and then retained only models with cross validation significance at an FDR of 0.05 across all peaks. The genetic association between predicted peak cQTL activity or AI and GWAS risk was calculated by FUSION, accounting for linkage disequilibrium[50,53]. FUSION considers the Z-score for genetic peak-trait association as

$$Z_{peak\ \rightarrow\ trait} = W Z_{snps\ \rightarrow\ trait}$$

where $Z_{snps\ \rightarrow\ trait}$ is a vector of snp-trait association Z-scores from GWAS summary statistics

$$Z_{snps\ \rightarrow\ trait} = \begin{bmatrix} Z_{snp1\ \rightarrow\ trait} \\ \vdots \\ Z_{snpn\ \rightarrow trait} \end{bmatrix}$$

and W is a weight matrix defined as

$$W = \sum_{p,\,s} \sum_{s,\,s}^{-1}$$

$\sum_{p,\,s}$ is the peak-snp covariance matrix, and $\sum_{s,\,s}$ is snp-snp covariance matrix, representing linkage disequilibrium. In practice, W is learned from the data through penalized regression. Assuming a normal distribution of $Z_{peak\ \rightarrow\ trait}$ around 0, then Z-score for a peak-trait CWAS association is

$$Z_{CWAS} = \frac{W Z_{snps\ \rightarrow trait}}{var\left(W Z_{snps\ \rightarrow\ trait}\right)} = \frac{W Z_{snps\ \rightarrow trait}}{\left(W \Sigma_{s,\,s} W^{T}\right)^{1/2}}$$

and the corresponding two-sided p-value is obtained from the normal distribution $N(0,1)$. CWAS associations were considered significant if p < 0.05 after Bonferonni correction for all peaks of a given type tested (N=5,580 for AR and 17,199 for H3K27ac).

GWAS datasets used in this study are listed in Table S1.

Overlap of GWAS, TWAS, and CWAS results

Genome-wide significant SNPs ($p < 5 \text{ x } 10^{-8}$) were obtained form published GWAS summary data[52], assigned hg19 coordinates buffered with 1Mb windows on either side, and merged where windows overlap to obtain 98 prostate cancer GWAS risk regions. Each region was evaluated for overlap with one or more high-confidence CWAS peaks (AR or H3K27ac) or TWAS genes (from prostate tumor reference panels, or panels incorporating all available tissues, and including splicing eQTLs). High-confidence peaks and genes were defined as those where or was greater than for the most significant GWAS SNP in the region. We elected not to threshold based on statistical colocalization because: (1) no colocalization method currently incorporates allele-specific signal; (2) colocalization methods are highly dependent on the molecular study size and underpowered for hundreds of samples[90]; and (3) colocalization probabilities are highly conservative even in large GWAS[91]. Our high-confidence regions should thus be interpreted as being consistent with explaining the majority of the GWAS variance at the locus.

Prostate cancer risk loci with significant CWAS associations but no significant GWAS associations were evaluated in a large prostate cancer GWAS that was published after this manuscript was prepared[56]. The 269 independent risk variants reported in the ref.[56] were buffered with 1Mb windows. AR and H3K27ac CWAS peaks were evaluated for overlap with these windows to identify peaks with nearby SNPs that were significant only in the larger GWAS.

Androgen deprivation therapy GWAS

Men who received androgen deprivation (ADT) for metastatic hormone-sensitive prostate cancer (N=687) from two cohorts were evaluated. 265 of these patients were from the control arm of the CHAARTED clinical trial (E3805)[72]. The remaining 422 were patients treated at Dana-Farber Cancer Institute. The study was performed under IRB-approved protocols that included informed consent for genotyping. These patients were selected to match enrollment criteria for CHAARTED. Subjects were genotyped at approximately one million SNPs with minor allele frequency ≥ 0.05 on Affymetrix 6.0 arrays. Genotypes for SNPs interrogated on the array were called using the Birdsuite algorithm. Alignment to the hg19 genome build was checked using tools provided in SHAPEIT. Strands were flipped using plink when necessary. SHAPEIT was used to pre-phase the SNPs using 1000 Genomes Phase 3 panel as the reference, followed by imputation using IMPUTE (v2.3.1). Time to progression, as assessed in the trial, was evaluated for association with genotypes with the Cox proportional hazards model implemented by the ProbABEL R package[92]. The square root of the corresponding χ^2 statistics were used as the GWAS summary statistics for CWAS analysis.

To limit hypothesis testing, we restricted CWAS association testing to CWAS AR peaks within 1Mb of the top 200 GWAS SNPs by significance (N=789 peaks).

CRISPRi suppression of ARBS

The gRNA sequences used to target CWAS enhancers were identified using the CRISPick algorithm (https://portals.broadinstitute.org/gppx/crispick/public). The highest scoring gRNAs near the center of a given peak were selected. The gRNA sequences (Table S5) were synthesized as single stranded oligonucleotides (IDT DNA) with compatible sticky ends (for detailed protocol see https://www.broadinstitute.org/rnai/public/resources/protocols). Annealed oligonucleotides were cloned into lenti_U6sg-KRAB-dCas9-puro using Esp3I. Insert sequences were confirmed by Sanger sequencing performed by the CCR Genomics Core at the National Cancer Institute.

Lentivirus was produced by transfecting 293T cells with the gRNA and KRAB-dCas9 expression plasmid together with the packaging plasmids VsVg (Addgene 12259) and psPax2 (Addgene 12260) using TransIT-LT1 transfection reagent (Mirus). Supernatant containing virus was harvested 48 hours following transfection and used to transduce the LNCaP cell line in the presence of 4 mg/ml polybrene and media exchanged after 24 hours. Conditions were optimized to ensure > 95% transduction as assessed by selection with puromycin. RNA was isolated 4 days after transduction using QIAGEN RNeasy Plus Kit and cDNA synthesized using NEB Protoscript II First Strand cDNA Synthesis Kit. Quantitative PCR was performed on a Quantstudio 6 using SYBR green. Primers used for qRT-PCR are listed in Table S5. A nontargeting gRNA and gRNA targeting an intergenic region were used as negative controls. Gene expression was normalized to GAPDH and DDCt values were calculated using the nontargeting gRNA as the control sample. Data from three independent biological replicates were used to determine average fold change and data represent the average and standard deviation with significance determined by Student's t test.

LNCaP DHT stimulation

LNCaP cells (ATCC CRL-1740) were cultured in phenol red free RPMI (#11835030, Gibco) with 10% charcoal stripped FBS (#100-119, Gembio) for 3 days. then were stimulated with either 10 nM DHT (5α-Androstan-17β-ol-3-one, Dihydrotestosterone, A8380, Sigma) or EtOH (Vehicle) for 16 hours. Subsequently cells were collected for further analysis accordingly. LNCaP cells were authenticated by comparing short tandem repeats to parental LNCaP cells in the ATCC database. Prior to experiments, cells tested for several strains of mycoplasma contamination using LookOut Mycoplasma PCR Detection Kit (Sigma-Aldrich #D9307).

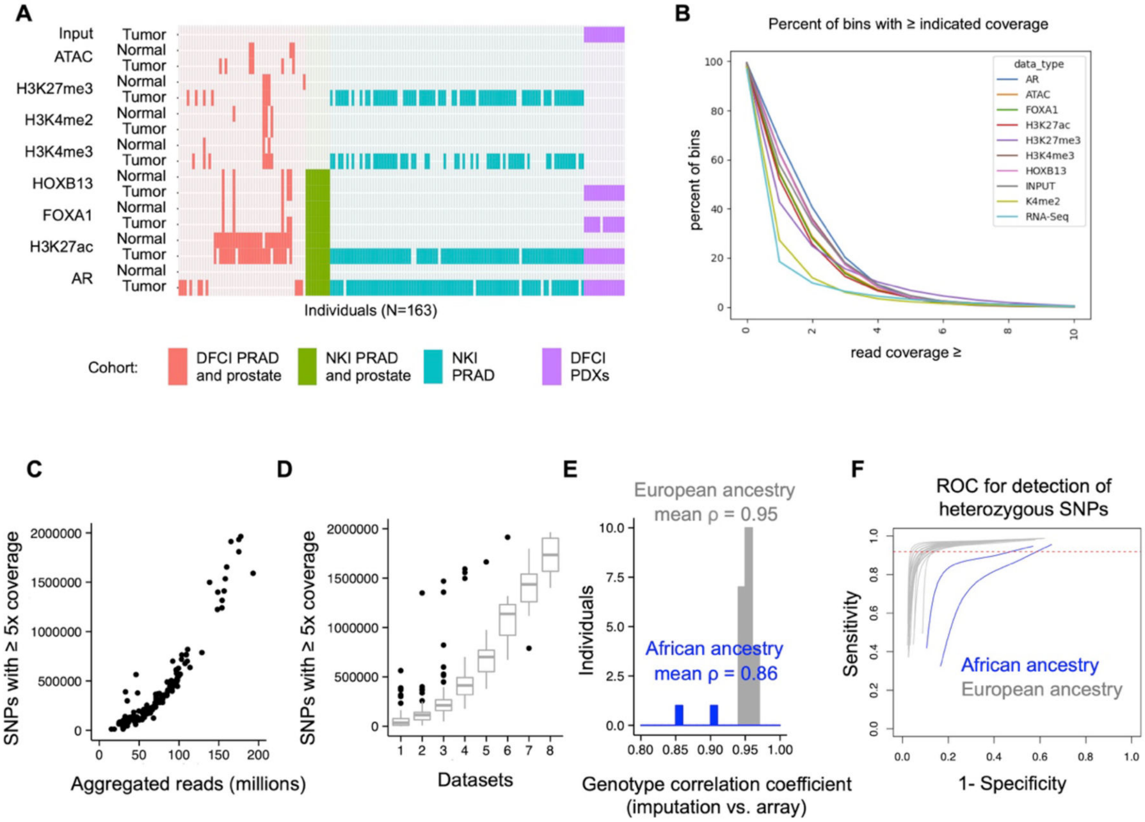
ChIP-seq in LNCaP was performed as previously described[77]. Briefly, Ten million cells were fixed with 1 % formaldehyde at room temperature for 10 minutes and quenched with 0.25M glycine, Harvested cells in lysis buffer (1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS and protease inhibitor (#11873580001, Roche) in PBS) were sheared to 300–800 bp chromatin using a Covaris E220 sonicator (140 watt peak incident power, 5% duty cycle, 200 cycleburtst). Sonicated chromatin was subjected to H3K27ac antibody (C15410196, Diagenode) coupled with Dynabeads protein A/G (Life Technology # 10001D, 10003D) overnight at 4 °C. Chromatin was washed in LiCl wash buffer (100 mM Tris pH 7.5, 500 mM LiCl, 1% NP-40, 1% sodium deoxycholate) 6 times for 10 minutes sequentially. Immuno-precipitated chromatin and input were treated with RNase A at 37 °C for 30 minutes and decrosslinked in elution buffer (1% SDS, 0.1 M NaHCO3) with proteinase K for 6–12 hours at 65 °C with gentle rocking. DNA was purified using Qiagen Qiaquick columns (#28104). Libraries were prepared using SMARTer ThruPLEX DNA-Seq Kit (Takara Bio # R400675)

ATAC seq libraries were prepared using Omni-ATAC protocol[93]. Freshly collected 50,000 nuclei in cold lysis buffer (10 mM Tris-HCl, pH 7.5, 10 mM NaCl, 3 mM MgCl2, 0.1% NP-40, 0.1% Tween20, 0.01% Digitonin) were fragmented in 50 μl of transposition mix (25 μl 2× TD buffer, 16.5 μl PBS, 0.5 μl 1% digitonin, 0.5 μl 10% Tween-20, 5 μl water) with 2.5 μl transposase (Illumina 20034197) for 30 min at 37 °C with shaking at 1000 r.p.m. in a thermomixer. DNA was purified using Qiagen MinElute (#28004) and libraries were amplified up to the cycle number determined by 1/3rd maximal qPCR fluorescence

Total mRNA was collected from 300,000 cells using RNA easy kit (Qiagen 74044) with RNase-Free DNase Set (Qiagen no. 79254) according to the manufacturer instructions. RNA purity and concentration were determined on 2100 Bioanalyzer (Agilent) using Agilent RNA 6000 Nano Kit # 5067-1511). 400ng RNA samples were submitted to Novogene for RNA library preparation.

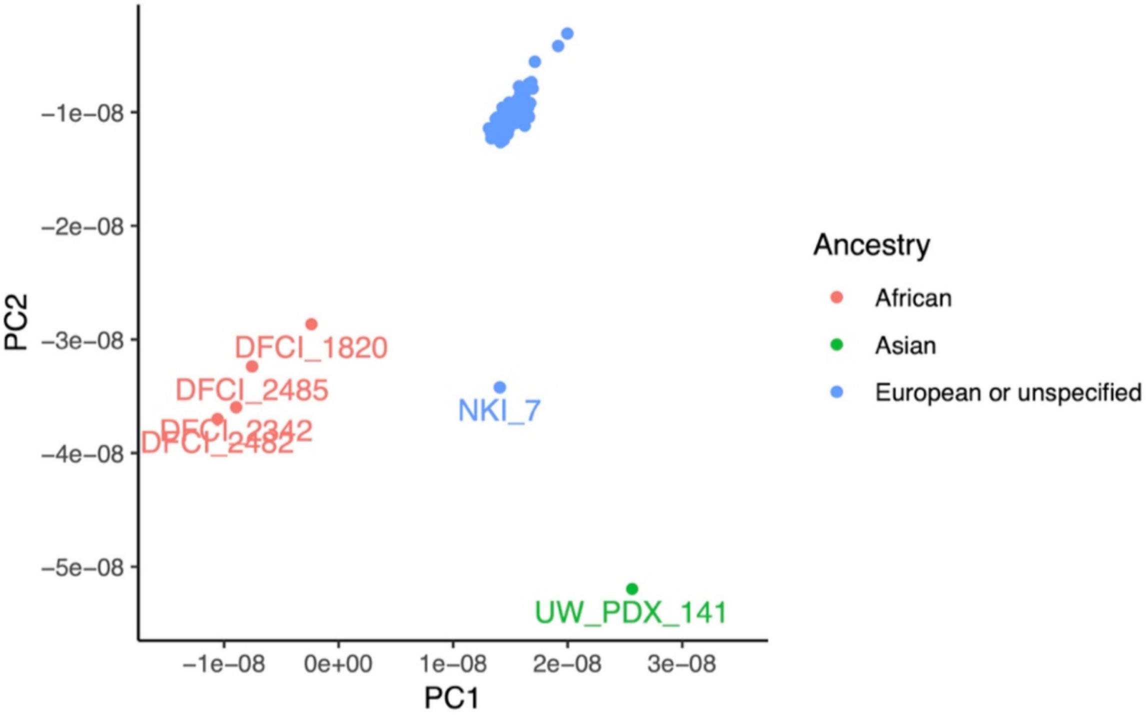
ChIP-seq, RNA-seq, and ATAC-seq libraries and sequenced with 150bp paired-end reads on a HiSeq 250 instrument (Novogene). ChIP-seq and ATAC-seq peaks were called using MACS2 as described above and allelic imbalance in peaks and gene expression was evaluated using stratAS[40].

Extended Data



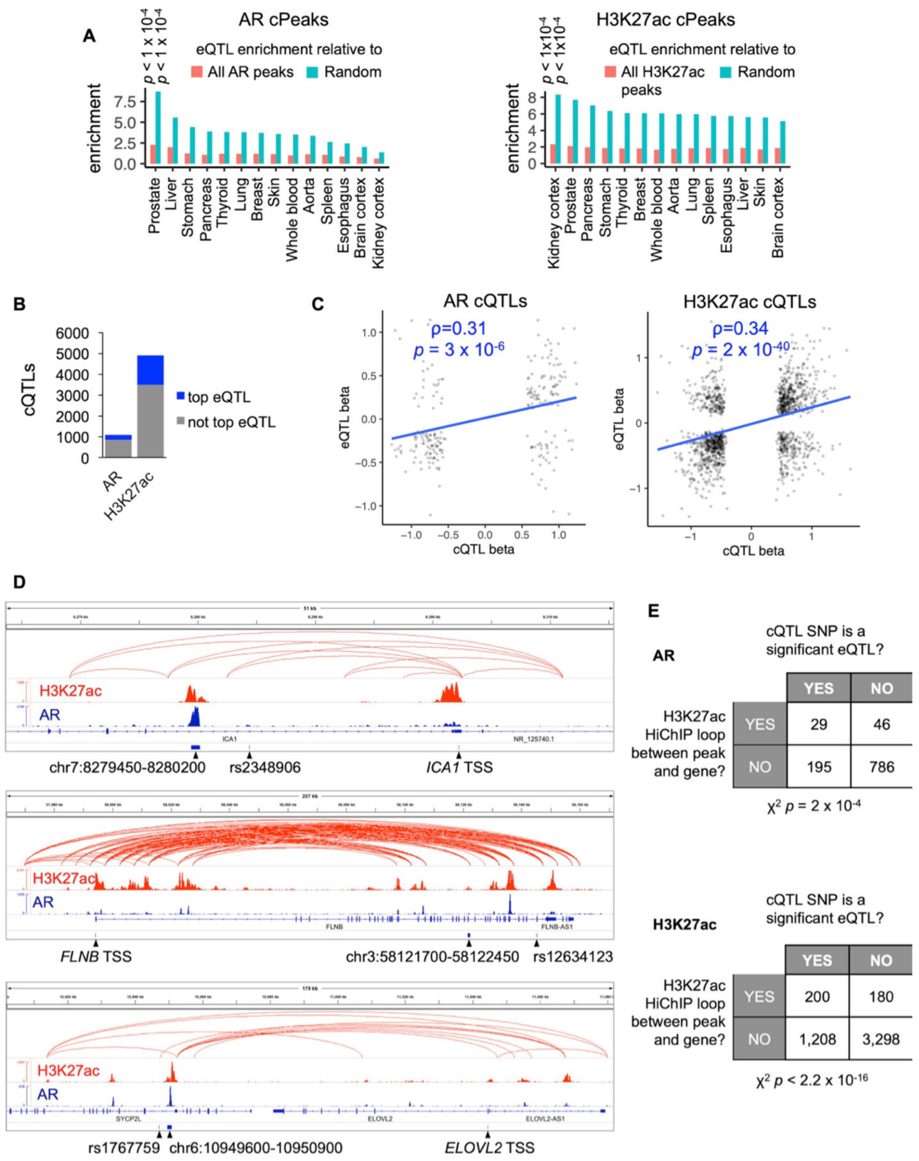
Extended Data Figure 1. Accurate genotyping of SNPs from epigenomic data.

(**A**) Overview of 575 epigenomic datasets merged across 163 individuals for genotyping. Datasets are colored by cohort (See Table S1). (**B**) Genomic distribution of reads in ChIP-seq, RNA-seq and input control (whole genome) data. The genome was divided into non-overlapping 500 base-pair windows and cumulative read counts for each bin were summed. For each datatype, five samples were randomly selected and down-sampled to 8.4 million reads for uniformity. The mean percentage of bins with the indicated number of read counts is shown for each datatype. (**C**) Number of covered SNPs (≥ 5 reads) versus total aggregated reads for each individual. (**D**) Number of covered SNPs (≥ 5 reads) for each individual (n=165) as the indicated number of datasets are merged. Datasets were added in random order for a given individual. For boxplots, lower and upper hinges indicate 25th and 75th percentiles; whiskers extend to 1.5 x the inter-quartile ranges (IQR). (**E**) Correlation of imputed versus array-based genotype dosages across 24 individuals. (**F**) Receiver operating characteristic curve for detection of heterozygous SNPs using sequencing and imputation, with array-based genotypes as ground truth. Dotted red line indicates a mean sensitivity of 0.92 at a specificity of 0.9 in individuals of European ancestry.



Extended Data Figure 2. Inferred ancestry of individuals in the study.

Projection of imputed genotypes onto the first two principal components of continental ancestry from ref.[78]. Individual identifiers for outlier samples (with values > 2 x standard deviation) are labeled. Self-reported ancestry is coded by color.



AR

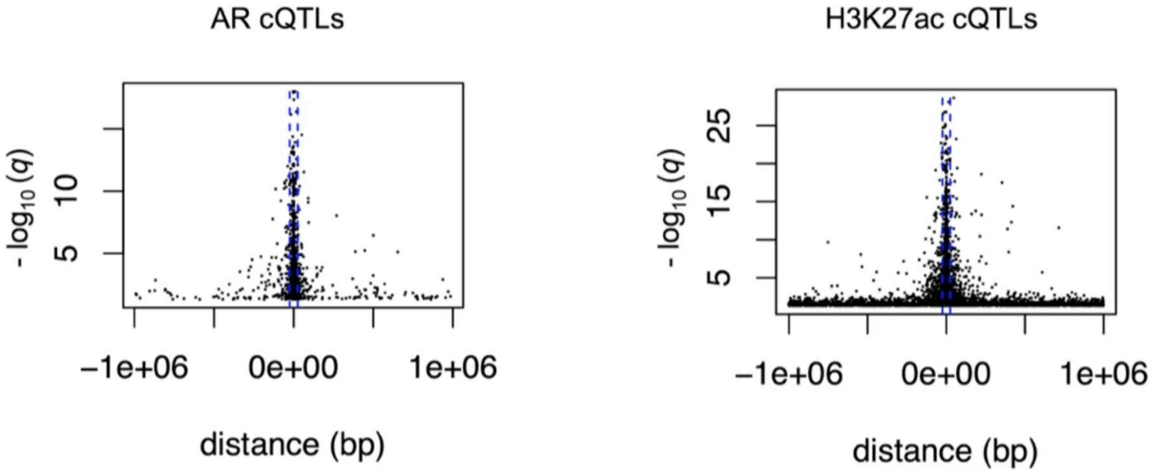
H3K27ac HiChIP loop between peak and gene?	YES	NO
YES	29	46
NO	195	786

H3K27ac

H3K27ac HiChIP loop between peak and gene?	YES	NO
YES	200	180
NO	1,208	3,298

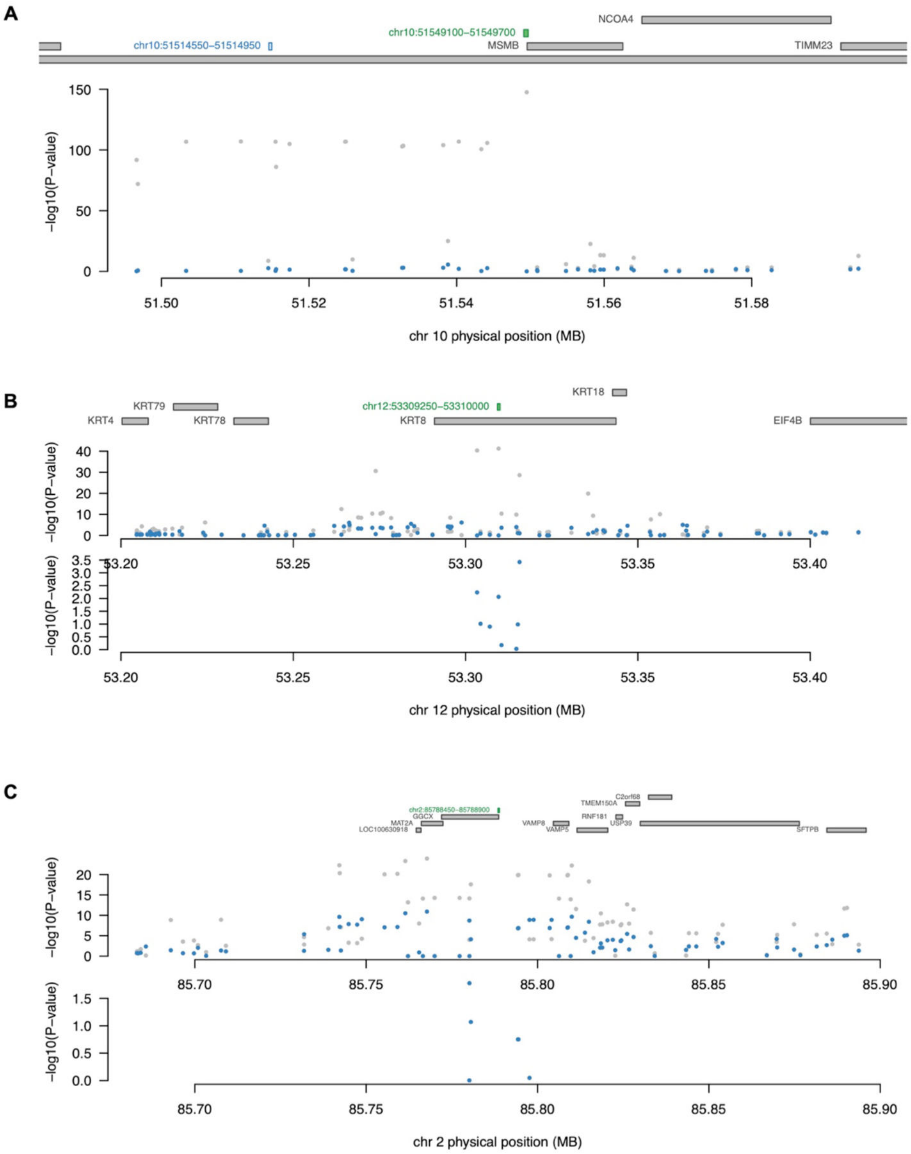
Extended Data Figure 3. Overlap of cQTLs with prostate tissue eQTLs.

(**A**) Enrichment of genetically determined AR peaks (left) and H3K27ac peaks (right) for overlap with GWAS risk SNPs eQTLs across various tissues. Empiric p values are derived 10,000 from permutations. (**B**) number of AR and H3K27ac cQTLs that are also the top eQTL for a gene in prostate tissue. (**C**) correlation of cQTL and eQTL effect size (β) for cQTL SNPs; p-value for Pearson correlation test is indicated. (**D**) Examples of SNPs (labeled with rs identifier) that are both AR cQTLs and eQTLs where the corresponding cPeak and eGene are connected by an H3K27ac HiChIP loop in LNCaP. cPeak coordinates are shown and eGene transcriptional start sites (TSS) is denoted. (**E**) Contingency table showing enrichment of H3K27ac HiChIP looping between the corresponding cPeak and eGene for cQTLs that are also eQTLs. Chi-square test p-values are indicated.



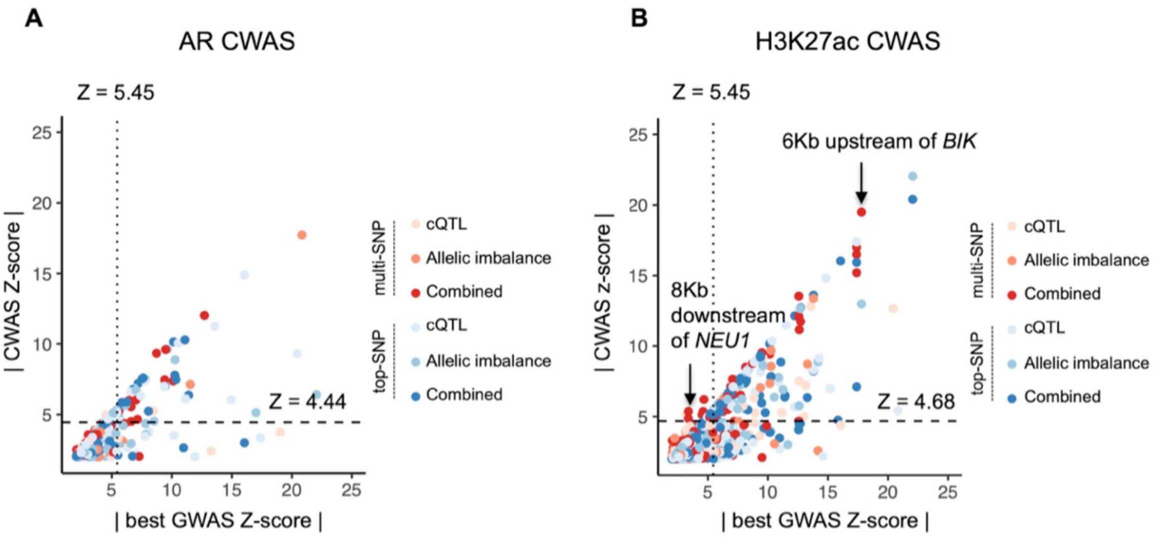
Extended Data Figure 4. Distribution of cQTLs around cPeaks.

cQTL SNP significance versus distance to the center of the corresponding cPeak for significant cQTLs (permutation-based q-value < 0.05). Dashed blue lines indicate ± 25Kb from the peak center.



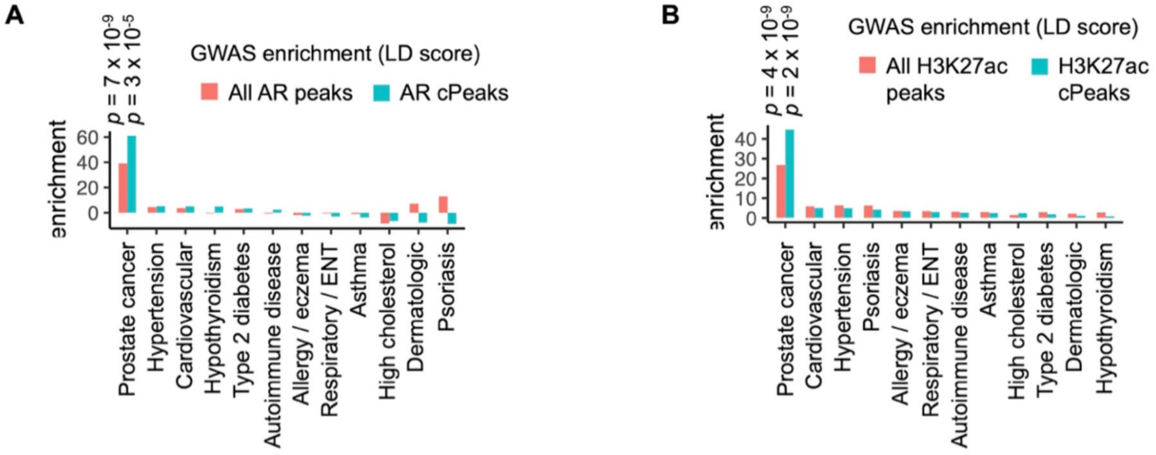
Extended Data Figure 5. Conditioning of GWAS SNP significance on genetically predicted CWAS AR binding.

Genomic context of AR CWAS ARBS (depicted in green) that are significantly associated with prostate cancer risk. Manhattan plots indicate significance of SNP associations with prostate cancer before and after conditioning on genetically predicted CWAS ARBS activity. (**A**) and (**B**) show representative examples where ARBS explain most of the nearby *cis*-SNP GWAS significance. (**C**) CWAS ARBS at the promoter of *GGCX*, where residual GWAS significance remains after conditioning on ARBS, suggesting additional mechanisms underlying risk conferred by SNPs in this region.



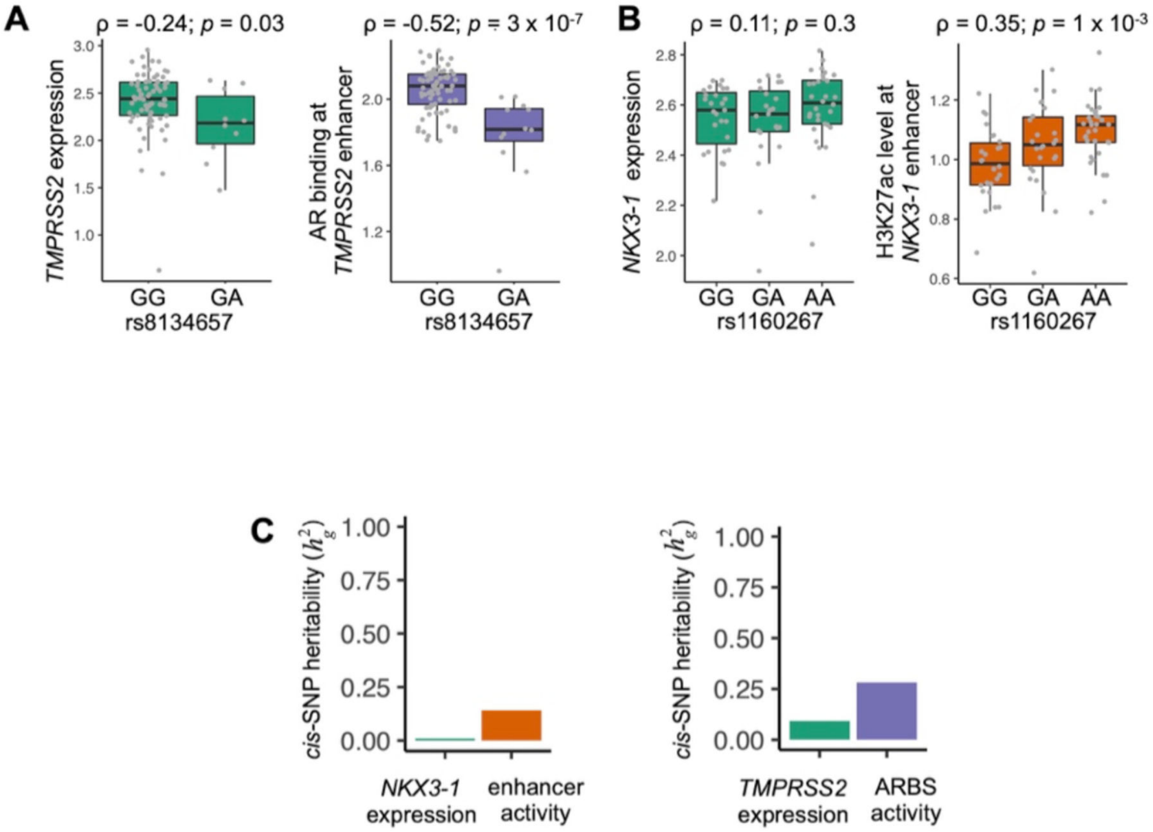
Extended Data Figure 6. Comparison of CWAS and GWAS significance for tested ARBS and H3K27ac peaks.

The absolute value of the association Z-score is plotted for CWAS peak-trait associations (y-axis) and GWAS SNP-trait associations for the most significant nearby SNP (x-axis). (**A**) shows ARBS and (**B**) shows H3K27ac peaks. Dashed horizontal lines indicate genome-wide significance thresholds for CWAS. Vertical dotted lines indicate the GWAS significance threshold of $z = 5.45$.



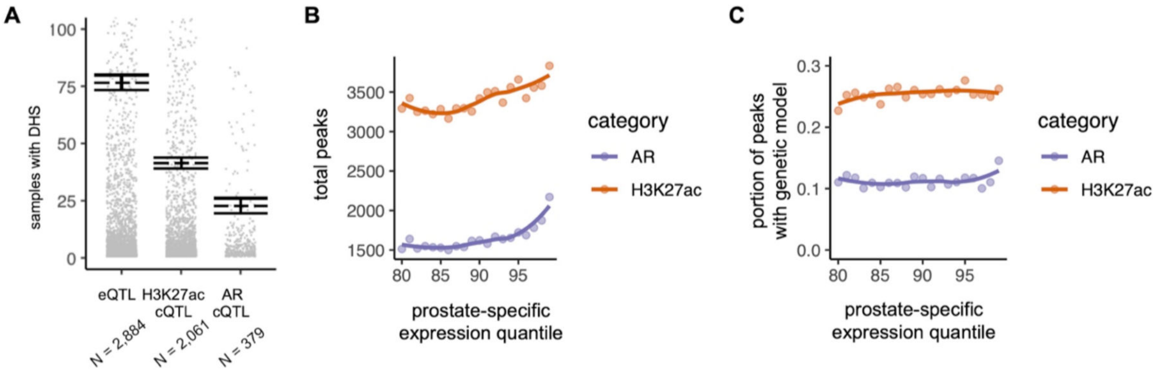
Extended Data Figure 7.

Enrichment of prostate cancer GWAS risk SNPs in genetically determined AR peaks (**A**) and H3K27ac peaks (**B**). Enrichment and p-values are derived from linkage disequilibrium score regression[5].



Extended Data Figure 8. cQTL *vs.* eQTL activity at *TMPRSS2* and *NKX3-1* loci.

(**A**) Normalized AR ChIP-seq reads at the *TMPRSS2* enhancer and *TMPRSS2* expression stratified by genotype of the indicated SNP. (**B**) Normalized H3K27ac ChIP-seq reads at the *NKX3-1* enhancer and *NKX3-1* expression stratified by genotype of the indicated SNP. ρ and *p*-values indicates Pearson correlation coefficient for (A) and (B). (**C**) Estimated *cis*-SNP heritability for the indicated epigenomic features and corresponding genes. For boxplots, lower and upper hinges indicate 25th and 75th percentiles; whiskers extend to 1.5 x the inter-quartile ranges (IQR).



Extended Data Figure 9. CWAS identifies associations not marked by a steady-state eQTL

(related to Figure 5). (**A**) Number of ENCODE samples (N=733, representing 438 cell types/states) with DNAse hypersensitivity at cQTL and eQTL SNPs. The data shown are from Fig. 5E. The scale is adjusted and mean ± standard error shown to better visualize differences between the groups. (**B**) Total AR or H3K2a7c peaks within 100kb of a gene as a function of prostate-specific gene expression, as quantified in Fig. 5F. (**C**) Portion of peaks in (B) with a CWAS model.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work is supported by grants from the PhRMA Foundation and the Kure It Cancer Research foundation (S.C.B). The androgen deprivation GWAS was supported in part by the U.S. National Cancer Institute of the National Institutes of Health under the award numbers U10CA180820, U10CA180794, and UG1CA233180 (C.J.S.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We are grateful to the E3805: CHAARTED Investigators; the patients who participated in the trial; the Prostate Cancer Foundation Mazzone Awards; and Sanofi for partial financial support and supplying docetaxel for early use (C.J.S.). In addition, we acknowledge Public Health Service Grants CA180794, CA180820, CA23318, CA66636, CA21115, CA49883, CA16116, CA21076, CA27525, CA13650, CA14548, CA35421, CA32102, CA31946, CA04919, CA107868, and CA184734 (C.J.S.). We are grateful for the generous support of Rebecca and Nathan Milikowsky and Debbie and Bob First.

Data availability

Public datasets used in this study are listed in Table S1. Data generated for this study are available in GEO (accession number GSE205885).

References

1. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science 2012;337(6099):1190–1195. doi:10.1126/science.1222794 [PubMed: 22955828]
2. Trynka G, Sandor C, Han B, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet 2013;45(2):124–130. doi:10.1038/ng.2504 [PubMed: 23263488]
3. Pickrell JK. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. Am J Hum Genet 2014;94(4):559–573. doi:10.1016/j.ajhg.2014.03.004 [PubMed: 24702953]
4. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 2014;42(Database issue):D1001–D1006. doi:10.1093/nar/gkt1229 [PubMed: 24316577]
5. Finucane HK, Bulik-Sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet 2015;47(11):1228–1235. doi:10.1038/ng.3404 [PubMed: 26414678]
6. Gusev A, Lee SH, Trynka G, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet 2014;95(5):535–552. doi:10.1016/j.ajhg.2014.10.004 [PubMed: 25439723]
7. Hormozdiari F, Gazal S, van de Geijn B, et al. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. Nat Genet 2018;50(7):1041–1047. doi:10.1038/s41588-018-0148-2 [PubMed: 29942083]
8. Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. Am J Hum Genet 2018;102(5):717–730. doi:10.1016/j.ajhg.2018.04.002 [PubMed: 29727686]

9. Wainberg M, Sinnott-Armstrong N, Mancuso N, et al. Opportunities and challenges for transcriptome-wide association studies. Nature Genetics 2019;51(4):592–599. doi:10.1038/s41588-019-0385-z [PubMed: 30926968]

10. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 2017;169(7):1177–1186. doi:10.1016/j.cell.2017.05.038 [PubMed: 28622505]

11. Wray NR, Wijmenga C, Sullivan PF, Yang J, Visscher PM. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. Cell 2018;173(7):1573–1580. doi:10.1016/j.cell.2018.05.051 [PubMed: 29906445]

12. Consortium TGte. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 2015;348(6235):648–660. doi:10.1126/science.1262110 [PubMed: 25954001]

13. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, et al. Genetic effects on gene expression across human tissues. Nature 2017;550(7675):204–213. doi:10.1038/nature24277 [PubMed: 29022597]

14. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 2020;369(6509):1318–1330. doi:10.1126/science.aaz1776 [PubMed: 32913098]

15. Kim J, Ghasemzadeh N, Eapen DJ, et al. Gene expression profiles associated with acute myocardial infarction and risk of cardiovascular death. Genome Medicine 2014;6(5):40. doi:10.1186/gm560 [PubMed: 24971157]

16. Singh T, Levine AP, Smith PJ, Smith AM, Segal AW, Barrett JC. Characterization of expression quantitative trait loci in the human colon. Inflamm Bowel Dis 2015;21(2):251–256. doi:10.1097/MIB.0000000000000265 [PubMed: 25569741]

17. Ram R, Mehta M, Nguyen QT, et al. Systematic Evaluation Of Genes And Genetic Variants Associated With Type 1 Diabetes Susceptibility. J Immunol 2016;196(7):3043–3053. doi:10.4049/jimmunol.1502056 [PubMed: 26912320]

18. Gong J, Mei S, Liu C, et al. PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. Nucleic Acids Res 2018;46(D1):D971–D976. doi:10.1093/nar/gkx861 [PubMed: 29036324]

19. Liu B, Gloudemans MichaelJ, Rao AS, Ingelsson E, Montgomery SB. Abundant associations with gene expression complicate GWAS follow-up. Nat Genet 2019;51(5):768–769. doi:10.1038/s41588-019-0404-0 [PubMed: 31043754]

20. Strober BJ, Elorbany R, Rhodes K, et al. Dynamic genetic regulation of gene expression during cellular differentiation. Science 2019;364(6447):1287–1290. doi:10.1126/science.aaw0040 [PubMed: 31249060]

21. Knowles DA, Davis JR, Edgington H, et al. Allele-specific expression reveals interactions between genetic variation and environment. Nat Methods 2017;14(7):699–702. doi:10.1038/nmeth.4298 [PubMed: 28530654]

22. Ward MC, Banovich NE, Sarkar A, Stephens M, Gilad Y. Dynamic effects of genetic variation on gene expression revealed following hypoxic stress in cardiomyocytes. Elife 2021;10. doi:10.7554/eLife.57345

23. Kumasaka N, Knights AJ, Gaffney DJ. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat Genet 2016;48(2):206–213. doi:10.1038/ng.3467 [PubMed: 26656845]

24. Wang AT, Shetty A, O'Connor E, et al. Allele-Specific QTL Fine Mapping with PLASMA. The American Journal of Human Genetics 2020;106(2):170–187. doi:10.1016/j.ajhg.2019.12.011 [PubMed: 32004450]

25. Kim-Hellmuth S, Bechheim M, Pütz B, et al. Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. Nat Commun 2017;8(1):266. doi:10.1038/s41467-017-00366-1 [PubMed: 28814792]

26. Umans BD, Battle A, Gilad Y. Where Are the Disease-Associated eQTLs? Trends in Genetics 2020;0(0). doi:10.1016/j.tig.2020.08.009

27. Yao DW, O'Connor LJ, Price AL, Gusev A. Quantifying genetic effects on disease mediated by assayed gene expression levels. Nature Genetics 2020;52(6):626–633. doi:10.1038/s41588-020-0625-2 [PubMed: 32424349]

28. Chun S, Casparino A, Patsopoulos NA, et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nat Genet 2017;49(4):600–605. doi:10.1038/ng.3795 [PubMed: 28218759]

29. Li YI, van de Geijn B, Raj A, et al. RNA splicing is a primary link between genetic variation and disease. Science 2016;352(6285):600–604. doi:10.1126/science.aad9417 [PubMed: 27126046]

30. McVicker G, van de Geijn B, Degner JF, et al. Identification of genetic variants that affect histone modifications in human cells. Science 2013;342(6159):747–749. doi:10.1126/science.1242429 [PubMed: 24136359]

31. Chen L, Ge B, Casale FP, et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell 2016;167(5):1398–1414.e24. doi:10.1016/j.cell.2016.10.026 [PubMed: 27863251]

32. Waszak SM, Delaneau O, Gschwind AR, et al. Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. Cell 2015;162(5):1039–1050. doi:10.1016/j.cell.2015.08.001 [PubMed: 26300124]

33. del Rosario RCH, Poschmann J, Rouam SL, et al. Sensitive detection of chromatin-altering polymorphisms reveals autoimmune disease mechanisms. Nat Methods 2015;12(5):458–464. doi:10.1038/nmeth.3326 [PubMed: 25799442]

34. Grubert F, Zaugg JB, Kasowski M, et al. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell 2015;162(5):1051–1065. doi:10.1016/j.cell.2015.07.048 [PubMed: 26300125]

35. Gate RE, Cheng CS, Aiden AP, et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nat Genet 2018;50(8):1140–1150. doi:10.1038/s41588-018-0156-2 [PubMed: 29988122]

36. Degner JF, Pai AA, Pique-Regi R, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 2012;482(7385):390–394. doi:10.1038/nature10808 [PubMed: 22307276]

37. Maurano MT, Haugen E, Sandstrom R, et al. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. Nature Genetics 2015;47(12):1393–1401. doi:10.1038/ng.3432 [PubMed: 26502339]

38. Deplancke B, Alpern D, Gardeux V. The Genetics of Transcription Factor DNA Binding Variation. Cell 2016;166(3):538–554. doi:10.1016/j.cell.2016.07.012 [PubMed: 27471964]

39. Alasoo K, Rodrigues J, Mukhopadhyay S, et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat Genet 2018;50(3):424–431. doi:10.1038/s41588-018-0046-7 [PubMed: 29379200]

40. Gusev A, Spisak S, Fay AP, et al. Allelic imbalance reveals widespread germline-somatic regulatory differences and prioritizes risk loci in Renal Cell Carcinoma. bioRxiv Published online May 8, 2019:631150. doi:10.1101/631150

41. Benaglio P, D’Antonio-Chronowska A, Ma W, et al. Allele-specific NKX2–5 binding underlies multiple genetic associations with human electrocardiographic traits. Nature Genetics 2019;51(10):1506–1517. doi:10.1038/s41588-019-0499-3 [PubMed: 31570892]

42. Jiang X, Finucane HK, Schumacher FR, et al. Shared heritability and functional enrichment across six solid cancers. Nat Commun 2019;10(1):431. doi:10.1038/s41467-018-08054-4 [PubMed: 30683880]

43. Davies RW, Flint J, Myers S, Mott R. Rapid genotype imputation from sequence without reference panels. Nat Genet 2016;48(8):965–969. doi:10.1038/ng.3594 [PubMed: 27376236]

44. Stelloo S, Nevedomskaya E, Kim Y, et al. Integrative epigenetic taxonomy of primary prostate cancer. Nat Commun 2018;9. doi:10.1038/s41467-018-07270-2

45. Pomerantz MM, Qiu X, Zhu Y, et al. Prostate cancer reactivates developmental epigenomic programs during metastatic progression. Nature Genetics 2020;52(8):790–799. doi:10.1038/s41588-020-0664-8 [PubMed: 32690948]

46. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics 2016;48(10):1279–1283. doi:10.1038/ng.3643 [PubMed: 27548312]

47. Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams RM Jr.. The American Soldier: Adjustment during Army Life. (Studies in Social Psychology in World War II), Vol. 1. Princeton Univ. Press; 1949:xii, 599.

48. Castel SE, Aguet F, Mohammadi P, et al. A vast resource of allelic expression data spanning human tissues. Genome Biology 2020;21(1):234. doi:10.1186/s13059-020-02122-z [PubMed: 32912332]

49. Liang Y, Aguet F, Barbeira AN, Ardlie K, Im HK. A scalable unified framework of total and allele-specific counts for cis-QTL, fine-mapping, and prediction. Nat Commun 2021;12(1):1424. doi:10.1038/s41467-021-21592-8 [PubMed: 33658504]

50. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 2016;48(3):245–252. doi:10.1038/ng.3506 [PubMed: 26854917]

51. Emami NC, Kachuri L, Meyers TJ, et al. Association of imputed prostate cancer transcriptome with disease risk reveals novel mechanisms. Nature Communications 2019;10(1):3107. doi:10.1038/s41467-019-10808-7

52. Schumacher FR, Al Olama AA, Berndt SI, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nature Genetics 2018;50(7):928–936. doi:10.1038/s41588-018-0142-8 [PubMed: 29892016]

53. Mancuso N, Gayther S, Gusev A, et al. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. Nature Communications 2018;9(1):4079. doi:10.1038/s41467-018-06302-1

54. Pomerantz MM, Shrestha Y, Flavin RJ, et al. Analysis of the 10q11 Cancer Risk Locus Implicates MSMB and NCOA4 in Human Prostate Tumorigenesis. PLoS Genet 2010;6(11). doi:10.1371/journal.pgen.1001204

55. Meuleman W, Muratov A, Rynes E, et al. Index and biological spectrum of human DNase I hypersensitive sites. Nature 2020;584(7820):244–251. doi:10.1038/s41586-020-2559-3 [PubMed: 32728217]

56. Conti DV, Darst BF, Moss LC, et al. Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. Nature Genetics 2021;53(1):65–75. doi:10.1038/s41588-020-00748-0 [PubMed: 33398198]

57. Wang X, Goldstein DB. Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. Am J Hum Genet 2020;106(2):215–233. doi:10.1016/j.ajhg.2020.01.012 [PubMed: 32032514]

58. Kasowski M, Grubert F, Heffelfinger C, et al. Variation in Transcription Factor Binding Among Humans. Science 2010;328(5975):232–235. doi:10.1126/science.1183621 [PubMed: 20299548]

59. Koh CM, Bieberich CJ, Dang CV, Nelson WG, Yegnasubramanian S, De Marzo AM. MYC and Prostate Cancer. Genes Cancer 2010;1(6):617–628. doi:10.1177/1947601910379132 [PubMed: 21779461]

60. Zhang B, Ci X, Tao R, et al. Klf5 acetylation regulates luminal differentiation of basal progenitors in prostate development and regeneration. Nature Communications 2020;11(1):997. doi:10.1038/s41467-020-14737-8

61. Bhatia-Gaur R, Donjacour AA, Sciavolino PJ, et al. Roles for Nkx3.1 in prostate development and cancer. Genes Dev 1999;13(8):966–977. [PubMed: 10215624]

62. Drobnjak M, Osman I, Scher HI, Fazzari M, Cordon-Cardo C. Overexpression of cyclin D1 is associated with metastatic prostate cancer to bone. Clin Cancer Res 2000;6(5):1891–1895. [PubMed: 10815912]

63. Economides KD, Capecchi MR. Hoxb13 is required for normal differentiation and secretory function of the ventral prostate. Development 2003;130(10):2061–2069. doi:10.1242/dev.00432 [PubMed: 12668621]

64. Wu D, Sunkel B, Chen Z, et al. Three-tiered role of the pioneer factor GATA2 in promoting androgen-dependent gene expression in prostate cancer. Nucleic Acids Res 2014;42(6):3607–3622. doi:10.1093/nar/gkt1382 [PubMed: 24423874]

65. Ahmed M, Soares F, Xia JH, et al. CRISPRi screens reveal a DNA methylation-mediated 3D genome dependent causal mechanism in prostate cancer. Nat Commun 2021;12(1):1781. doi:10.1038/s41467-021-21867-0 [PubMed: 33741908]

66. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779 [PubMed: 25826379]

67. Kim SM, Kim JY, Choe NW, et al. Regulation of mouse steroidogenesis by WHISTLE and JMJD1C through histone methylation balance. Nucleic Acids Res 2010;38(19):6389–6403. doi:10.1093/nar/gkq491 [PubMed: 20530532]

68. Jin G, Sun J, Kim ST, et al. Genome-wide association study identifies a new locus JMJD1C at 10q21 that may influence serum androgen levels in men. Human Molecular Genetics 2012;21(23):5222–5228. doi:10.1093/hmg/dds361 [PubMed: 22936694]

69. Levasseur A, St-Jean G, Paquet M, Boerboom D, Boyer A. Targeted Disruption of YAP and TAZ Impairs the Maintenance of the Adrenal Cortex. Endocrinology 2017;158(11):3738–3753. doi:10.1210/en.2017-00098 [PubMed: 28938438]

70. Hawley JR, Zhou S, Arlidge C, et al. Cis-regulatory Element Hijacking Overshadows Topological Changes in Prostate Cancer. bioRxiv Published online January 6, 2021:2021.01.05.425333. doi:10.1101/2021.01.05.425333

71. Sáez C, González-Baena AC, Japón MA, et al. Expression of basic fibroblast growth factor and its receptors FGFR1 and FGFR2 in human benign prostatic hyperplasia treated with finasteride. The Prostate 1999;40(2):83–88. doi:10.1002/(SICI)1097-0045(19990701)40:2&lt;83::AID-PROS3&gt;3.0.CO;2-N [PubMed: 10386468]

72. Sweeney CJ, Chen YH, Carducci M, et al. Chemohormonal Therapy in Metastatic Hormone-Sensitive Prostate Cancer. New England Journal of Medicine 2015;373(8):737–746. doi:10.1056/NEJMoa1503747 [PubMed: 26244877]

73. Pomerantz M, Wang XV, Kantoff PW, et al. Genome-wide association study (GWAS) of response to androgen deprivation therapy (ADT) and survival in metastatic prostate cancer (PCa). JCO 2016;34(15_suppl):1540–1540. doi:10.1200/JCO.2016.34.15_suppl.1540

74. Whitaker HC, Shiong LL, Kay JD, et al. N-acetyl-L-aspartyl-L-glutamate peptidase-like 2 is overexpressed in cancer and promotes a pro-migratory and pro-metastatic phenotype. Oncogene 2014;33(45):5274–5287. doi:10.1038/onc.2013.464 [PubMed: 24240687]

75. African Ancestry Prostate Cancer GWAS Consortium, Berndt SI, Wang Z, et al. Two susceptibility loci identified for prostate cancer aggressiveness. Nat Commun 2015;6(1):6889. doi:10.1038/ncomms7889 [PubMed: 25939597]

76. Zhang Z, Chng KR, Lingadahalli S, et al. An AR-ERG transcriptional signature defined by long-range chromatin interactomes in prostate cancer cells. Genome Res 2019;29(2):223–235. doi:10.1101/gr.230243.117 [PubMed: 30606742]

Additional References

77. Baca SC, Takeda DY, Seo JH, et al. Reprogramming of the FOXA1 cistrome in treatment-emergent neuroendocrine prostate cancer. Nature Communications 2021;12(1):1979. doi:10.1038/s41467-021-22139-7

78. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 2009;10(3):R25. doi:10.1186/gb-2009-10-3-r25 [PubMed: 19261174]

79. Zhang Y, Liu T, Meyer CA, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biology 2008;9(9):R137. doi:10.1186/gb-2008-9-9-r137 [PubMed: 18798982]

80. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative Genomics Viewer. Nat Biotechnol 2011;29(1):24–26. doi:10.1038/nbt.1754 [PubMed: 21221095]

81. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. Nature 2015;526(7571):68–74. doi:10.1038/nature15393 [PubMed: 26432245]

82. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 2012;28(19):2520–2522. doi:10.1093/bioinformatics/bts480 [PubMed: 22908215]

83. Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet 2016;48(11):1443–1448. doi:10.1038/ng.3679 [PubMed: 27694958]

84. van de Geijn B, McVicker G, Gilad Y, Pritchard JK. WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods 2015;12(11):1061–1063. doi:10.1038/nmeth.3582 [PubMed: 26366987]

85. Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. Tools and best practices for data processing in allelic expression analysis. Genome Biology 2015;16(1):195. doi:10.1186/s13059-015-0762-6 [PubMed: 26381377]

86. Kuilman T, Velds A, Kemper K, et al. CopywriteR: DNA copy number detection from off-target sequence data. Genome Biology 2015;16(1):49. doi:10.1186/s13059-015-0617-1 [PubMed: 25887352]

87. Delaneau O, Ongen H, Brown AA, Fort A, Panousis NI, Dermitzakis ET. A complete tool set for molecular QTL discovery and analysis. Nature Communications 2017;8(1):15452. doi:10.1038/ncomms15452

88. Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. Journal of Evolutionary Biology 2005;18(5):1368–1373. doi:10.1111/j.1420-9101.2005.00917.x [PubMed: 16135132]

89. Gusev A, Lawrenson K, Lin X, et al. A transcriptome-wide association study of high grade serous epithelial ovarian cancer identifies novel susceptibility genes and splice variants. Nat Genet 2019;51(5):815–823. doi:10.1038/s41588-019-0395-x [PubMed: 31043753]

90. Hukku A, Pividori M, Luca F, Pique-Regi R, Im HK, Wen X. Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations. Am J Hum Genet 2021;108(1):25–35. doi:10.1016/j.ajhg.2020.11.012 [PubMed: 33308443]

91. Barbeira AN, Bonazzola R, Gamazon ER, et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. Genome Biol 2021;22. doi:10.1186/s13059-020-02252-4 [PubMed: 33413586]

92. Aulchenko YS, Struchalin MV, van Duijn CM. ProbABEL package for genome-wide association analysis of imputed data. BMC Bioinformatics 2010;11(1):134. doi:10.1186/1471-2105-11-134 [PubMed: 20233392]

93. Corces MR, Trevino AE, Hamilton EG, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods 2017;14(10):959–962. doi:10.1038/nmeth.4396 [PubMed: 28846090]

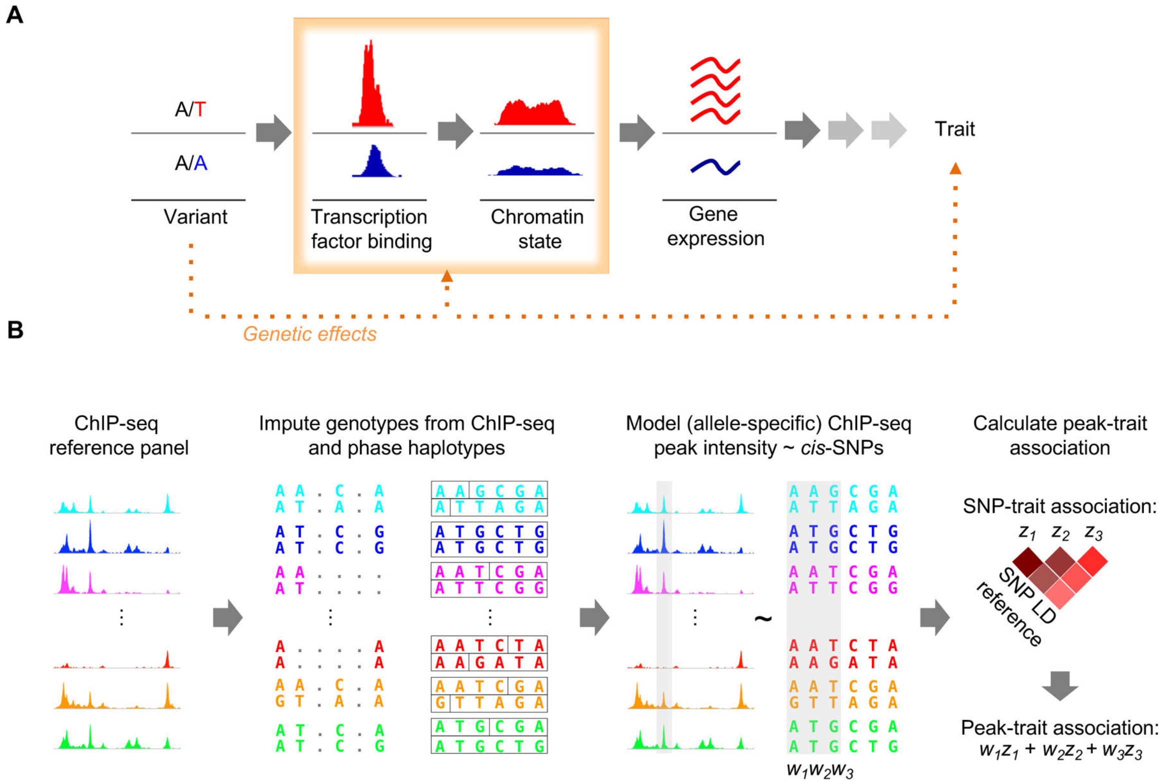


Figure 1. Overview of the method.

(A) Cistrome-wide association studies identify epigenomic features that are genetically associated with a trait. **(B)** Epigenomic sequencing reads (ChIP-seq and ATAC-seq) are merged on a per-individual basis and used to impute SNP genotypes. Haplotypes are then phased based on reference panels. Normalized read abundance and allele-specific reads at heterozygous SNPs are modeled as a function of *cis*-SNP genotypes. The resulting models capture the genetic determinants of peak intensity.

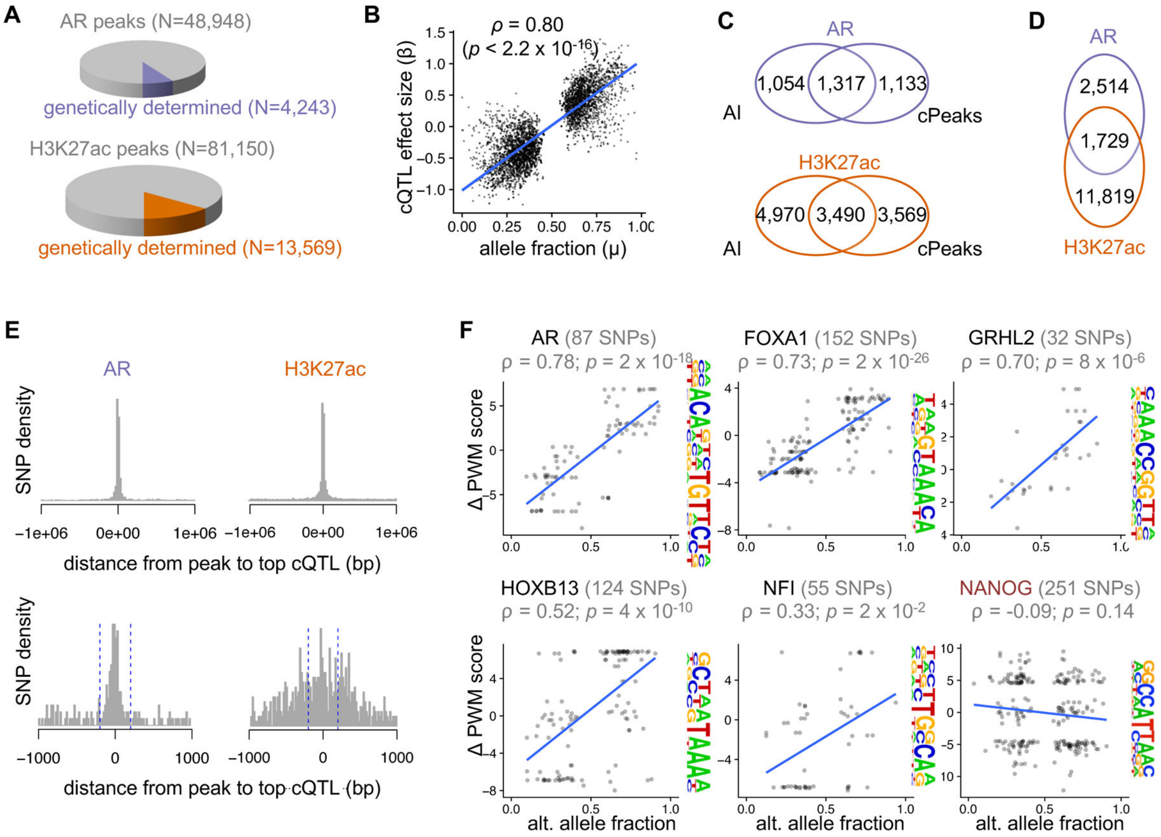


Figure 2. Genetic variation creates abundant chromatin QTLs and allelically imbalanced regulatory elements.

(**A**) Portion of all AR and H3K27ac peaks with evidence of genetic determination, defined as a significant combined test for allelic imbalance and cQTL with Q < 0.05 (methods). (**B**) cQTL effect size (*β*) versus allele fraction (*μ*) for peaks with allelic imbalance. *μ* for one SNP per peak is shown. ρ indicates Pearson correlation coefficient. (**C**) Overlap of allelically imbalanced (AI) and chromatin QTL (cQTL) peaks. (**D**) Overlap of genetically determined AR and H3K27ac peaks in (A). (**E**) Distance from the center of significant AR cQTL peaks (permutation-based q value < 0.05) to the corresponding SNP. Blue dashed lines mark ±200bp from the peak center. (**F**) For all heterozygous SNPs overlapping the indicated motif, the difference in the motif position weight matrix (PWM) score for alternate *vs*. reference alleles is plotted against the allele fraction observed in AR ChIP-seq reads. The top five motifs inferred *de novo* from 10,000 randomly selected AR binding peaks are shown. The NANOG motif (red) is included as a negative control. p-values for Pearson correlation are indicated.

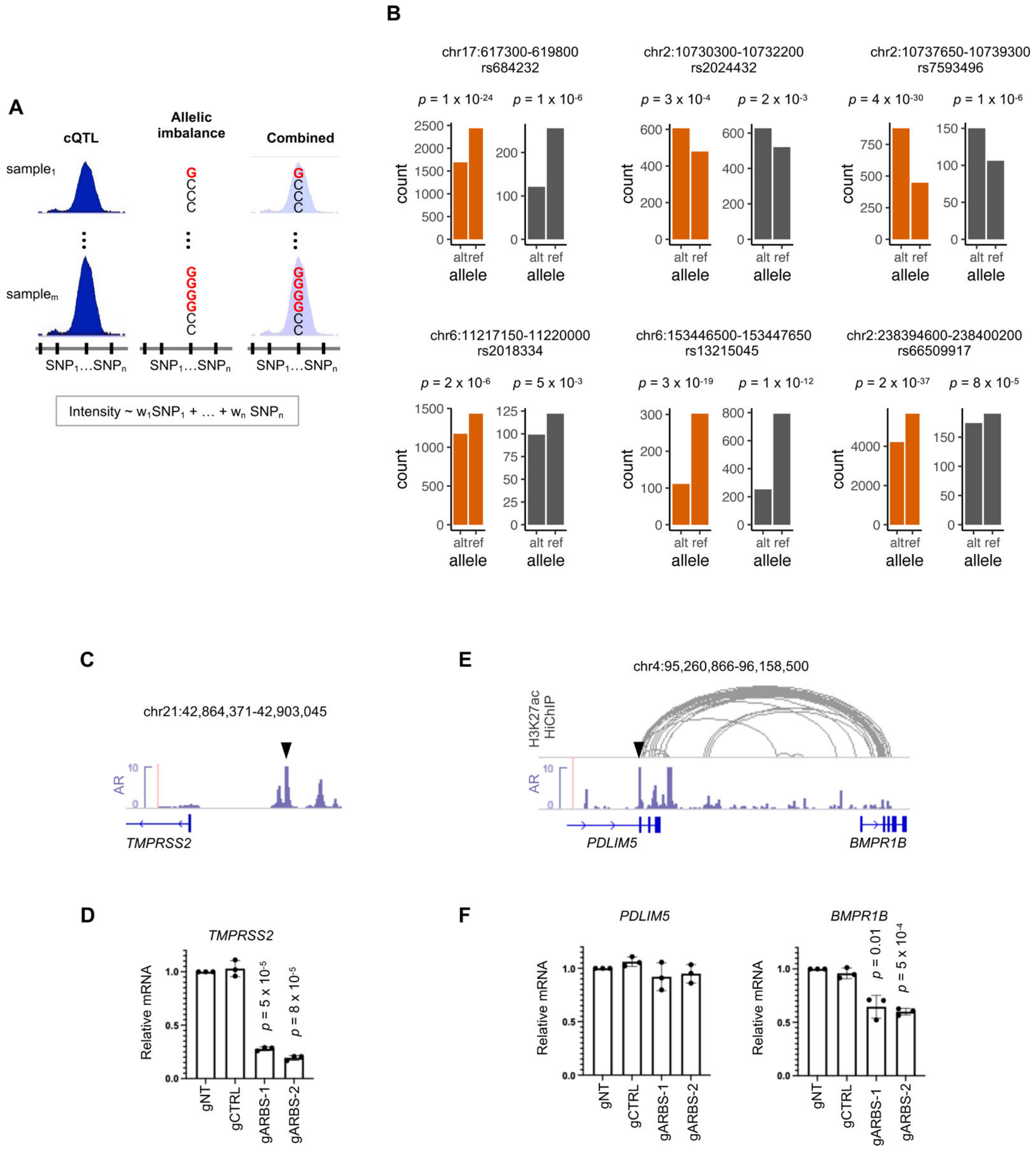


Figure 3. Integrative cistrome models identify genetic determinants of gene regulation.

(**A**) Total peak intensity, allele-specific activity, or both are modeled based on *cis*-SNP genotypes. Models include either linear combinations of SNPs ("multi-SNP"), or the single most significantly predictive SNP ("top SNP"; methods). (**B**) *In vitro* validation of allelically imbalanced regulatory element SNPs. Regulatory elements containing SNPs were assessed for enhancer activity *in vitro* using SNP STARR-seq (Methods). Bar plots indicate reads from reference or alternate haplotypes in H3K27ac ChIP-seq data (orange) and normalized transcript counts for each SNP genotype from SNP STARR-seq (gray). p-values for allelic imbalance under the beta-binomial model are indicated (Methods) (**C**) Prostate cancer-associated ARBS (black triangle) upstream of *TMPRSS2*. (**D**) Effect on *TMPRSS2* transcript expression with CRISPRi suppression of ARBSs shown in (C) (n=3 independent experiments). gNT and gCTRL indicate two non-targeting control guide RNAs. (**E**) Prostate cancer-associated ARBS (black triangle) within *BMPR1B*. (**F**) Effect on *BMPR1B* and

PDLIM5 expression with CRISPRi suppression of ARBSs shown in (E) (n=3 independent experiments). For (D) and (F), error bars indicate median and range in (D) and (F); p-values are calculated with the Wilcoxon rank-sum test.

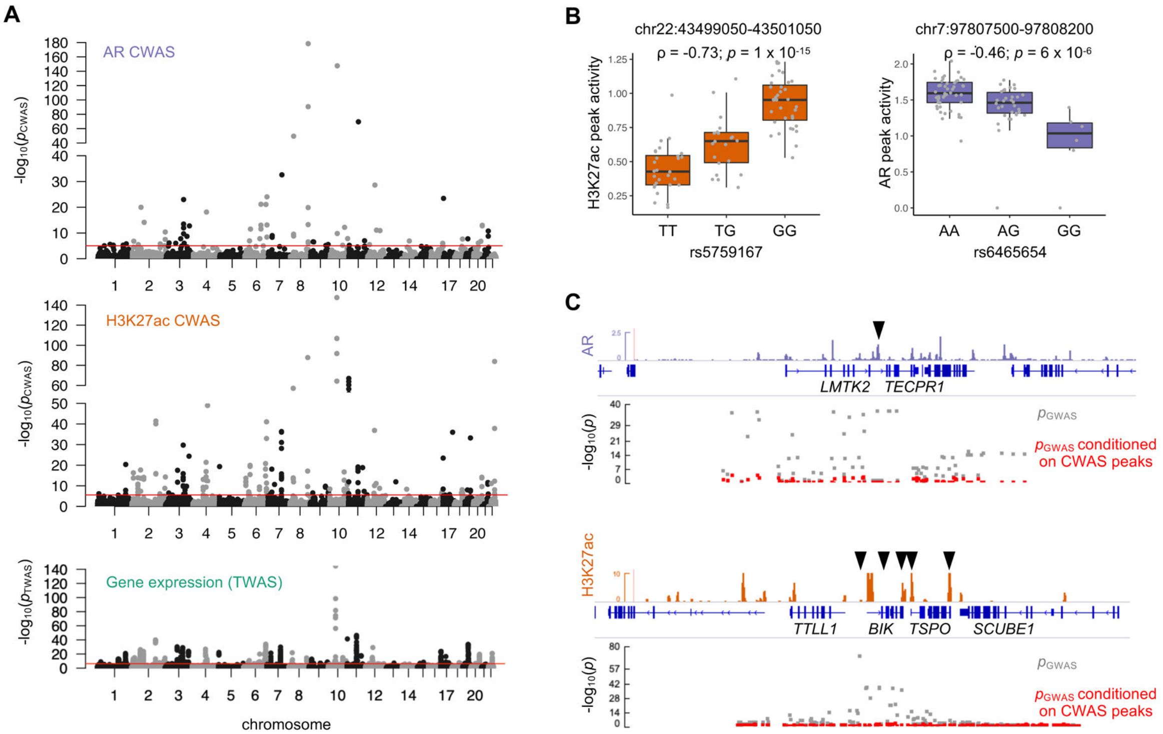
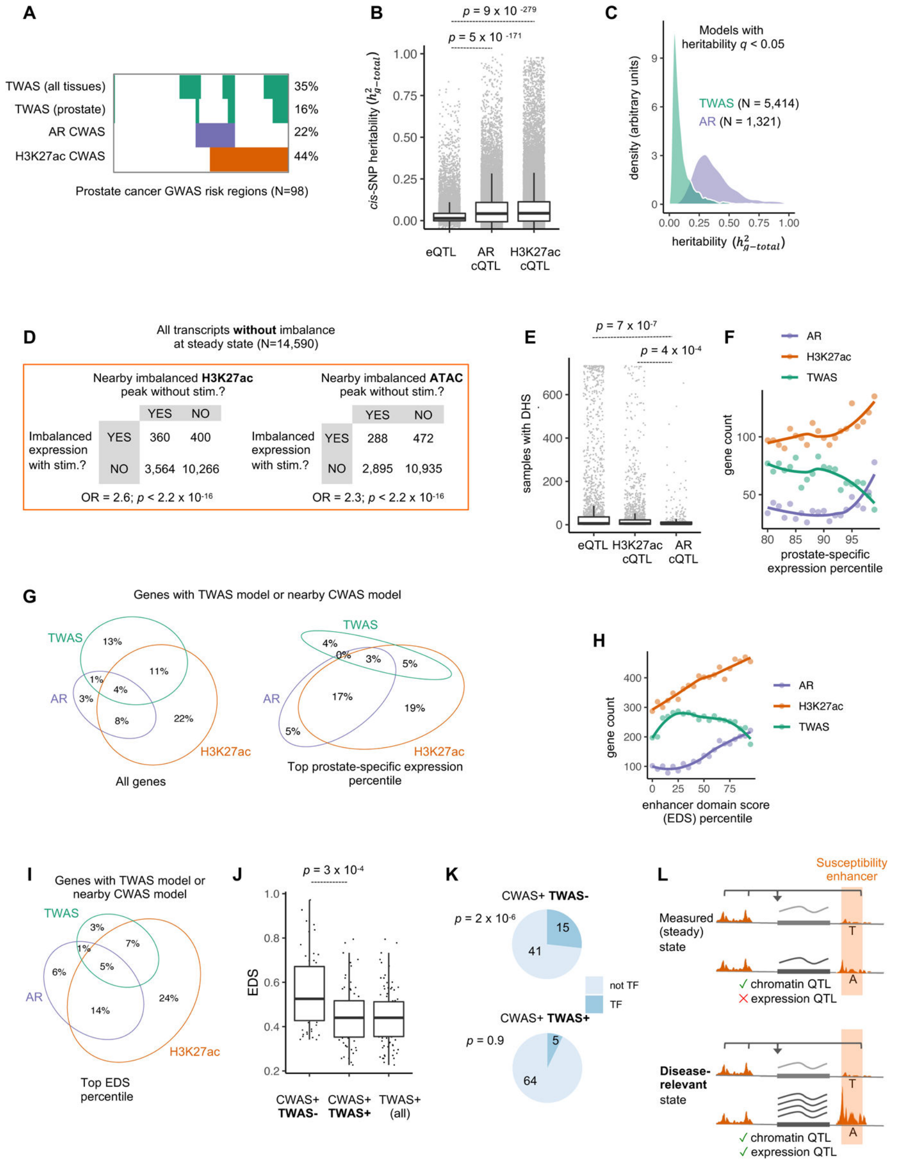


Figure 4. CWAS identifies prostate cancer risk mediated by genetic variation in AR binding and regulatory element activity.

(**A**) Manhattan plot showing significant genetic associations with prostate cancer for AR CWAS, H3K27ac CWAS, and TWAS. Red lines indicate genome-wide significance thresholds. (**B**) Normalized read counts at the indicated peaks stratified by genotype of the indicated SNP. Lower and upper hinges indicate 25th and 75th percentiles; whiskers extend to 1.5 x the inter-quartile ranges (IQR). p-values for Pearson correlation are indicated (**C**) GWAS SNP significance in the vicinity of the peaks shown in (H), with and without conditioning on genetically predicted activity. The CWAS peaks are marked by a black triangle.



Nearby imbalanced **H3K27ac** peak without stim.?

Imbalanced expression with stim.?		YES	NO
	YES	360	400
	NO	3,564	10,266

OR = 2.6; $p < 2.2 \times 10^{-16}$

Nearby imbalanced **ATAC** peak without stim.?

Imbalanced expression with stim.?		YES	NO
	YES	288	472
	NO	2,895	10,935

OR = 2.3; $p < 2.2 \times 10^{-16}$

Figure 5. CWAS identifies associations not marked by a steady-state eQTL.

(**A**) Prostate cancer risk loci were defined as genome-wide significant SNPs ± 1Mb and assessed for overlap with a high-confidence CWAS or TWAS peak. TWAS results using reference panels with only prostate tissue or all tissues are shown separately. (**B**) Estimated *cis*-SNP heritability for assessable genes (n=16,634), AR peaks (n=32,434), or H3K27ac peaks(n=54,262). (**C**) Distribution of heritability estimates for genes or AR peaks with significant heritability (q<0.05). (**D**) Steady-state chromatin measurements revealing context-dependent genetic effects on gene regulation. H3K27ac ChIP-seq, ATAC-seq, and RNA-seq data from LNCaP were generated at baseline and after 16 hours of stimulation with dihydrotestosterone (DHT) and assessed for allelic imbalance[40]. Contingency tables show all transcripts that do not exhibit allelically imbalanced expression at baseline, stratified by (1) whether they demonstrate imbalanced expression with DHT treatment and (2) whether they are within 100kb of an ATAC-seq or H3K27ac peak with allelic imbalance at baseline. Odds ratio (OR) that a transcript with stimulation-induced imbalance falls within 100kb of a peak that is imbalanced at baseline, compared to transcripts without stimulation-induced imbalance. *p*-values from chi-square tests are indicated. (**E**) Number of ENCODE samples (n=733, representing 438 cell types/states)[55] with DNAse hypersensitivity at cQTL SNPs (n=379 and 2,061 for AR and H3K27ac, respectively) and eQTL SNPs (n=2,884). (**F**) Number of genes with a TWAS model or AR/H3K27ac CWAS model (within 100kb) as a function of prostate-specific expression. Expression in prostate was compared to mean across all GTEx tissues to obtain a z-scores, which were binned by percentiles. (**G**) Percent of genes with TWAS models or CWAS models (within 100kb) for all genes (left) and the top percentile of prostate-specific expression (right). (**H**) Data from (F) grouped by enhancer domain score (EDS) percentile. (**I**) Percent of genes with TWAS models or nearby CWAS models for genes in the top EDS percentile. (**J**) Boxplots of EDS scores for genes (n=224) within central 100kb of the indicated category of GWAS risk regions. (**K**) Number of genes in indicated category of GWAS risk regions that encode TFs. p-value from chi-square test is indicated. (**L**) Model demonstrating how latent eQTLs are observable as steady-state cQTLs. p-values indicate Wilcoxon rank-sum tests for (B), (E), and (J). For boxplots, lower and upper hinges indicate 25th and 75th percentiles; whiskers extend to 1.5 x the inter-quartile ranges (IQR).

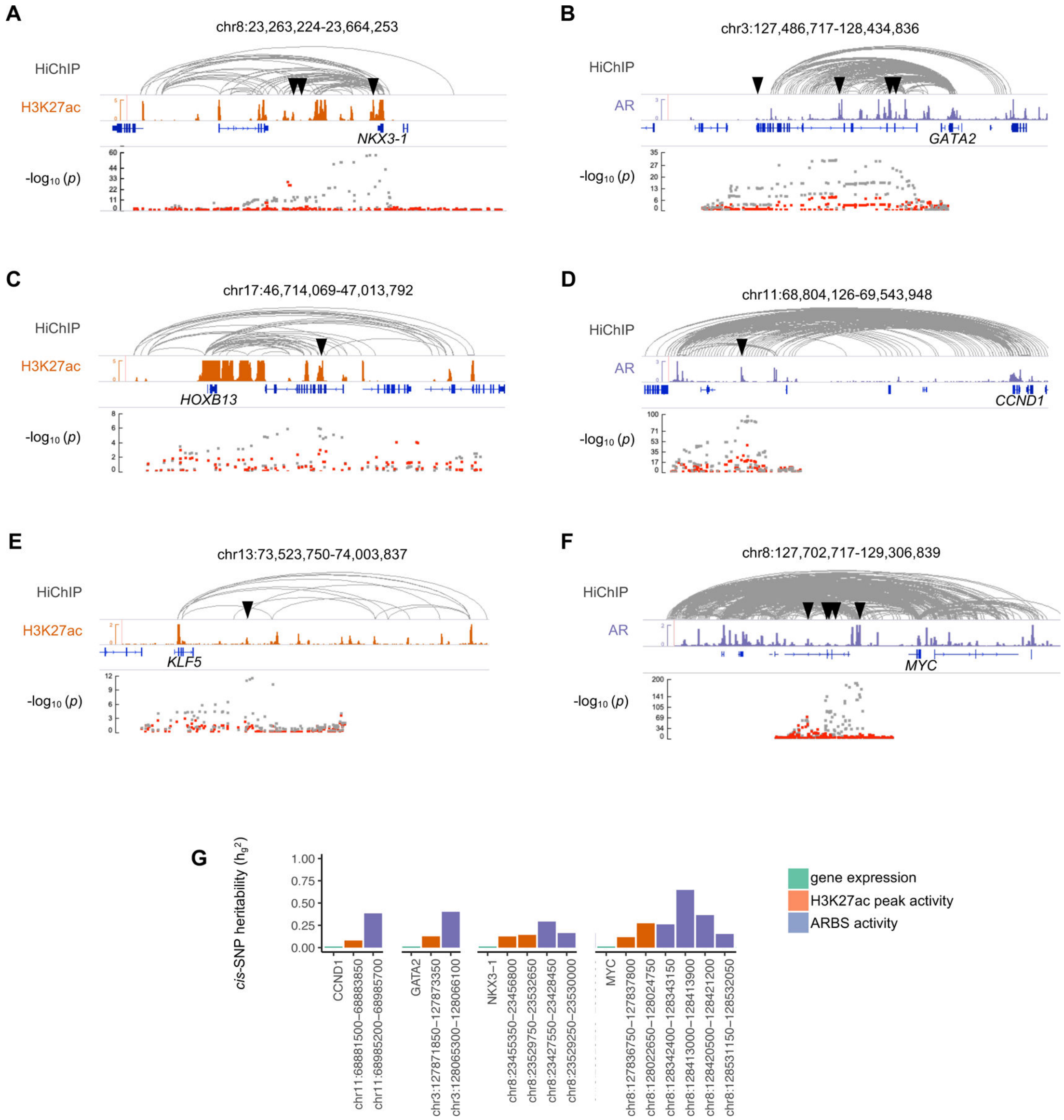


Figure 6. CWAS associations linked to selected prostate developmental genes and proto-oncogenes.

(A-F) Panels show the genomic context for CWAS ARBS or H3K27ac peaks near select genes with biological relevance to prostate cancer: NKX3-1 (**A**), GATA2 (**B**), HOXB13 (**C**), CCND1 (**D**), KLF5 (**E**), and MYC (**F**). For each panel, tracks from top to bottom show H3K27ac HiChIP loops in LNCaP (gray), Normalized read counts for H3K27ac (orange) or AR (purple) ChIP-seq in LNCaP, gene annotations, and significant CWAS H3K27ac peaks or CWAS ARBS (indicated by black triangles). The bottom track shows prostate cancer GWAS SNP significance in the vicinity of the CWAS peaks in gray, and the residual significance after conditioning upon the CWAS H3K27ac peak or ARBS in red. (**G**) *cis*-SNP heritability of indicated genes and CWAS peaks within the regions shown in A-F. Only CWAS peaks with significant *cis*-SNP heritability ($p < 0.05$) are shown.

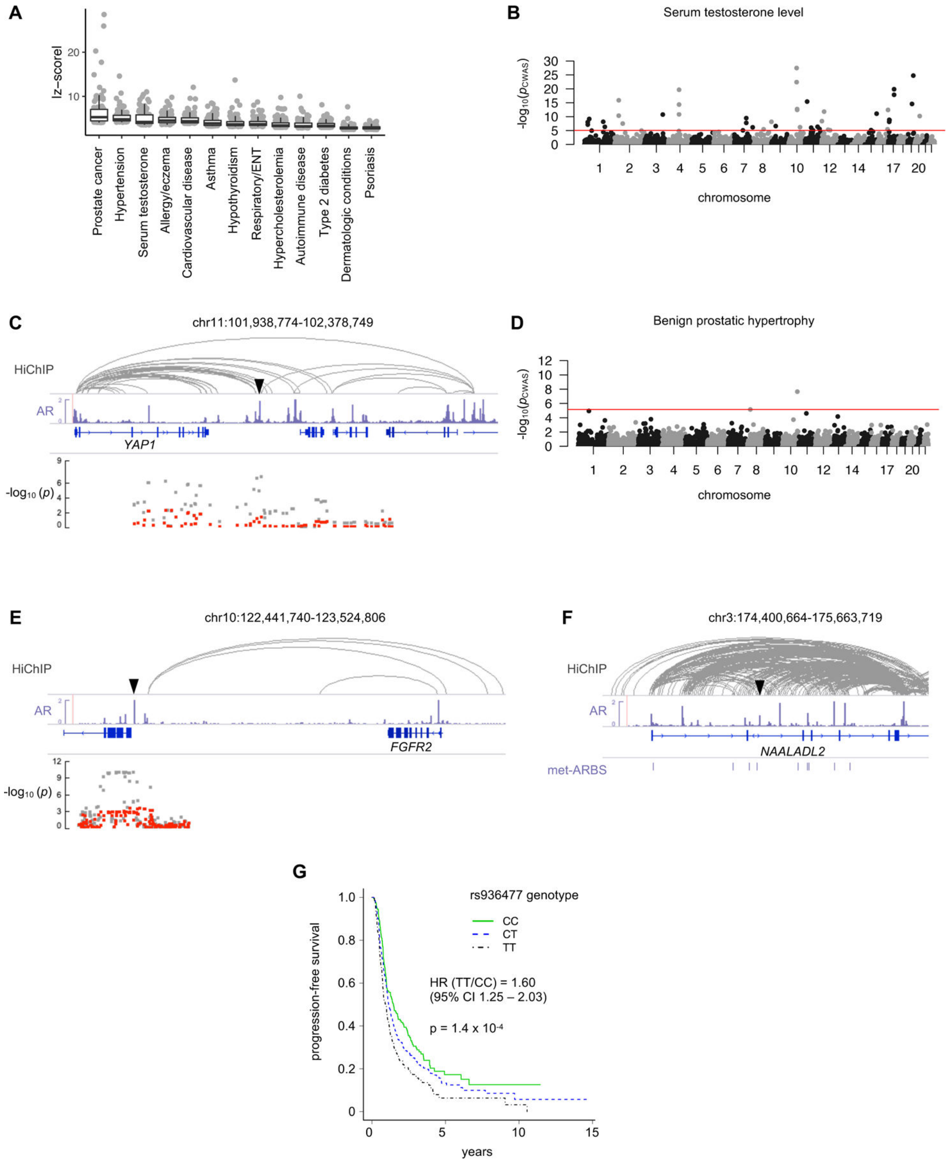


Figure 7. CWAS identifies ARBS underlying heritability of multiple androgen-regulated phenotypes.

(**A**) AR CWAS was performed on GWAS for the indicated phenotypes. The absolute value of the effect size Z was calculated for ARBS associations and the top 100 are displayed for each phenotype. (**B**) Manhattan plot showing significance of ARBS associations with testosterone levels among individuals in the UK Biobank[66]. (**C**) Epigenomic context of a significant CWAS ARBS for testosterone near *YAP1.* Tracks from top to bottom show H3K27ac HiChIP loops in LNCaP (gray), normalized AR ChIP-seq read counts in LNCaP (purple), gene annotations, and the location of the significant CWAS ARBS (black triangle). The bottom track shows testosterone GWAS SNP significance in the vicinity of the CWAS peaks in gray and the residual significance after conditioning upon predicted activity of the ARBS in red. (**D**) Manhattan plot showing significance of ARBS associations with BPH among individuals in the UK Biobank. (**E**) Epigenomic context of a significant CWAS ARBS for BPH near *FGFR2.* Tracks are as described for (B). (**F**) Epigenomic context of

CWAS ARBS within *NAALADL2* associated with response to androgen deprivation therapy among men with prostate cancer from a clinical trial[72]. Met-ARBs (purple) signify AR binding sites that are enriched in metastatic castration-resistant prostate cancer compared to prostate-localized tumors[45]. (**G**) Kaplan-Meier curve showing progression-free survival on androgen deprivation therapy stratified by patient genotype at rs936477, the SNP that determines intensity of the ARBS within *NAALADL2* shown in (F).