nature genetics

ARTICLES

https://doi.org/10.1038/s41588-022-01168-y

Check for updates

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[1,3,8,17] ✉ and Matthew L. Freedman[1,2,3,17] ✉

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 are also associated with disease. In prostate cancer, CWAS identified regulatory elements and androgen receptor-binding sites that explained the association at 52 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 noncoding genetic variants contribute to diseases and complex phenotypes has proven to be difficult[8–11]. To address this challenge, large-scale efforts have cataloged many thousands of *cis*-acting expression quantitative trait loci (eQTLs)[12–14]. At these loci, the genotype of a 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 or in differentiated tissues, but they can be observed 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, transcription factor binding or chromatin accessibility[36–39]. In a complementary manner to cQTLs, allelic imbalance 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]. Use of cQTLs and allelic imbalance 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.

[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, British Columbia, Canada. [6]Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA, USA. [7]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, 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]Department of Biomedical Engineering, Laboratory of Chemical Biology and Institute for Complex Molecular Systems, Eindhoven University of Technology, Eindhoven, The Netherlands. [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. [17]These authors jointly supervised this work: Alexander Gusev, Matthew L. Freedman. ✉e-mail: alexander_gusev@dfci.harvard.edu; mfreedman@partners.org

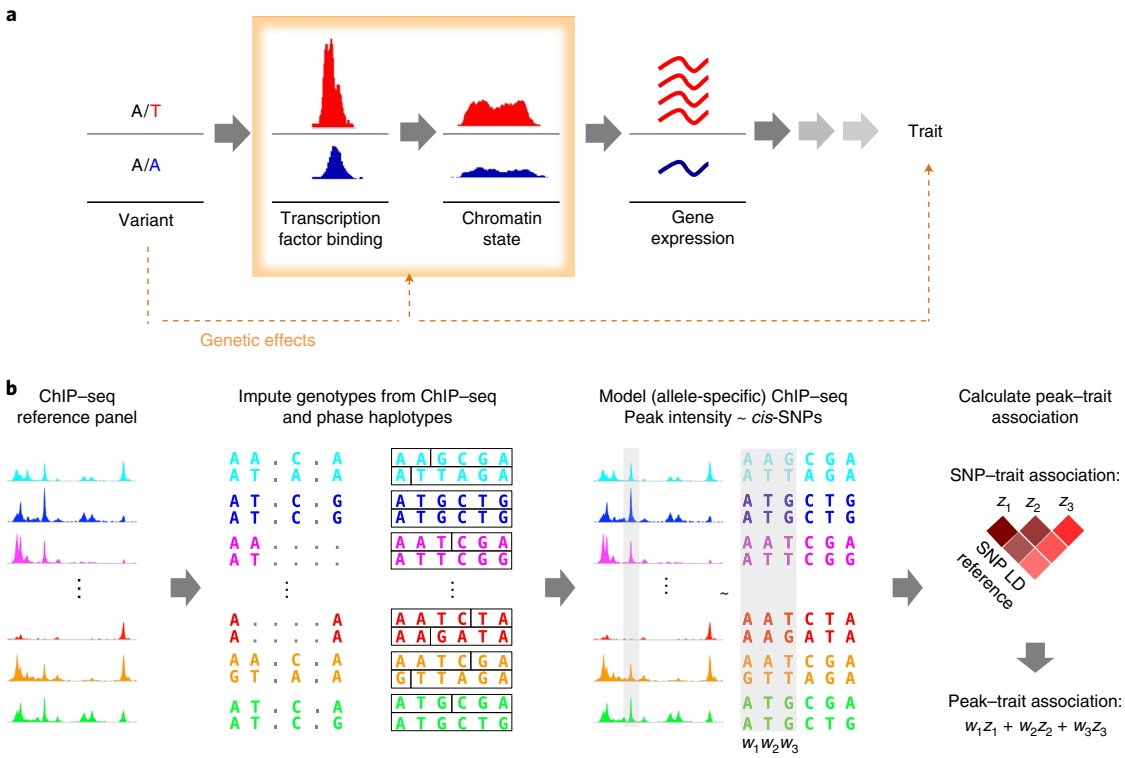


Fig. 1 | Overview of the method. a, CWAS identify epigenomic features that are genetically associated with a trait. **b**, Epigenomic sequencing reads (ChIP–seq and ATAC–seq) are merged on a per-individual basis and used to impute SNP genotypes. Haplotypes are then phased based on reference panels. Normalized 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.

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 transcription factor-binding sites). We introduce the cistrome-wide association study (CWAS), which identifies the genetic determinants of transcription factor 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 found that heritable variation in the cistrome of the androgen receptor (AR)—a critical transcription factor 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 in 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.

Results

Overview of the methods. We developed a systematic approach that links genetic variation in transcription factor binding or the chromatin state to trait variation (Fig. 1a). We leveraged 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 the 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 identified genetic determinants of epigenomic features (for example, AR binding or H3K27ac) by jointly relating allelic imbalance and peak intensity to nearby SNPs. These models identify SNPs that correlate with epigenomic peak intensity. Integrating this information with summary statistics from GWAS, we identified peaks whose genetic determinants were associated with the trait of interest. The result was a 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 used data from two recent studies of prostate cancer epigenomes, which performed ChIP–seq for transcription factors and histone modifications across a combined cohort of 163 men of predominantly European ancestry[44,45] (Supplementary Tables 1 and 2 and Extended Data Fig. 2). The dataset comprised 131 ChIP–seq experiments for AR and 176 for H3K27ac. Because these samples had 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 Supplementary Note).

By analyzing both allelic imbalance 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 allelic imbalance[47] identified 4,243 AR-binding sites (ARBS; 9% of total) and 13,569 H3K27ac peaks (17% of total) where the genotype of a 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 allelic imbalance, 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

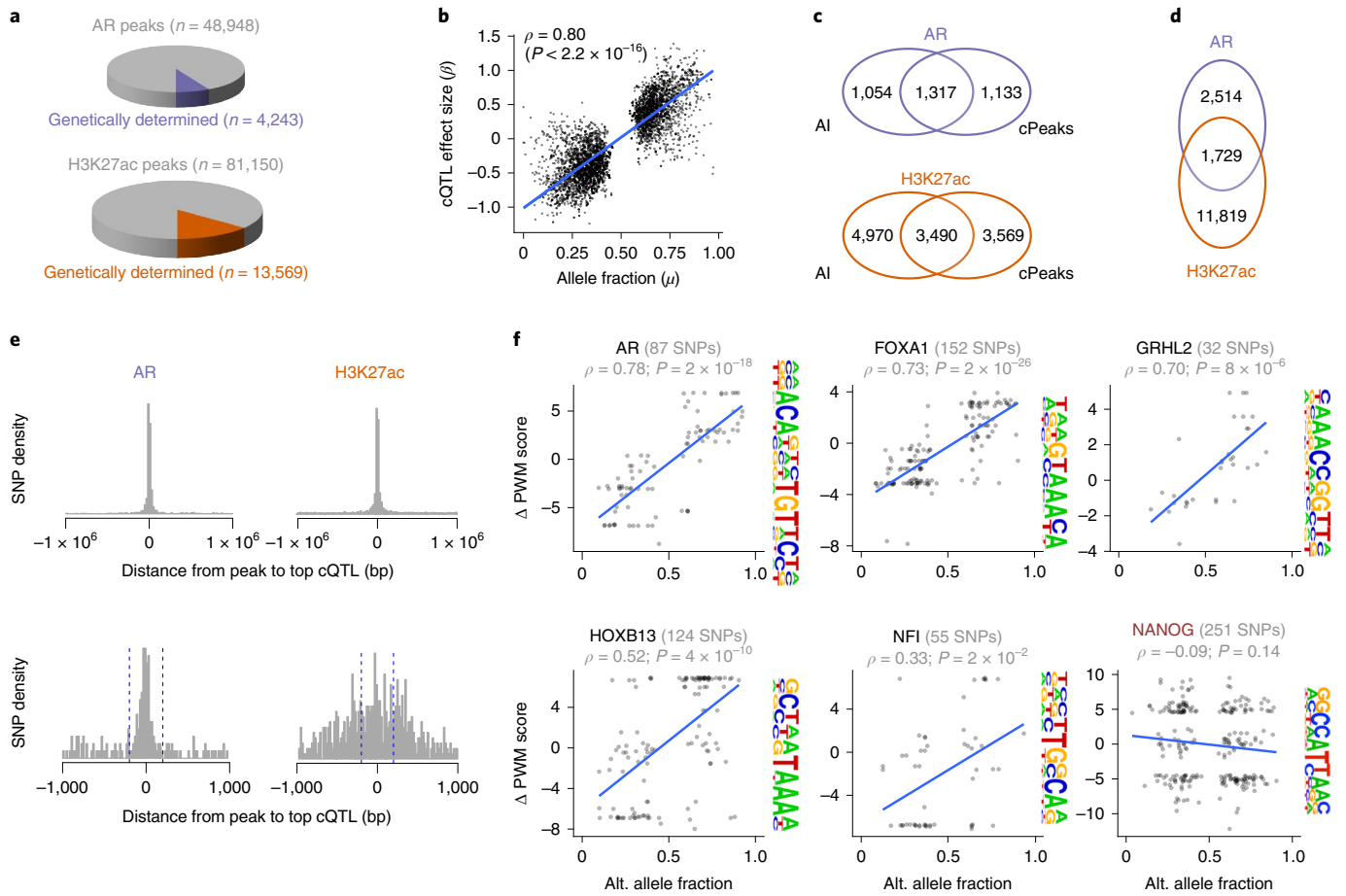


Fig. 2 | Genetic variation creates abundant cQTLs and allelically imbalanced regulatory elements. a, Portion of all AR and H3K27ac peaks with evidence of genetic determination, defined as a significant combined test for allelic imbalance and cQTLs with $Q < 0.05$ (Methods). **b**, cQTL effect size (β) versus allele fraction (μ) for peaks with allelic imbalance. μ for one SNP per peak is shown. ρ indicates Pearson's correlation coefficient. **c**, Overlap of allelically imbalanced (AI) and 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 ±200 bp 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 versus 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.

to that in a larger study of gene expression allelic imbalance and cQTLs[48]. Measuring both allelic imbalance 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 transcription factor and H3K27ac ChIP–seq data captured overlapping but distinct genetic regulation. cQTLs overlapped significantly with eQTLs from an independent Genotype-Tissue Expression (GETx) study[14] and demonstrated correlated effects on chromatin and gene expression (Extended Data Fig. 3 and Supplementary Note).

cQTL SNPs tended to reside in or near peaks: 50% of AR cQTLs and 35% of H3K27ac cQTLs were within 10 kb of the corresponding peak center (Fig. 2e and Extended Data Fig. 4). Ten percent of AR cQTLs fell within 200 bp of the peak center, suggesting that these SNPs directly affect binding of core transcriptional machinery. Accordingly, 450 heterozygous SNPs within binding motifs of AR and its cofactors demonstrated allelic imbalance, with AR preferentially binding to the allele that was 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 allelic imbalance 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 25 kb of a peak[50], an interval that contained 84% of the top 5% of AR cQTLs by significance (Extended Data Fig. 4 and Supplementary Note). Fivefold cross-validation demonstrated that 5,580 of 48,948 AR peaks (11%) and 17,199 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 ($q < 0.05$; Supplementary Tables 3 and 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 and Methods). In addition, suppression of genetically determined ARBS in LNCaP prostate cancer cells using CRISPR interference

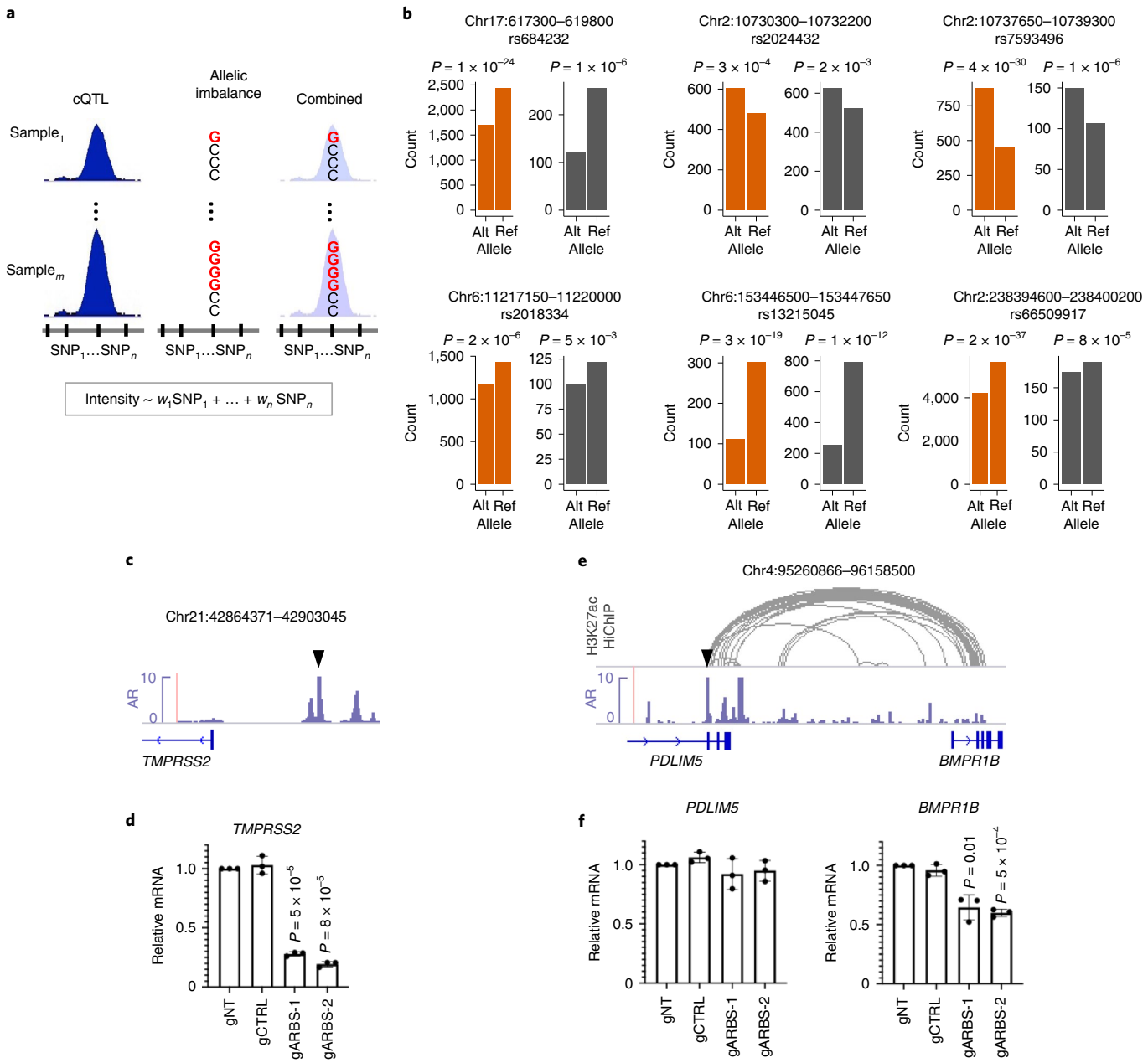


Fig. 3 | Integrative cistrome models identify genetic determinants of gene regulation. a, Total peak intensity, allele-specific activity or both modeled based on *cis*-SNP genotypes. Models include either linear combinations of SNPs ('multi-SNP') or the single most significantly predictive SNP ('top 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 the ARBS shown in **c** ($n = 3$ independent experiments). gNT and gCTRL indicate two nontargeting control guide RNAs. Data are shown as the mean and s.e.m.; *P* values were calculated with two-tailed *t*-tests. **e**, Prostate cancer-associated ARBS (black triangle) within *BMPR1B*. **f**, Effect on *BMPR1B* and *PDLIM5* expression with CRISPRi suppression of the ARBS shown in **e** ($n = 3$ independent experiments).

(CRISPRi) suppressed the expression of genes linked to these ARBS by H3K27ac HiChIP loops. For instance, suppression of an ARBS that was 14 kb-upstream markedly reduced *TMPRSS2* expression (Fig. 3c,d and Supplementary Table 5), 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*; 134 kb away) with no effect on the gene containing the ARBS (*PDLIM5*; Fig. 3e,f and Supplementary Table 5). 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 CWAS to associate genetically predicted peak intensity with prostate cancer risk, using summary

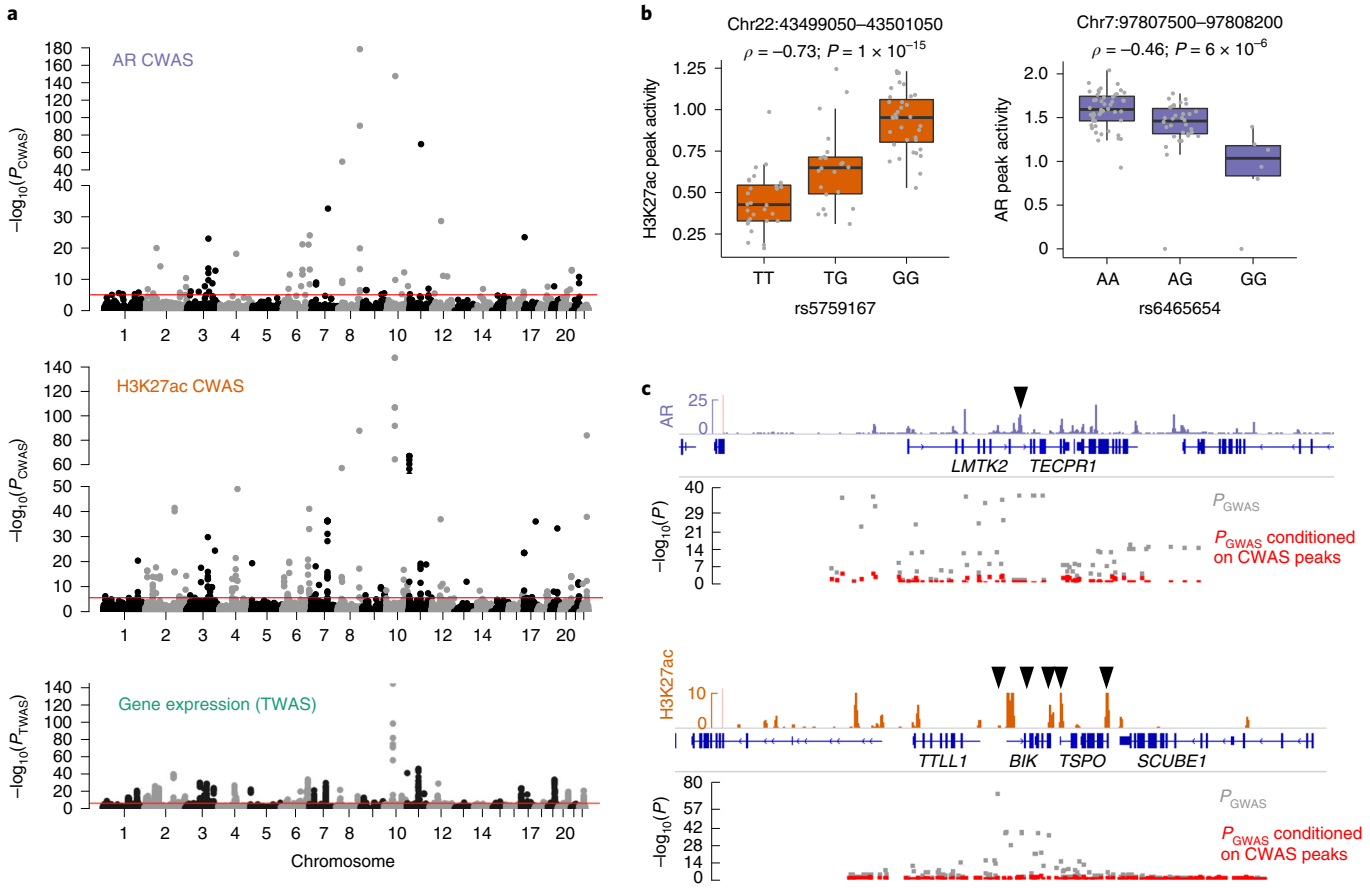


Fig. 4 | CWAS identifies prostate cancer risk mediated by genetic variation in AR binding and regulatory element activity. a, Manhattan plot showing significant genetic associations with prostate cancer for AR CWAS, H3K27ac CWAS and TWAS. Red lines indicate genome-wide significance thresholds. **b**, 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 times the interquartile ranges (IQR). *P* values for Pearson correlation are indicated. **c**, GWAS SNP significance in the vicinity of the peaks shown in **b**, with and without conditioning on genetically predicted activity. The CWAS peaks are marked by a black triangle.

statistics from a prostate cancer GWAS of 140,306 males[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 using summary statistics from GWAS, CWAS takes advantage of the large size of GWAS studies without requiring participant-level information.

CWAS identified 74 ARBS (of 5,580 ARBS with genetic models) and 199 H3K27ac peaks (of 17,199) that were significantly associated with prostate cancer risk after Bonferroni correction for multiple hypotheses tested (Fig. 4a and Supplementary Tables 6 and 7). CWAS association explained >90% of the GWAS signal for 41% of AR CWAS regions and 52% of H3K27ac CWAS regions (Fig. 4b,c, Extended Data Fig. 5a,b and Supplementary Tables 6 and 7). For instance, a single intragenic ARBS within *LMTK2* accounted for the significant GWAS association at this region (Fig. 4c). Similarly, H3K27ac 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 new peak–trait associations across 17 regions without a nearby genome-wide-significant GWAS SNP (Extended Data Fig. 6 and Supplementary Tables 6 and 7). CWAS enabled these discoveries by limiting hypothesis testing to SNPs with a high prior likelihood of affecting phenotypes—that is, 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). Notably, GWAS associations were confirmed in 12 of 17 regions with new CWAS associations after this manuscript was prepared in a larger GWAS incorporating an additional ~94,000 individuals[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 eQTLs

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* (refs. [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 accounts for most of the GWAS association signal, allowing risk to be linked to 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 ±1-Mb 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 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 idea, *cis*-SNPs explained a significantly greater portion of the heritability of AR and H3K27ac total peak intensity ($h^2_{g\text{ - total}}$) than the heritability of gene expression levels ($P=5\times10^{-171}$ and $P=9\times10^{-279}$ 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 (measured with ATAC–seq), H3K27ac and gene expression data in LNCaP cells at baseline and after 16 h 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 (odds ratio (OR), 2.3 and 2.6 for ATAC–seq and H3K27ac, respectively; $P<2.2\times10^{-16}$; 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 ($P=7\times10^{-7}$ and 4×10^{-4}, for eQTLs versus 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 the 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).

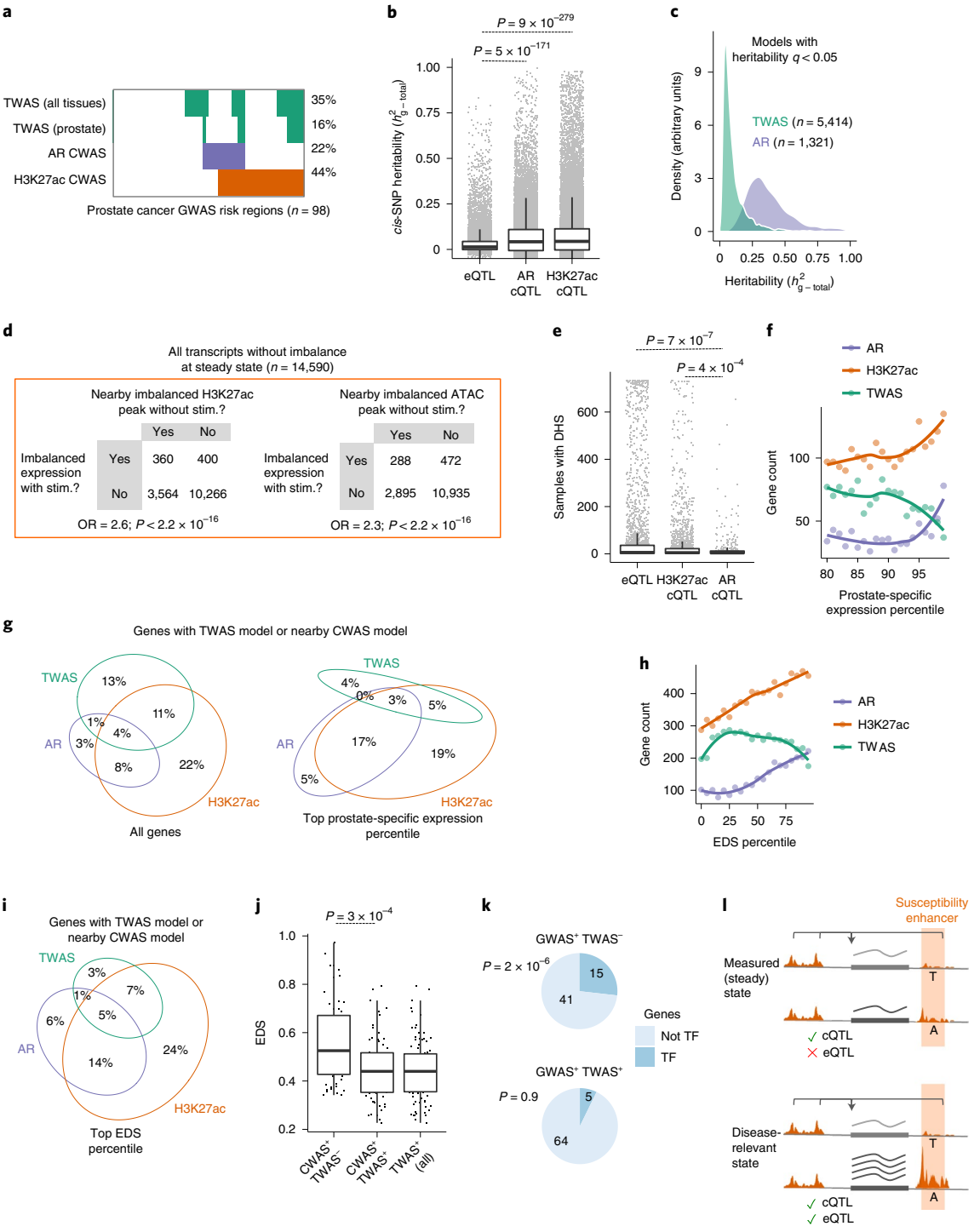
Third, consistent with prior work[56], we found that TWAS models were depleted among genes with the highest degree of regulation, as assessed by the enhancer domain score (EDS; Fig. 5h,i). In contrast, high-EDS genes were the most likely to have nearby CWAS models (Fig. 5h,i). A known limitation of steady-state eQTLs is that they are depleted around highly regulated (high-EDS) genes, which include transcription factors, 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 transcription factor 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).

Fig. 5 | CWAS identifies associations not marked by a steady-state eQTL. a, Prostate cancer risk loci were defined as genome-wide-significant SNPs ±1 Mb 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 ($n=54{,}262$). *P* values are from Wilcoxon rank-sum tests. Boxplot lower and upper hinges indicate 25th and 75th percentiles; whiskers extend to 1.5 times the IQR. **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, assay for transposase-accessible chromatin and sequencing (ATAC–seq) and RNA-seq data from LNCaP cells were generated at baseline and after 16 h of stimulation with dihydrotestosterone (DHT) and assessed for allelic imbalance[40]. Contingency tables show all transcripts that did not exhibit allelically imbalanced expression at baseline, stratified by (1) whether they demonstrated imbalanced expression with DHT treatment and (2) whether they are within 100 kb of an ATAC–seq or H3K27ac peak with allelic imbalance at baseline. The OR is shown that a transcript with stimulation-induced imbalance falls within 100 kb of a peak that is imbalanced at baseline, compared to transcripts without stimulation-induced imbalance. *P* values from chi-squared tests are indicated. **e**, Number of Encyclopedia of DNA Elements (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 ($n=2{,}884$). Boxplots and *P* values are as described for **b**. DHS, DNase hypersensitivity site. **f**, Number of genes with a TWAS model or AR/H3K27ac CWAS model (within 100 kb) as a function of prostate-specific expression. Expression in prostate was compared to the mean across all GTEx tissues to obtain *z*-scores, which were binned by percentile. **g**, Percent of genes with TWAS models or CWAS models (within 100 kb) for all genes (left) and the top percentile of genes with prostate-specific expression (right). **h**, Data from **f** grouped by EDS percentile. **i**, Percentage 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 the central 100 kb of the indicated category of GWAS risk regions. Boxplots and *P* values are as described for **b**. **k**, Number of genes in the indicated category of GWAS risk regions that encode transcription factors (TFs). *P* value from chi-squared tests are indicated. **l**, Model demonstrating how latent eQTLs are observable as steady-state cQTLs.

CWAS implicates developmental genes in prostate cancer

The advantages of the 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* (ref. [59]), *KLF5* (ref. [60]), *NKX3-1* (ref. [61]), *CCND1* (ref. [62]), *HOXB13* (ref. [63]) and *GATA2* (ref. [64]), physically interacted with CWAS ARBS and/or H3K27ac peaks, as assessed by H3K27ac HiChIP (Fig. 6a–f). Conditioning GWAS SNP associations on 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.

Notably, 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 ~220 kb centromeric to *MYC* containing the variant rs11986220 was recently shown to impair *MYC* expression, proliferation and tumorigenesis 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



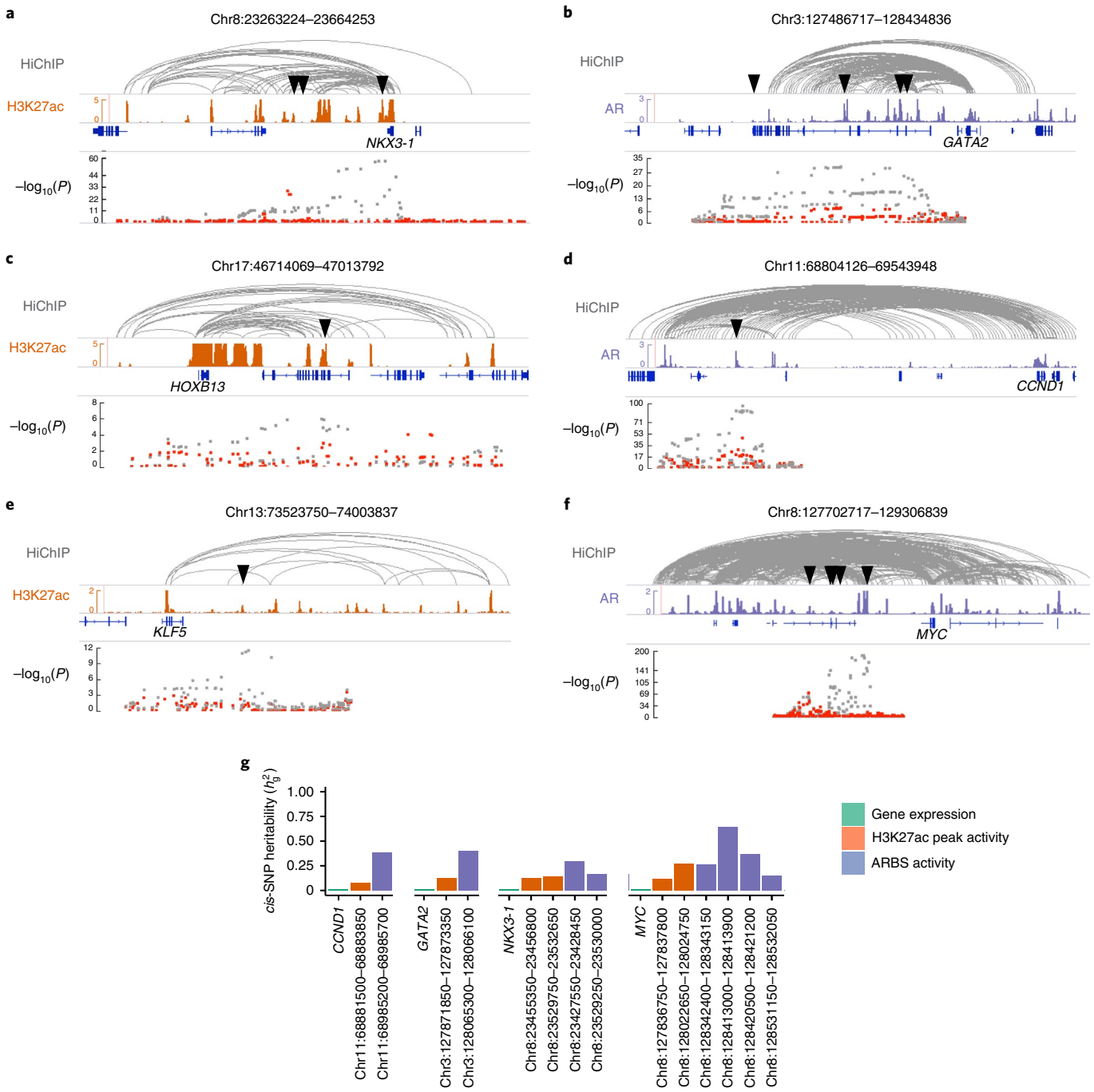


Fig. 6 | CWAS associations linked to selected prostate developmental genes and proto-oncogenes. a–**f**, Genomic context for CWAS ARBS or H3K27ac peaks near select genes with biological relevance to prostate cancer: *NKX3-1* (**a**), *GATA2* (**b**), *HOXB13* (**c**), *CCND1* (**d**), *KLF5* (**e**) and *MYC* (**f**). For each panel, tracks from top to bottom show H3K27ac HiChIP loops in LNCaP cells (gray), normalized read counts for H3K27ac (orange) or AR (purple) ChIP-seq in LNCaP cells, 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 on the CWAS H3K27ac peak or ARBS in red. **g**, *cis*-SNP heritability of the 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.

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 new regions ($n = 45$) associated with testosterone levels among male UK Biobank participants[66] (Fig. 7b and Supplementary Table 8). The most significant ($P = 3 \times 10^{-28}$) was 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 a sex hormone-binding globulin. For

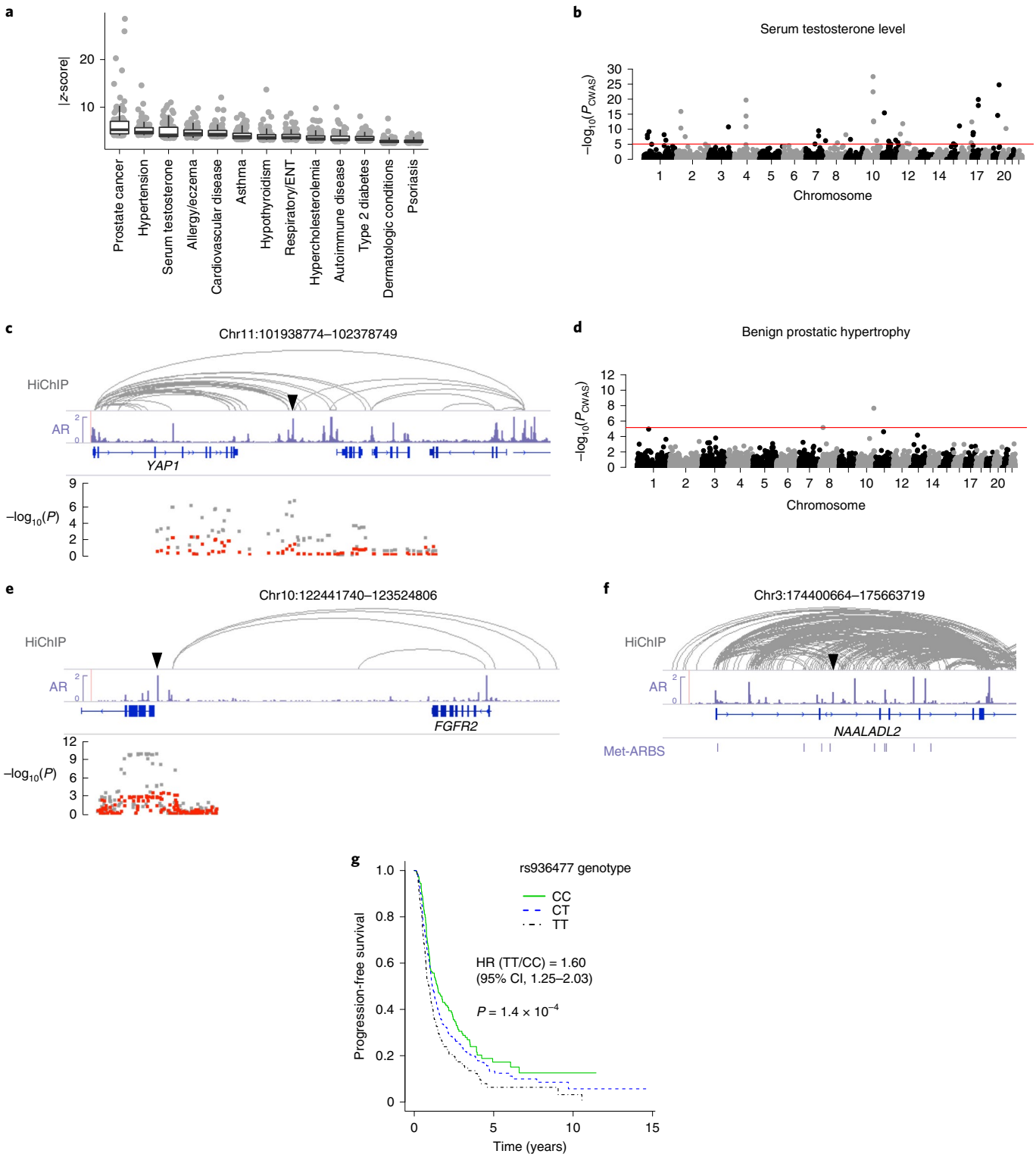


Fig. 7 | CWAS identifies ARBS underlying heritability of multiple androgen-regulated phenotypes. a, AR CWAS was performed on GWAS for the indicated phenotypes. The absolute value of the effect size *Z* was calculated for ARBS associations, and the top 100 are displayed for each phenotype. ENT, ear nose throat. **b**, Manhattan plot showing the significance of ARBS associations with testosterone levels among individuals in the UK Biobank[66]. The red line indicates the cistrome-wide significance threshold. **c**, Epigenomic context of a significant CWAS ARBS for testosterone near *YAP1*. Tracks from top to bottom show H3K27ac HiChIP loops in LNCaP cells (gray), normalized AR ChIP–seq read counts in LNCaP cells (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 on the predicted activity of the ARBS in red. **d**, Manhattan plot showing the 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 ADT among men with prostate cancer from a clinical trial[72]. The Met-ARBS track (purple) shows ARBS that are enriched in metastatic castration-resistant prostate cancer compared to prostate-localized tumors[45]. **g**, Kaplan–Meier curve showing progression-free survival on ADT stratified by patient genotype at rs936477, the SNP that determines intensity of the ARBS within *NAALADL2* shown in **f**.

seven of these peaks (16%), a significant GWAS association was not detectable within 1 Mb. These new 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 and Supplementary Table 9). The most significantly associated ARBS ($P=2\times10^{-8}$) 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 transcription factor gene *NKX3-1* (ref. [45]), which has not been implicated in BPH previously. These results demonstrate that CWAS identifies ARBS that accounts 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 cohorts 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 the genome-wide significance threshold of $P<5\times10^{-8}$. To increase power, we applied CWAS to regions within 1 Mb 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 ($P=7.8\times10^{-5}$; hazard ratio (HR), 1.29; 95% confidence interval (CI), 1.13–1.46; Fig. 7f and Supplementary Table 10). 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 ($P=4.18\times10^{-8}$)[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 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 ARBS 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 and 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 transcription factors, 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. Notably, 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.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgments, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-022-01168-y.

Received: 9 April 2022; Accepted: 19 July 2022;
Published online: 07 September 2022

References

1. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
2. Trynka, G. et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
3. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
4. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP–trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
5. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

6. Gusev, A. et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
7. Hormozdiari, F. et al. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018).
8. Gallagher, M. D. & Chen-Plotkin, A. S. The post-GWAS era: from association to function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
9. Wainberg, M. et al. Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
10. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
11. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common disease is more complex than implied by the core gene omnigenic model. *Cell* **173**, 1573–1580 (2018).
12. GTex Consortium.The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
13. GTEx Consortium.Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
14. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
15. Kim, J. et al. Gene expression profiles associated with acute myocardial infarction and risk of cardiovascular death. *Genome Med.* **6**, 40 (2014).
16. Singh, T. et al. Characterization of expression quantitative trait loci in the human colon. *Inflamm. Bowel Dis.* **21**, 251–256 (2015).
17. Ram, R. et al. Systematic evaluation of genes and genetic variants associated with type 1 diabetes susceptibility. *J. Immunol.* **196**, 3043–3053 (2016).
18. Gong, J. et al. PancanQTL: systematic identification of *cis*-eQTLs and *trans*-eQTLs in 33 cancer types. *Nucleic Acids Res.* **46**, D971–D976 (2018).
19. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).
20. Strober, B. J. et al. Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).
21. Knowles, D. A. et al. Allele-specific expression reveals interactions between genetic variation and environment. *Nat. Methods* **14**, 699–702 (2017).
22. Ward, M. C., Banovich, N. E., Sarkar, A., Stephens, M. & Gilad, Y. Dynamic effects of genetic variation on gene expression revealed following hypoxic stress in cardiomyocytes. *eLife* **10**, e57345 (2021). 2021).
23. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **48**, 206–213 (2016).
24. Wang, A. T. et al. Allele-specific QTL fine mapping with PLASMA. *Am. J. Hum. Genet.* **106**, 170–187 (2020).
25. Kim-Hellmuth, S. et al. Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nat. Commun.* **8**, 266 (2017).
26. Umans, B. D., Battle, A. & Gilad, Y. Where are the disease-associated eQTLs?. *Trends Genet.* **37**, 109–124 (2021).
27. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).
28. Chun, S. et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).
29. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
30. McVicker, G. et al. Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747–749 (2013).
31. Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414 (2016).
32. Waszak, S. M. et al. Population variation and genetic control of modular chromatin architecture in humans. *Cell* **162**, 1039–1050 (2015).
33. del Rosario, R. C. H. et al. Sensitive detection of chromatin-altering polymorphisms reveals autoimmune disease mechanisms. *Nat. Methods* **12**, 458–464 (2015).
34. Grubert, F. et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* **162**, 1051–1065 (2015).
35. Gate, R. E. et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* **50**, 1140–1150 (2018).
36. Degner, J. F. et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
37. Maurano, M. T. et al. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.* **47**, 1393–1401 (2015).
38. Deplancke, B., Alpern, D. & Gardeux, V. The genetics of transcription factor DNA binding variation. *Cell* **166**, 538–554 (2016).
39. Alasoo, K. et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **50**, 424–431 (2018).
40. Gusev, A. et al. Allelic imbalance reveals widespread germline-somatic regulatory differences and prioritizes risk loci in renal cell carcinoma. Preprint at *bioRxiv* https://doi.org/10.1101/631150 (2019).
41. Benaglio, P. et al. Allele-specific NKX2-5 binding underlies multiple genetic associations with human electrocardiographic traits. *Nat. Genet.* **51**, 1506–1517 (2019).
42. Jiang, X. et al. Shared heritability and functional enrichment across six solid cancers. *Nat. Commun.* **10**, 431 (2019).
43. Davies, R. W., Flint, J., Myers, S. & Mott, R. Rapid genotype imputation from sequence without reference panels. *Nat. Genet.* **48**, 965–969 (2016).
44. Stelloo, S. et al. Integrative epigenetic taxonomy of primary prostate cancer. *Nat. Commun.* **9**, 4900 (2018).
45. Pomerantz, M. M. et al. Prostate cancer reactivates developmental epigenomic programs during metastatic progression. *Nat. Genet.* **52**, 790–799 (2020).
46. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
47. Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A. & Williams R. M. Jr. *The American Soldier: Adjustment During Army Life* Vol. 1 (Princeton University Press,1949).
48. Castel, S. E. et al. A vast resource of allelic expression data spanning human tissues. *Genome Biol.* **21**, 234 (2020).
49. Liang, Y., Aguet, F., Barbeira, A. N., Ardlie, K. & Im, H. K. A scalable unified framework of total and allele-specific counts for *cis*-QTL, fine-mapping, and prediction. *Nat. Commun.* **12**, 1424 (2021).
50. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
51. Emami, N. C. et al. Association of imputed prostate cancer transcriptome with disease risk reveals novel mechanisms. *Nat. Commun.* **10**, 3107 (2019).
52. Schumacher, F. R. et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).
53. Mancuso, N. et al. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nat. Commun.* **9**, 4079 (2018).
54. Pomerantz, MM. et al. Analysis of the 10q11 cancer risk locus implicates *MSMB* and *NCOA4* in human prostate tumorigenesis. *PLoS Genet.* **6**, e1001204 (2010).
55. Meuleman, W. et al. Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
56. Conti, D. V. et al. Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nat. Genet.* **53**, 65–75 (2021).
57. Wang, X. & Goldstein, D. B. Enhancer domains predict gene pathogenicity and inform gene discovery in complex disease. *Am. J. Hum. Genet.* **106**, 215–233 (2020).
58. Kasowski, M. et al. Variation in transcription factor binding among humans. *Science* **328**, 232–235 (2010).
59. Koh, C. M. et al. MYC and prostate cancer. *Genes Cancer* **1**, 617–628 (2010).
60. Zhang, B. et al. Klf5 acetylation regulates luminal differentiation of basal progenitors in prostate development and regeneration. *Nat. Commun.* **11**, 997 (2020).
61. Bhatia-Gaur, R. et al. Roles for Nkx3.1 in prostate development and cancer. *Genes Dev.* **13**, 966–977 (1999).
62. Drobnjak, M., Osman, I., Scher, H. I., Fazzari, M. & Cordon-Cardo, C. Overexpression of cyclin D1 is associated with metastatic prostate cancer to bone. *Clin. Cancer Res.* **6**, 1891–1895 (2000).
63. Economides, K. D. & Capecchi, M. R. Hoxb13 is required for normal differentiation and secretory function of the ventral prostate. *Development* **130**, 2061–2069 (2003).
64. Wu, D. et al. Three-tiered role of the pioneer factor GATA2 in promoting androgen-dependent gene expression in prostate cancer. *Nucleic Acids Res.* **42**, 3607–3622 (2014).
65. Ahmed, M. et al. CRISPRi screens reveal a DNA methylation-mediated 3D genome dependent causal mechanism in prostate cancer. *Nat. Commun.* **12**, 1781 (2021).
66. Sudlow, C. 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 Med.* **12**, e1001779 (2015).
67. Kim, S. M. et al. Regulation of mouse steroidogenesis by WHISTLE and JMJD1C through histone methylation balance. *Nucleic Acids Res.* **38**, 6389–6403 (2010).
68. Jin, G. et al. Genome-wide association study identifies a new locus *JMJD1C* at 10q21 that may influence serum androgen levels in men. *Hum. Mol. Genet.* **21**, 5222–5228 (2012).
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* **158**, 3738–3753 (2017).
70. Hawley, J. R. et al. Reorganization of the 3D genome pinpoints noncoding drivers of primary prostate tumors. *Cancer Res.* **81**, 5833–5848 (2021).

71. Sáez, C. et al. Expression of basic fibroblast growth factor and its receptors FGFR1 and FGFR2 in human benign prostatic hyperplasia treated with finasteride. *Prostate* **40**, 83–88 (1999).
72. Sweeney, C. J. et al. Chemohormonal therapy in metastatic hormone-sensitive prostate cancer. *N. Engl. J. Med.* **373**, 737–746 (2015).
73. Pomerantz, M. et al. Genome-wide association study (GWAS) of response to androgen deprivation therapy (ADT) and survival in metastatic prostate cancer (PCa). *JCO* **34**, 1540 (2016).
74. Whitaker, H. C. 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* **33**, 5274–5287 (2014).
75. Berndt, S. I. et al. Two susceptibility loci identified for prostate cancer aggressiveness. *Nat. Commun.* **6**, 6889 (2015).
76. Zhang, Z. et al. An AR-ERG transcriptional signature defined by long-range chromatin interactomes in prostate cancer cells. *Genome Res.* **29**, 223–235 (2019).

Methods

This research complies with all relevant ethical regulations. Tumor specimens were collected and profiled previously under protocols approved by the institutional review boards of the Dana-Farber Cancer Institute[45] and the Netherlands Cancer Institute[44].

ChIP–seq peak calling. ChIP–seq fastq files from ref. [44] were downloaded from the Sequence Read Archive (SRA) using SRA toolkit fastq dump v2.10.0. 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 the Gene Expression Omnibus (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]. Nonuniquely 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 >10-fold over background) and percent of peak overlap with DHS peaks derived from 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 = 1,240, niterations = 40, method = diploid (https://hub.docker.com/r/stefangroha/stitch_gcs/tags). The imputation reference panel contained haplotypes of 2,505 individuals in phase 3 of the 1000 Genomes Project[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 the bcftools gtcheck function to test pairwise correlation of homozygous SNPS across all files. Samples were clustered based on correlation. Six bam files of 581 that clustered in a cluster of a different individual were excluded from the analysis.

Twenty-four samples were subjected 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 versus the 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 50-bp windows and including any window with peaks in >5% of samples. Windows were buffered by 100 bp 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 the power and accuracy of allelic imbalance detection. Imputed SNP genotypes were phased with Eagle2 (ref. [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 $R_{\text{alt},i} \mid R_{\text{ref},i}\ \text{BetaBin}(\pi_j, \rho_{ij})$, where π is the mean allelic ratio and ρ is a locally defined, per-individual sequence read correlation parameter reflecting overdispersion.

Copy number profiles were estimated from off-target ChIP–seq reads with CopywriteR[86] and used in modeling of the overdispersion parameter ρ, to account for overdispersion 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 using the 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.

Allelic imbalance P values were false discovery rate (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 transcription factor-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 versus alternate alleles.

cQTL detection. QTLtools (v1.2)[87] was used for cQTL detection. Reads per kilobase per million mapped reads (RPKM) values for each sample at AR and H3K27ac consensus peaks were 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 1-Mb window were assessed for cQTLs with QTLtools cis (--normal--permute: 1,000) after regressing out the first six 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 cQTL SNPs were within 25 kb of the corresponding peak, we also took a focused approach and calculated nominal P values for *cis*-SNP pairs within 25 kb, forgoing permutation, which was often not possible at a distance of 25 kb due to a limited number 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 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 equally sized samplings of the complete set of AR or H3K27ac peaks to generate a null distribution. We reported the ratio of peak territory containing 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 $\mathbf{y}_{\text{total}}$, with each row containing an individual; the vector of allelic chromatin activity $\mathbf{y}_{\text{allelic}}$, defined as $\log(\boldsymbol{N}_{\text{p}}/\boldsymbol{N}_{\text{m}})$, where $\boldsymbol{N}_{\cdot}$ 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 $\boldsymbol{H}_{\text{p}}$ and $\boldsymbol{H}_{\text{m}}$, 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 allelic reads. In model 1 ('cQTL model'), the relationship between total chromatin activity and genotype is modeled as $\mathbf{y}_{\text{total}} \sim \boldsymbol{X}_{\text{total}} + \epsilon$, where $\boldsymbol{X}_{\text{total}} = \boldsymbol{H}_{\text{p}} + \boldsymbol{H}_{\text{m}}$ 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 refs. [24,49], the relationship between allelic chromatin activity and haplotype is modeled as $\mathbf{y}_{\text{allelic}} \sim \boldsymbol{X}_{\text{allelic}} + \epsilon$, where $X_{\text{allelic}} = H_{\text{p}} - H_{\text{m}}$ and corresponds to the −1/0/1 allele phase. Finally, in model 3 ('combined model'), we define a 'combined' model as $\begin{bmatrix} \tilde{\mathbf{y}}_{\text{total}} \\ \tilde{\mathbf{y}}_{\text{allelic}} \end{bmatrix} \sim \begin{bmatrix} \tilde{\boldsymbol{X}}_{\text{total}} \\ \tilde{\boldsymbol{X}}_{\text{allelic}} \end{bmatrix} + \epsilon$, 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 perform comparably to other penalization schemes[89]). Predictive accuracy was evaluated by fivefold cross-validation and quantified as the Pearson correlation to the true $\mathbf{y}_{\text{total}}$ or $\mathbf{y}_{\text{allelic}}$ 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 the window set to 25 kb to include SNPs within 25 kb of the peak center.

CWAS analysis. Integrative models of cQTL and allelic imbalance were built as described above for each consensus AR or H3K27ac peak based on the genotypes of *cis*-SNPs within 25 kb (the number of significant models was largely insensitive to the window size; Supplementary 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 allelic imbalance and GWAS risk was calculated by FUSION, accounting for linkage disequilibrium[50,53]. FUSION considers the *Z*-score for genetic peak–trait association as

$$\mathbf{Z}_{\text{peak}\rightarrow\text{trait}} = \mathbf{W}\mathbf{Z}_{\text{SNPs}\rightarrow\text{trait}}$$

where $\mathbf{Z}_{\text{SNPs}\rightarrow\text{trait}}$ is a vector of SNP–trait association *Z*-scores from GWAS summary statistics

$$\mathbf{Z}_{\text{SNPs}\rightarrow\text{trait}} = \begin{bmatrix} \mathbf{Z}_{\text{SNP1}\rightarrow\text{trait}} \\ \vdots \\ \mathbf{Z}_{\text{SNP}n\rightarrow\text{trait}} \end{bmatrix}$$

and *W* is a weight matrix defined as

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

$\mathbf{\Sigma}_{\text{p,s}}$ is the peak–SNP covariance matrix and $\mathbf{\Sigma}_{\text{s,s}}$ is the SNP–SNP covariance matrix, representing linkage disequilibrium. In practice, $\boldsymbol{W}$ is learned from the data through penalized regression. Assuming a normal distribution of $\mathbf{Z}_{\text{peak}\rightarrow\text{trait}}$ around 0, then *Z*-score for a peak–trait CWAS association is

$$\mathbf{Z}_{\text{CWAS}} = \frac{\mathbf{W}\mathbf{Z}_{\text{SNPs}\rightarrow\text{trait}}}{\text{var}\left(\mathbf{W}\mathbf{Z}_{\text{SNPs}\rightarrow\text{trait}}\right)} = \frac{\mathbf{W}\mathbf{Z}_{\text{SNPs}\rightarrow\text{trait}}}{\left(\mathbf{W}\mathbf{\Sigma}_{\text{s,s}}\mathbf{W}^{\mathbf{T}}\right)^{1/2}}$$

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 Bonferroni 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 Supplementary Table 1.

Overlap of GWAS, TWAS and CWAS results. Genome-wide-significant SNPs ($P < 5 \times 10^{-8}$) were obtained from published GWAS summary data[52], assigned hg19 coordinates buffered with 1-Mb windows on either side and merged where windows overlapped 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 of the following: (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 ref. [56] were buffered with 1-Mb 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.

ADT GWAS. Men who received ADT for metastatic hormone-sensitive prostate cancer ($n = 687$) from two cohorts were evaluated. Overall, 265 of these patients were from the control arm of the CHAARTED clinical trial (E3805)[72]. The remaining 422 patients were treated at Dana-Farber Cancer Institute. The study was performed under institutional review board-approved protocols that included informed consent for genotyping. These patients were selected to match enrollment criteria for CHAARTED. Participants were genotyped at approximately 1 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 prephase the SNPs using the 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 roots 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 1 Mb 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 (Supplementary Table 5) 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 collected 48 h following transfection and used to transduce the LNCaP cell line in the presence of 4 mg ml^{-1} polybrene, and the medium was exchanged after 24 h. Conditions were optimized to ensure >95% transduction as assessed by selection with puromycin. RNA was isolated 4 d after transduction using the QIAGEN RNeasy Plus kit, and cDNA was synthesized using the 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 Supplementary Table 5. A nontargeting gRNA and gRNA targeting an intergenic region were used as negative controls. Gene expression was normalized to that of *GAPDH* and △△Ct 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 d, and they were then stimulated with either 10 nM DHT (5α-androstan-17β-ol-3-one, dihydrotestosterone, A8380, Sigma) or ethanol (vehicle) for 16 h. Subsequently, cells were collected for further analysis. LNCaP cells were authenticated by comparing short tandem repeats to those of parental LNCaP cells in the ATCC database. Before experiments, cells were tested for several strains of mycoplasma contamination using the LookOut Mycoplasma PCR Detection kit (Sigma-Aldrich, D9307).

ChIP–seq in LNCaP cells was performed as previously described[77]. Briefly, 10 million cells were fixed with 1% formaldehyde at room temperature for 10 min and quenched with 0.25 M glycine. Cells collected in lysis buffer (1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS and protease inhibitor (11873580001, Roche) in PBS) were sheared to chromatin of 300–800 bp in size using a Covaris E220 sonicator (140-W peak incident power, 5% duty cycle, 200 cycleburtst). Sonicated chromatin was subjected to anti-H3K27ac antibody (C15410196, Diagenode) coupled with Protein A/G Dynabeads (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) six times for 10 min sequentially. Immunoprecipitated chromatin and input were treated with RNase A at 37 °C for 30 min and de-cross-linked in elution buffer (1% SDS, 0.1 M $NaHCO_3$) with proteinase K for 6–12 h at 65 °C with gentle rocking. DNA was purified using Qiagen Qiaquick columns (28104). Libraries were prepared using the SMARTer ThruPLEX DNA-Seq kit (Takara Bio, R400675).

ATAC–seq libraries were prepared using the 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 $MgCl_2$, 0.1% NP-40, 0.1% Tween-20, 0.01% digitonin) were fragmented in 50 µl of transposition mix (25 µl of 2× TD buffer, 16.5 µl PBS, 0.5 µl of 1% digitonin, 0.5 µl of 10% Tween-20, 5 µl water) with 2.5 µl transposase (Illumina, 20034197) for 30 min at 37 °C with shaking at 1,000 r.p.m. in a thermomixer. DNA was purified using Qiagen MinElute (28004), and libraries were amplified up to the cycle number determined by one-third maximal qPCR fluorescence.

Total mRNA was collected from 300,000 cells using the RNA easy kit (Qiagen, 74044) with the RNase-Free DNase Set (Qiagen, 79254) according to the manufacturer's instructions. RNA purity and concentration were determined on a 2100 Bioanalyzer (Agilent) using the Agilent RNA 6000 Nano kit (5067-1511). Then, 400-ng RNA samples were submitted to Novogene for RNA library preparation.

ChIP–seq, RNA-seq and ATAC–seq libraries were sequenced with 150-bp 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].

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Public datasets used in this study are listed in Supplementary Table 1. Data generated for this study are available in GEO (accession number GSE205885).

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.

References

77. Baca, S. C. et al. Reprogramming of the FOXA1 cistrome in treatment-emergent neuroendocrine prostate cancer. *Nat. Commun.* **12**, 1979 (2021).
78. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
79. Zhang, Y. et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008). (2008).
80. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
81. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
82. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
83. Loh, P. R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
84. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
85. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).
86. Kuilman, T. et al. CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol.* **16**, 49 (2015).
87. Delaneau, O. et al. A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).
88. Whitlock, M. C. Combining probability from independent tests: the weighted *Z*-method is superior to Fisher's approach. *J. Evolut. Biol.* **18**, 1368–1373 (2005).
89. Gusev, A. et al. A transcriptome-wide association study of high grade serous epithelial ovarian cancer identifies novel susceptibility genes and splice variants. *Nat. Genet.* **51**, 815–823 (2019).
90. Hukku, A. et al. Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations. *Am. J. Hum. Genet.* **108**, 25–35 (2021).
91. Barbeira, A. N. et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22**, 49 (2021). 2021)
92. Aulchenko, Y. S., Struchalin, M. V. & van Duijn, C .M. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinf.* **11**, 134 (2010).
93. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).

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 National Cancer Institute of the National Institutes of Health under 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 (S.C.B.).

Author contributions

S.C.B., A.G. and M.L.F. conceived 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 cells. 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 performed on samples and data provided by C.J.S. J.A. assisted with implementation of the CWAS pipeline.

Competing interests

The authors declare no competing interests.

Additional information

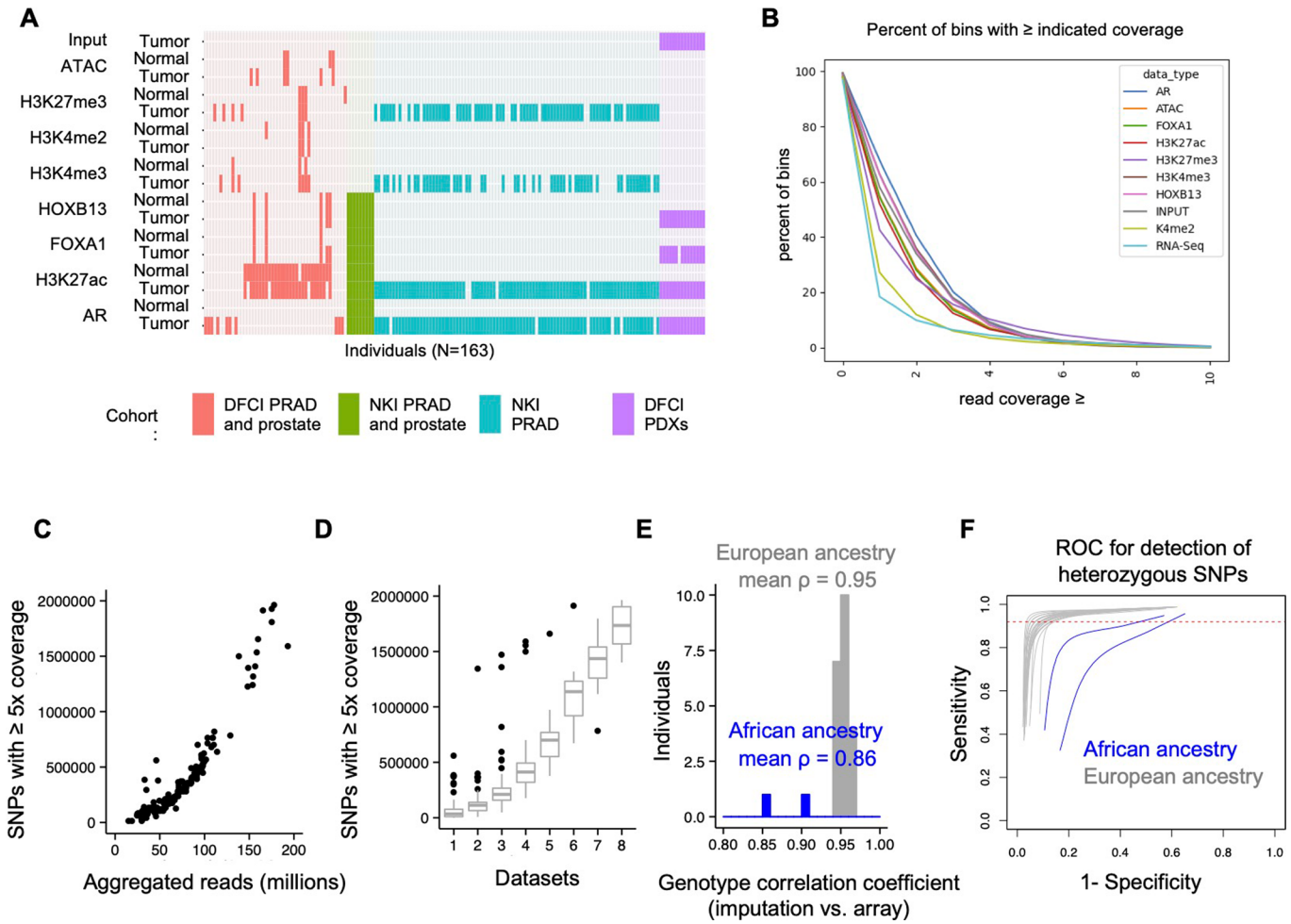
Extended data is available for this paper at https://doi.org/10.1038/s41588-022-01168-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01168-y.

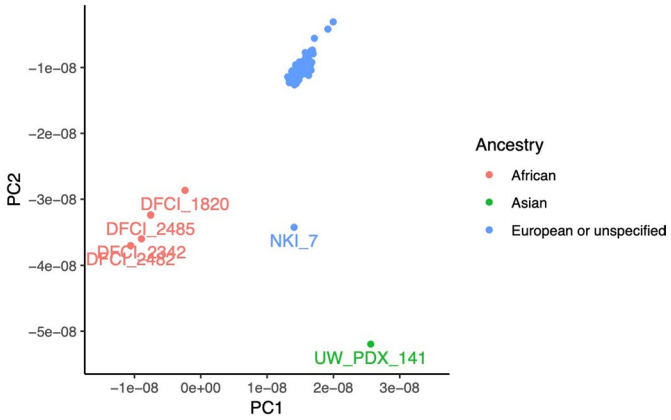
Correspondence and requests for materials should be addressed to Alexander Gusev or Matthew L. Freedman.

Peer review information *Nature Genetics* thanks Jason Stein and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

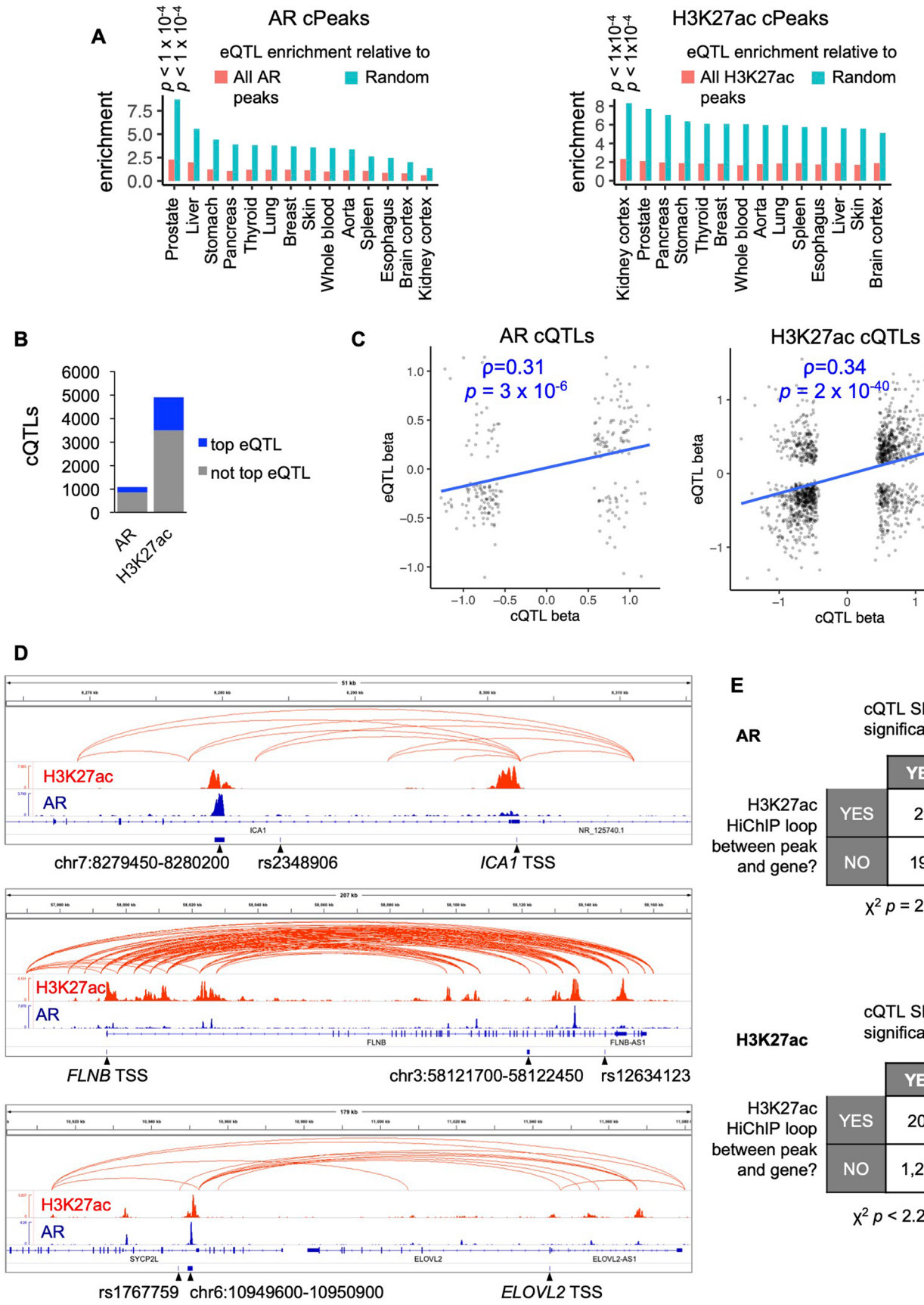
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 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 Supplementary Table 1). (**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.

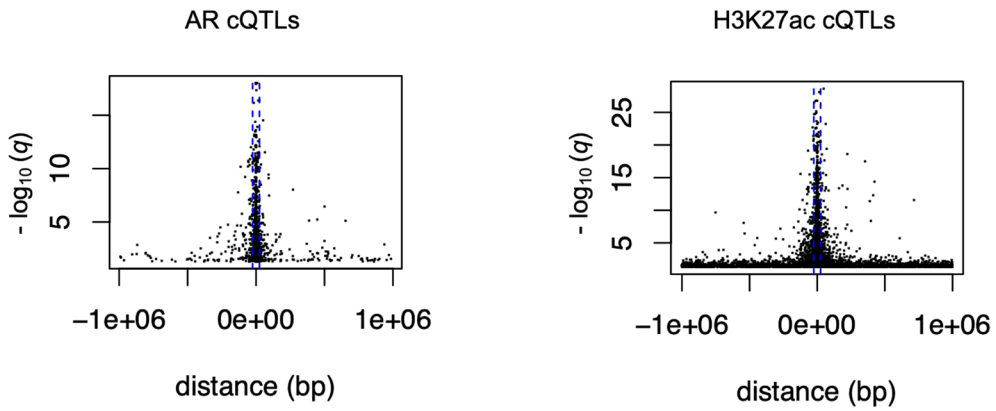


Extended Data Fig. 2 | Inferred ancestry of individuals in the study. Projection of imputed genotypes onto the first two principal components of continental ancestry from ref. [78]. Individual identifiers for outlier samples (with values > 2 x standard deviation) are labeled. Self-reported ancestry is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

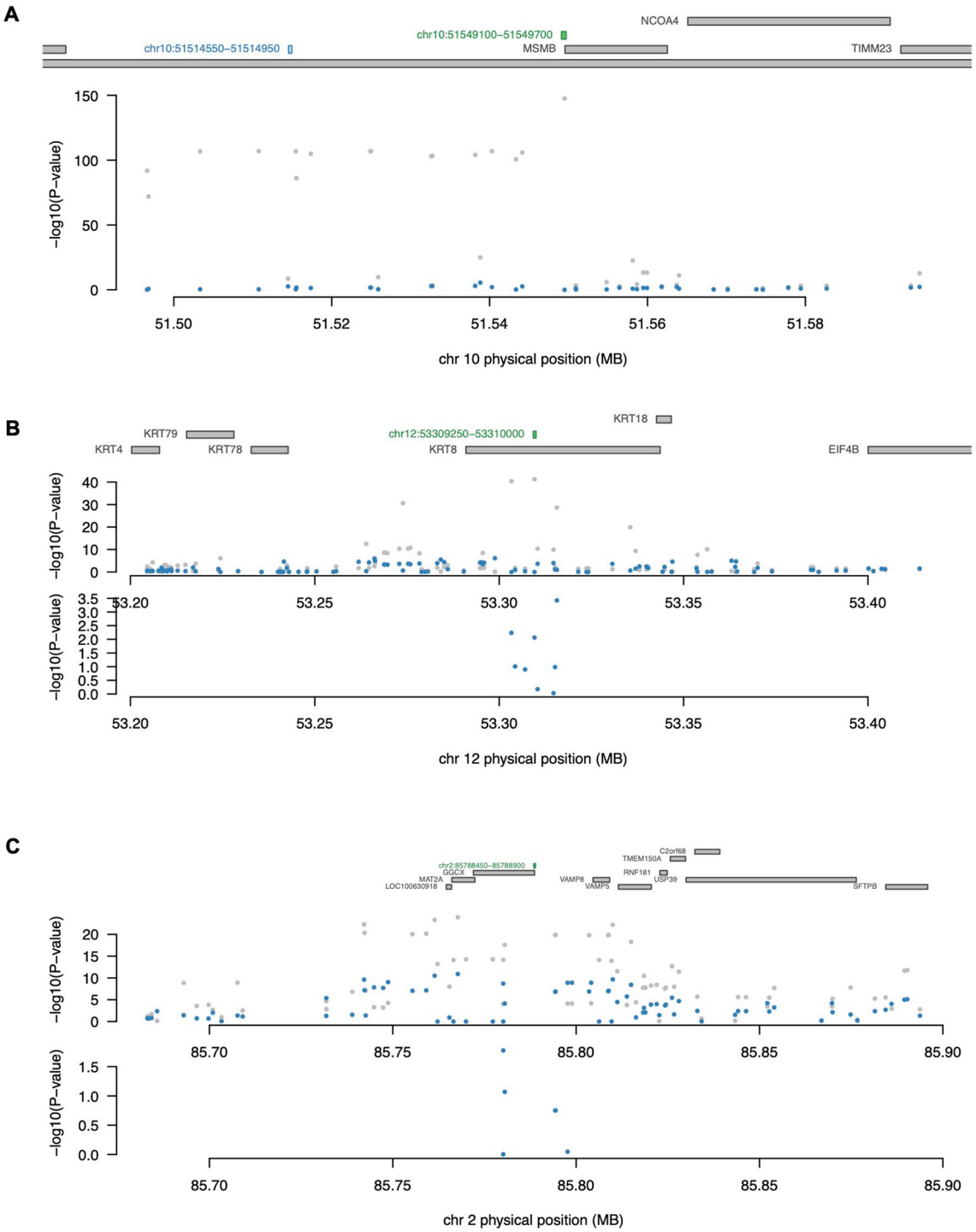


Extended Data Fig. 3 | See next page for caption.

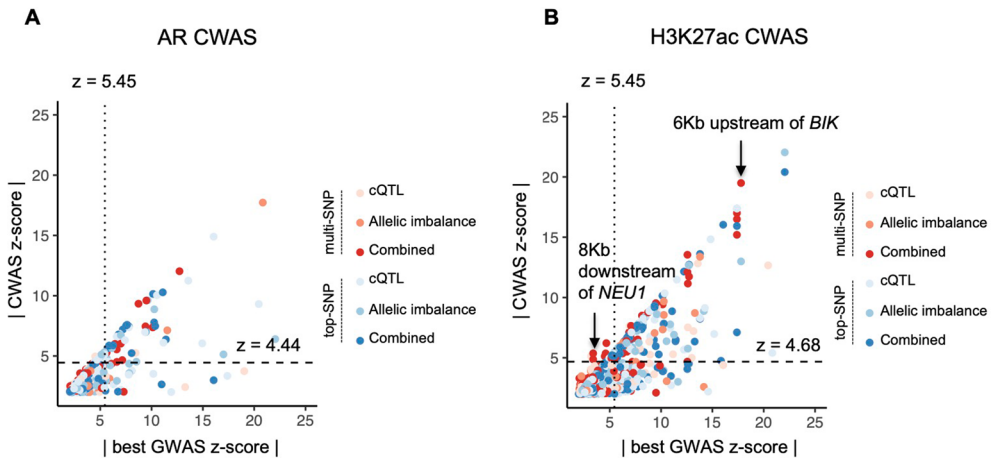
Extended Data Fig. 3 | Overlap of cQTLs with prostate tissue eQTLs. (**a**) Enrichment of genetically determined AR peaks (left) and H3K27ac peaks (right) for overlap with GWAS risk SNPs eQTLs across various tissues. Empiric p values are derived 10,000 from permutations. (**b**) number of AR and H3K27ac 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.



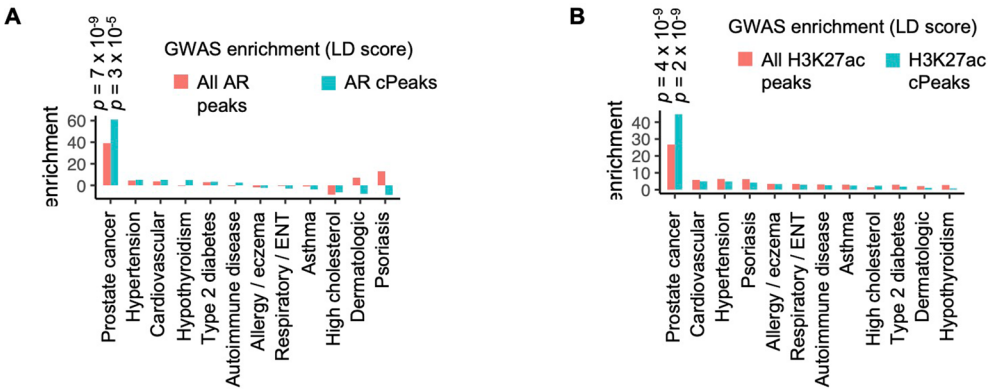
Extended Data Fig. 4 | Distribution of cQTLs around cPeaks. cQTL SNP significance versus distance to the center of the corresponding cPeak for significant cQTLs (permutation-based q-value < 0.05). Dashed blue lines indicate ± 25 Kb from the peak center.



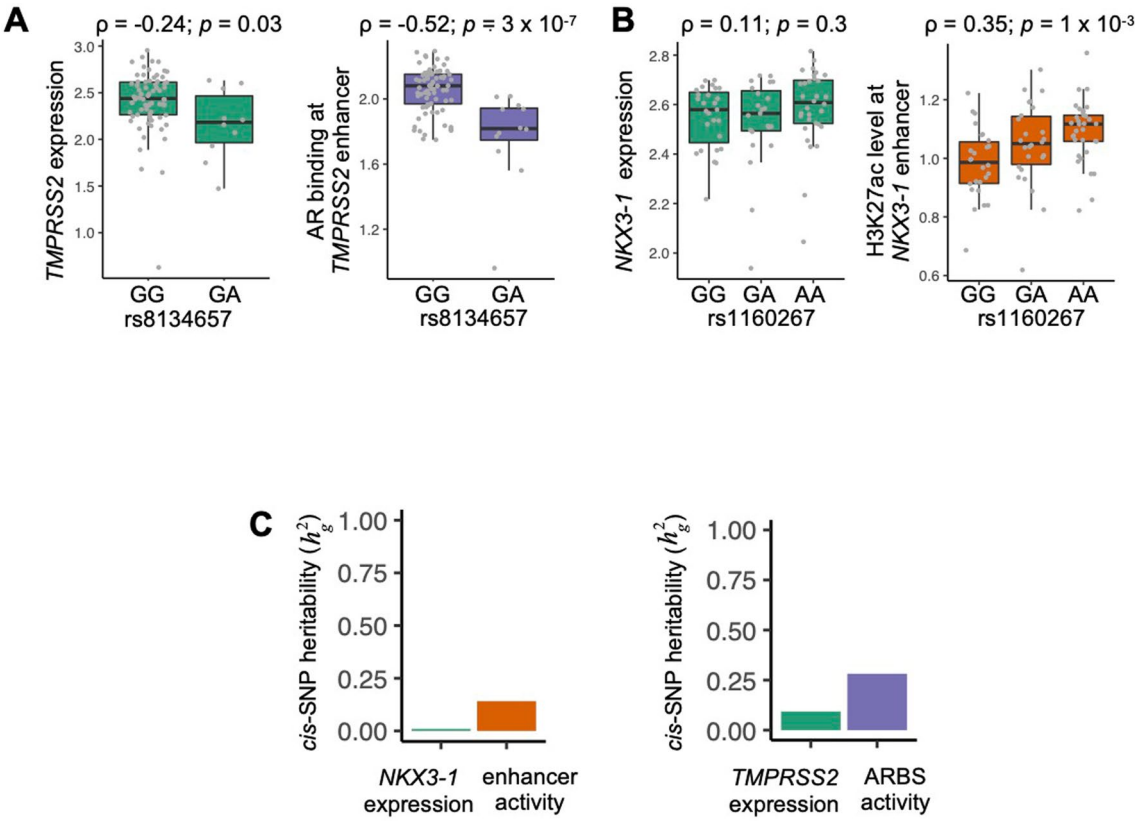
Extended Data Fig. 5 | Conditioning of GWAS SNP significance on genetically predicted CWAS AR binding. Genomic context of AR CWAS ARBS (depicted in green) that are significantly associated with prostate cancer risk. Manhattan plots indicate significance of SNP associations with prostate cancer before 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 Fig. 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 Fig. 7 | Enrichment of prostate cancer GWAS risk SNPs in genetically determined AR peaks and H3K27ac peaks. Enrichment and p-values for AR peaks (**a**) and H3K27ac peaks (**b**) derived from linkage disequilibrium score regression[5].



Extended Data Fig. 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 indicate 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 75— percentiles; whiskers extend to 1.5 x the inter-quartile ranges (IQR).

nature research

Corresponding author(s): Matthew Freedman MD
Sasha Gusev PhD

Last updated by author(s): June 18, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	NA
Data analysis	The following computational tools were used in this manuscript: MACS v2.1.1.20140616, Burrows-Wheeler Aligner (BWA) version 0.7.17, BEDTools v2.26.0., SRA toolkit fastq dump v 2.10.0, bcftools v1.9, plink2, QTLtools v1.2, Homer v4.10, IMPUTE (v2.3.1). Source code for analyses in this manuscript is available at: https://github.com/scbaca/cwas, https://github.com/gusevlab/stratAS, https://github.com/scbaca/chip_imputation, and https://doi.org/10.5281/zenodo.6666796. Please see the methods section of the manuscript for additional information.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data generated for this study are available in GEO (accession number GSE205885). Public datasets used in this study are listed in Table S1 and were obtained from SRA (https://www.ncbi.nlm.nih.gov/sra), GEO (GSE130408; GSE120738; GSE120741; GSE161948), GTEx (https://console.cloud.google.com/storage/browser/gtex-resources; https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz; https://storage.googleapis.com/gtex_analysis_v8/reference/gencode.v26.GRCh38.genes.gtf) and ENCODE (https://zenodo.org/record/3838751/files/

DHS_Index_and_Vocabulary_hg19_WM20190703.txt.gz?download=1; https://zenodo.org/record/3838751/files/DHS_Index_and_Vocabulary_hg38_WM20190703.txt.gz?download=1).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size was determined by the availability of public AR and H3K27ac prostate cancer ChIP-seq datasets. We used all available public ChIP-seq data from large-scale studies of prostate cancer epigenomes. Our identification of abundant cQTLs from this dataset demonstrates that this sample size was adequate to detect the effect of common genetic variants on chromatin features.
Data exclusions	All sequencing datasets were clustered based on genotypes of homozygous SNPs, as inferred from read pileups. Any samples that did not cluster with other samples originating from the same individual were excluded. Six samples out of 581 were excluded for this reason. See manuscript text for further details.
Replication	CRISPRi experiments were performed in two independent experiments, each time with a set of three replicates, to confirm reproducibility. cQTL effects were validated through analysis of allelic imbalance and vice versa. To replicate findings of associations between cQTLs and prostate cancer loci, we used an established method (colocalization analysis) to confirm that the peaks and prostate cancer risk tends to colocalize to the same genetic variants.
Randomization	Randomization was not used in this study. Effects of covariates were account for in cQTL calculation using a standard approach (adjustment for the first several principal components from epigenomic data).
Blinding	Blinding was not relevant to this study because it does not involve an intervention that could be subject to bias.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	H3K27ac antibody (C15410196, Diagenode; 1:1000 dilution)
Validation	This antibody has been validated and extensively used (eg, https://www.diagenode.com/en/p/h3k27ac-polyclonal-antibody-premium-50-mg-18-ml).

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	LNCaP cells and 293T cells were obtained from the American Type Culture Collection.
Authentication	LNCaP cells and 239T cells were authenticated through analysis of short tandem repeats

Mycoplasma contamination	Prior to experiments, cells tested negative for mycoplasma contamination using LookOut Mycoplasma PCR Detection Kit (Sigma-Aldrich #D9307).
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this study.